

IMPERIAL

MODELING TRUST IN ASSISTIVE HUMAN-ROBOT INTERACTION: A DATA-DRIVEN APPROACH

NICO LINGG

Thesis submitted for the degree of Doctor of Philosophy

Supervised by PROFESSOR YIANNIS DEMIRIS
Personal Robotics Lab
Department of Electrical and Electronic Engineering
Imperial College London

February 2025

COPYRIGHT DECLARATION

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a **Creative Commons Attribution Non-Commercial No Derivatives** License (CC BY-NC).

Under this license, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the license terms clear to others by naming the license and linking to the license text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this license or permitted under UK Copyright Law.

ORIGINALITY DECLARATION

I hereby declare that this thesis and the work herein detailed, was composed and originated by myself, except where appropriately referenced and credited.

London, February 2025

Nico Lingg

ABSTRACT

In human-robot interaction, particularly in assistive robotics where users rely on autonomous or semi-autonomous systems for physical support, trust is central to safe and effective collaboration. While research on trust in robotics has grown, many studies use simplified scenarios that fail to capture the complexities of real-world interactions. This thesis addresses these gaps by investigating trust through experimental user studies, introducing novel measurement tools, and developing machine learning models tailored to realistic assistive robotics applications.

Through reviewing empirical studies, we establish that existing work often uses constrained methodological approaches missing critical aspects of embodied interaction. Building on these insights, we develop and validate experimental platforms for studying trust in naturalistic settings. Using an autonomous wheelchair system, we show that trust is strongly influenced by robot performance in real navigation tasks: good performance enhances trust and attitudes, while poor performance harms perception and engagement.

To capture dynamic trust changes, we introduce *Trusty*, a handheld device enabling continuous trust measurement without disrupting interaction flow. In a user study with our wheelchair system, *Trusty* is validated against established questionnaires and shown to provide reliable real-time trust feedback.

Leveraging continuous trust data alongside video, gaze tracking, and physiological signals, we design a transformer-based model to predict user trust. Fusing these modalities with auxiliary perceptual tasks improves classification accuracy, with eye-gaze data proving particularly valuable.

We also present AMIGA (Assistive Mobile Interactive Grasping Agent), a mobile manipulation platform combining a UR10e arm with a powered wheelchair base for robust, cost-effective assistive tasks. Using AMIGA, we develop Trust-ACT, an imitation learning approach that incorporates trust annotations for policy refinement. Trust-based trajectory selection improves success rates and execution times in block-stacking tasks.

This work advances reliable, user-centered assistive robotics by deepening the understanding of trust and providing practical tools and frameworks.

ACKNOWLEDGEMENTS

I want to thank my advisor, Yiannis, for your guidance and support over these four years. Your mentorship has been essential to my research and growth.

A special thank you to my wife, Clare, for being by my side through it all. Your support kept me going through the ups and downs.

To my labmates, working together has been both fun and educational, and I'll always remember our time building robots and sharing ideas. To my friends and family, thank you for your encouragement. Finally, to everyone who contributed to this thesis - your help and support made it all possible!

CONTENTS

1	Introduction	23
1.1	Motivation	23
1.2	Limitations of Existing HRI Trust Research	25
1.3	Research Objectives and Approach	26
1.4	Key Contributions	27
1.5	Thesis Roadmap	28
1.6	Publications	30
2	Background	33
2.1	Defining Trust in Human-Robot Interaction	33
2.2	Trust Factors	35
2.3	Trust Measurements	36
2.3.1	Subjective Measures	38
2.3.2	Objective Measures	39
2.4	Trust Models	40
2.4.1	Qualitative Models	41
2.4.2	Quantitative Models	44
2.4.3	Performance-Centric Algebraic Models	44
2.4.4	Time-Series Models	46
2.4.5	MDP and POMDP-Based Models	47
2.4.6	Gaussian-Based Models	49
2.4.7	Dynamic Bayesian Network (DBN)-Based Models	49
2.4.8	Machine Learning Models for Trust Estimation	50
2.5	Real-time Feedback for Internal States	51
2.6	Transformers: Technical Foundations for Sequence Modeling	52
2.6.1	Transformer Architecture	52

2.6.2	Multi-Head Self-Attention	53
2.6.3	Positional Encodings	54
2.7	Conclusion	55
3	Trust in Human-Robot Interaction: A Review and Analysis of Em- pirical Studies	56
3.1	Introduction	56
3.2	Methodology	57
3.2.1	Search Strategy and Selection	57
3.2.2	Inclusion and Exclusion Criteria	58
3.2.3	Categorizing the HRI Trust Literature	58
3.3	Results	66
3.3.1	Experimental Context and Setup	66
3.3.2	Robot Characteristics	68
3.3.3	Trust Factor Manipulations	69
3.3.4	Statistical Analysis Methods	72
3.4	Discussion	73
3.5	Conclusion	74
4	Wheelchair Experimentation Setup: Platform and Methodology	75
4.1	Overview of the Experimental Approach	75
4.2	Autonomous Wheelchair Platform	76
4.2.1	Hardware Components	76
4.2.2	Software Architecture	76
4.3	Performance Modulation	77
4.3.1	Parameter Tuning	78
4.3.2	Performance Validation	79
4.4	Trusty: Continuous Trust Measurement Device	79
4.5	Measures	80
4.5.1	Subjective Measures	80
4.5.2	Objective Measures	80

4.6	Experimental Procedure	82
4.6.1	Environment Setup	82
4.6.2	Experimental Steps	83
4.7	Participants	84
4.8	Ethical Considerations	85
4.9	Conclusion	85
5	Building Trust in Assistive Robotics: Insights From a Real-World Mobile Navigation Experiment	86
5.1	Introduction	86
5.2	Research Questions and Hypotheses	87
5.3	Methods	87
5.3.1	Experimental Design	87
5.3.2	Trust Measurement Framework	88
5.4	Effects of Robot Performance on Trust	88
5.4.1	Data Preparation	88
5.4.2	Data Analysis	89
5.5	Discussion	92
5.6	Conclusion	94
6	Beyond Self-Report: A Continuous Trust Measurement Device for HRI	95
6.1	Introduction	95
6.2	Trusty: Design and Implementation	95
6.3	Experimental Validation	97
6.4	Objectives and Hypotheses	97
6.5	Trusty Validation Results	98
6.5.1	Statistical Approach	98
6.5.2	Trust Measurement Validation	98
6.6	Discussion	102
6.7	Conclusion	103

7	Trust Prediction in Assistive Robotics using Multi-Modal Video Transformers	104
7.1	Introduction	104
7.2	Initial Modeling Attempts and Limitations	105
7.3	Background and Motivation	106
7.4	Dataset Description	107
7.5	Multi-Modal Video Transformer	108
7.5.1	Architecture	108
7.5.2	Training Methodology	110
7.6	Experimental Results	112
7.6.1	Multi-Task Contribution	112
7.6.2	Multi-Modal Contribution	113
7.6.3	Model Architecture and Data Impact	113
7.6.4	Comprehensive Performance Evaluation	114
7.7	Discussion	116
7.8	Comparison with Existing Trust Models	117
7.9	Conclusion	118
8	Multi-Modal Trust-Aware Imitation Learning for Robotic Manipulation with AMIGA	119
8.1	Introduction	119
8.2	AMIGA Mobile Manipulation Platform	120
8.2.1	Introduction, Motivation, and Goals	120
8.2.2	Design Methodology and System Configuration	121
8.2.3	Software and Control Architecture	122
8.2.4	Control and User Interaction Examples	126
8.3	Trust-ACT: Integrating Trust in Imitation Learning	129
8.3.1	Motivation	130
8.3.2	Task and Data Collection	130
8.3.3	Trust-ACT: Trust-Conditioned Policy Architecture	132

8.3.4	Experiment Results	135
8.4	Discussion & Conclusions	138
9	Conclusions and Future Work	140
9.1	Summary of Contributions	140
9.2	Limitations	142
9.3	Future Directions	143
9.4	Concluding Thoughts	143
A	Human-Robot Interaction Trust Questionnaires	145
B	Training Hyperparameters for Multi-Modal Transformer	148
C	AMIGA System Drawings	150
D	Training Hyperparameters for Trust-ACT	153
E	iPad Controller Interface Prototype	155
F	Repository and Video Resources	157
	Bibliography	159

LIST OF FIGURES

Figure 1.1	Autonomous Wheelchair in an Office Environment . . .	24
Figure 1.2	Thesis Roadmap	29
Figure 2.1	Key components influencing trust in HRI	34
Figure 2.2	Dimensionality of Trust Measurement	36
Figure 2.3	Categorization of Trust Measurement Methods	37
Figure 2.4	Trust models in research	41
Figure 2.5	Transformer Architecture	53
Figure 2.6	Scaled dot-product and multi-head attention	54
Figure 3.1	Distribution of Studies and Participants Across HRI Dimensions	66
Figure 3.2	Distribution of Robot Autonomy in HRI Studies	68
Figure 3.3	Most Common Robot Platforms in HRI Trust Research	69
Figure 3.4	Impact of Different Factors on Robot Trust	70
Figure 3.5	Statistical Methods in HRI Trust Studies	72
Figure 4.1	Autonomous wheelchair system architecture	77
Figure 4.2	Wheelchair performance comparison: Good vs. Bad . .	78
Figure 4.3	Trusty continuous trust measurement device	79
Figure 4.4	Experimental wheelchair navigation setup	82
Figure 4.5	Autonomous wheelchair experiment protocol	83
Figure 5.1	Trust questionnaire results: Before and after comparison	90
Figure 5.2	Trust score comparisons: Good vs. bad-performing robot conditions	91
Figure 6.1	Continuous trust measurement device designs	96
Figure 6.2	Trusty: Trust measurement device and output	97

Figure 6.3	Comparison of Trusty and verbal self-report distributions	101
Figure 7.1	Limitations of Standard Linear Regression Models for Trust Prediction	106
Figure 7.2	Multi-modal video transformer architecture	109
Figure 7.3	Time-series analysis of trust prediction model performance	115
Figure 7.4	Confusion matrices for trust and performance classification	116
Figure 8.1	AMIGA mobile manipulation platform	120
Figure 8.2	Block diagram of AMIGA's software and control architecture	125
Figure 8.3	ROS-based with MoveIt	126
Figure 8.4	Vision-based teleoperation using Apple Vision Pro	127
Figure 8.5	GELLO-based direct control interface	128
Figure 8.6	Experimental setup and trust-conditioned trajectories	129
Figure 8.7	Block stacking sequence: Grasping, retracting, and stacking	131
Figure 8.8	Proposed Trust-Aware Imitation Learning Architecture	132
Figure 8.9	Trust prediction model validation against measured trust scores	136
Figure 8.10	Latent space trajectories showing trust-level behavior patterns	136
Figure E.1	Basic control mode with face tracking and virtual joystick control.	155
Figure E.2	Enhanced navigation mode with environmental mapping and object detection.	156

LIST OF TABLES

Table 2.1	Factors affecting trust in HRI	35
Table 2.2	Subjective trust scales in automation and robotics . . .	38
Table 2.3	Objective trust measures	39
Table 2.4	Muir's Integrated Model of Trust applied to autonomous wheelchairs	42
Table 2.5	Hancock et al.'s Framework for trust in autonomous wheelchairs	43
Table 2.6	Hoff and Bashir's Trust Model for autonomous wheelchairs	44
Table 3.1	Categories and Dimensions of Human-Robot Trust . .	64
Table 3.2	Distribution of participants and studies across task domains	67
Table 3.3	Trust factors and experimental environments relationship	71
Table 3.4	Trust factors across interaction intensity levels	71
Table 3.5	Trust factors relationship with task types	71
Table 4.1	Data collection streams in wheelchair experiment . . .	81
Table 5.1	Trust questionnaire sub-scales categorization	88
Table 5.2	Changes in trust factors before and after robot interaction	89
Table 5.3	T-test: trust factors for good vs bad robot performance	92
Table 5.4	Impact of robot interaction on trust factors	93
Table 6.1	Differences in post-questionnaire scores between conditions	99
Table 6.2	Comparison of trust measures across conditions	99

Table 6.3	Correlation: verbal trust vs Trusty scores	99
Table 6.4	Correlations: trust scores vs questionnaire results . . .	100
Table 6.5	Correlation: verbal checkpoints vs Trusty segments . .	101
Table 7.1	Dataset composition and trust-labeled data distribution	108
Table 7.2	Trainable Parameters of Model Components	110
Table 7.3	Experimental results summary Multi-Modal Transformer	112
Table 8.1	Morphological analysis matrix showing three system configurations for AMIGA	123
Table 8.2	Cost comparison of AMIGA mobile manipulation con- figurations	125
Table 8.3	Performance comparison between sampling strategies across trust levels	137
Table A.1	Trust in Automated Systems Test Questionnaire by Wojton et al. (2020)	145
Table A.2	Trust in Automation Questionnaire by Körber (2019) .	146
Table A.3	Trust Perception Scale by Schaefer (2016)	147
Table B.1	Training data augmentation parameters	148
Table B.2	Model Training Hyperparameters	148
Table D.1	Image Augmentation Parameters and Their Configur- ation Values	153
Table D.2	Model Architecture and Training Configuration	154
Table F.1	GitHub repositories for thesis code	157
Table F.2	Video demonstrations of key thesis technologies	158

ACRONYMS

AMCL	Adaptive Monte Carlo Localization
ACT	Action Chunking with Transformers
AMIGA	Assistive Mobile Interactive Grasping Agent
AUTC	Area Under The Trust Curve
CVAE	Conditional Variational Autoencoder
ECG	Electrocardiogram
EDA	Electrodermal Activity
EEG	Electroencephalogram
fEMG	Facial electromyogram
HRI	Human-Robot Interaction
HR	Heart Rate
HRV	Heart Rate Variability
IMU	Inertial Measurement Unit
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
MDP	Markov Decision Process
MLP	Multi-Layer Perceptron
DBN	Dynamic Bayesian Network
POMDP	Partially Observable Markov Decision Processes
ROS	Robot Operating System
TEB	Timed Elastic Band
ViT	Vision Transformer
TOAST	Trust in Automated Systems Test
TiA	Trust in Automation
TPS	Trust Perception Scale

STATISTICAL TESTS

Throughout this thesis, we employ statistical tests suited to the nature of our data, which are often non-normal. Below, we describe the tests used in our analyses, along with brief explanations and examples of their application within the thesis.

PAIRED SAMPLES T-TEST: Compares continuous measurements from the same participants at different time points, assuming normally distributed differences. Applied in Chapter 5 to analyze changes in trust-related scores pre- and post-robot interaction.

INDEPENDENT SAMPLES T-TEST: Compares means of a continuous variable between two independent groups, assuming normality. Used in Chapter 5 to evaluate trust sub-scale differences between good- and bad-performing robot groups.

MANN-WHITNEY U TEST: Non-parametric alternative for comparing two independent groups without assuming normal distribution. Applied in Chapter 6 to assess differences in continuous trust measurements and verbal ratings between robot performance conditions.

SPEARMAN'S RANK CORRELATION (ρ): Non-parametric measure of monotonic relationships between continuous or ordinal variables. Used in Chapter 6 to evaluate alignment between continuous trust signals, verbal self-reports, and questionnaire-based trust scores.

Statistical significance was evaluated at ($p < 0.05$). While we acknowledge limitations of null hypothesis significance testing [Ioannidis \(2005\)](#), we report both significant results and those approaching significance to provide a complete picture of the observed patterns in our data.

INTRODUCTION

Robotic systems are increasingly moving beyond controlled industrial environments into our daily lives - appearing in hospitals, homes, offices, and public spaces. This transition is particularly evident in assistive robotics, where autonomous systems are being developed to support mobility, manipulation, and daily tasks. Unlike traditional industrial robots that operate in segregated spaces, these assistive robots work in close physical proximity with users, often in unstructured environments with changing conditions and unexpected obstacles. The success of these systems depends not only on their technical capabilities but also on their ability to establish and maintain appropriate levels of user trust.

1.1 MOTIVATION

In these real-world settings, trust becomes a critical factor determining the success of Human-Robot Interaction ([HRI](#)). When we trust a robot appropriately, we can benefit from its capabilities while maintaining awareness of its limitations. However, both insufficient and excessive trust can lead to problems. For example, a user who lacks trust in an autonomous wheelchair might constantly override its navigation assistance, negating the benefits of automation and experiencing unnecessary physical strain. Conversely, excessive trust could lead users to rely on the system in situations beyond its capabilities, potentially creating unsafe conditions.

Effectively solving these issues of miscalibrated trust requires framing it not as a passive outcome, but as a dynamic state the robot should actively manage. A truly adaptive robot should maintain an estimation of a human's trust from available signals and then, based on that estimate, take actions to alter that trust to an appropriate level. For instance, it might act more cautiously to moderate over-trust or more transparently to build under-trust. This closed-loop paradigm represents the ultimate goal for safe and effective



Figure 1.1: An autonomous wheelchair navigating a real-world office environment. Such routine tasks, like passing through doorways, present continuous challenges that directly influence and test user trust in the system.

human-robot collaboration. The challenge of understanding trust becomes particularly complex in assistive robotics applications, where users depend directly on the robot’s capabilities. Unlike simpler automation scenarios, these interactions involve continuous physical coupling between user and robot, often in dynamic and unpredictable environments. Traditional approaches to studying trust in [HRI](#) - typically relying on post-interaction questionnaires or simplified laboratory studies - fail to capture the rich, evolving nature of trust in these real-world scenarios.

Consider an autonomous wheelchair navigating through an office environment, as shown in [Figure 1.1](#). Even these routine scenarios present continuous challenges that can affect user trust - navigating through doorways, avoiding obstacles, and maneuvering in confined spaces. Our trust in the robot might fluctuate rapidly based on its performance: hesitation near obstacles, smooth navigation through tight spaces, or unexpected movements can all influence trust levels. Traditional trust measurement methods cannot capture these moment-to-moment changes, limiting our understanding of how trust develops and evolves during physical [HRI](#).

This thesis investigates trust as a metric for evaluating and improving [HRI](#) in assistive robotics through a data-driven approach. We collect and analyze rich multi-modal data from real-world interactions to examine how trust relates to objective performance metrics and develop tools for continuous trust measurement. Using transformer-based models, we demonstrate that trust can be predicted from multi-modal sensor data. Further, we develop an imitation learning framework that leverages trust annotations to condition robot policy generation, resulting in trajectories that exhibit distinct characteristics corresponding to varying levels of human trust during interactions.

1.2 LIMITATIONS OF EXISTING HRI TRUST RESEARCH

Existing literature on trust in [HRI](#) (reviewed extensively in Chapter 3) reveals several frequent methodological patterns:

- **Passive Observation Approaches:** Many studies instruct participants to watch pre-recorded videos or simulations of robots operating in specific tasks. These designs, though informative in certain respects, seldom address the complexities of physical user-robot co-location, such as shared space, direct contact, or unexpected perturbations.
- **Limited Autonomous Capabilities:** Most studies rely heavily on pre-programmed behaviors or Wizard of Oz methods where a hidden operator controls the robot. Rather than developing robots capable of robust autonomous operation, these approaches sidestep the engineering challenges of building systems that can reliably execute complex tasks in real-world settings.
- **Tightly Controlled Lab Studies:** When real robots are involved, the experiments often occur in structured, low-risk environments where the robot's behavior is heavily choreographed. While this enables systematic testing of certain variables, it may not capture the varied and sometimes unpredictable nature of everyday environments (e.g., narrow hallways, cluttered rooms).
- **Discrete or Post-Task Trust Measurements:** Commonly, trust is gauged only via pre-task and post-task questionnaires. Although such measures can confirm whether participants ended up "trusting" the system or not, they do not explain how trust changed during the interaction.
- **Limited Physical Interaction:** Even when participants interact with robots, the degree of physicality (e.g., physically riding a wheelchair or handling objects in tandem with a manipulator) is often minimal. This can miss crucial aspects of embodied risk, which frequently shape real user impressions of trust.

While none of these approaches are inherently problematic, controlled methods can yield important foundational insights, the combined reliance on them suggests there is a shortage of studies reflecting the dynamic, physically coupled, and real-world nature of many modern robotics applications. This context motivates our efforts to design and evaluate trust in more naturalistic scenarios, gathering continuous feedback that can inform how trust might form over time.

1.3 RESEARCH OBJECTIVES AND APPROACH

Building on these observations, we formulate our overarching research goal: to design, implement, and evaluate data-driven methods for measuring and modeling trust in assistive robots operating in unstructured or semi-unstructured real-world environments, with an emphasis on continuous user feedback and multi-modal data fusion. Concretely, we address three interlinked objectives:

O1. Real-World Experimental Investigations of Trust. We set out to create two primary experimental platforms that encourage rich physical [HRI](#):

- **Autonomous Wheelchair:** Navigating an office-like environment, requiring participants to actually ride the wheelchair as it moves.
- **AMIGA (Mobile Manipulator):** A custom-built research platform combining a robotic arm with a mobile base, which we used to develop trust-aware manipulation strategies through imitation learning.

Our focus on real-world tasks aims to capture aspects of trust that might not appear in more constrained or purely observational settings.

O2. Development of Continuous Trust Measurement Methods. Recognizing the limits of purely discrete trust measurements, we explore how to capture trust in real-time using both direct user input and physiological/behavioral signals. Notably, we introduce the Trusty device, a handheld slider enabling participants to report their trust in a non-disruptive way throughout the interaction. We also collect data such as Heart Rate ([HR](#)), Electrodermal Activity ([EDA](#)), eye gaze, and robot telemetry for complementary insights.

O3. Computational Modeling and Trust-Aware Policies. To leverage trust measurements in robotics, we develop deep learning models for trust prediction from multi-modal data. First, we train transformer models to predict trust from ego-centric video, eye gaze, and physiological signals in wheelchair navigation. Second, we apply trust prediction to robotic manipulation, where a trust prediction model scores and selects high-performing trajectories generated through imitation learning.

Rather than claiming to conclusively solve all trust-related challenges in robotics, we aim to advance the field by offering empirical evidence and modeling approaches that highlight the interplay between real-world conditions, continuous trust assessment, and robot adaptation.

1.4 KEY CONTRIBUTIONS

Within the framework of these objectives, we highlight the following contributions, acknowledging that they are steps toward broader integration of trust modeling in robotics:

- C1. Empirical Review of HRI Trust Studies:** A systematic investigation of trust literature in [HRI](#), underscoring how many studies rely on scenarios where physical interaction is limited. This helps clarify the need for deeper, more immersive experimentation. (Chapter [3](#))
- C2. Real-World Wheelchair Trust Experiments:** Implementation and testing of an autonomous wheelchair in a standard office environment, allowing participants to ride the robot. We examine how varied performance profiles affect user trust over the course of actual navigation tasks. (Chapters [4–5](#))
- C3. Continuous Trust Measurement Tools:** Introduction and validation of the Trusty device, alongside multi-modal physiological data collection. These methods yield detailed time-series trust data that supplement or refine standard questionnaires, providing richer insight into the evolving nature of trust. (Chapter [6](#))
- C4. Multi-Modal Deep Learning for Trust Prediction:** Development of transformer-based models that integrate video, sensor, and physiological signals. We explore how these models can infer trust states in near real-time, potentially supporting adaptive behaviors in the future. (Chapter [7](#))
- C5. AMIGA Platform Development:** Design and implementation of a versatile mobile manipulation platform, the Assistive Mobile Interactive Grasping Agent ([AMIGA](#)), which combines a UR10e robotic arm with a powered wheelchair base. [AMIGA](#) serves as an open research platform in the Personal Robotics laboratory, facilitating a wide range of studies to advance assistive robotics research. (Chapter [8](#))
- C6. Trust-Aware Imitation Learning (Trust-ACT):** Development of Trust-ACT, an imitation learning approach that uses trust annotations to improve manipulation policies. The method extends Action Chunking with Transformers ([ACT](#)) with trust conditioning and prediction, generating trajectories reflecting varying trust levels and enabling best-of-n selection via a trust prediction model. Validated on [AMIGA](#) (Chapter [8](#)).

1.5 THESIS ROADMAP

The thesis is structured into three overarching sections (see Figure 1.2), followed by concluding thoughts and future directions:

1. Foundations:

- **Chapter 2 (Background):** Provides an overview of key trust definitions, models, and conceptual frameworks drawn from robotics, psychology, and automation literature.
- **Chapter 3 (Trust Review - Analysis of Empirical Studies):** Reviews current approaches to studying trust in [HRI](#), noting that many rely on simulations, media, or minimal physical interactions, and highlights the need to explore trust in real-world settings.

2. Wheelchair Trust Studies:

- **Chapter 4 (Experimental Platform - Autonomous Wheelchair):** Describes the design, experimental setup, and technical implementation of our autonomous wheelchair platform for real-world user studies, demonstrating robust navigation in unstructured settings.
- **Chapter 5 (Building Trust - Trust Evolution Analysis):** Presents our preliminary experimental findings, revealing that varying performance profiles significantly influence user trust during direct physical interaction with the autonomous wheelchair.
- **Chapter 6 (Beyond Self-Report - Continuous Trust Measure):** Introduces and validates the Trusty device for continuous trust measurement, showing that its readings align well with traditional trust questionnaires and verbal trust scores.
- **Chapter 7 (Trust Prediction - Multi-Modal Deep Learning):** Explores the integration of video, eye gaze, physiological signals, and robot telemetry through a transformer-based deep learning framework to predict user trust states during interactions.

3. Robot Manipulation:

- **Chapter 8 (AMIGA Platform - Trust-ACT for Imitation Learning):** First details the design and implementation of our custom mobile manipulation platform, followed by the development of Trust-ACT, an imitation learning method that uses trust annotations to improve trajectory selection during policy execution.

The thesis concludes with Chapter 9, where we synthesize key findings across all experimental work, discuss broader implications for trust in HRI, and outline promising directions for future research.

By integrating multiple lines of inquiry — literature review, real-world wheelchair experiments, continuous measurement tools, and advanced modeling methods — we seek to provide a more holistic perspective on trust within assistive robotics. Each component of the thesis is intended as a step toward deepening our understanding of how trust arises, changes, and can be used to enhance human-robot collaboration in tangible, real-world contexts.

Ultimately, we aim to offer insights that may help roboticists, designers, and researchers develop assistive systems that better respond to the subtleties of user trust and expectations. We do not claim to provide a definitive solution for all trust-related questions. Rather, we hope this work contributes empirically grounded methods, data, and frameworks that can inspire further exploration and refinement of trust-aware approaches in HRI.

1.6 PUBLICATIONS

Throughout this PhD, we have generated several peer-reviewed publications. The key publications that serve as the foundation of this thesis are highlighted first, along with their respective thesis chapters:

Lingg, N., and Demiris, Y. (2023), “Building Trust in Assistive Robotics: Insights From a Real-World Mobile Navigation Experiment,” *Proceedings of the First International Symposium on Trustworthy Autonomous Systems (TAS ’23)*, Article 37, 1–7, doi: [10.1145/3597512.3597519](https://doi.org/10.1145/3597512.3597519).

- This study investigates how trust is influenced by the performance of an autonomous wheelchair in a real-world mobile navigation task involving 27 participants. The findings reveal that good performance enhances trust and improves attitudes towards the robot, while poor performance negatively affects perception and attitudes, underscoring the importance of robot performance in building user trust.
- **Chapter 5** is based on this article.

Lingg, N., and Demiris, Y. (2023), “Beyond Self-Report: A Continuous Trust Measurement Device for HRI,” in *Proceedings of the 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 2220–2225, doi: [10.1109/RO-MAN57019.2023.10309660](https://doi.org/10.1109/RO-MAN57019.2023.10309660).

- The research introduces Trusty, a handheld device for continuously measuring trust levels in HRI, and validates its effectiveness through a

study with 29 participants using an autonomous wheelchair for mobile navigation tasks. The study finds that Trusty’s measurements align with traditional trust questionnaires and self-reported trust scores.

- **Chapter 6** is based on this article.

Lingg, N., and Demiris, Y. (2024), “Trust Prediction in Assistive Robotics using Multi-Modal Video Transformers,” in *16th International Conference on Social Robotics (ICSR)*, pp. 128–140, doi: [10.1007/978-981-96-3519-1_13](https://doi.org/10.1007/978-981-96-3519-1_13).

- This research presents a proof-of-concept study for transformer-based trust prediction in assistive robotics. The study introduces a novel multi-modal video transformer to predict user trust in human-robot interactions with an autonomous wheelchair. The model achieves approximately 65% accuracy in predicting trust levels (high, medium, low), with eye-gaze data significantly enhancing performance.
- **Chapter 7** is based on this article.

Lingg, N., and Demiris, Y. (2025), “Trust-ACT: Integrating Trust in Imitation Learning,” in *Proceedings of the 2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, (to appear).

- This paper introduces Trust-ACT, a framework that integrates trust as a conditioning variable into an imitation learning policy using a conditional variational autoencoder (CVAE). The work addresses posterior collapse with a partial masking technique and uses a trust prediction model to select optimal trajectories, improving task success rates.
- **Chapter 8.3** is based on this article.

In addition, the internship at Apple’s Machine Learning Group led to the following paper presented at a NeurIPS workshop:

Lingg, N., Sarabia, M., Zappella, L., Theobald, B. J. (2022), “Contrastive Self-Supervised Learning for Skeleton Representations,” *The Workshop on Self-Supervised Learning - Theory and Practice (NeurIPS 2022)*, arXiv preprint arXiv:2211.05304, doi: [10.48550/arXiv.2211.05304](https://doi.org/10.48550/arXiv.2211.05304).

- This research paper investigates using SimCLR, a self-supervised learning technique, to create strong representations of human skeleton point clouds. The study finds that spatial and temporal augmentations are crucial for achieving favorable results in tasks like skeleton reconstruction and activity classification. Additionally, using a graph-based encoder like ST-GCN produces superior quality representations.

The following publications, though not central to the core thesis, reflect additional research contributions made during the PhD program:

Luo, H., Chacon Quesada, R., Casado, F. E., **Lingg, N.**, Demir, Y. (2025), "Interface Matters: Comparing First and Third-Person Perspective Interfaces for Bi-Manual Robot Shoe Lacing Behavioural Cloning," *2025 IEEE International Conference on Robotics and Automation (ICRA)* (Accepted).

- This study investigates the role of user interface design in imitation learning ([ACT](#)), comparing the effectiveness of first and third-person perspective interfaces for robot shoe-lacing, a highly dexterous, bi-manual manipulation task. The research highlights how first-person perspectives improve execution time and consistency, potentially enhancing expert demonstrations and learning outcomes

Yarici, M.C., Von Rosenberg, W., Hammour, G., Davies, H., Amadori, P., **Lingg, N.**, Demir, Y., and Mandic, D. (2022), "Hearables: Feasibility of Recording Cardiac Rhythms from Single Ear Locations," *bioRxiv*, pp.2022-12., doi: [10.1101/2022.12.11.519949](https://doi.org/10.1101/2022.12.11.519949).

- This work explores the feasibility of hearables for health monitoring, specifically examining Electrocardiogram ([ECG](#)) signals from one ear. The study demonstrates that one-ear [ECG](#) signals are reliable and show resilience to external noise in real-world situations.

Yarici, M. C., Amadori, P., Davies, H., Nakamura, T., **Lingg, N.**, Demir, Y., and Mandic, D. P. (2022), "Hearables: Ear [EEG](#) Based Driver Fatigue Detection," *arXiv*, doi: [10.48550/arXiv.2301.06406](https://doi.org/10.48550/arXiv.2301.06406).

- This paper discusses the use of Electroencephalogram ([EEG](#)) based monitoring systems for detecting driver fatigue and drowsiness. It explores the potential of in-ear [EEG](#) sensors as a wearable and unobtrusive solution for continuous monitoring of driver vigilance.

With this context in place, we now turn to Chapter 2. This chapter reviews the foundational concepts of trust from robotics, psychology, and automation literature, laying the theoretical groundwork for the experiments and findings discussed in the following chapters.

BACKGROUND

In this chapter, we review current research on trust, drawing insights from robotics, automation, machine learning, and psychology. This interdisciplinary approach helps us build a solid foundation for understanding trust in Human-Robot Interaction (HRI) — a key element for achieving effective collaboration between humans and robots. Our review lays the theoretical and technical groundwork that supports the experimental work presented in later chapters. We begin by defining trust in the context of HRI (Section 2.1), discuss the factors influencing trust formation (Section 2.2), established methods for measuring trust (Section 2.3), computational models of trust (Section 2.4), and current limitations in real-time trust assessment (Section 2.5). Finally, we provide technical foundations on transformer architectures (Section 2.6), which serve as a key building block for the deep learning approaches we develop in later chapters.

2.1 DEFINING TRUST IN HUMAN-ROBOT INTERACTION

Defining trust within the nuanced context of HRI is crucial for developing effective collaborations between humans and robots, ensuring safety, and enhancing user acceptance of robotic systems. Existing definitions from various fields offer valuable insights, but require careful consideration for application to HRI. Several definitions focus on the role of uncertainty and vulnerability, such as that proposed by Lee and See (2004), who define trust as ‘the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability.’ Similarly, Muir (1994) offers a relevant perspective, emphasizing that ‘trust is the intervening variable that mediates supervisors’ intervention behavior’. Other definitions emphasize the role of expectations, including that of de Vries et al. (2003): ‘the expectation of a user about the system, that the system will perform a certain task...while the outcome...is uncertain.’ In addition, Biros et al. (2004) and Madhavan

and Wiegmann (2007) highlight the role of confidence in their definitions of trust. Biros et al. (2004) define it as ‘having confidence in and entrusting the system automation to do the appropriate action’, while Madhavan and Wiegmann (2007) define it as ‘the expectation of, or confidence in, another and is based on the probability that one party attaches to co-operative or favourable behavior by other parties’.

Building upon the existing literature and acknowledging the unique aspects of HRI, we characterize our approach to understanding trust in this thesis. We see trust as arising in situations involving uncertainty and vulnerability for the trustor. As a dynamic, multidimensional construct, it evolves over time and across different contexts, and cannot be directly observed or measured. Trust is inherently task-specific, with a person’s trust in a robot varying depending on the activity or situation. It also acts as a mediator between past experiences and future reliance, influencing how past interactions affect future behavior. Finally, trust demonstrates individual differences, varying significantly based on the trustor’s personality, experiences, and characteristics.

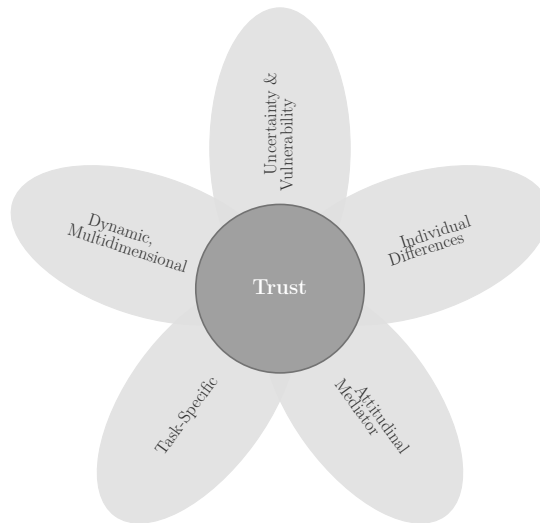


Figure 2.1: Conceptualization of trust in HRI as a dynamic construct emerging from uncertainty and vulnerability, influenced by individual traits and mediating between past experiences and future reliance.

This view of trust in HRI, as illustrated in Figure 2.1, highlights the importance of real-world, embodied interactions. The dynamic and multidimensional nature of trust becomes particularly evident when humans interact with robots in physical environments. The robot’s embodied presence allows for non-verbal communication—such as gestures, eye contact, and proxemics—which can enhance user trust. For example, a robot that maintains appropri-

ate eye contact and responds to human gestures may appear more trustworthy. Additionally, the unpredictable nature of real-world environments requires robots to adapt and respond to unexpected events, demonstrating reliability and competence, which further influences trust formation. The task-specific nature of trust is also best understood in real-world settings; for instance, a user’s trust in a robot may vary when the robot is tasked with simple navigation versus handling fragile objects. The specific demands and user expectations of different tasks strongly influence how trust evolves over time.

In summary, we view trust in [HRI](#) as a task-specific and dynamic state. It is a complex construct influenced by individual traits and past experiences, forming the basis for future reliance on robots, particularly in uncertain and vulnerable contexts.

2.2 TRUST FACTORS

Trust in [HRI](#) is shaped by multiple interconnected factors that influence how users develop and maintain confidence in robotic systems. A foundational meta-analysis by [Hancock et al. \(2011\)](#) identified and categorized these factors into three main groups: human-related factors, robot-related factors and environmental factors. Each of these categories contains two subcategories with more specific factors. For example, robot-related factors are divided into performance- (e.g., reliability, predictability, failure rate and transparency) and attribute- (e.g., proximity, robot personality, adaptability and anthropomorphism) based factors. A quantitative examination of these factors has

1. Robot-Related		2. Human-Related	3. Task & Environment-Related
(a) Performance-Related	(c) Appearance-Related	Personality	Nature of task
Dependability, reliability and error	Similarity with operator	Culture	Physical presence of robot in task site
Level of automation	Gender	System understanding	In-group membership
Situation awareness, feedback and Transparency	Harmony of appearance with task	Demographics	Duration of interaction
	Anthropomorphism	Subjective feeling	Task site
(b) Behavior-Related		Experience with robots	Revocability
Dialogues	Proximity		Rationality
Likeability & friendliness	Personality (harmony with task)		Risk
Confess to reliability	Apologize for failure		Workload, complexity and required level of multitasking

Table 2.1: Factors affecting trust in [HRI](#) (adapted from [Khavas et al. \(2020\)](#))

shown that trust development is most affected by robot-related factors, followed by environmental factors. The least significant factors influencing trust were found to be human-related factors, but the authors attribute this result to a lack of empirical data from existing research. A recent paper by [Khavas et al. \(2020\)](#) updated the findings from [Hancock et al. \(2011\)](#) and excluded some factors while adding others due to a lack of increased interest in recent studies. Furthermore, the authors modified the classification for some factors to reflect the current state of trust research compared to [Hancock et al. \(2011\)](#). Table 2.1 shows the factors affecting trust in HRI according to [Khavas et al. \(2020\)](#).

2.3 TRUST MEASUREMENTS

As established in Section 2.1, trust is a complex, latent, and multi-dimensional internal state influenced by the numerous factors discussed in Section 2.2. This inherent complexity makes direct and complete measurement challenging. For empirical study and practical application, we often rely on assessing a simplified representation of this state.

We suggest that when individuals are asked to report their trust (e.g., "How much do you trust this robot?"), they undertake a cognitive process that effectively reduces the dimensionality of their internal assessment. This could involve integrating the various salient factors from the high-dimensional trust space — such as perceived performance, appearance, task context, and vulnerability (Figure 2.2, left panel) — into a single, reportable judgment or feeling of trust. The outcome of such a process can be conceptualized as a

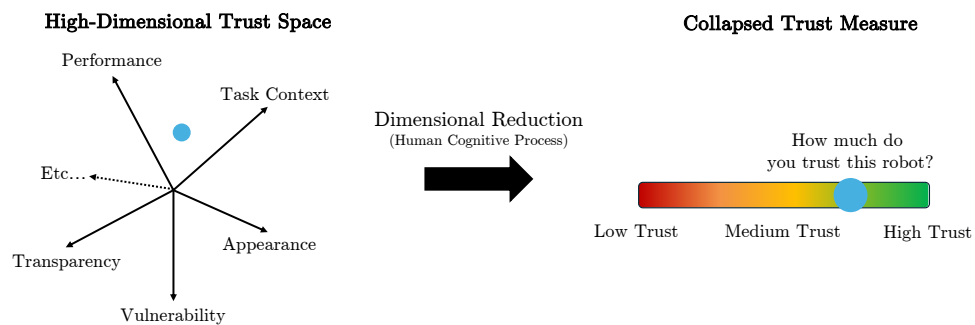


Figure 2.2: Trust measurement concept: A complex internal trust state, influenced by multiple factors (left), is cognitively processed by humans into a single 'collapsed' trust level (right) for reporting, a process conceptualized as dimensional reduction.

'collapsed' trust measure (Figure 2.2, right panel), which is reflected in the single values often obtained through self-report instruments.

It is typically this 'collapsed' representation that we aim to capture using various measurement techniques. Therefore, throughout this thesis, while acknowledging the multi-faceted nature of trust defined earlier (Section 2.1), when we refer to measuring, predicting, or conditioning on trust, we are specifically referring to operationalized indicators or reported levels of this collapsed representation. These serve as valuable, tractable proxies for quantitative analysis, allowing us to leverage the practical utility of quantitative scales and models while remaining grounded in the acknowledged complexity of trust.

With this understanding of how complex trust states are translated into reportable, collapsed measures, we now turn to the specific methods researchers use to capture or infer these trust indicators. Broadly, these approaches can be categorized into subjective and objective measures. Subjective measures aim to directly measure the user's reported level of trust, often through questionnaires or self-report scales. Objective measures, on the other hand, track physiological or behavioral correlates that are thought to reflect the underlying trust state, though they typically require further interpretation. Figure 2.3 provides an overview of these primary categories and their sub-types.

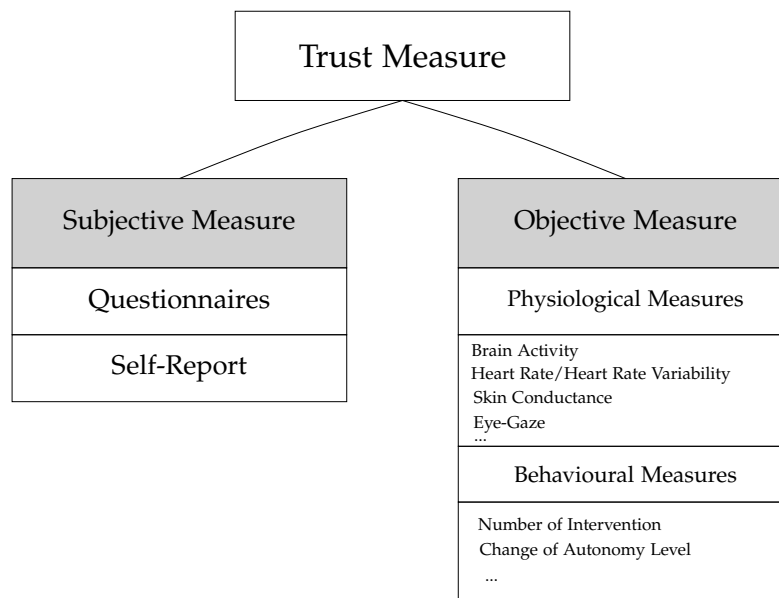


Figure 2.3: Common trust measurement methods in automation and robotics, categorized into subjective approaches (e.g., questionnaires, self-reports) and objective approaches. Objective methods include physiological indicators (e.g., heart rate, eye-gaze) and behavioral indicators (e.g., interventions, changes in autonomy level).

Paper	Description	Domain
Jian et al. (2000)	A 7-point scale consisting of 12 items that measure global trust in a system. Items include ‘The system is dependable’ and ‘I can trust the system.’	HRI
Lee and Moray (1994)	A 10-point scale comprising five questions that measure trust and self-confidence in the automated system. Questions include ‘How much did you trust the automatic controller?’	Automation
Madsen and Gregor (2000)	A 5-point Likert scale (strongly disagree to strongly agree) consisting of 25 items measuring human–computer cognition- and affective-based trust. Items include ‘The system performs reliably.’	Human-Computer Interaction
Chien et al. (2014)	A 5-point Likert scale (strongly disagree to strongly agree) across three main constructs (performance expectancy, etc.) and three types of moderators (cultural differences, etc.) for a cross-cultural trust scale with 9 items.	Automation
Desai et al. (2013)	Trust level changes measured every 25 seconds during experiments (increase, decrease, or no change). <i>AUTC</i> used to infer trust changes.	HRI
Merritt (2011)	A 5-point Likert scale (strongly disagree to strongly agree) consisting of six items. Items include ‘For the most part, I distrust machines.’	Automation
Körber (2019)	A 5-point Likert scale (strongly disagree to strongly agree) comprising 19 items measuring trust in automation. Questions include ‘The system state was always clear to me.’	Autonomous Vehicle
Wojton et al. (2020)	A 7-point Likert scale (strongly disagree to strongly agree) with nine items capturing two trust dimensions: system understanding and system performance.	Automation
Schaefer (2016)	A 40-item scale where participants estimate the percentage of time that the robot meets certain criteria or exhibits certain attitudes.	Robotics

Table 2.2: List of commonly used subjective scales for trust measurements in automation and robotics

2.3.1 Subjective Measures

Various subjective surveys and self-report scales have been proposed and used in the past in robotics and automation research. A selection of examples is shown in Table 2.2. Most of these subjective measures use a questionnaire with a Likert scale to quantify trust. However, there are several limitations of self-report scales. First, they do not produce a dynamic measurement, meaning they only capture the user’s perceived trust at one particular time (typically before and after the experiment) rather than measuring continuously over the duration of the experiment. Second, not all humans are sufficiently self-aware to understand their current level of trust accurately. In addition, Likert scales have been found to be susceptible to acquiescence bias (Hinz et al., 2007), which could cause an incorrect perception of trust. Another approach, developed by Desai et al. (2013), uses the Area Under The Trust Curve (*AUTC*) measure. At regular time intervals, participants are asked to report trust changes (trust increased, trust decreased, trust unchanged), which are then integrated over time to obtain the *AUTC* measure. Researchers selecting a

Objective Measure	Description	Example Studies
<i>Physiological Measures</i>		
Heart Rate (HR) / Heart Rate Variability (HRV)	Measures cardiac activity; changes indicate stress or relaxation levels associated with trust	Van Lange et al. (2011)
Electrodermal Activity (EDA)	Assesses skin conductance; increases with sweating due to stress or arousal	Walker, Wang, Martens and Verwey (2019)
Electroencephalogram (EEG)	Records brain's electrical activity; detects cognitive and emotional states linked to trust	de Visser et al. (2018)
Eye-gaze tracking	Monitors eye movements and blink rates; indicates attention, arousal, and trust levels	Walker, Wang, Martens and Verwey (2019)
Facial expression analysis	Interprets facial muscle movements to detect emotional states	Khalid et al. (2016)
Facial electromyogram (fEMG)	Measures facial muscle activity; sensitive to subtle emotional expressions	Ben Ajen-aghughrure et al. (2021)
Skin temperature	Observes changes in skin temperature reflecting emotional states	Chauhan et al. (2024)
Voice/speech analysis	Analyzes vocal characteristics; may reflect emotions related to trust	Elkins and Derrick (2013)
<i>Behavioral Measures</i>		
Number of interventions	Counts how often a human intervenes in the robot's actions; higher frequency may indicate lower trust	Xu and Dudek (2015)
Change of autonomy level	Observes human-initiated changes in the robot's autonomy level	Kaniarasu et al. (2012)
Compliance or reliance	Measures the extent to which humans rely on the robot's suggestions or actions	De Visser et al. (2017); Weerawardhana et al. (2024)
Hesitation or response time	Analyzes delays in human responses as indicators of uncertainty or trust levels	Maehigashi et al. (2024); Maehigashi and Yamada (2023)

Table 2.3: Objective physiological and behavioral measures commonly employed for assessing trust in HRI. Physiological measures capture unconscious bodily responses, while behavioral measures quantify observable human actions and decisions during robot interactions.

subjective scale to use in their experiment must consider that, depending on the experiment, different scales might not give the same results; most subjective questionnaires are not interchangeable (Kessler et al., 2017).

2.3.2 Objective Measures

Physiological signals can be used to estimate the stress levels and cognitive load of humans (Conway et al., 2013). For example, in stressful situations, human body functions such as HR, skin conductance and brain signals are

affected, and physical signals such as changes in facial expression, speech or behavior can be detected. These physiological responses can then be used to predict a person's stress level (Conway et al., 2013). The same premise is true for trust. Researchers have proposed various objective measures for trust estimation. Table 2.3 shows commonly used objective measures that have been used in trust research.

The advantages of objective measures are that they can be used for real-time assessment in any environment, and any personal biases or falsifications that are often present in subjective measures can be ruled out. According to Walker, Wang, Martens and Verwey (2019), a combination of eye gaze and EDA signals predicts trust better than either method would individually. Another study showed that interventions in a robot's autonomy level could be used as a real-time indication of a change in trust (Kaniarasu et al., 2012). Van Lange et al. (2011) found that measures of HR predict trust, behavioural trust and self-reported trust. The premise is that people tend to feel less anxious and more relaxed around people or systems they trust. Furthermore, several studies (Ajenaghughrure et al., 2019; de Visser et al., 2018) suggest that EEG is a highly promising measure of trust, but this approach requires further research. In a recent paper, Ben Ajenaghughrure et al. (2021) investigated which objective measures are most suitable for real-time trust assessment. Using an autonomous driving simulator, the researchers collected EEG, EDA, Electrocardiogram (ECG), eye-tracking and fEMG data and trained a variety of trust classifier models individually for each objective measure and as an ensemble. The authors found that EEG is most reliable in predicting trust as a single measure, but an ensemble approach outperformed all other models. Combining multiple signals to improve trust estimates seems particularly promising.

2.4 TRUST MODELS

Trust models have developed across various categories, each tailored to meet specific research needs and diverse contexts. As shown in Figure 2.4, these models can be broadly classified into qualitative and quantitative approaches, both offering distinct methodologies for understanding, measuring, and effectively predicting trust outcomes.

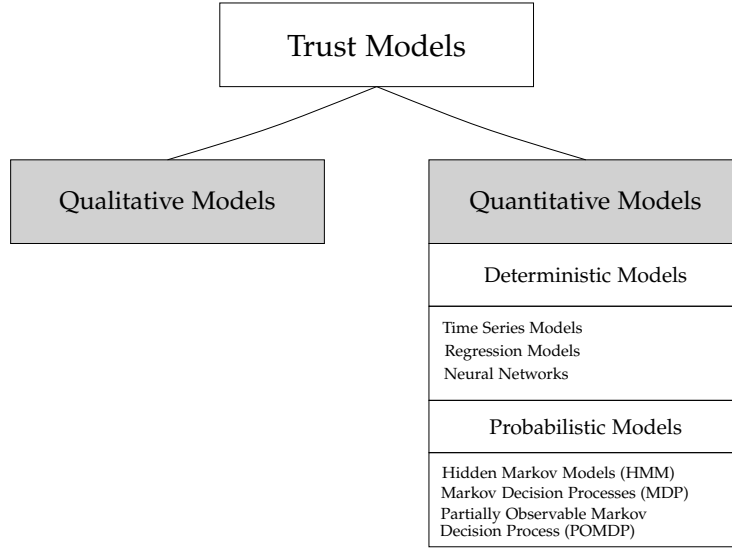


Figure 2.4: Common trust models: Illustrating the distinction between qualitative (descriptive) and quantitative (computational) approaches.

2.4.1 Qualitative Models

Various qualitative models have been proposed to understand the factors influencing trust in automation (Hancock et al., 2011; Hoff and Bashir, 2015; Muir, 1994). These models offer valuable insights and serve as heuristics to guide research but are not designed for quantitative predictions.

Muir (1994) developed a qualitative model of trust in automation by adapting models of interpersonal trust to the human-machine relationship. This integrated model draws upon two key frameworks: Barber’s (1983) taxonomy of trust expectations (persistence, technical competence, and fiduciary responsibility) and Rempel et al.’s (1985) model of trust dynamics (predictability, dependability, and faith). The core of Muir’s model lies in defining trust (T_i) held by an individual i in an automation system j as a composite of these expectations:

$$T_i = E_i(P_n + P_m) + E_i(TCP_j) + E_i(FR_j) \quad (2.1)$$

where E_i represents the expectations held by individual i ; P_n and P_m represent the expectations of persistence of the natural physical and moral social orders, respectively; TCP_j is the expectation of technical competence from system j ; and FR_j is the expectation of fiduciary responsibility from system j . To illustrate the application of this model, consider the development of trust in an autonomous wheelchair. Table 2.4 outlines how these expectations intersect with the stages of trust development.

Table 2.4: Examples of Muir’s Integrated Model of Trust in an Autonomous Wheelchair (Adapted from Muir, 1994). This table illustrates how trust may develop through stages of Predictability, Dependability, and Faith (columns, adapted from Rempel et al., 1985) across three key expectations users may have of automation: Persistence, Technical Competence, and Fiduciary Responsibility (rows, adapted from Barber, 1983). Each cell provides a concise example of how the corresponding expectation may manifest in user experience with the wheelchair at each stage of trust development.

Expectation	Predictability	Dependability	Faith
Persistence	Moves straight, stops for obstacles.	Navigates unfamiliar rooms, avoids obstacles.	Navigates crowded environments safely.
Technical Competence	Follows pre-programmed routes.	Adjusts to uneven surfaces, narrow corridors.	Adapts to unpredictable environments (airports).
Fiduciary Responsibility	Stops safely for obstacles, avoids collisions.	Avoids pedestrians, adjusts route for timely arrival.	Makes safe decisions in emergencies (sudden crossings, power outages).

Muir further proposes a hypothetical regression model to capture the potential interactions between these expectations:

$$T_i = B_0 + B_1P + B_2TCP + B_3FR + B_4P \cdot TCP + B_5P \cdot FR + B_6TCP \cdot FR + B_7P \cdot TCP \cdot FR \quad (2.2)$$

where P represents Persistence, TCP represents Technical Competence, FR represents Fiduciary Responsibility, B_0 is the intercept, and $B_1, \dots, B_7$ are coefficients representing the weights and interaction effects. This equation highlights the potential complexity of trust formation, suggesting that it’s not a simple additive function of individual expectations. However, it remains a hypothetical model due to the challenges in measuring these constructs.

Similarly, [Hancock et al. \(2011\)](#) developed a qualitative framework for [HRI](#) trust through a meta-analysis of 19 empirical studies. They categorized influential factors into three groups: human-related (H), robot-related (R), and environment-related (E), expressed as:

$$T = f(H, R, E) \quad (2.3)$$

where T represents trust. These categories encompass various sub-factors influencing trust formation, as detailed in Section 2.2 (Table 2.1). To illustrate, consider an automated wheelchair user relying on the technology (Table 2.5).

Table 2.5: Examples of [Hancock et al. \(2011\)](#) framework applied to an autonomous wheelchair scenario. This table demonstrates how various human-related, robot-related, and environmental factors may influence trust in an autonomous wheelchair for a user with a disability who relies on it for mobility.

Category	Example Factors
Human-related factors	User's disability level, prior experience with assistive technology, technical proficiency, anxiety about technology failures, confidence in their ability to handle unexpected situations, familiarity with the wheelchair's functionalities.
Robot-related factors	<i>Performance:</i> Navigation accuracy and reliability, responsiveness to user commands, battery life and reliability, safety mechanisms (e.g., obstacle avoidance, emergency stops). <i>Attributes:</i> Comfort and ergonomics of the wheelchair, ease of use and customization of controls, aesthetic design, perceived "intelligence" and adaptability of the navigation system.
Environmental factors	Accessibility of the built environment (e.g., ramps, elevators, smooth surfaces), presence of obstacles and crowds, weather conditions, social stigma related to disability and assistive technology, support and training available from healthcare professionals.

[Hancock et al. \(2011\)](#)'s analysis, discussed further in Section 2.2, revealed that robot-related factors (R), especially performance-based ones, had the most significant impact on trust. While human-related (H) and environmental (E) factors also contribute, their influence was found to be smaller. Unlike Muir's theoretical model, this framework is descriptive, empirically-based, and avoids specific quantitative formulations.

Likewise, [Hoff and Bashir \(2015\)](#), based on a systematic review of 127 studies, proposed a three-layered model of trust in automation: dispositional, situational, and learned trust. This model can be represented as:

$$T = f(D, S, L) \quad (2.4)$$

where T represents overall trust, and D , S , and L denote dispositional, situational, and learned trust, respectively.

Dispositional trust (D) reflects an individual's inherent tendency to trust automation, influenced by factors like culture, age, gender, and personality. Situational trust (S) depends on the interaction context, including workload, risk, organizational setting, and the user's transient internal state (e.g., mood, self-confidence). Learned trust (L) arises from past experiences with the specific automation or similar systems, shaped by the system's performance, design, and the user's prior knowledge. Table 2.6 illustrates this model applied to an autonomous wheelchair user.

Table 2.6: Hoff and Bashir’s (2015) Three-Layered Trust Model Applied to an Autonomous Wheelchair. This table demonstrates how dispositional, situational, and learned trust may interact to influence a user’s overall trust in an autonomous wheelchair.

Trust Layer	Example Factors
Dispositional (D)	The user’s personality (e.g., general tendency to trust technology), cultural background, age, and gender may influence their initial and overall trust in the wheelchair.
Situational (S)	The user’s current workload (e.g., navigating a busy environment), perceived risk of using the wheelchair in specific situations (e.g., crossing a busy street), social context (e.g., presence of other people), and their current mood and self-confidence may affect their trust level.
Learned (L)	Prior experiences with the specific wheelchair model (e.g., past failures or successes), experiences with other similar assistive technologies, knowledge about the wheelchair’s functionalities and limitations, and awareness of its reliability and safety features may shape the user’s trust.

2.4.2 Quantitative Models

Quantitative trust models provide mathematical representations of trust, enabling robots to predict human trust levels and adapt their behavior accordingly to maintain effective human-robot collaboration. These models have been developed using various approaches, which are commonly categorized into five classes: performance-centric algebraic models, time-series models, Markov Decision Process (MDP) and Partially Observable Markov Decision Processes (POMDP) models, Gaussian-based models, and Dynamic Bayesian Network (DBN)-based models.

Each category offers unique perspectives and methodologies for understanding, measuring, and predicting trust. The choice of model depends on factors such as desired interpretability, data availability, the need to capture temporal dynamics, computational resources, and integration with robot control systems.

2.4.3 Performance-Centric Algebraic Models

Performance-centric algebraic models are deterministic models that quantify trust based primarily on observable performance metrics of the robot or automation system. These models often use explicit mathematical equations that directly link trust to performance indicators, assuming a direct and immediate relationship between the two.

An example of such a model is the RoboTrust algorithm proposed by Mikulski et al. (2012), designed for trust assessment in autonomous convoy operations. The trust at time step k , denoted as $\text{RoboTrust}(k)$, is computed using a conservative update mechanism:

$$\text{RoboTrust}(k) = \min \left(\begin{array}{c} \frac{\sum_{j=k-\tau_T}^k \text{OB}(j)}{\tau_T+1} \\ \frac{\sum_{j=k-\tau_T-1}^k \text{OB}(j)}{\tau_T+2} \\ \vdots \\ \frac{\sum_{j=k-c_T}^k \text{OB}(j)}{c_T+1} \end{array} \right) \quad (2.5)$$

The model uses two key parameters: tolerance (τ_T) and confirmation (c_T) defining minimum and maximum observation window lengths respectively, with binary observations $\text{OB}(j) \in \{0,1\}$ indicating success/failure at each time step. The confirmation parameter describes how many consecutive positive observations are needed to reach complete trust, while tolerance defines how many consecutive negative observations lead to complete distrust. By computing averages over progressively longer windows (from τ_T to c_T) and taking their minimum, the model maintains sensitivity to recent negative experiences while requiring sustained good performance to build trust. Originally developed for convoy scenarios where follower vehicles must infer trustworthiness from leader performance without knowledge of convoy goals, this conservative approach prioritizes safety in failure-critical situations. However, the model has notable limitations: it cannot learn behavior patterns over time, applies parameters uniformly across agents regardless of their characteristics, and lacks mechanisms for dynamic parameter adjustment based on experience.

Another performance-centric algebraic model was proposed by Xu and Dudek (2016) for asymmetric human-robot collaborations. Their model quantifies changes in trust through a weighted combination of performance metrics and assessments:

$$\begin{aligned} \Delta T_{\text{session}}(W, Q) &= \frac{1}{Q} [\text{round}(Q \cdot T^{\text{post-session}}) - \text{round}(Q \cdot T^{\text{pre-session}})] \\ &= \sum_{\forall i,j,k} (\omega_0 + \omega_i E_i^{\text{post-event}}(W) + \omega_j E_j^{\text{session}} + \omega_k A_k^{\text{session}}) \end{aligned} \quad (2.6)$$

where W represents the post-event window duration, Q is the trust quantization level, $E_i^{\text{post-event}}$ are experience-based metrics within window W , E_j^{session} are session-level metrics, A_k^{session} are user assessments, and ω terms are weights learned from data. The model considers performance at

multiple timescales, from immediate post-event reactions to session-wide experiences. Validated in visual navigation tasks, this model predicted trust changes with personality factors being stronger predictors than experience metrics.

Other performance-centric algebraic models include [Floyd et al. \(2015\)](#), where trust in cooperative patrol tasks updates based on cumulative success or failure of robot actions, $\text{Trust}_B = \sum_{i=1}^n w_i c_i$, with c_i as action success (1) or failure (−1) and w_i as weights. [Itoh and Tanaka \(2000\)](#) introduced a step-function model defining trust as $t(x) = 1$ for dependable conditions ($x \in D$), $\alpha_x \in [0, 1]$ for unpredictable conditions ($x \in UP$), and 0 otherwise, linking trust to performance thresholds. Similarly, [Rabby et al. \(2020\)](#) proposed a step function where trust $T(t)$ is 0 if robot performance $R_p(t)$ is below f_p , transitions to ϵ near dependability f_D , and reaches 1 at a maximum threshold f_F , reflecting relative performance.

Performance-centric algebraic models are efficient in scenarios where robot performance is the dominant consideration, and trust adjustments based on performance changes are immediate and direct. However, they are limited in their ability to capture the dynamic and multifaceted nature of trust. They often overlook factors such as human individual differences, contextual influences, and the history of interactions. Moreover, they assume immediate and linear effects of performance on trust, which may not reflect the complexity observed in real-world [HRI](#) scenarios.

2.4.4 Time-Series Models

Time-series models capture the temporal evolution of trust by modeling it as a dynamic process influenced by past trust levels and other time-dependent variables. These models account for the inertia in trust dynamics, recognizing that trust changes gradually over time rather than instantaneously.

An influential time-series model is the Autoregressive Moving Average Vector (ARMAV) model proposed by [Lee and Moray \(1992\)](#). As a general form, an AR model can be represented as:

$$T(k) = \phi_1 T(k-1) + A_1 P(k) + A_2 F(k) + a(k) \quad (2.7)$$

In this equation, $T(k-1)$ is the trust level at time $k-1$, $P(k)$ represents the performance of the automation at time k , and $F(k)$ denotes fault occurrences at time k . The coefficients ϕ_1 , A_1 , and A_2 quantify the impact of previous trust, performance, and faults, respectively. The term $a(k)$ is a random noise

capturing unexplained variations. Lee and Moray (1992) found that, for their specific application, a model incorporating past values of performance and faults provided a better fit, and the noise term was found to be negligible.

This model reflects that current trust is influenced by prior trust and current system performance and faults. The inclusion of the noise term acknowledges the stochastic nature of trust dynamics. However, the model assumes linear relationships between variables, which may not capture more complex, non-linear trust dynamics.

Sadrifaridpour et al. (2016) extended this approach by incorporating human physical and cognitive performance into the trust model. The trust at time k , $T(k)$, is given by:

$$\begin{aligned} T(k) = & AT(k-1) + B_1 P_R(k) + B_2 P_R(k-1) \\ & + C_1 P_H(k) + C_2 P_H(k-1) \\ & + D_1 F(k) + D_2 F(k-1) \end{aligned} \quad (2.8)$$

Here, $P_R(k)$ is the robot's performance at time k , $P_H(k)$ is the human's performance at time k , and $F(k)$ represents fault occurrences at time k . Coefficients A , B_i , C_i , and D_i are parameters estimated from data. This model acknowledges that both human and robot performances, as well as faults, influence trust over time, capturing the mutual adaptation in human-robot teams.

Time-series models are valuable for modeling trust in scenarios where interactions occur over extended periods, and trust dynamics are influenced by past experiences. They can capture temporal patterns and inertia in trust changes. However, they may assume linear relationships and require careful parameter estimation, which can be challenging due to the complexity of trust dynamics. Additionally, they may not fully capture non-linearities or account for unobserved variables influencing trust.

2.4.5 MDP and POMDP-Based Models

MDP and POMDP models represent trust as a state within a decision-making framework, allowing for optimization of robot policies that consider trust dynamics. In POMDPs, trust is a partially observable state, acknowledging that it cannot be directly measured but can be inferred from observations.

Mangalindan et al. (2023) developed a POMDP framework for trust-aware assistance-seeking in human-supervised object collection tasks. They model

trust as a binary hidden state (high T^H or low T^L), with state $s_k = (T_k, E_k, C_k)$, where T_k represents the trust level, E_k represents the human's experience (reliable E^+ or faulty E^-), and C_k denotes task complexity (low C^L or high C^H). The reward function is defined as:

$$R_{o,E}^a = \begin{cases} +2, & \text{if } (a, o, E) = (a^-, o^+, E^+), \\ +1, & \text{if } a = a^+, \\ 0, & \text{if } (a, o) = (a^-, o^-), \\ -3, & \text{if } (a, o, E) = (a^-, o^+, E^-) \end{cases} \quad (2.9)$$

where a^- represents autonomous operation, a^+ represents asking for assistance, o^+ represents relying on the robot, o^- represents human intervention.

Wu et al. (2017) developed a framework for trust-based task allocation in manufacturing systems by modeling trust through parallel composition of multiple MDPs. Their framework models trust as:

$$M = M^w | M^r | M^t | M^f \quad (2.10)$$

where M^w represents the assembly plan with human and robot behaviors, M^r is the robot performance MDP, M^t models human trust, and M^f represents human fatigue. The human trust MDP is defined as $M^t = S^t, s_0^t, A^t, T^t, c^t$ where S^t is the set of trust levels, s_0^t is the initial trust level, $A^t = A^r \cup \text{repair}$ is the action set, T^t represents transition probabilities, and $c^t : S^t \times A^t \rightarrow \mathbb{R}^+$ is the cost function. The parallel composition captures the interactions between trust, fatigue, and performance, where transition probabilities depend on both actions and system states. The framework was validated on an industrial assembly process.

Other approaches include the work of Chen et al. (2018), who developed a trust-POMDP approach that integrates human trust into robot decision making by modeling trust as a latent variable within a POMDP to enable trust-seeking robot behaviors. Their model was validated in a table-clearing task where the robot learned to modify its behaviors based on inferred human trust levels.

MDP and POMDP models are useful tools for modeling sequential decision making under uncertainty. They are particularly useful when trust influences the outcomes of interactions and when the robot can take actions to affect trust. However, these models can be computationally intensive, especially with large state and action spaces. Defining accurate transition and observation models requires extensive data and may be challenging due to the partially observable nature of trust.

2.4.6 Gaussian-Based Models

Gaussian-based models provide a statistical framework for representing trust as a continuous random variable, often utilizing Gaussian processes to model trust dynamics. These models offer flexibility by incorporating both the trust's mean and its uncertainty, making them particularly suited for capturing complex, non-linear relationships between influencing factors.

[Soh et al. \(2020\)](#) introduced a Gaussian process-based trust model where trust $\tau_t^a(\mathbf{x})$ is a latent function mapped from task features $\mathbf{x}$ to a value in $[0, 1]$. The model expresses trust as:

$$\tau_t^a(\mathbf{x}_k) = \int P(c_k^a = 1 | f^a, \mathbf{x}_k) p_k(f^a) df^a \quad (2.11)$$

where $f^a(\mathbf{x}_k)$ represents a latent function, and $p_k(f^a)$ is the human's belief over f^a at time step k , modeled as a Gaussian Process. $P(c_k^a = 1 | f^a, \mathbf{x}_k)$ is the likelihood of successful robot performance given the latent function and task features. The integral computes the expected probability of success. A probit likelihood, combined with the GP prior, enables approximate Bayesian inference, maintaining an approximately Gaussian posterior. This approach was validated in physical and simulated experiments (Fetch robot object manipulation and autonomous driving). Results showed GPs effectively captured trust dynamics and task-specific adaptation. Gaussian-based models capture non-linearities and provide uncertainty estimates, ideal for high-dimensional, dynamic contexts. However, they can be computationally intensive, require careful optimization, rely on Gaussian priors (which may not always hold), and often focus on performance, overlooking contextual or individual factors. Despite these limitations, their probabilistic framework and adaptability make them valuable for trust modeling.

2.4.7 Dynamic Bayesian Network (DBN)-Based Models

[DBNs](#) offer a probabilistic graphical approach to model trust, incorporating multiple interacting factors and their temporal influence. [DBNs](#) represent the dependencies between trust and influencing factors over time, allowing for the inference of trust states based on observed variables. The OPTIMo model by [Xu and Dudek \(2015\)](#) is a prominent example. The model captures how trust evolves based on prior trust and changes in robot performance, incorporating uncertainty through the Gaussian distribution.

Trust at time k , t_k , is modeled as:

$$P(t_k|t_{k-1}, p_k, p_{k-1}) = \mathcal{N}(t_k; t_{k-1} + \omega_{tb} + \omega_{tp}p_k + \omega_{td}(p_k - p_{k-1}), \sigma_t^2) \quad (2.12)$$

In this equation, t_{k-1} is the trust at time $k-1$, p_k and p_{k-1} are the robot's performance at times k and $k-1$, respectively. The parameters ω_{tb} , ω_{tp} , and ω_{td} are weights representing bias, the influence of current performance, and the influence of performance change, respectively. The term σ_t^2 is the variance capturing uncertainty. The OPTIMO model also includes observations such as human interventions and trust feedback, allowing the model to update trust estimates based on observable actions.

In multi-robot settings, [Fooladi Mahani et al. \(2021\)](#) extended the DBN framework by treating the trust in each robot as a discrete variable within a latent, unobservable state. The trust state of robot r at time k depends on its previous trust level and on the robot's performance, with transition probabilities parameterized by a categorical Boltzmann machine (CBM) - a type of neural network suitable for modeling discrete distributions. Observations such as operator interventions and trust feedback are also incorporated, enabling real-time inference of the human's trust in multiple robots. Validated in collaborative search-and-rescue tasks, this DBN approach demonstrated high accuracy in predicting trust distributions across robot teams.

DBN-based models are effective in capturing the dynamic and uncertain nature of trust, accounting for multiple factors and their temporal relationships. They are particularly suitable for complex HRI scenarios where trust is influenced by various interacting elements. However, DBNs can become computationally complex as the number of variables grows. They require substantial data for parameter estimation and may be sensitive to the choice of initial conditions and network structure.

2.4.8 Machine Learning Models for Trust Estimation

Building on the trust modeling approaches discussed in previous sections, recent advances in machine learning have enabled a shift from traditional computational models to data-driven approaches that capture complex, non-linear relationships between physiological signals and trust states. Deep learning architectures in particular have demonstrated success in processing multi-modal data streams including [EDA](#), [EEG](#), facial expressions, and kinematic patterns for real-time trust estimation.

Studies using ensemble methods for trust classification have shown promising results. Analysis of electrodermal activity using XGBoost achieved 68.99% accuracy and an AUC score of 0.73 in distinguishing between high and low trust states during human-robot handover tasks [Campagna, Chrysostomou and Rehm \(2024b\)](#). Research on facial feature analysis employed a CNN architecture with three convolutional layers (32, 48, 64 filters), achieving 78.61% accuracy for handover tasks [Campagna, Chrysostomou and Rehm \(2024a\)](#). Motion-based trust assessment using CNN architectures applied to Laban Movement Analysis features demonstrated even higher accuracy, reaching 98% for transport tasks and 95% for draping tasks [Rehm et al. \(2024\)](#).

These data-driven machine learning approaches represent promising early efforts in modeling trust, with the potential to uncover nuanced, non-linear relationships between physiological signals and dynamic trust states. Tools like Trusty (Chapter 6) and growing datasets are expanding the use of machine learning to improve trust estimation and robotic interactions.

Future work will likely focus on end-to-end architectures capable of learning directly from raw multi-modal data. This thesis explores this direction by leveraging continuous trust signals, multi-modal data fusion, and advanced learning techniques, aiming to address diverse scenarios in real-world HRI.

2.5 REAL-TIME FEEDBACK FOR INTERNAL STATES

Capturing dynamic internal states such as trust in real time is a growing area of interest in HRI research. While techniques for real-time feedback have been applied in related domains, their use specifically for trust measurement in robotics remains limited. Insights from fields like emotion recognition provide a useful foundation for exploring these possibilities.

For example, a systematic literature review by [Fuentes et al. \(2017\)](#) identified a need for more research and interfaces focused on self-reporting emotions. Tools such as the AffectButton ([Broekens and Brinkman, 2009](#)) and digital sliders for emotional assessment ([Betella and Verschure, 2016](#)) have demonstrated flexible and efficient ways to capture real-time feedback. Similarly, linear potentiometers have been used in studies to signal real-time pedestrian intentions and perceived safety ([Walker, Dey, Martens, Pfleging, Eggen and Terken, 2019](#)). These applications illustrate the potential of real-time feedback systems to provide valuable insights into dynamic states.

2.6 TRANSFORMERS: TECHNICAL FOUNDATIONS FOR SEQUENCE MODELING

Transformers (Vaswani et al., 2017) have become one of the most influential advances in deep learning for sequence modeling. Unlike Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) architectures, Transformers process input sequences in parallel and rely on attention mechanisms to learn long-range dependencies. This design not only accelerates training (by avoiding sequential recurrence) but also captures contextual relationships in a more flexible manner. Transformers have demonstrated remarkable success in domains such as natural language processing, computer vision (e.g., Vision Transformer (ViT) (Dosovitskiy et al., 2020) and TimeSformer (Bertasius et al., 2021)).

For trust modeling in robotics - which often requires integrating multi-modal, time-series data e.g., video frames, eye gaze, physiological signals - Transformers can be particularly powerful. They excel at learning cross-modal relationships and capturing temporal dependencies over long sequences. This section details the the core technical elements of Transformers, including the multi-head self-attention mechanism, the standard encoder-decoder structure, and the design choices that make them relevant to real-time trust prediction.

2.6.1 Transformer Architecture

The original Transformer (Vaswani et al., 2017), as depicted in Figure 2.5, follows an encoder-decoder paradigm. Many subsequent variants (e.g., BERT, ViT, TimeSformer) use only the encoder for representation learning.

The encoder consists of N identical layers. Each layer contains a multi-head self-attention sub-layer and a position-wise feedforward network (FFN). Residual connections and layer normalization wrap around these sub-layers, ensuring stable training. The encoder takes a sequence of tokens $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, which are first embedded into vectors (possibly with positional encodings added). The encoder outputs a sequence of hidden states $\{\mathbf{h}_1, \dots, \mathbf{h}_m\}$, where each $\mathbf{h}_i$ is a contextualized embedding of $\mathbf{x}_i$.

The decoder mirrors the encoder but includes a cross-attention sub-layer in addition to self-attention and feedforward sub-layers. It takes the encoder outputs as inputs, enabling it to attend to the encoder's representations. In language tasks, the decoder predicts output tokens autoregressively. For other applications, such as multi-modal fusion or classification, the decoder can be adapted or replaced with downstream heads.

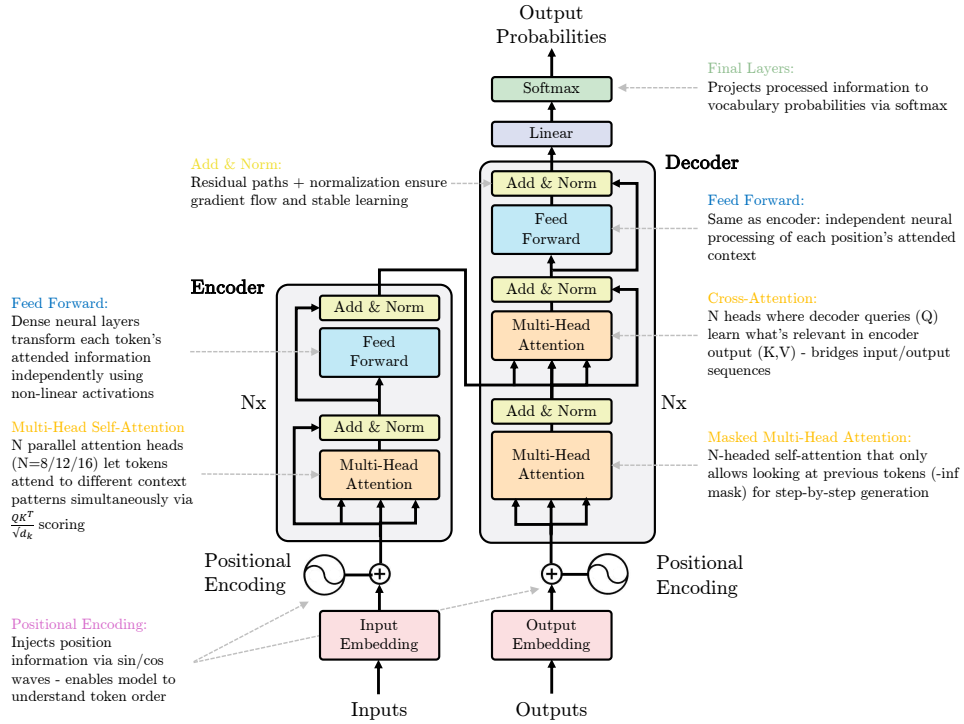


Figure 2.5: A schematic of the original Transformer (Vaswani et al., 2017), highlighting multi-head self-attention, feedforward layers, and residual connections. The encoder stack processes input sequences in parallel, while the decoder uses self-attention and cross-attention with encoder outputs.

2.6.2 Multi-Head Self-Attention

The self-attention mechanism allows the model to determine how each position in the sequence relates to every other position, capturing long-range dependencies effectively. In scaled dot-product attention, queries (Q), keys (K), and values (V) are learned projections of the input embeddings:

$$\mathbf{Q} = \mathbf{XW}_Q, \quad \mathbf{K} = \mathbf{XW}_K, \quad \mathbf{V} = \mathbf{XW}_V, \quad (2.13)$$

where $\mathbf{X}$ is a batch of input embeddings and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices. The attention operation is given by:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right) \mathbf{V}, \quad (2.14)$$

where d_k is the dimensionality of $\mathbf{K}$ (often the hidden dimension divided by the number of attention heads).

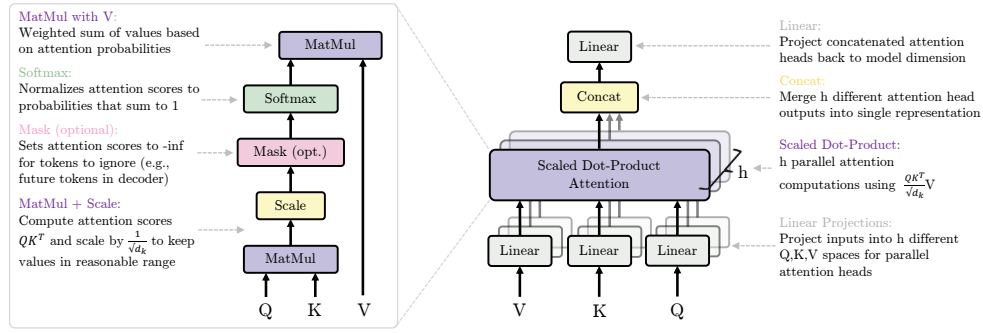


Figure 2.6: (Left) Scaled dot-product attention, where Q, K, V are projected from the input sequence. (Right) Multi-head attention repeats the attention operation in parallel and concatenates the results.

Instead of using a single attention function, Transformers apply h parallel attention heads, each with its own projection matrices:

$$\text{head}_i = \text{Attention}(QW_Q^{(i)}, KW_K^{(i)}, VW_V^{(i)}), \quad (2.15)$$

for $i = 1, \dots, h$. The outputs of these heads are concatenated and projected:

$$\text{MultiHead}(Q, K, V) = [\text{head}_1 \parallel \dots \parallel \text{head}_h] W_O. \quad (2.16)$$

Figure 2.6 (left) depicts scaled dot-product attention, while (right) shows how multiple heads combine.

2.6.3 Positional Encodings

Because Transformers do not process inputs sequentially like RNNs, positional encodings are added to token embeddings to inject information about the ordering. The original Transformer used sine-cosine position encodings:

$$\begin{aligned} \text{PE}(\text{pos}, 2i) &= \sin(\text{pos}/10^{4i/d_{\text{model}}}), \\ \text{PE}(\text{pos}, 2i+1) &= \cos(\text{pos}/10^{4i/d_{\text{model}}}). \end{aligned} \quad (2.17)$$

These sinusoidal encodings allow the model to extrapolate to sequence lengths longer than those seen during training and provide position information at different scales through their wavelengths. The relative positions of any two tokens can be expressed as a linear function of their encodings, facilitating the learning of position dependent features. For images and videos, 2D or 3D positional embeddings can be used (Bertasius et al., 2021; Dosovitskiy et al., 2020). These adaptations enable Transformers to process spatial and temporal information while preserving structural relationships in the data.

2.7 CONCLUSION

In this chapter, we examined the multifaceted concept of trust in HRI by drawing on literature from robotics, automation, machine learning, and psychology. We defined trust in HRI as a dynamic, context-specific phenomenon that is shaped by uncertainty, vulnerability, and individual factors. We then explored trust formation by discussing how human-, robot-, and environment-related elements collectively influence trust, with robot performance consistently emerging as a key determinant. Building on these insights, we investigated various methods for measuring trust, highlighting the complementary value of subjective questionnaires alongside objective physiological and behavioral metrics. We also reviewed computational models of trust — ranging from qualitative frameworks to advanced machine learning techniques — which underscored both the complexity of modeling trust and the importance of considering real-time factors. Finally, we provided the technical foundations on Transformer architectures that will serve as a key building block for the deep learning approaches we develop in subsequent chapters.

Taken together, our findings underscore that trust in HRI is an interplay of both personal and contextual elements. While we acknowledge that traditional questionnaires have proven valuable for understanding users' perceptions, the dynamic nature of trust suggests that incorporating more continuous and objective measures could be beneficial. We see the growing potential of real-time feedback systems and modern deep learning architectures, particularly transformers, as promising directions for modeling trust in complex scenarios.

To ground these theoretical insights in practice, the next chapter examines how trust has been studied empirically in HRI research through a review of 689 studies. In doing so, we identify both the strengths of existing approaches and opportunities for new methods, thereby providing more context for the experimental work we develop in later chapters.

TRUST IN HUMAN-ROBOT INTERACTION: A REVIEW AND ANALYSIS OF EMPIRICAL STUDIES

3.1 INTRODUCTION

Building upon the theoretical foundations of trust in Chapter 2, in this chapter, we examine how trust has been empirically studied in Human-Robot Interaction (HRI). Our focus shifts from theory to a synthesis of 689 empirical studies published between 2015 and 2024, each involving human participants interacting with robots where trust was a key investigative focus¹.

While the fundamental importance of trust in HRI is well-established, understanding how it manifests in practice remains challenging. Appropriately calibrated trust enables effective use of robotic capabilities while acknowledging their limitations. Miscalibrated trust can lead to underutilization or overreliance, highlighting the need to understand the perceptual, behavioral, and situational factors that influence trust in robotic systems Hancock et al. (2011). While theoretical frameworks offer a foundation, empirical studies are essential for validating these models in real-world contexts. Trust in HRI is a dynamic, multidimensional construct that evolves through experience, particularly in situations involving uncertainty de Vries et al. (2003); Hoff and Bashir (2015). The following analysis examines these studies through three key dimensions: experimental environment (media, simulation, real-world), interaction level (passive, minimal, direct contact), and task type (social, manipulation, navigation, game, supervision, evaluation). By looking at the methods, tasks, and the role of robot performance, we highlight patterns and opportunities in existing trust research approaches. This analysis not only suggests the value of studying trust in complex environments, but also helps shape our subsequent experimental design.

Our findings reveal that many current studies rely on simulation-based or media-based designs - often involving passive or minimal interaction through

¹ The dataset is available at: <https://github.com/NicoLingg/trust-in-hri-review-data>

surveys and questionnaires - that allow for greater analytical precision and larger participant pools. However, they may underrepresent the complexities of real-world human-robot trust, particularly in scenarios requiring physical interaction. One notable observation is the distinct discrepancy in sample sizes across different interaction formats: media-based and simulation studies often recruit hundreds of participants, while real-world, direct-contact settings tend to have a significantly lower number of participants. This disparity underscores how the logistical and ethical complexities of closer-proximity interaction complicate collecting robust datasets. These insights help shape our approach in subsequent chapters, from the design of real-world experiments (Chapter 4) to the development of new measurement tools (Chapter 6).

3.2 METHODOLOGY

3.2.1 Search Strategy and Selection

This review targeted empirical research on trust in [HRI](#) published between January 2015 and December 2024. The initial literature search for this review began in October 2020, with paper collection and analysis occurring over the subsequent years. The literature search was conducted across the Google Scholar, IEEE Xplore, and ACM Digital Library databases. We also examined the citations of the initially identified papers to uncover additional relevant studies. We used broad and inclusive search terms ("*trust*," "*trustworthiness*," "*human-robot interaction*," "*robot*," "*robotics*," "*autonomous systems*," "*autonomous vehicles*"), combined with Boolean operators (AND, OR) to ensure comprehensive coverage of relevant work. For example, search queries included combinations such as ("*trust*" AND "*robot*" AND "*interaction*"), or ("*trust*" AND "*robotics*"). We refined our queries iteratively based on the keywords that were more common in the initially collected literature to ensure a thorough collection of empirical research on the topic.

We applied a two-stage screening process to the selected articles. In the first stage, we automatically filtered articles by analyzing the frequency of key terms (e.g., "*participants*," "*robot*") to assess whether they likely described a user study involving robotics. In the second stage, we evaluated the remaining articles in detail to confirm their eligibility based on our inclusion and exclusion criteria, while also extracting relevant information for further analysis. To help with the analysis of this large corpus, we used large language models to chat with the papers individually, helping us understand and standardize the extraction of relevant information, with categorizations being manually verified.

3.2.2 *Inclusion and Exclusion Criteria*

We considered journal or conference papers written in English and published between January 2015 and December 2024. A key requirement was that each study presented empirical research involving human participants. This could be in the form of an experiment where participants interacted with a physical robot, a high-fidelity simulation of a robotic system, or through media such as videos or images depicting robot behavior. The central focus of each study needed to be trust within the context of human-robot interaction. This criterion was used to distinguish these works from research primarily focused on artificial intelligence systems or general automation that lacked a physical or simulated robotic embodiment. If a paper included multiple distinct studies, we treated each study as an independent data point.

Studies that did not meet these criteria were excluded. For example, non-empirical works, such as conceptual papers, tutorials, or theses, were not included. Similarly, studies examining trust in artificial intelligence without a clear link to robotics, including those focused on chatbots or purely conversational agents, were excluded, as were those on automation that did not emphasize robotic systems. From an initial pool of over 1,000 articles (after removing duplicates), 616 papers remained following the full-text review. Several papers contained multiple experiments, resulting in a total of 689 distinct studies, each coded along multiple dimensions, summarized below.

3.2.3 *Categorizing the HRI Trust Literature*

To analyze the diverse approaches in [HRI](#) trust studies, we categorized the literature along five dimensions: Environment, Interaction Intensity, Task, Trust Factors, and Robot Autonomy Level. These dimensions help organize the various ways researchers have studied trust in [HRI](#), from the context of interaction to the level of robot independence in decision-making.

3.2.3.1 *Environment*

Studies were grouped by how and where participants encountered the robot:

- **Media:** Studies where interaction with the robot is mediated through screen-based or textual materials. This includes studies using pre-recorded videos, static images, or text descriptions of robot behavior. In these studies, participants do not interact with a physical robot or

simulation in real-time, but rather observe or evaluate robot behavior through media representations.

- **Simulation:** These studies utilize interactive virtual or augmented environments that allow real-time engagement with simulated robots. This category includes both computer-based simulations where participants can control or interact with virtual robots through interfaces, and augmented reality setups where virtual robots are overlaid on the physical environment. The key distinction from media-based studies is that participants can actively influence the robot's behavior in real-time.
- **Real-world:** Studies are conducted with physical robots in actual environments. These studies involve direct interaction between participants and robots, whether through physical contact, verbal communication, or spatial interaction. Real-world studies may take place in controlled laboratory settings or naturalistic environments, but crucially involve a physical robot presence rather than virtual or media representations.

3.2.3.2 *Interaction Intensity*

We classified the degree of engagement between human participants and the robot into:

- **Passive Observation:** Participants only watch the robot (through media, simulation, or live setting) without direct engagement. This includes watching pre-recorded videos or live demonstrations, viewing static images, or observing real-time robot operations from a distance. The flow of information is unidirectional—from robot to human—with no opportunity for the participant to influence or respond to the robot's actions. Common examples include watching videos of robots performing tasks, observing robots interacting with others, or evaluating robot behavior through recorded scenarios.
- **Minimal Interaction:** Some interaction, often verbal or at a distance, but without physical contact. This includes giving voice commands, providing instructions through an interface, or making decisions about robot interventions. While participants can influence robot behavior, they maintain physical distance and do not directly manipulate the robot. Examples include using speech recognition to direct robot movement, supervising robot task execution with the ability to interrupt operations, interacting through gesture recognition systems, or providing real-time feedback through computer interfaces.

- **Direct-Contact Interaction:** Physical contact or close-proximity collaboration between the participant and robot. This ranges from simple actions like pressing a robot-mounted button to complex collaborative tasks requiring coordinated movement. Examples include physical human-robot handovers of objects, direct manipulation of robot limbs for demonstration learning, shared workspace collaboration in assembly tasks, physical guidance of robot movement, and interaction with assistive robots such as robotic wheelchairs or walking aids.

3.2.3.3 Task

We recorded the primary task of the study. Tasks typically included:

- **Social:** Tasks focused on social interaction and communication between humans and robots. This includes engaging in dialogue (e.g., conversational agents, question-answering), emotional expression (e.g., robots displaying or responding to emotions), or social cues (e.g., gaze direction, gesture recognition). Examples include robots acting as coaches providing feedback, companions for elderly care, teaching assistants in educational settings, or social support agents. These tasks often evaluate the robot's ability to maintain appropriate social distance, interpret human social signals, and provide contextually appropriate responses.
- **Manipulation:** Tasks involving physical interaction with objects, requiring precise control and coordination. This encompasses object assembly (e.g., collaborative furniture building, industrial assembly), robot-human handovers (e.g., tool passing, package delivery), sorting tasks (e.g., organizing items by category), and precision manipulation (e.g., pouring liquids, handling delicate objects). These tasks often focus on the robot's dexterity, timing, and ability to adapt to human movements while maintaining safety protocols.
- **Navigation:** Tasks centered on robot movement through space, either independently or while interacting with humans. This includes autonomous path-following (e.g., delivery robots in hospitals), guiding participants through environments (e.g., museum tour guides, shopping assistants), obstacle avoidance in dynamic environments, and collaborative navigation where robots must coordinate their movement with human partners. These tasks evaluate the robot's spatial awareness, path planning abilities, and capacity to navigate safely around humans.

- **Game:** Structured activities involving defined rules and objectives, often used to study trust development in controlled settings. This includes economic trust games (e.g., investment games, prisoner's dilemma), cooperative games requiring teamwork (e.g., collaborative puzzle-solving, team-based challenges), competitive games (e.g., board games, strategic competitions), and educational games designed to study human-robot interaction patterns. These tasks often incorporate elements of strategy, decision-making, and social dynamics.
- **Supervision:** Tasks where participants take an overseer role in monitoring robot operations. This includes system monitoring (e.g., observing autonomous manufacturing processes), error detection and intervention (e.g., identifying when robots make mistakes), performance evaluation (e.g., assessing task completion quality), and safety supervision (e.g., emergency stop decisions). These tasks often study how humans maintain situation awareness and make intervention decisions when overseeing autonomous systems.
- **Evaluation:** Tasks focused on human assessment of robot characteristics or capabilities, typically through structured evaluation protocols. This includes performance rating (e.g., assessing task completion efficiency), appearance evaluation (e.g., rating robot design elements), behavior assessment (e.g., judging appropriateness of robot actions), and trust calibration studies (e.g., measuring perceived reliability). These tasks often use standardized scales, surveys, or comparative ranking systems to collect systematic feedback on robot attributes and capabilities.

3.2.3.4 *Robot Autonomy Level*

We classified each study based on the robot's level of autonomy using six standardized categories:

- **Not autonomous:** Studies where robot actions are simulated, hypothetical, or purely descriptive without any real autonomy and no physical robot present, do not involve physical robot systems. This includes text-based descriptions of robots, survey responses about robots, or scenarios where no actions are carried out by a robot.
- **Pre-programmed (non-adaptive):** The robot follows a fixed sequence of pre-set actions or behaviors without any adaptation to the specific situation or user. While the robot may respond to sensors, there is no user influence on the robot's behavior.

- **Wizard of Oz (directly controlled):** Robot actions are directly controlled by a human operator, either in real-time or through pre-scripted choices. This includes scenarios where the operator controls the robot with or without participants' knowledge.
- **Shared control (fixed rules):** Robot operates independently but uses fixed rules or pre-determined parameters to respond to the human user. The coordination between robot and human is based on fixed parameters set by designers.
- **Shared control (adaptive):** The robot operates independently and adapts its behavior based on real-time interaction with the human, including responses to explicit commands or implicit cues like gaze or gestures.
- **Fully autonomous:** Robot operates without direct human assistance and can navigate the environment and respond to tasks, but mostly with limited adaptability to new or unexpected scenarios.

3.2.3.5 *Trust Factors*

Most studies intentionally manipulated at least one element hypothesized to impact trust. We tracked five broad categories of manipulated trust factors:

- **Performance:** Adjusting success/failure rates, accuracy, or reliability of the robot's task performance. This includes manipulating task completion success rates (e.g., varying the percentage of successful object grasps), timing accuracy (e.g., adjusting response delays or execution speed), precision in movement or positioning (e.g., placement accuracy in assembly tasks), and system reliability (e.g., frequency of technical malfunctions). Researchers often systematically varied these parameters to understand how different levels of performance influence trust development and maintenance. Performance manipulations typically focused on objective, measurable aspects of robot capability rather than subjective interpretations.
- **Behavior:** Changing the robot's strategy, level of autonomy, or moral decision-making without directly altering performance outcomes. This encompasses variations in decision-making approaches (e.g., conservative vs. risk-taking strategies), autonomy levels (e.g., fully autonomous vs. shared control), moral choices in dilemma situations (e.g., utilitarian vs. deontological approaches), and behavioral adaptability (e.g., learn-

ing from human preferences). These manipulations maintain consistent task success rates while varying how the robot achieves its objectives.

- **Communication:** Manipulating verbal or non-verbal cues, including emotional displays or social attitudes. This category includes variations in verbal communication style (e.g., formal vs. casual language), explanation provision (e.g., detailed vs. minimal task descriptions), emotional expression (e.g., empathetic vs. neutral responses), non-verbal cues (e.g., gaze patterns, proxemic behavior), and social presence indicators (e.g., personality traits, social roles).
- **Design:** Modifying the robot's appearance, user interface, or form factor. This encompasses changes to physical appearance (e.g., anthropomorphic vs. mechanical design), interface design elements (e.g., visualization styles, control layouts), aesthetic features (e.g., color schemes, material choices), and form factor variations (e.g., size, shape, proportions). These manipulations investigate how visual and interactive design elements influence trust and subsequent trust development, particularly in first-encounter scenarios.
- **Task:** Adjusting the task structure while maintaining consistent robot behavior and performance. This includes varying task complexity (e.g., simple vs. multi-step procedures), cognitive load requirements (e.g., attention demands, decision-making complexity), interaction dynamics (e.g., cooperative vs. competitive settings), time pressure (e.g., strict vs. relaxed deadlines), and environmental conditions (e.g., noise levels, lighting conditions).

Additional examples of each category and their typical implementations, along with representative papers from the literature, can be found in Table 3.1. It is worth noting that these categories are not strictly bounded; in practice, individual studies often span multiple domains or occupy gray areas, making the classification somewhat subjective. Nonetheless, we adopt this framework as a practical tool for organizing the diverse range of HRI trust research, while acknowledging that some overlap or interpretation is inevitable.

Table 3.1: Categories for analyzing HRI trust, including environment, interaction, task, trust factors, and robot autonomy, with representative papers.

Environment	
Media	Interaction via text, images, or videos. Typically, participants watch a short clip or read a scenario, then rate trust. Campanozzi et al. (2019) ; Kraus et al. (2024) ; Lyons et al. (2021)
Simulation	Virtual or augmented environments where participants observe or control a simulated robot to complete tasks. Albayram et al. (2020) ; Chi and Malle (2023) ; Esterwood and Robert (2023)
Real-world	Physical presence of the robot in the same space as participants, enabling real-time, embodied interactions. Campagna, Dadgostar, Chrysostomou and Rehm (2024) ; Goubard and Demiris (2024) ; Hald et al. (2021)
Interaction Intensity	
Passive observation	Participants only watch the robot (live or recorded) with no direct engagement or control. Ayoub et al. (2021) ; Romeo et al. (2023) ; Wald et al. (2024)
Minimal interaction	Some limited engagement (often verbal), such as providing commands or decisions, but no physical contact. Guznov et al. (2020) ; Hulcelle et al. (2021) ; Loizaga et al. (2024)
Direct-contact	Physical collaboration or touch, for example pressing a robot button, handing objects, or riding in a robotic platform. Hopko et al. (2023) ; Lingg and Demiris (2024) ; Nikolaidis et al. (2017)
Task	
Social	Emphasis on communication or emotional exchange (e.g., conversation, social coaching, companionship). Giorgi et al. (2022) ; Kraus et al. (2022) ; Shidujaman et al. (2020)
Manipulation	Tasks involving physical handling of objects (e.g., assembling items with the robot, passing objects). Cogurcu et al. (2023) ; Elias and Trigo (2024) ; Hopko et al. (2024)
Navigation	Robot motion or guidance tasks such as path-following or leading participants through a space. Dyck et al. (2023) ; Lingg and Demiris (2023b) ; Seet et al. (2020)
Game	Structured play or competition with the robot (e.g., cooperative board games, trust-based economic games). García-Corretjer et al. (2023) ; Mota et al. (2016) ; Pompe et al. (2022)
Supervision	Participants oversee or monitor the robot’s actions, often deciding whether to intervene. Luo et al. (2021) ; Okamura and Yamada (2020) ; Palmer et al. (2020)

Continued on next page

Table 3.1 – continued from previous page

Evaluation	Participants mainly rate or assess the robot's actions or attributes, typically via questionnaires or interviews. Chi et al. (2023) ; Emaminejad et al. (2024) ; Malle and Ullman (2021)
Trust Factors	
Performance	Manipulating the robot's reliability, accuracy, or success/failure rates to influence perceived trust. Akash et al. (2018) ; Parron et al. (2023) ; Zhang et al. (2024)
Behavior	Altering strategy or autonomy in ways that do not directly affect task performance outcomes. Abd et al. (2017) ; Bridgwater et al. (2020) ; Khavas et al. (2021)
Communication	Adjusting verbal or non-verbal cues, emotional displays, or social attitudes to affect trust. Calvo-Barajas et al. (2020) ; Lee et al. (2023) ; Leichtmann et al. (2024)
Design	Changing the robot's physical appearance or interface (e.g., humanoid vs. mechanical form factor). Adamik et al. (2021) ; Bernotat et al. (2021) ; Song and Luximon (2024)
Task	Modifying the task structure itself (e.g., level of complexity, cooperative vs. competitive conditions). Akash et al. (2020) ; Azevedo-Sa et al. (2021) ; Nenna et al. (2024)
Robot Autonomy	
Not autonomous	Robot actions are mostly hypothetical, or purely descriptive, without real robot interaction. Aliasghari et al. (2023) ; Biermann et al. (2020) ; Pinney et al. (2022)
Pre-programmed	The robot follows a fixed sequence without adaptation. Calvo-Barajas et al. (2020) ; Cucciniello et al. (2021) ; Zhang et al. (2024)
Wizard of Oz	Robot actions are directly controlled by a human operator, who makes all decisions about the robot's behavior. Ahmad et al. (2019) ; Graf et al. (2022) ; Miller et al. (2021)
Shared control (fixed)	Robot operates independently with fixed rules in response to the human. Alhaji et al. (2024) ; Chauhan et al. (2024) ; Menolotto et al. (2024)
Shared control (adap.)	The robot operates independently and adapts its behavior based on real-time interaction. Che et al. (2020) ; Lagomarsino et al. (2023) ; Sun et al. (2020)
Fully autonomous	Robot operates without direct human assistance, and with full or limited adaptability. Babel et al. (2022) ; Correia et al. (2018) ; Lingg and Demiris (2023a)

3.3 RESULTS

In this section, we present findings from our examination of empirical studies on trust in [HRI](#). Our review revealed notable patterns across key methodological and experimental dimensions, including experimental context, robot characteristics, interaction intensity, trust factor manipulations, and statistical methods. We organize the findings below according to these dimensions, with all data, figures, and tables reflecting the complete dataset.

3.3.1 Experimental Context and Setup

The distribution of studies across experimental environments shows a balanced spread, with simulation-based research being the most common ($n=271$, $\approx 39\%$), followed by real-world experiments ($n=221$, $\approx 32\%$) and media-based studies ($n=197$, $\approx 29\%$). As illustrated in Figure 3.1, sample sizes varied considerably across these environments: media-based studies averaged 219 participants, simulation-based studies averaged 70 participants, and real-world studies had the lowest average with 41 participants.

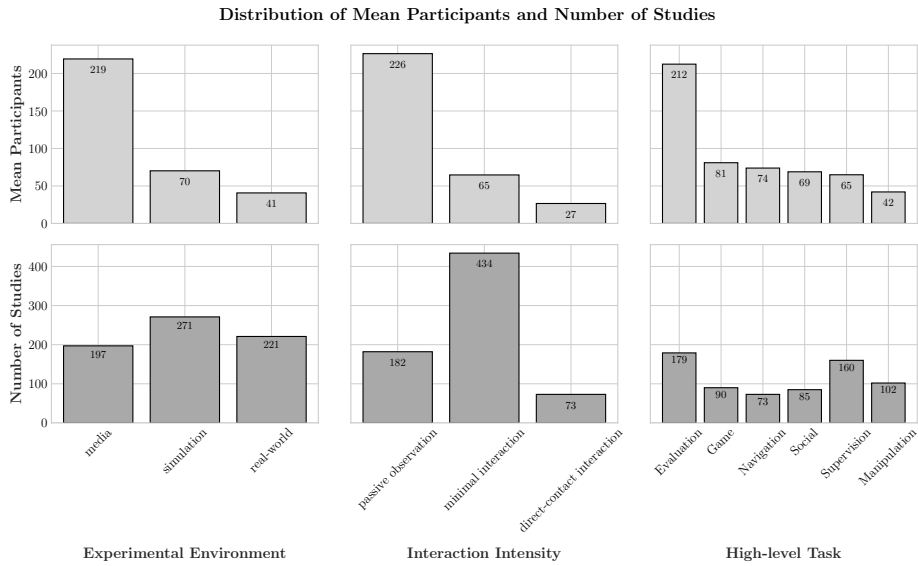


Figure 3.1: Distribution of mean participants and studies across three key dimensions in [HRI](#) trust research. The left panel shows participant counts across experimental environments (media, simulation, real-world). The middle panel shows interaction intensity levels (passive observation, minimal interaction, and direct-contact), revealing minimal interaction’s dominance. The right panel shows the distribution across task types (Evaluation, Game, Navigation, Social, Supervision, and Manipulation), showcasing the diversity of research contexts.

A similar pattern of decreasing participant numbers with increasing interaction complexity emerged when considering interaction intensity. While passive observation studies averaged 226 participants, minimal interaction studies averaged 65 participants, and direct-contact studies averaged only 27 participants. The distribution of study counts was notably skewed, with minimal interaction dominating the literature ($n=434$, $\approx 63\%$), followed by passive observation ($n=182$, $\approx 26\%$), and direct-contact studies being the least common ($n=73$, $\approx 11\%$).

Turning to the nature of the tasks, evaluation tasks remained the most frequently studied ($n=179$) and featured the highest mean participant count (212). Supervision tasks accounted for $n=160$ studies with an average of 65 participants each. Game-based studies, though fewer in total ($n=90$), showed a higher mean participant count (81), suggesting that structured, rule-based tasks are often conducted with relatively large sample sizes. Navigation, social interaction, and manipulation tasks featured a moderate range of mean participants (74, 69, and 42, respectively), with the number of studies being $n=73$ for navigation, $n=85$ for social, and $n=102$ for manipulation.

Table 3.2 provides a more granular view of these patterns, revealing notable interactions between task types and interaction intensity. Passive observation studies showed particularly high participant numbers for evaluation (245), game (380), and navigation (301) tasks while maintaining moderate numbers

Type	Eval	Game	Nav	Social	Sup	Manip
Mean Participants						
Passive	245	380	301	127	128	98
Minimal	98	70	59	62	59	53
Direct	—	24	27	33	—	26
Number of Studies						
Passive	139	4	5	12	13	9
Minimal	40	80	64	66	147	37
Direct	—	6	4	7	—	56

Table 3.2: Distribution of valid participants and study counts across high-level task domains (Eval: Evaluation, Game: Games, Nav: Navigation, Social: Social Interaction, Sup: Supervision, Manip: Manipulation). Passive interaction studies have the highest participant counts, particularly in Evaluation, Games and Navigation, whereas minimal interaction studies are more frequent, notably in Games and Supervision. Direct interaction studies are the least frequent and involve the fewest participants.

(98–128) across the remaining tasks. Minimal interaction studies demonstrated more consistent participant numbers across all task types (59–98), alongside higher study counts in supervision (147). Direct-contact studies, while having the lowest participant numbers (24–33), were more frequent in manipulation tasks (56 studies), reflecting the necessity of physical interaction.

3.3.2 Robot Characteristics

Analysis of experimental robot autonomy levels highlights how varying degrees of autonomy are deployed in trust research. As shown in Figure 3.2, pre-programmed (non-adaptive) robots were most prevalent ($n \approx 258$), followed by shared control (fixed rules) ($n \approx 179$) and wizard of oz setups ($n \approx 110$). The remaining categories - not autonomous ($n = 40$), shared control (adaptive) ($n \approx 45$), and fully autonomous ($n \approx 39$) - were less commonly employed. This distribution underscores the continued reliance on partial or pre-programmed autonomy and direct human control in HRI studies investigating trust.

Figure 3.3 further illustrates the distribution of specific robot platforms used across studies, with the 15 most frequently adopted robots highlighted. Humanoid platforms such as Nao (≈ 50 studies) and Pepper (≈ 40 studies) were especially popular in social interaction and evaluation contexts. Industrial arms (e.g. UR10, Franka Emika Panda, Baxter) were more common in manipulation tasks involving direct or minimal human-robot interaction.

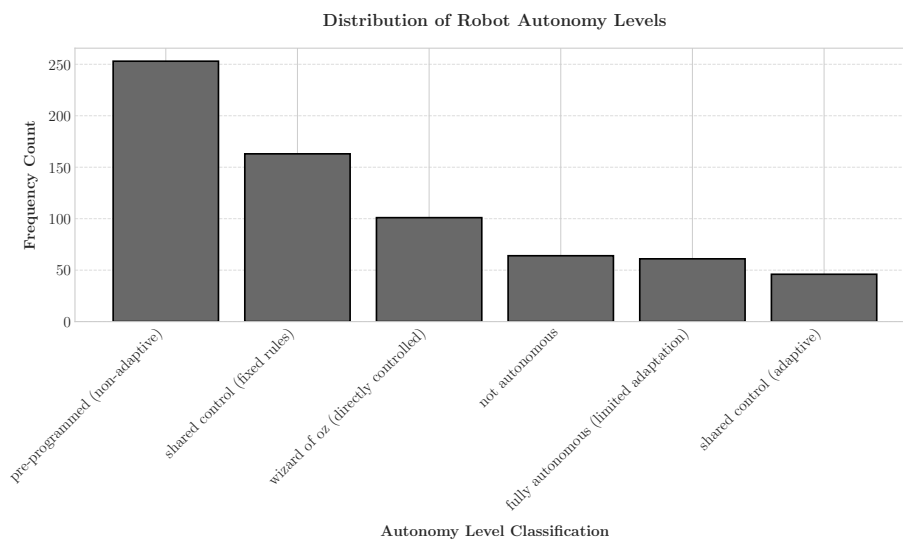


Figure 3.2: Distribution of robot autonomy levels across the analyzed HRI studies, showing a prevalence of pre-programmed (non-adaptive) and shared control (fixed rules) approaches, while wizard of oz, fully autonomous, and adaptive systems are less frequently used.

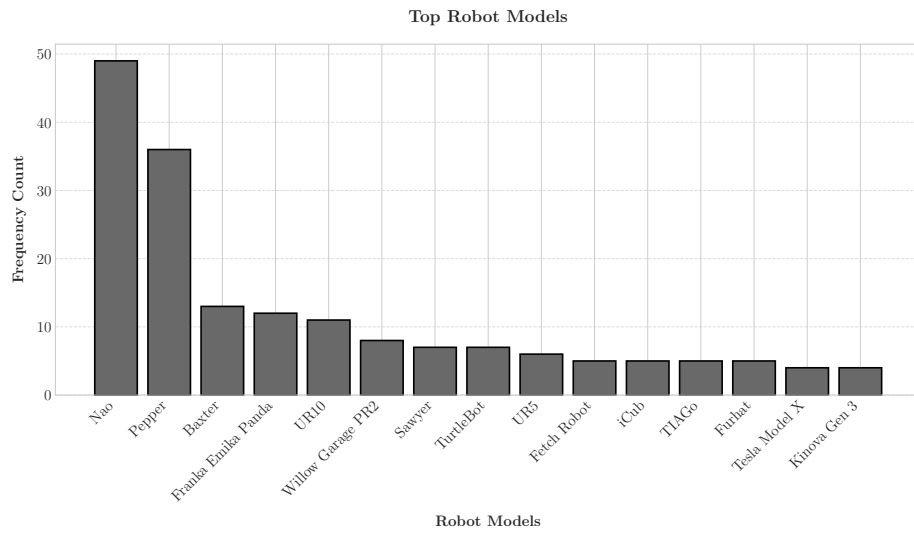


Figure 3.3: Frequency distribution of robot models used in [HRI](#) trust studies. The bar chart presents the 15 most commonly used robot platforms, showing the distribution of different robot types across the reviewed studies.

Many studies, especially those relying on surveys and questionnaires, described robots in more general and nonspecific terms without specifying a particular platform, while others employed unique, custom-built robots.

3.3.3 Trust Factor Manipulations

Analysis of trust factors across the studies suggests some interesting patterns in both how frequently different factors were studied and their effectiveness in influencing trust. Performance appeared most frequently in the literature (approximately 250 studies) and showed particularly strong results, with 231 studies reporting significant effects and only 17 indicating no impact. Communication was studied almost as often (around 248 studies) and proved similarly influential, with 198 studies showing trust effects compared to 49 without impact. Studies examining behavior (187 total) revealed more varied outcomes, with 138 showing positive impacts and 48 reporting no effects. Task-oriented manipulations demonstrated moderate effectiveness, explored in about 160 studies, with 79 reporting significant effects and 80 showing no impact. Finally, design elements, studied least frequently (85 studies), showed mixed results, with 58 reporting positive impacts and 28 indicating no effects. As illustrated in Figure 3.4, these patterns suggest that while multiple factors influence trust, performance and communication emerge as particularly consistent contributors across the literature.

To explore the role of these factors in greater depth, we analyzed their effectiveness across different research contexts using heat maps (Tables 3.3,

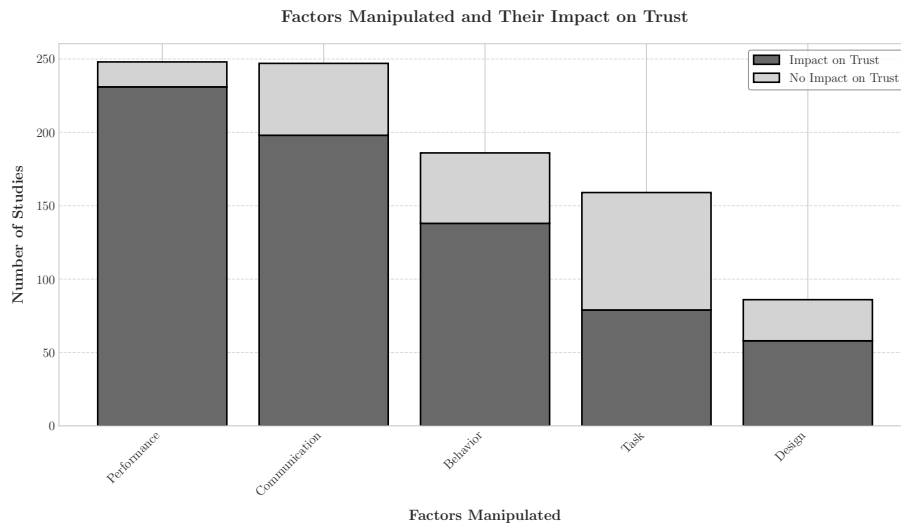


Figure 3.4: Distribution of studies showing significant impact versus no impact on trust across different factor categories. Performance and communication factors show the highest proportion of significant effects, while design factors show more mixed results.

3.4, and 3.5). These heat maps visualize the proportion of studies where a particular factor successfully influenced trust, calculated as the ratio of studies reporting significant trust impacts to the total number of studies manipulating that factor. Values range from 0 to 1, where higher values (closer to 1) indicate the factor reliably influenced trust in most studies, while lower values (closer to 0) suggest it often failed to affect trust even when explicitly manipulated. For example, a value of 0.90 means that 90% of studies manipulating that factor found significant effects on trust, while a value of 0.30 indicates only 30% of studies found meaningful impacts.

Table 3.3 examines factor effectiveness across different experimental environments (media, real-world, and simulation). Performance demonstrates remarkably consistent high effectiveness (0.86-0.91) across all environments, establishing it as a robust driver of trust regardless of context. Communication shows strong but slightly more varied effects (0.75-0.82), with its highest impact in simulation studies (0.82) and lowest in real-world settings (0.75). Notably, task-related factors show a clear environmental dependency, with substantially higher effectiveness in media-based studies (0.67) compared to simulation (0.31) or real-world settings (0.34).

When analyzing interaction intensity (Table 3.4), we find interesting variations in factor effectiveness. While performance maintains high impact across all interaction levels (0.83-0.91), behavior is notably more influential in direct-contact scenarios (0.79) compared to passive observation (0.55). Conversely, design factors show stronger effects in passive observation settings (0.75) than

Table 3.3: Distribution showing the relationship between trust factors and experimental environments (media, real-world, simulation). Darker colors indicate stronger relationships. Performance factors show consistently high influence across all environments, while task-related factors show greater variability.

Environment	Performance	Behavior	Communication	Design	Task
Media	0.91	0.58	0.80	0.68	0.67
Real-world	0.86	0.70	0.75	0.70	0.34
Simulation	0.91	0.71	0.82	0.55	0.31

Table 3.4: Distribution displaying the relationship between trust factors and interaction intensity levels. Performance and communication factors maintain high influence across all interaction types, while behavioral factors show more variation.

Factor	Performance	Behavior	Communication	Design	Task
Direct	0.83	0.79	0.64	0.71	0.24
Minimal	0.91	0.69	0.81	0.58	0.33
Passive	0.90	0.55	0.78	0.75	0.69

Table 3.5: Distribution illustrating the relationship between trust factors and high-level task types. The data reveals task-specific patterns in how different factors influence trust, with evaluation tasks showing the most consistent high-value relationships across factors.

Factor	Performance	Behavior	Communication	Design	Task
Evaluation	0.94	0.50	0.79	0.73	0.53
Game	0.94	0.81	0.74	0.55	0.46
Manipulation	0.86	0.72	0.76	0.67	0.39
Navigation	0.91	0.77	0.75	0.64	0.21
Social	0.92	0.70	0.81	0.50	0.28
Supervision	0.86	0.65	0.84	0.57	0.33

minimal interaction contexts (0.58), suggesting visual design elements are more influential when participants aren't actively engaging with the robot.

The task-specific analysis (Table 3.5) reveals nuanced patterns across different contexts. Performance consistently shows the highest effectiveness (0.86-0.94) across all tasks, but is particularly strong in evaluation (0.94) and game-based scenarios (0.94). Communication demonstrates notably high effectiveness in supervision tasks (0.84), possibly reflecting the importance of clear information exchange in oversight situations. Interestingly, task-related manipulations show relatively low effectiveness across most categories (0.21-0.53), with their strongest impact appearing in evaluation contexts (0.53).

3.3.4 Statistical Analysis Methods

An analysis of the statistical approaches used across these studies highlights clear methodological patterns. As depicted in Figure 3.5, Analysis of Variance (ANOVA) remained the most common, appearing in about 300 studies, followed by t-tests (≈ 200 studies) and Pearson correlation (≈ 90 studies). These three methods collectively represent the core of statistical analyzes in contemporary HRI trust research. In addition to these dominant methods, non-parametric techniques such as Mann-Whitney U and Chi-squared tests were each used in multiple studies, contributing to the diversity of analytical approaches. Less common but still noteworthy methods (e.g., Bayesian inference, structural equation modeling, and various machine learning-based analyzes) each appeared in a small subset of papers ($< 10\%$ overall).

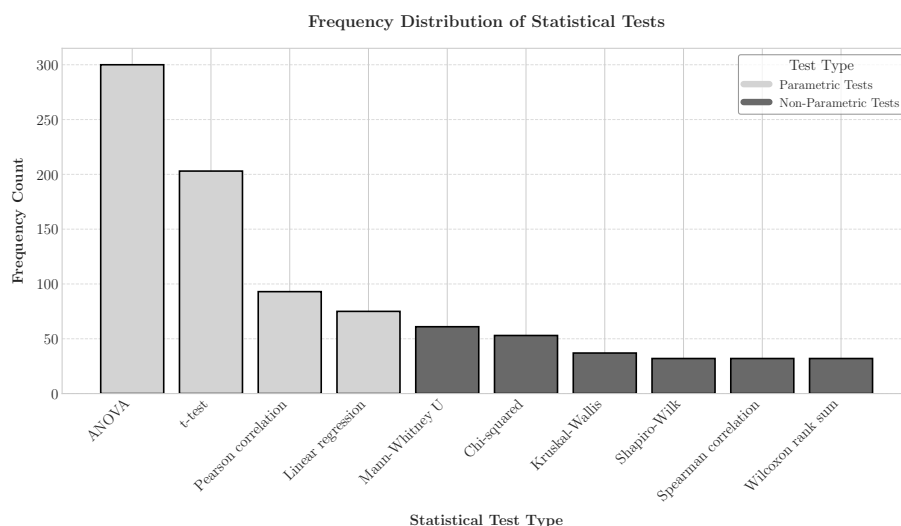


Figure 3.5: Frequency distribution of statistical tests employed in HRI trust research. The histogram shows the relative usage of different statistical methods, separated into parametric tests (light gray) and non-parametric tests (dark gray), illustrating the methodological patterns in the field.

3.4 DISCUSSION

This review of empirical studies on trust in [HRI](#) reflects the field's ongoing challenge in balancing rigorous experimental control with the authenticity of real-world contexts. Many current approaches rely on simulation-based or media-based designs, often involving passive or minimal interaction through surveys and questionnaires, which allow for greater analytical precision and larger participant pools. However, they may underrepresent the complexities of real-world human-robot trust, particularly in scenarios requiring physical or prolonged interactions. One notable observation is the distinct discrepancy in sample sizes across different interaction formats: media-based and simulation studies often recruit hundreds of participants, while real-world, direct-contact settings tend to have a significantly lower number of participants. This disparity underscores how the logistical and ethical complexities of closer-proximity interaction complicate collecting robust datasets, potentially imbalancing our empirical understanding of how trust evolves during more immersive and demanding tasks.

It's also noticeable that researchers tend to use certain common robot platforms for social interaction studies, with platforms like Nao or Pepper being frequently selected. This preference can often be attributed to their availability and relative ease of programming for social tasks, which in turn helps provide consistency across experiments. However, a focus on these specific robots can have a downside, as they are often limited in their capabilities beyond social applications. This can hinder the exploration of more advanced robotics required for real-world [HRI](#) scenarios like navigation and manipulation. Furthermore, as our analysis across different task types reveals (Section 3.3, Table 3.5), the factors influencing trust appear distinct depending on the interaction's nature. For instance, the elements shaping trust in a primarily social context likely differ from those in a collaborative, task-focused scenario like manipulation or navigation. This suggests that studies where people just watch a robot or interact with it briefly are likely to miss out on how trust really builds over time, when people use robots for longer periods. Our analysis demonstrate that while features like communication style and design influence trust in the interaction, their impact is ultimately dependent on the robot's basic ability to demonstrate reliable performance. In situations where performance is inadequate, any positive influence from communication or design is quickly undermined by user frustration or safety concerns.

Crucially, our analysis demonstrates a robust and consistent pattern: robot performance is the most consistently influential factor in building trust across diverse environments, interaction styles, and tasks. This finding aligns with previous meta-analyses, such as that by [Hancock et al. \(2011\)](#), reinforcing the idea that the foundation of trust in HRI is significantly rooted in demonstrated competence and reliability, not merely perceived social cues or likeability. The engineering challenges associated with achieving robust, reliable robot performance in a multitude of tasks, therefore, represent an area where progress could greatly further the field of trust in robotics.

Overall, the current landscape suggests that while controlled, isolated studies remain valuable, there is a growing need to examine trust in more multifaceted scenarios, such as navigation and manipulation tasks where robots can operate with greater autonomy. Expanding research in these directions may help address the “realism gap” by better reflecting the nuanced ways users form, maintain, and lose trust as they engage more deeply with robots in everyday conditions.

3.5 CONCLUSION

This review of empirical studies highlights key findings in trust research for HRI, with robot performance consistently identified as the strongest factor shaping trust. While controlled studies have provided valuable insights, their reliance on simulation-based and passive methods underscores the need for research in real-world, autonomous settings where trust develops through actual HRIs.

The next chapters build on these observations to address these gaps. Chapter 4 introduces our experimental platform and methodology, featuring an autonomous wheelchair system operating in an office-like environment. Chapter 5 then investigates how robot performance affects trust in these scenarios. Chapter 6 presents Trusty, a handheld device designed for continuously measuring trust during physical interaction. Chapter 7 follows with a multi-modal transformer framework for real-time trust prediction using diverse data streams. Finally, Chapter 8 showcases AMIGA, a custom-built mobile manipulation platform, and introduces a trust-conditioned imitation learning approach (Trust-ACT).

WHEELCHAIR EXPERIMENTATION SETUP: PLATFORM AND METHODOLOGY

Building on the insights we discussed in Chapter 3, which highlighted the need for more real-world studies of trust in Human-Robot Interaction (HRI), we now present the experimental portion of this thesis (see Figure 1.2). In this chapter, we introduce the autonomous wheelchair platform and methodology that form the foundation for the subsequent chapters. We outline a detailed experimental approach that integrates traditional questionnaire-based assessments with continuous trust measurements and multi-modal sensor data. This methodology allows us to address some of the limitations of passive or simulation-based studies discussed in our review, providing a valuable dataset that supports the trust measurement, modeling, and prediction work we present later in this thesis.

4.1 OVERVIEW OF THE EXPERIMENTAL APPROACH

Our experimental framework centers on three key elements identified as critical for trust research: systematic manipulation of robot performance, continuous trust measurement, and comprehensive data collection across multiple modalities. Our platform combines an autonomous wheelchair navigating through an office environment with an array of measurement tools, including eye-tracking, physiological sensors, and the custom-designed Trusty device for continuous trust measurements. This setup enables us to examine trust formation during human-robot interactions, with participants actively navigating alongside the robot and engaging with its behaviors rather than relying solely on remote observation or control. While this dynamic interaction better reflects real-world assistive robotics applications, it also presents unique challenges for data collection and analysis — challenges we address through integrated sensor systems and synchronized data streams.

4.2 AUTONOMOUS WHEELCHAIR PLATFORM

4.2.1 *Hardware Components*

The experimental platform is based on the ARTA smart wheelchair platform (Carlson and Demiris, 2012), an electric-powered wheelchair enhanced with additional sensors and computational capabilities to operate autonomously. The key hardware components included a control interface, sensors, and a computing unit.

An Arduino Uno microcontroller served as the control interface between the wheelchair's joystick and a laptop, allowing interception and injection of joystick inputs. This setup enabled seamless switching between manual and autonomous modes, providing participants with the option to intervene if necessary.

The wheelchair is equipped with several sensors to facilitate autonomous navigation and data collection. Two front-mounted Hokuyo URG-04LX-UG01 scanning laser rangefinders provided two-dimensional range data for obstacle detection, while a rear SICK LMS200 laser rangefinder monitored obstacles behind the wheelchair. A Phidgets spatial 3/3/3 Inertial Measurement Unit (IMU) provided orientation and acceleration data essential for localization and navigation tasks.

The computing unit, a laptop running Robot Operating System (ROS), handled sensor processing, navigation algorithms, and device communication. ROS was chosen for its modularity and scalability, enabling integration of software components for autonomous operation.

4.2.2 *Software Architecture*

The wheelchair's software architecture was built upon ROS, leveraging its capabilities for modular integration and communication between different components. The ROS navigation stack facilitated real-time path planning and obstacle avoidance, utilizing both global and local planners to generate safe and efficient trajectories.

Localization: We employed Adaptive Monte Carlo Localization (AMCL) for estimating the wheelchair's pose within a pre-mapped environment, effectively handling uncertainties in sensor data and robot motion (Fox et al., 1999). The AMCL algorithm uses a particle filter to represent the robot's belief about its position, updating this belief based on sensor observations and motion models.

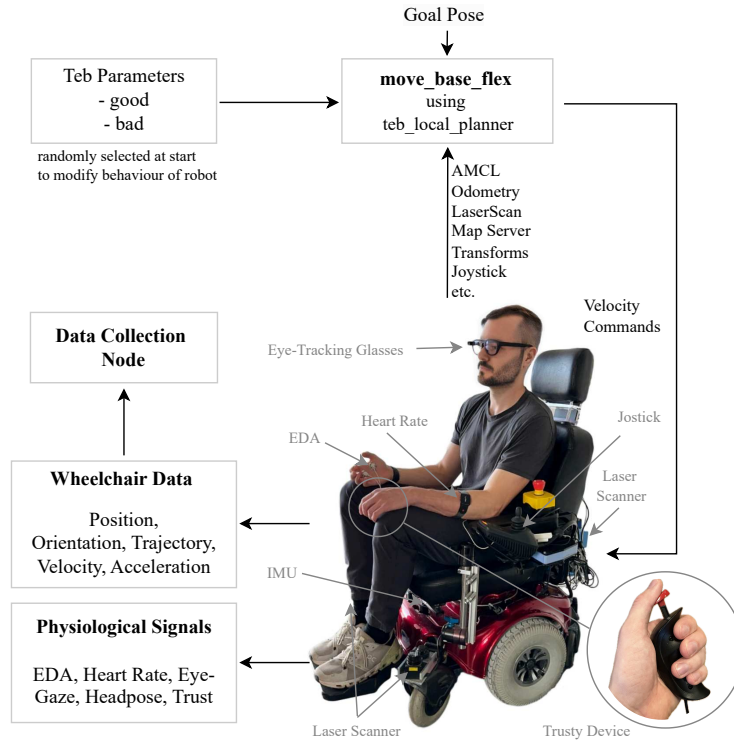


Figure 4.1: High-level architecture of the autonomous wheelchair system, showing sensor inputs, control interfaces, and data collection modules.

Path Planning: For local trajectory optimization, we utilized the Timed Elastic Band (TEB) local planner (Roesmann et al., 2012, 2017). The TEB planner generates feasible trajectories that consider dynamic constraints and obstacles, optimizing the path with respect to time and safety. By adjusting parameters within the TEB planner, we created different navigation behaviors to study their impact on trust, as detailed in Section 4.3.

All devices and sensors were synchronized using a private Network Time Protocol server, ensuring consistent timestamps across data streams.

4.3 PERFORMANCE MODULATION

To investigate the effect of the robot’s performance on user trust, we developed two distinct profiles: *good* and *bad* performance. In the good condition, the wheelchair navigated smoothly and efficiently, closely following planned paths. In contrast, the bad condition featured erratic behavior, including abrupt stops and swerves, while maintaining safety by avoiding collisions.¹

¹ A video demonstration comparing the wheelchair’s navigation behavior for both performance profiles during the first checkpoint segment can be viewed at: <https://youtu.be/kGKKXGT33QU>

4.3.1 Parameter Tuning

Specific parameters in the [TEB](#) local planner were adjusted to achieve these performance profiles. The `weight_obstacle` parameter, which influences the robot's behavior when near obstacles, was tuned to alter the robot's responsiveness to nearby objects. Velocity limits (`max_vel_x` for linear speed and `max_vel_theta` for angular speed) were modified to control the wheelchair's movement speeds. The `min_obstacle_dist` parameter set the minimum allowed distance to obstacles, affecting how cautiously the robot navigated around objects. Additionally, the `global_plan_viapoint_sep` parameter, which affects how closely the robot follows the global plan, was adjusted to induce deviations from the optimal path in the bad performance condition.

We empirically determined these parameters through iterative testing and fixed them for all participants to ensure consistency across trials. By systematically manipulating these parameters, we created controlled variations in the robot's performance, enabling us to study their impact on user trust.

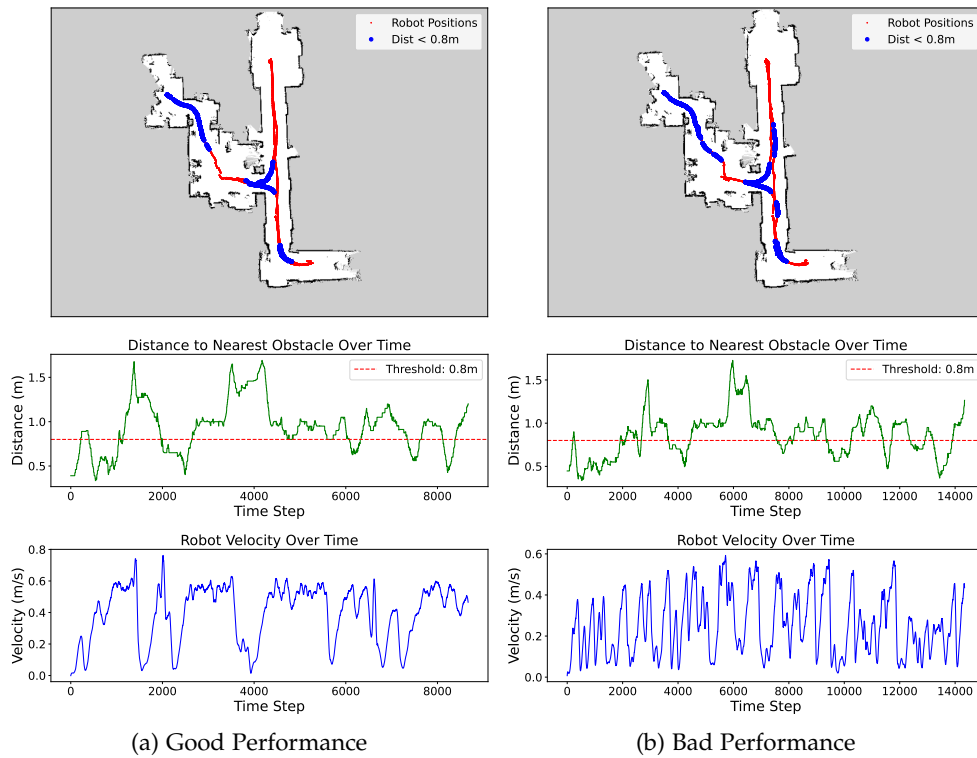


Figure 4.2: Wheelchair performance comparison: (a) Good performance – smoother velocity, consistent obstacle distance. (b) Bad performance – erratic velocity, fluctuating obstacle distance.

4.3.2 Performance Validation

We validated the performance profiles by analyzing velocity smoothness and obstacle distance consistency. Figure 4.2 compares the wheelchair's behavior under the two profiles. In the good performance condition, the wheelchair maintained a smooth velocity profile and consistent distances to obstacles. In the bad performance condition, the wheelchair exhibited fluctuations in velocity and varying distances to obstacles, reflecting less predictable and less comfortable navigation behavior.

4.4 TRUSTY: CONTINUOUS TRUST MEASUREMENT DEVICE

The Trusty device, illustrated in Fig. 4.3, was developed to capture real-time trust fluctuations. This handheld controller is based on a modified slot-car controller and includes a $10k\Omega$ linear slide potentiometer connected to an Arduino Nano microcontroller. The Arduino reads analog signals from the potentiometer and publishes them to ROS using the roserial package.

Participants adjusted the slider to reflect their trust level, ranging from low to high, without requiring visual attention to the device. This design minimized interference with participants' natural interaction with the robot and allowed continuous trust measurement throughout the experiment.



Figure 4.3: The Trusty device: (a) A modified slot-car controller repurposed for trust measurement. (b) Example usage - pressing down indicates higher trust, allowing continuous feedback during navigation without visual attention.

Before the main experiment, participants completed a brief training session with Trusty to familiarize themselves with its operation. They practiced adjusting the slider, ensuring they understood how to indicate their trust levels continuously.

At each checkpoint, we compared participants' Trusty measurements with their verbal trust ratings to validate the device's effectiveness. This dual measurement approach provided a basis for analyzing the relationship between discrete and continuous trust measurements, which we explore in detail in Chapter 6.

4.5 MEASURES

To assess trust within our experimental framework, we integrated both subjective assessments and objective measurements. This approach allowed us to capture not only participants' self-reported trust levels but also their physiological responses throughout the interaction.

4.5.1 *Subjective Measures*

Participants completed established trust questionnaires before and after the experiment to assess their initial trust levels and any changes resulting from the interaction. The questionnaires included the Trust in Automated Systems Test (TOAST) (Wojton et al., 2020), Trust in Automation (TiA) (Körber, 2019), and the Trust Perception Scale (TPS) (Schaefer, 2016). These instruments are widely used in HRI research and assess various dimensions of trust, including reliability, predictability, and perceived understanding of the robot.

At predefined checkpoints during the navigation task, participants also provided verbal trust ratings on a scale from 1 to 7. These ratings offered immediate insights into participants' trust levels at specific moments during the interaction. Trust levels were also continuously recorded throughout the experiment using the Trusty device, described in Section 4.4. The continuous trust measurement allowed us to capture fluctuations in trust in response to the robot's behavior in real-time.

4.5.2 *Objective Measures*

Physiological data were collected to examine participants' physiological responses to the robot's behavior. Heart Rate (HR) was measured at 1 Hz

Table 4.1: Overview of multi-modal data collection streams in the wheelchair trust study, including sampling rates, data types, and their characteristics.

Sensor/Device	Sampling Rate	Data Type	Purpose/Usage
HR	1 Hz	Continuous physiological data	Assess physiological changes associated with robot actions
EDA	60 Hz	Continuous physiological data	Assess stress and emotional arousal levels
Eye Gaze	250 Hz	2D gaze coordinates	Examine attention patterns and focus areas
Head Pose	200 Hz	3D orientation angles	Analyze head movement and alignment with robot actions
Egocentric Video	60 Hz	Sequential image frames	Provide visual context for navigation events
Trusty Device	60 Hz	Continuous trust levels	Capture real-time fluctuations in trust
Verbal Ratings	At checkpoints	Discrete ratings (1–7 scale)	Validate continuous trust measures at specific points
Wheelchair Telemetry	60 Hz	Position, orientation, velocity, acceleration	Monitor robot navigation dynamics and state
Trust Questionnaires	Pre/Post	Categorical survey responses	Establish baseline and final trust levels

using a Polar Verity Sense device, providing data on cardiovascular responses. Electrodermal Activity (EDA) was measured at 60 Hz using a BITalino device (Da Silva et al., 2014), offering insights into sympathetic nervous system activity associated with stress and arousal. All data streams were time-stamped and synchronized via ROS, ensuring accurate alignment of events across different modalities for subsequent analysis. Table 4.1 provides an overview of all data collection streams, sampling rates, formats, and purposes.

Eye-tracking data were collected using Pupil Labs eye-tracking glasses (Tonsen et al., 2020), capturing gaze direction at 250 Hz, blink and fixation events, and head pose data (roll, pitch, yaw) at 200 Hz. This data provided information on where participants directed their attention and how they physically responded to the robot’s movements.

Wheelchair telemetry data were recorded, including joystick inputs to monitor manual control interventions, position and orientation data from the localization algorithms, velocity and acceleration data for movement dynamics, and distances to obstacles measured using LIDAR sensors.

All data streams were time-stamped and synchronized via ROS, ensuring alignment of events across different modalities for subsequent analysis.

4.6 EXPERIMENTAL PROCEDURE

4.6.1 Environment Setup

We conducted the experiments in an unstructured office environment featuring typical obstacles such as desks, chairs, and narrow passages. We specifically chose the environment to include various navigation challenges, including open spaces, corridors of different widths, and doorways. Figure 4.4 shows both the environment map with planned trajectories and corresponding real-world views at each checkpoint. ²

The navigation course consisted of five segments between checkpoints ($Start \rightarrow A \rightarrow B \rightarrow C \rightarrow D \rightarrow End$), each presenting different challenges. For example, the initial segment required navigation through an open office space with furniture, while later segments involved traversing corridors of varying widths and negotiating doorways. This variety of scenarios was essential for studying trust development across different navigation contexts.

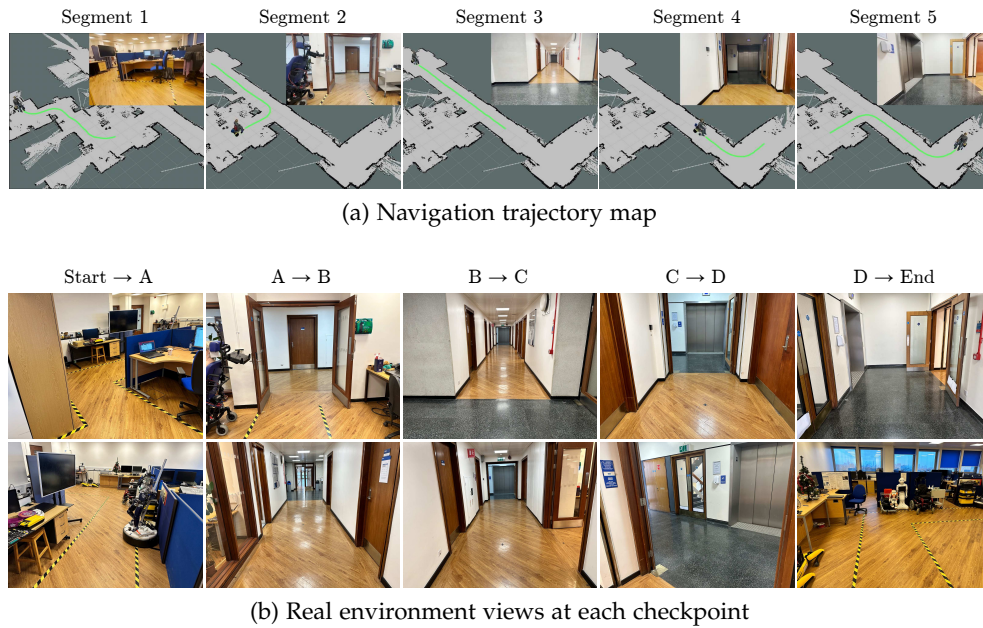


Figure 4.4: The experimental environment: (a) Map showing five navigation segments with the wheelchair's planned trajectories (green) overlaid on the occupancy grid map. (b) Real-world views showing the approach (top) and arrival (bottom) at each checkpoint. The course progresses through distinct areas including open spaces, corridors, and doorways, representing common navigation challenges in assistive robotics applications.

² A video demonstration of the autonomous wheelchair system navigating through this environment can be viewed at: <https://youtu.be/b6Q6ef0KtnU>

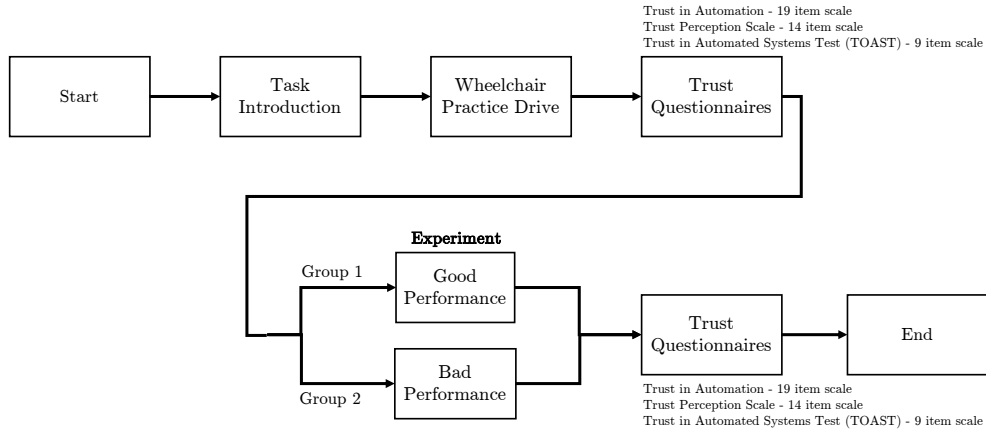


Figure 4.5: Diagram of the experiment protocol detailing participant interaction with the autonomous wheelchair. Steps include task introduction, optional wheelchair test drive, and trust measurement before and after the experiment. Two robot performance conditions are tested as in-between study. Duration: 30-45 minutes.

4.6.2 Experimental Steps

We designed the experimental procedure to provide a consistent experience for all participants while capturing the necessary data for analysis. A high-level overview of the procedure is illustrated in Fig. 4.5. The steps were as follows:

1. **Informed Consent and Introduction:** Participants provided written consent and were briefed on the study objectives and procedures. They were informed that the purpose was to investigate interactions with an autonomous wheelchair, without revealing the specific focus on trust.
2. **Pre-Experiment Questionnaires:** Participants completed the trust questionnaires to assess their initial trust levels. This baseline measurement was essential for evaluating changes resulting from the interaction.
3. **Introduction to the Wheelchair and Trusty:** Participants received instructions on operating the wheelchair and using the Trusty device. They were informed that they could take manual control at any time, but that the wheelchair would move 60% slower in manual mode, encouraging reliance on autonomous navigation.
4. **Familiarization:** Participants were offered an optional test drive to familiarize themselves with manual control of the wheelchair and to become comfortable with the Trusty device.

5. **Sensor Setup:** Participants were fitted with physiological sensors and eye-tracking glasses. Care was taken to ensure that the equipment was comfortable and did not hinder movement.
6. **Performance Profile Assignment:** Using a between-subjects design, a performance profile (good or bad) was randomly assigned to the wheelchair, unknown to the participants.
7. **Navigation Task:**
 - a) Participants navigated the wheelchair to five predefined checkpoints autonomously. The route included various challenges such as narrow passages and turns.
 - b) At each checkpoint, participants provided verbal trust ratings on a scale from 1 to 7.
 - c) Participants could switch to manual control at any time, though the speed reduction served as a deterrent.
 - d) The Trusty device continuously recorded trust levels throughout the task.
8. **Post-Experiment Questionnaires:** Upon completion of the navigation task, participants filled out the trust questionnaires again to assess any changes in their trust levels.
9. **Debriefing:** Participants were informed about the purpose of the study and the performance manipulations used. They were given the opportunity to ask questions and provide feedback on their experience.

The entire experiment, including setup, navigation task, and questionnaires, lasted approximately 45 minutes per participant.

4.7 PARTICIPANTS

A total of 34 participants (17 female, 16 male, 1 undisclosed), aged between 22 and 45 years (mean age: 27.6 years), were recruited from the university community. Participants were nearly equally split between those with prior robotics experience (47%) and those without (53%), as determined by their history of programming or building robots. This balance was intended to mitigate potential biases in trust perception, as prior familiarity with robots can influence initial trust levels.

4.8 ETHICAL CONSIDERATIONS

This study was conducted in accordance with ethical guidelines for research involving human participants. All participants provided written informed consent prior to participation, ensuring they understood the study objectives, procedures, and their right to withdraw at any time without penalty.

The study received ethical approval from the Imperial College Science, Engineering and Technology Research Ethics Committee (SETREC). Participants were debriefed at the end of the study, during which the purpose and methodology, including performance manipulations, were explained in full. This ethical framework ensured that the research upheld the principles of respect, transparency, and participant well-being.

4.9 CONCLUSION

In this chapter, we introduced our autonomous wheelchair platform and the methodology that forms the basis of the experimental work presented in the first part of this thesis (see Chapters 5–7). Our platform combines multiple data collection methods, including traditional questionnaires, verbal trust ratings, and continuous feedback from the custom Trusty device, with physiological sensors and eye-tracking. By systematically varying the robot’s performance and gathering rich, multi-modal data in a realistic, unstructured environment, we have built a robust framework for exploring trust in HRI.

The dataset generated through this approach will support several key analyses in the following chapters: examining how performance influences trust development (Chapter 5), validating the Trusty device as an effective continuous trust measurement tool (Chapter 6), and developing predictive models using multi-modal transformer architectures (Chapter 7).

BUILDING TRUST IN ASSISTIVE ROBOTICS: INSIGHTS FROM A REAL-WORLD MOBILE NAVIGATION EXPERIMENT

5.1 INTRODUCTION

The interplay between robot performance and user trust remains a central question within Human-Robot Interaction ([HRI](#)) research, especially in the domain of assistive robotics. In these settings, individuals rely on robots for essential tasks such as mobility or everyday support, making the development of trust both critical and difficult to assess. This chapter draws on findings from our published work, *Building Trust in Assistive Robotics: Insights From a Real-World Mobile Navigation Experiment*, to investigate how direct experience with an autonomous system affects user trust across multiple dimensions.

Building on the experimental methods introduced in Chapter 4, we conducted a study with participants who experienced either “good” or “bad” robot performance conditions during a navigation task. We then used validated trust questionnaires—administered before and after the interaction—to examine how these contrasting levels of performance, coupled with hands-on exposure to the robot, influence key facets of trust.

Our results reveal two primary findings. First, the quality of robot performance strongly influences participants’ trust-related perceptions and attitudes: reliable performance enhances confidence and acceptance, while poor performance undermines them. Second, direct interaction with the robot significantly improves users’ understanding of its capabilities, regardless of performance quality, highlighting the importance of hands-on engagement in building trust. These insights demonstrate that successful human-robot interaction depends on both reliable robot performance and opportunities for direct user experience, providing practical guidance for developing more effective assistive robotic systems while laying the groundwork for advanced approaches we explore in the following chapters.

5.2 RESEARCH QUESTIONS AND HYPOTHESES

Previous research indicates that trust often emerges from direct experience (Lee and See, 2004; Muir, 1994). In the context of assistive robotics, this prompts two central questions: how does the quality of robot performance influence trust, and does hands-on engagement shape how users perceive and understand a robot’s capabilities? To explore these questions, we tested the following five hypotheses:

- **H1:** Participants’ perception of the robot’s behavior will improve while interacting with a good-performing robot but deteriorate while interacting with a bad-performing wheelchair.
- **H2:** Participants’ understanding of the robot will improve regardless of the robot’s performance during the interaction.
- **H3:** Participants’ attitudes towards the robot will improve if they interact with a good-performing robot and decrease if they interact with a bad-performing robot.
- **H4:** Participants’ attitudes towards the developer will improve if they interact with a good-performing robot and decrease if they interact with a bad-performing robot.
- **H5:** Participants’ sense of familiarity with robotic systems will not be affected by the interaction with the robot, regardless of the robot’s performance.

These hypotheses examine how performance quality (H1, H3, H4) and interaction itself (H2, H5) influence different aspects of trust. Hypothesis H4 specifically probes whether users attribute the robot’s performance success or failure to its developers, rather than the robot itself.

5.3 METHODS

5.3.1 *Experimental Design*

As detailed in Chapter 4, we employed a between-subjects design where participants interacted with an autonomous wheelchair under either good or bad performance conditions. The experimental task involved package delivery to predetermined locations within an unstructured office environment, illustrated in Figure 4.4.

5.3.2 Trust Measurement Framework

This chapter focuses on the analysis of three validated trust questionnaires: the Trust in Automated Systems Test (TOAST) (Wojton et al., 2020), Trust in Automation (TiA) (Körber, 2019), and Trust Perception Scale (TPS) (Schaefer, 2016). Participants completed these questionnaires before and after interacting with the wheelchair. While additional data including continuous trust measurements and physiological signals were also collected, those will be explored in Chapters 6 and 7.

5.4 EFFECTS OF ROBOT PERFORMANCE ON TRUST

Our analysis shows that exposure to an autonomous wheelchair influences trust-related factors. Direct interaction improved participants' understanding of the robot and, for those encountering good performance, enhanced their perceptions of and attitudes toward the robot. Meanwhile, bad performance conditions negatively impacted perceptions and attitudes related to the robot's behavior. These findings establish that trust is sensitive to performance quality, consistent with the broader conceptual framework of trust in automation discussed in the thesis background.

5.4.1 Data Preparation

We categorized the trust questionnaire sub-scales into five themes corresponding to our hypotheses: the robot's behavior (i.e., the different performance profiles presented in the previous chapter), participants' understanding of the robot, participants' attitudes towards the robot, participants' attitudes towards the developer, and participants' familiarity with the robot. The groupings are displayed in Table 5.1. Following established procedures, we normalized scores on a 0–1 scale for each participant and sub-scale.

Table 5.1: Sub-scales of trust questionnaires categorized into five groups in the study.

TiA	TOAST	TPS
Predictability ①	Performance ①	Understanding ②
Reliability/Competence ①	Understanding ②	
Trust in the System ③		
Propensity to Trust ③		
Developer ④		
Familiarity ⑤		

① Robot's behavior ③ Attitude towards the robot ⑤ Familiarity
 ② Understanding of the robot ④ Attitude towards the developer

5.4.2 Data Analysis

Broadly speaking, testing our hypotheses involved asking two questions for each trust factor:

1. How do trust scores change after the interaction?
2. Does this change depend on the performance of the robot during the interaction?

We addressed the first question by conducting a paired samples t-test to compare the sub-scales scores taken before and after the interaction, as outlined in Table 5.2. We found a significant improvement in scores related to the robot's behavior, participants' understanding of the robot, and their attitude towards the robot. This finding was supported by the visual representation of the data in Figure 5.1. However, there was no significant difference in the distributions of scores related to the attitude towards the developer and familiarity before and after the interaction. To address the second research question, we analyzed the change in scores for each sub-scale before and after the interaction of each participant. We employed an independent samples t-test to examine the potential differences in the mean change of scores between those who interacted with a good-performing robot and those who interacted with a bad-performing robot. The outcomes of this statistical analysis are presented in Table 5.3. Our findings indicate that solely the factors about the robot's behavior and the participant's attitude towards the robot were affected by the experimental condition. A visual depiction of the data is shown in Figure 5.2.

Table 5.2: Assessment of trust factor changes before and after the experiment using paired samples t-test. Significant improvements were found in participants' scores for the robot's behavior, understanding of the robot, and attitude towards the robot.

Measure 1		Measure 2	t	df	p	
Reliability/Comp. B	-	Reliability/Comp. A	-1.910	26	0.067	*
Predictability B	-	Predictability A	-0.706	26	0.486	
Performance B	-	Performance A	-2.679	26	0.013	**
Understanding B	-	Understanding A	-3.742	26	< .001	**
Trust Perception B	-	Trust Perception A	-2.028	26	0.053	*
Trust in the System B	-	Trust in the System A	-2.907	26	0.007	**
Propensity to Trust B	-	Propensity to Trust A	-2.326	26	0.028	**
Developers B	-	Developers A	-0.960	26	0.346	
Familiarity B	-	Familiarity A	-0.097	26	0.924	

Note. Student's t-test.

A/B denotes After/Before, * $p < 0.1$, ** $p < 0.05$

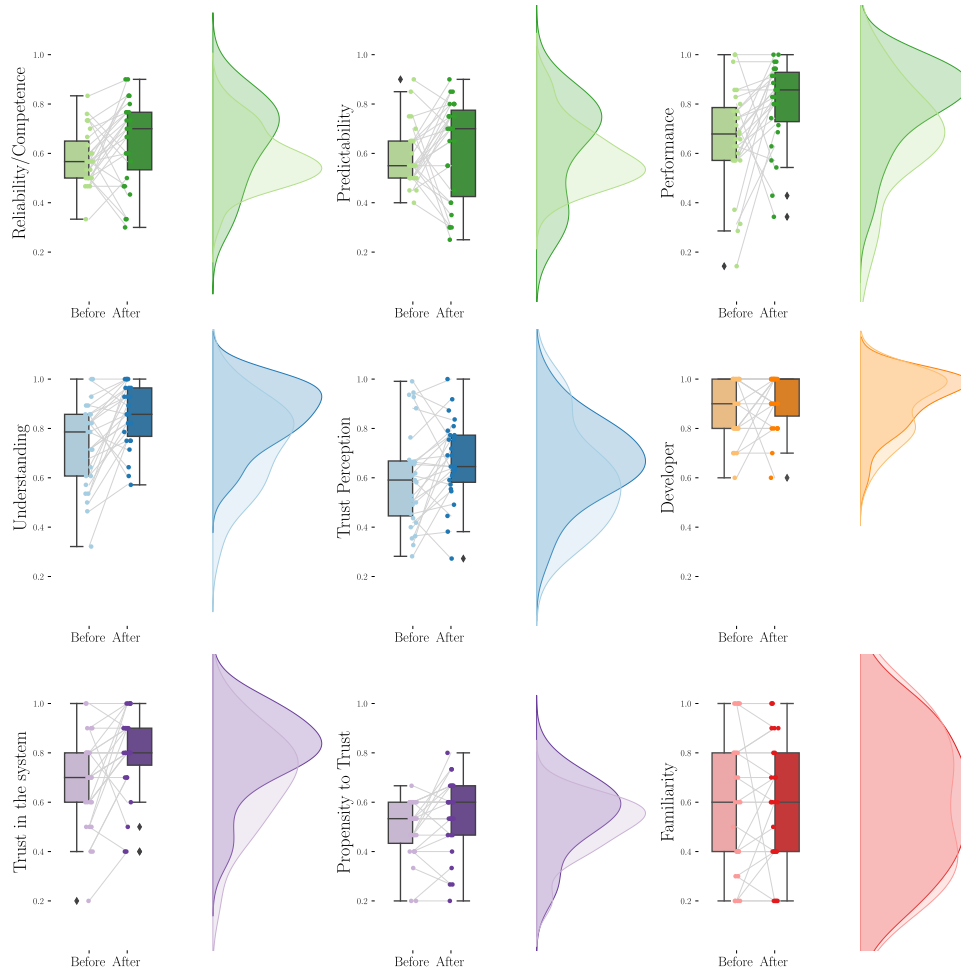


Figure 5.1: Visualizations of the results of three trust questionnaires, divided into before and after categories. Each sub-scale is shown, revealing an overall increase in all factors, except for familiarity with robots, which remained constant. These results demonstrate that direct interaction with the robot influenced multiple aspects of trust, though not uniformly across all measured dimensions.

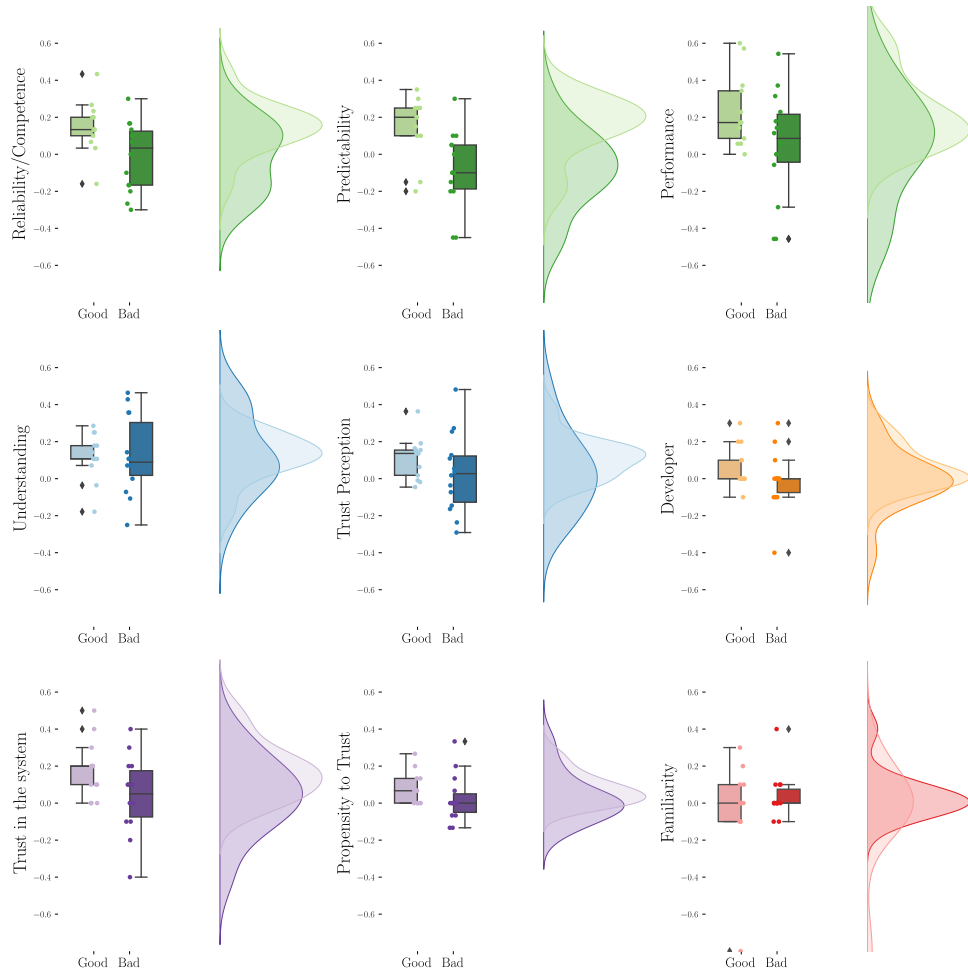


Figure 5.2: Visualizations depicting results split by good and bad-performing condition. As expected, the magnitude of score improvements for most factors was higher for those who interacted with a good-performing robot than those who interacted with a bad-performing robot.

Table 5.3: Results of independent samples t-test comparing trust factor changes for good- and bad-performing robot. The results indicate that the experimental condition is significantly affected only by the behavior of the robot and the participant's attitude towards the robot .

	t	df	p	
Reliability/Competence Change	-2.576	25	0.016	**
Predictability Change	-3.225	25	0.003	**
Performance Change	-1.837	25	0.078	*
Understanding Change	0.061	25	0.952	
Trust Perception Change	-1.202	25	0.241	
Trust in the System Change	-1.902	25	0.069	*
Propensity to Trust Change	-1.264	25	0.218	
Developers Change	-1.287	25	0.210	
Familiarity Change	0.863	25	0.396	

Note. Student's t-test.

* $p < 0.1$, ** $p < 0.05$

5.5 DISCUSSION

Our findings support H1, which states that participants' perception of the robot's behavior improved when interacting with a good-performing robot and deteriorated when interacting with a bad-performing wheelchair. Similarly, H3 was supported, with participants' attitudes towards the robot improving in the good-performing condition and decreasing in the bad-performing condition. However, H4 was not supported, as participants' attitudes towards the developer did not change in either condition. Regarding H2, we found that participants' understanding of the robot improved after the interaction, regardless of the robot's performance. This result suggests that interaction with a robot can enhance participants' understanding of the robot's capabilities and limitations. H5 was supported, and participants' sense of familiarity with robotic systems was not influenced by the interaction with the robot, regardless of the robot's performance. These findings are summarised in Table 5.4.

Our results suggest that the robot's performance during the interaction is critical in shaping participants' perceptions and attitudes towards it. The finding that participants' perception and attitude towards the robot were better in the good-performing condition is consistent with previous research

Table 5.4: Summary of Results: Illustrating the influence of interacting with a robot on five trust factors. The findings indicate whether each factor improved following the interaction or was contingent on the condition.

	Improved after Interaction	Depends on Condition
① Robot's behavior	✓	✓
② Understanding of the robot	✓	✗
③ Attitude towards the robot	✓	✓
④ Attitude towards the developer	✗	✗
⑤ Familiarity	✗	✗

on the effect of performance on trust in automation [Hancock et al. \(2011\)](#); [van den Brule et al. \(2014\)](#). This result suggests that designers of robotic systems should prioritize ensuring good performance to enhance users' perception and attitude towards the robot. Furthermore, our findings indicate that interacting with a robot can improve participants' understanding of the robot's capabilities and limitations. This result highlights the important role interactive robots can play in user education and consequent trust-building. Finally, our study found no evidence that brief interaction with a robot can increase participants' familiarity with robotic systems. Thus, more extensive exposure or training may be necessary for users to gain a general familiarity with robots.

These results also align well with how we understand trust in this thesis (Section 2.1). The wheelchair task created exactly the type of uncertain and vulnerable situation where trust becomes important - participants had to rely on the robot's navigation decisions for their safety. The changes we observed in trust scores confirm that trust is dynamic, shifting based on direct experience with the robot. Interestingly, while some aspects of trust (like behavior perception) depended heavily on performance, others (like understanding) improved regardless, supporting the idea that trust has multiple dimensions (Section 2.3). The fact that all our different trust questionnaires showed similar patterns also suggests that participants were integrating various factors into their overall trust judgments.

5.6 CONCLUSION

In conclusion, this chapter establishes two fundamental insights about trust in [HRI](#) through our autonomous wheelchair experiments. First, we demonstrate that robot performance strongly influences how users develop trust: good performance enhances confidence and acceptance, while poor performance undermines both. Second, we find that direct interaction with a robot improves users' understanding of its capabilities, regardless of performance quality. This highlights the value of studying trust through real-world interaction rather than passive observation.

Our results reveal that both the actual experience of direct robot interaction and its performance play significant roles in shaping trust. While poor performance negatively influences perception and attitudes, this experience consistently improves users' understanding of the robot's capabilities. These findings have important implications for developing autonomous assistive devices, emphasizing that successful human-robot interaction depends on both reliable performance and opportunities for direct user experience.

However, our reliance on pre- and post-interaction questionnaires reveals an important limitation: we can only capture trust at discrete moments, not throughout the interaction. This observation leads us to Chapter [6](#), where we explore complementary approaches to measuring trust during [HRI](#). By introducing more sensitive measurement techniques, we aim to more effectively capture the granular shifts in user trust, setting the stage for advanced modeling approaches we develop later in this thesis.

BEYOND SELF-REPORT: A CONTINUOUS TRUST MEASUREMENT DEVICE FOR HRI

6.1 INTRODUCTION

Measuring trust in Human-Robot Interaction (HRI) presents a significant challenge, especially in assistive robotics where trust can shift as the interaction unfolds. Traditional questionnaires offer a snapshot of trust at a particular moment, but they often miss the ongoing, gradual changes that may occur as people interact with robots. This limitation is especially critical in assistive settings, as understanding these subtle changes can offer valuable insights into user needs and expectations, ultimately guiding the creation of more intuitive and supportive robotic systems.

Based on our published work, *Beyond Self-Report: A Continuous Trust Measurement Device for HRI*, this chapter introduces and validates a novel approach to continuous trust measurement. Building on the insights from Chapter 5, where we established that trust evolves with robot exposure and performance, we present Trusty, a handheld device specifically designed to capture trust fluctuations without disrupting the natural flow of interaction.

Using the experimental platform detailed in Chapter 4, we evaluated Trusty in a user study where participants interacted with an autonomous wheelchair. Our results demonstrate strong correlations between Trusty measurements and established trust metrics, while providing richer temporal insights into how trust develops and changes during interaction. The continuous trust measurements we collect through Trusty provide valuable training data for the trust prediction model developed in Chapter 7.

6.2 TRUSTY: DESIGN AND IMPLEMENTATION

The primary design objective for Trusty was to develop an intuitive, non-intrusive device capable of capturing continuous trust signals during inter-



Figure 6.1: Three proposed continuous time trust measurement prototypes: a sliding scale (left), a rotary dial (center), and the final chosen device, a modified slot-car controller (right). The slot-car controller design allows participants to change their trust input intuitively and continuously, without the need for visual guidance.

action with a robot. In our iterative design process, we evaluated several interface options (see Figure 6.1), including a sliding scale and a rotary dial. Although these alternatives were capable of continuous input, they required users to divert visual attention to ensure precise operation. Consequently, we adopted a modified slot-car controller design, which leverages tactile feedback to allow for natural, eyes-free operation.

The final implementation of Trusty uses a $10k, \Omega$ linear slide potentiometer housed within a repurposed slot-car controller case. Users indicate their trust level by pressing the controller’s push bar - pressing further indicates higher trust, while releasing indicates lower trust. As illustrated in Figure 6.2(a), an Arduino Nano microcontroller acquires the analog signal from the potentiometer and streams the data into our Robot Operating System (ROS)-based experimental framework described in Chapter 4.

Trusty continuously records trust measurements throughout the interaction session, allowing for a more detailed analysis of trust over time. Figure 6.2(b) shows typical output from Trusty, including both the raw trust signal and its smoothed variants (e.g., via rolling mean or exponential moving average). This fine temporal resolution provides an advantage over traditional questionnaire methods by capturing subtle fluctuations in trust that might otherwise go unnoticed.

Trusty’s complete source code and detailed hardware specifications, including build instructions and component lists, are available on GitHub.¹

¹ Code available at: https://github.com/NicoLingg/trusty_trust_recorder

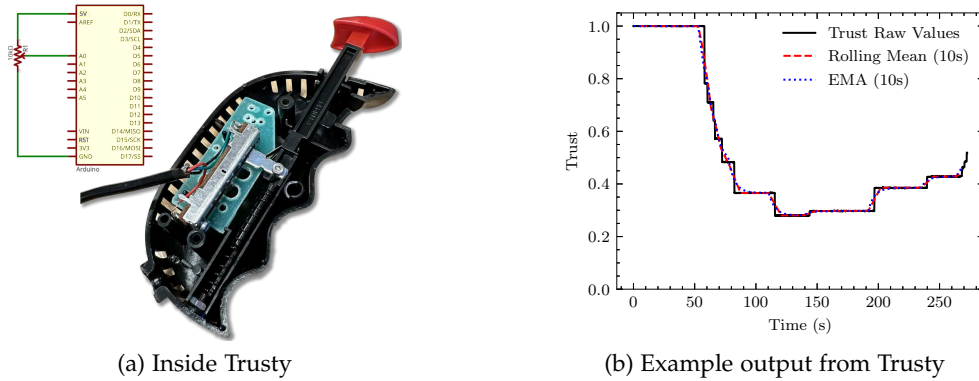


Figure 6.2: (a) Inside view of the Trusty device, showing the 10kΩ linear slide potentiometer and mechanical linkage to the push bar. The Arduino Nano reads the analog signal and sends it into ROS. (b) Example of continuous trust output recorded during an experiment. The raw trust values (solid line) are shown alongside smoothed versions (e.g., rolling mean, EMA) to illustrate how trust can be analyzed over time.

6.3 EXPERIMENTAL VALIDATION

Trusty was validated during the autonomous wheelchair experiment described in Chapter 4. In this study, participants collaborated with the wheelchair to deliver packages along a series of predetermined checkpoints. Participants continuously rated their trust using Trusty and provided additional verbal trust ratings at designated checkpoints. Moreover, participants completed established trust questionnaires (Trust in Automated Systems Test (TOAST), Trust in Automation (Tia), Trust Perception Scale (TPS)) both before and after the experiment, as detailed in Chapter 5.

6.4 OBJECTIVES AND HYPOTHESES

To assess Trusty's performance as a continuous trust measurement tool, we formulated the following hypotheses:

- **H1:** Trust scores for participants who interacted with a good-performing wheelchair will be significantly higher than those who interacted with a bad-performing wheelchair.
- **H2:** Post-experiment verbal self-report scores and Trusty measurements will show significant positive correlations with post-experiment trust questionnaire scores, whereas no significant correlations will be observed with pre-experiment trust questionnaire scores.
- **H3:** Trusty measurements will correspond strongly with concurrent verbal self-report trust scores

6.5 TRUSTY VALIDATION RESULTS

6.5.1 *Statistical Approach*

In our experiment, we collected trust-related data from multiple sources, including scores from reputable trust questionnaires ([TOAST](#), [TIA](#), [TPS](#)) administered both before and after the experiment, verbal self-report scores on a Likert scale, and continuous trust measurements from Trusty (with values ranging from 1 to 1024).

For statistical analysis, we employed non-parametric methods such as the Mann-Whitney U test and Spearman's rank correlations. We assumed that observations were independent, that the data were at least ordinal and possibly non-normally distributed, and that a monotonic relationship existed between variables. In order to maximize our sensitivity in this exploratory study, we opted not to apply multiple-comparison corrections (e.g., Bonferroni correction), thereby accepting a modest increase in the type I error rate.

In conducting our analysis, we recognize that the chosen statistical methods, while not necessarily the most robust or comprehensive, were selected for their straightforwardness and interpretability. As shown in our review of [HRI](#) trust research (Chapter 3), these methods align with the most common analytical approaches in the field. Given the exploratory nature of our study, these statistical tests serve as useful tools for illuminating potential relationships and correlations within our data. We note, however, that while these methods offer meaningful insights, they should be considered as a stepping stone towards more detailed and comprehensive analyses in future studies.

6.5.2 *Trust Measurement Validation*

To validate our experimental design, we first conducted Mann-Whitney U tests comparing post-experiment questionnaire scores of participants who interacted with a good-performing wheelchair against those who interacted with a bad-performing wheelchair (see Table 6.1). The results revealed statistically significant differences by performance condition ($p < 0.05$), particularly for the Performance component of the [TOAST](#) scale, while the Understanding component and overall scores remained consistent. Similar tests applied to Trusty data and verbal self-report scores indicated that both the continuous measurements and verbal ratings differed significantly between both performance conditions (see Table 6.2).

Table 6.1: Results of the Mann-Whitney U test assessing differences in post-questionnaire scores across experimental conditions.

	W	df	p
All After	53.000		0.023*
Trust Perception After	55.000		0.031*
TiA After	53.000		0.025*
TOAST After	80.000		0.283
TOAST After - Performance	52.000		0.022*
TOAST After - Understanding	128.500		0.313

Note. Mann-Whitney U test – * $p < 0.05$

Table 6.2: Results of the Mann-Whitney U test for comparing trust measures across different conditions, including verbal checkpoints and Trusty segments.

	W	df	p
Verbal Checkpoint 1	25.500		< .001*
Verbal Checkpoint 2	26.000		< .001*
Verbal Checkpoint 3	16.000		< .001*
Verbal Checkpoint 4	15.000		< .001*
Verbal Checkpoint 5	28.000		< .001*
Trusty Segment 1	41.000		0.008*
Trusty Segment 2	39.000		0.007*
Trusty Segment 3	20.000		< .001*
Trusty Segment 4	26.000		< .001*
Trusty Segment 5	25.000		< .001*

Note. Mann-Whitney U test – * $p < 0.05$

Table 6.3: Spearman's Correlation between changes in verbal trust scores and Trusty scores, showing the relationship between participants' self-reported trust levels and Trusty

	Spearman's rho	p
Verbal Change - Trusty Change	0.751	< .001*

Note. * $p < 0.05$

We further performed Spearman's rank correlations comparing the change scores of verbal self-report and Trusty measurements. As shown in Table 6.3, a strong positive correlation was observed (Spearman's $\rho = 0.751$, $p < 0.001$), indicating that increases (or decreases) in verbal trust ratings were accompanied by corresponding changes in Trusty scores.

Furthermore, we computed Spearman's rank correlations to evaluate the relationship between post-experiment verbal trust scores and Trusty measurements with the trust levels obtained from post-experiment questionnaires. As detailed in Table 6.4, both the verbal trust ratings and the continuous Trusty data show statistically significant positive correlations with the post-experiment questionnaire scores ($p < 0.05$, with Spearman's ρ values ranging from 0.462 to 0.653). These findings indicate that the trust levels recorded during the task - whether through verbal self-report or via Trusty — are in strong alignment with the overall trust assessments captured by the questionnaires administered at the end of the experiment. In line with our expectations, no significant correlations were observed between these measures and the pre-experiment questionnaire scores ($p > 0.05$), suggesting that participants' trust perceptions evolved as a result of their interaction with the wheelchair. Overall, all three measurement modalities consistently captured the dynamic changes in trust over the course of the experiment.

Table 6.4: Spearman's Correlations showing that post-verbal self-report scores and Trusty scores are significantly correlated with post-experiment trust questionnaire scores

			Spearman's ρ	p
Verbal Ckpt. 5	-	TiA After	0.653	0.001*
	-	TOAST After	0.541	0.002*
	-	Trust Perception After	0.474	0.009*
	-	TOAST Before	-0.056	0.772
	-	TiA Before	0.148	0.444
	-	Trust Perception Before	0.175	0.365
Trusty Seg. 5	-	TiA After	0.470	0.012*
	-	TOAST After	0.500	0.007*
	-	Trust Perception After	0.462	0.013*
	-	TOAST Before	-0.005	0.981
	-	TiA Before	0.175	0.373
	-	Trust Perception Before	0.202	0.303

Note. * $p < 0.05$

Table 6.5: Spearman's correlations between sequential pairs of verbal checkpoints and Trusty segments, demonstrating significant positive correlations across all five measurement points

	Spearman's rho	p
Verbal Checkpoint 1 - Trusty Segment 1	0.800	< .001*
Verbal Checkpoint 2 - Trusty Segment 2	0.745	< .001*
Verbal Checkpoint 3 - Trusty Segment 3	0.802	< .001*
Verbal Checkpoint 4 - Trusty Segment 4	0.808	< .001*
Verbal Checkpoint 5 - Trusty Segment 5	0.812	< .001*

Note. * $p < 0.05$

Similarly, we computed Spearman's rank correlations between verbal self-report scores and the corresponding Trusty measurements collected during the experiment. As detailed in Table 6.5, strong and statistically significant positive correlations were observed between each verbal checkpoint and its corresponding Trusty segment across the various time points. In essence, an increase in verbal trust ratings at a checkpoint is consistently accompanied by an increase in the Trusty measurement, and vice versa. Moreover, the violin plots in Figure 6.3 visually underscore the distributional similarity between the two measures, further confirming their concordance in capturing trust dynamics throughout the interaction.

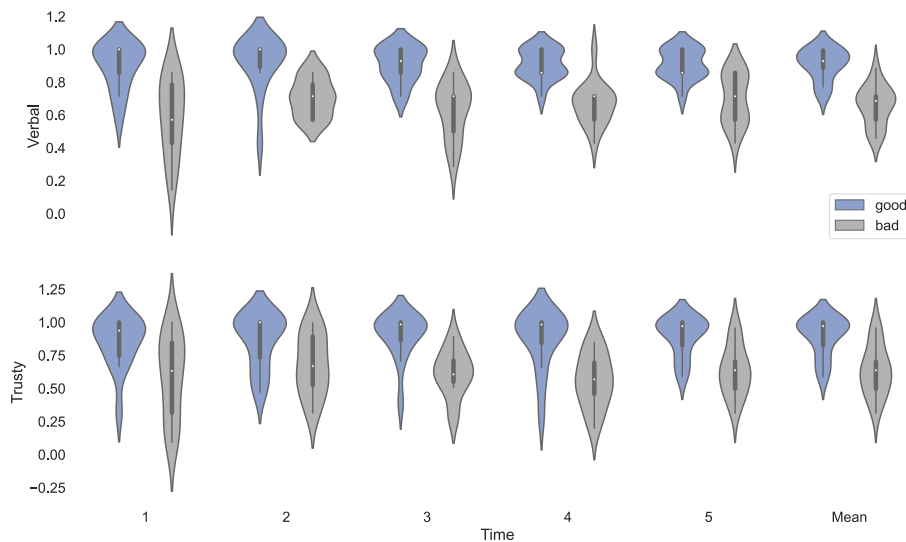


Figure 6.3: Comparison of Trusty and verbal self-report distributions across different conditions depicted through violin plots revealing their similarity. The results suggest that Trusty and verbal self-report exhibit comparable patterns across various conditions, as evidenced by the overlapping distributions in the violin plots.

6.6 DISCUSSION

In this study, we explored continuous trust measurement in [HRI](#) using an experimental setup involving an autonomous wheelchair. Our results provided strong support for our first hypothesis (H1): participants who interacted with the good-performing wheelchair demonstrated significantly higher post-experiment questionnaire scores and self-reported trust levels compared to those who experienced poor performance.

Additionally, we evaluated the validity of Trusty. Our analysis showed positive correlations between post-experiment verbal self-report, Trusty, and trust questionnaire scores, though we recognize the complexities involved in directly comparing continuous and discrete measurements. Despite this, our findings indicate that both verbal self-report and Trusty measurements effectively capture changes in trust during interactions with the autonomous wheelchair. As hypothesized, we found no significant correlations with pre-experiment trust questionnaire scores, suggesting that participants' trust perceptions evolved during the interaction (H2). This absence of significant correlations further supports the idea that trust develops as users engage with the autonomous system, highlighting the need for dynamic trust assessment methods. Furthermore, we observed strong correspondence between Trusty and verbal self-report measures at different time points during the interaction, indicating that Trusty shows promise as a valid measurement tool for assessing trust in this context (H3).

These findings suggest that Trusty can be an alternative to traditional questionnaires, with the added advantage of capturing more detailed trust dynamics during interactions. In addition to its use in research, Trusty can also be applied as a feedback mechanism in [HRI](#). It can provide researchers with quantitative data on trust levels, which can be used to evaluate the effectiveness of different [HRI](#) designs, interfaces, and control strategies. Moreover, Trusty could serve as a valuable labeling tool for machine learning models, as it can generate continuous time series of trust measurements during interactions. This feature enables the collection of large amounts of high-quality labeled data, which can be utilised for training and evaluating machine learning models. By incorporating Trusty's continuous trust measurements as the response variable, we can develop models that predict trust more accurately and reliably using other interaction data (e.g., [HR](#), eye gaze).

While Trusty offers valuable insights into trust dynamics and performance-based trust in [HRI](#), we must acknowledge its limitations. It may not be reliable for individuals with physical disabilities that prevent them from

using the push bar. Additionally, cultural differences in the expression of trust may not be accurately captured by the device. We also note that the device may struggle to distinguish between changes in trust and changes in fatigue or other factors affecting displacement, potentially leading to inaccurate measurements over extended periods. Furthermore, while our results are promising, we have only validated Trusty in the specific context of the autonomous wheelchair scenario, and further studies across diverse [HRI](#) contexts would be valuable to establish its broader applicability.

6.7 CONCLUSION

In this chapter, we introduced and validated Trusty, a novel, tactile device for continuous trust measurement in physical [HRI](#). Our experimental results show that Trusty strongly correlates with established trust metrics while capturing dynamic temporal variations that traditional questionnaire-based methods may overlook.

Trusty provides a practical alternative for assessing trust in [HRI](#), with potential applications in predictive modeling. Its continuous measurements offer rich insights into how trust develops and changes during human-robot interactions. While we need further validation across diverse contexts, we will apply this work to the machine learning approaches explored in [Chapter 7](#) and [Chapter 8](#), helping to develop more responsive, human-centered robotic systems.

TRUST PREDICTION IN ASSISTIVE ROBOTICS USING MULTI-MODAL VIDEO TRANSFORMERS

7.1 INTRODUCTION

In the preceding chapters, we established the importance of trust for effective Human-Robot Interaction (HRI) in assistive scenarios (Chapter 5) and introduced methodologies for measuring trust continuously during interaction (Chapter 6). We showed that trust levels change based on robot performance and can be measured in real-time. Building upon these insights, this chapter addresses the subsequent challenge: leveraging this understanding to develop predictive models capable of estimating user trust during interaction.

Predicting such a complex, dynamic state like trust, however, is non-trivial. The central problem addressed in this chapter is therefore: How can we effectively predict a user’s trust level during real-time interaction with an assistive robot? To address this challenge, we leverage rich, synchronized multi-modal data streams including egocentric video, gaze tracking, physiological signals, and robot data. As we detail in Section 7.2, initial attempts using standard machine learning techniques on non-visual data face significant hurdles, underscoring the need for approaches leveraging rich data streams.

This observation led to the development of the solution presented in this chapter, based on our published work, *Trust Prediction in Assistive Robotics using Multi-Modal Video Transformers*. We introduce a multi-modal, multi-task learning approach designed specifically for this trust prediction problem. Building on the dataset from Chapter 4, we combine trust ratings from the Trusty device (Chapter 6, discretized into levels) with these multi-modal data streams, including visual context from the egocentric video.

Our core approach uses a state-of-the-art video transformer that merges multiple, synchronized data streams. We also include additional tasks, such as estimating the robot’s distance to obstacles and its velocity, to help the model capture a more complete picture of the interaction context. This combination

of multi-task training and multi-modal fusion yields significantly improved trust prediction performance compared to both the standard approaches initially explored and models relying on single data types. All code used in this work is available online.¹ The repository includes our multi-modal video transformer architecture and training methods.

7.2 INITIAL MODELING ATTEMPTS AND LIMITATIONS

Before developing our multi-modal architecture that is the focus of this chapter, we first investigated the feasibility of predicting trust using standard machine learning techniques applied to readily available non-visual data streams. In particular, we explored models commonly used for regression and time-series analysis, such as Linear Regression, Random Forests, and Multi-Layer Perceptrons, when applied to the trust prediction task using non-visual data streams.

These models were trained using physiological signals (Electrodermal Activity (EDA), Heart Rate (HR)) and robot data (velocity, distance to obstacles, etc.) collected during the wheelchair experiment (Chapter 4). To incorporate temporal context, we engineered features like moving averages and lagged values from these signals. Importantly, these initial explorations did not utilize the raw egocentric video stream, focusing instead on what could be achieved with readily processed sensor and robot data alone. We experimented with both global models (trained on pooled data from all participants) and local models (trained individually per participant), using the first 80% of each participant’s interaction data for training and the final 20% for testing generalization. Similar patterns of overfitting and global averaging were observed across the different model types (Linear Regression, Random Forests, MLPs), reinforcing the conclusions drawn from the Linear Regression examples.

Evaluating these models revealed significant challenges related to generalization. These challenges, observed across the different model types explored, are illustrated using Linear Regression examples in Figure 7.1. Here, the local models (green lines), while adapting to individual training data, often struggled to produce stable or accurate predictions on the subsequent test data (right of the dashed line). This tendency towards overfitting, where models capture noise specific to the training segment rather than underlying trends, resulted in unreliable performance. The global models (red lines) presented a different issue, converging towards the mean trust level (purple dotted line) and thus failing to capture individual trust levels.

¹ Code available at: <https://github.com/NicoLingg/multi-modal-video-transformer>

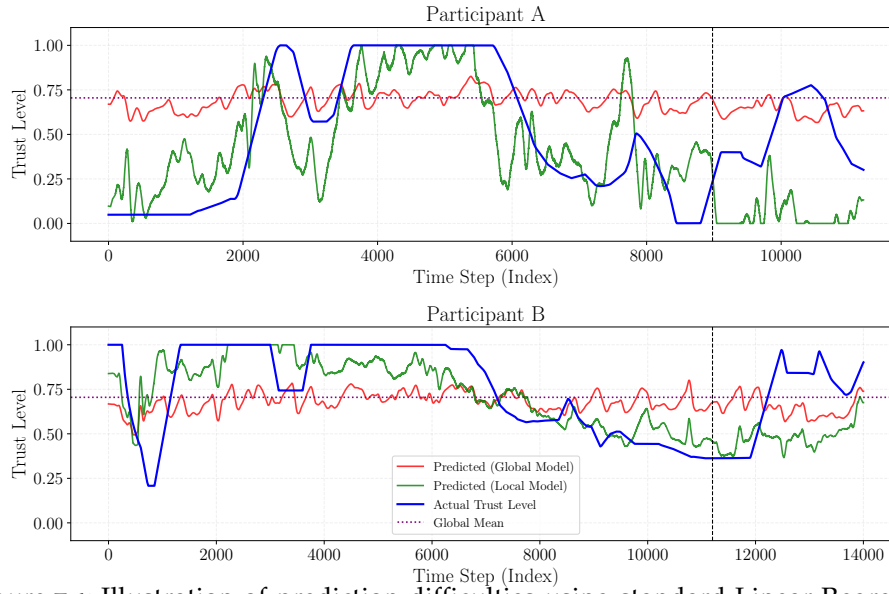


Figure 7.1: Illustration of prediction difficulties using standard Linear Regression models for Participant A (Top) and Participant B (Bottom). Actual trust is shown in blue. The horizontal purple dotted line represents the global mean trust across the dataset. Global model predictions (red) tend towards this mean, failing to capture individual dynamics. Local model predictions (green) show signs of overfitting; while fitting the training data (left of the dashed line), they exhibit high volatility and poor generalization on the test data (right of the dashed line).

The observed instability in the regression outputs, characterized by local overfitting and global underfitting (Figure 7.1), suggested limitations in using these standard regression models with non-visual features to capture the complex, individual dynamics of trust. This observation pointed towards the potential benefits of an architecture capable of integrating richer contextual information, particularly the visual stream, and handling sequential dependencies more effectively. Furthermore, the challenges in obtaining stable continuous predictions led us to explore approaching the problem as a classification task (predicting discrete trust levels) for increased robustness. These findings collectively motivated the development of the multi-modal transformer approach detailed later in this chapter (Section 7.5).

7.3 BACKGROUND AND MOTIVATION

Given the challenges highlighted by initial modeling attempts (Section 7.2), the need for more alternative approaches becomes evident. Trust in HRI is a multi-faceted concept that depends on multiple aspects of the interaction between human and robot. Successfully predicting this state in real-time

would offer significant benefits, allowing robots, for instance, to modify their actions based on inferred user trust levels to improve collaboration. While earlier work in trust prediction often relied on single data streams like physiological signals or interaction behaviors, modern robotic systems can collect various types of data simultaneously, from video and audio to sensor readings and performance metrics. This diversity of available signals, combined with recent advances in processing multiple modalities together, suggests an opportunity to develop richer models for trust prediction.

Transformer architectures have shown impressive results in handling and integrating diverse data streams, making them a promising solution for this challenge. Their self-attention mechanisms are particularly effective at revealing relationships between different types of interaction data. Given their recent successes in various robotics applications (from perception to control), these architectures hold promise for effectively predicting human factors such as trust. In our work, we explore how well these models can fuse multiple data sources to predict trust states during [HRI](#).

7.4 DATASET DESCRIPTION

Building on the experimental setup described in Chapter 4, we collected a comprehensive dataset comprising five synchronized data streams:

1. **Attention Patterns:** Eye gaze coordinates and head pose measurements from Pupil Labs glasses (250 Hz), revealing where participants focus during different trust states and how their attention patterns might signal changes in trust levels.
2. **Physiological Response:** Continuous [HR](#) (1 Hz) and [EDA](#) (60 Hz) measurements that may reflect unconscious responses to robot behavior, particularly during moments of uncertainty or changed trust.
3. **Robot State:** Wheelchair telemetry including position, orientation, and velocity (60 Hz), enabling analysis of how specific robot behaviors and movement patterns correlate with trust levels.
4. **Trust Labels:** Ground truth trust measurements from the Trusty device (60 Hz) introduced in Chapter 6, providing continuous labels for supervised learning.

For training, we synchronized all sensor data streams and resampled them to a fixed 60 Hz frequency, ensuring temporal alignment across modalities.

Table 7.1: Dataset composition. Trust-labeled data from 34 participants combined with supplementary data (1 subject) without trust labels. Supplementary data aims to enhance regression tasks, indirectly improving trust prediction. Both datasets include wheelchair performance, [EDA](#), [HR](#), head pose and eye gaze signals.

DATASET	FRAMES	SUBJECTS	SIGNALS
Supplementary Data	414,799	1	
Participants Data	444,537	34	+ Trust
Total	859,336		

To augment the dataset for performance and regression tasks, we collected additional supplementary data from a single subject who varied their behavior across multiple sessions. This supplementary data excludes trust signals and is used to enhance the model’s training on objective tasks, while the trust prediction task relies solely on data from the 34 participants. The loss functions for trust prediction are masked for supplementary samples during training. As shown in Table 7.1, the enriched dataset, combining supplementary and participant data, totals over 850,000 tokens. To create a validation set, we set aside data from two randomly selected participants.

7.5 MULTI-MODAL VIDEO TRANSFORMER

Motivated by the limitations of standard approaches detailed in Section 7.2, we propose a transformer-based architecture designed for sequential processing of multiple input streams to predict trust. The model’s design prioritizes efficient fusion of video, gaze, and physiological data while maintaining temporal relationships across all modalities, unlike methods relying on aggregated or point-in-time features. By processing multiple modalities simultaneously, we enable the model to capture subtle patterns that might be missed when analyzing individual data streams in isolation.

7.5.1 Architecture

The core architecture, shown in Figure 7.2, consists of five components that transform raw sensor data into trust predictions. Each component plays a specific role in processing and integrating the multi-modal data streams:

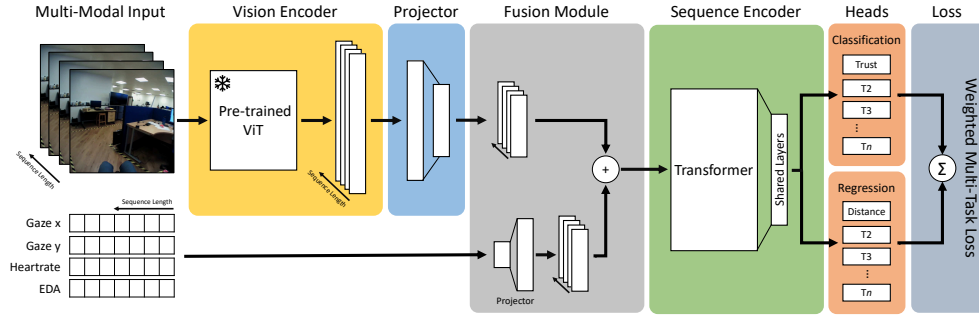


Figure 7.2: Illustration of our proposed multi-modal video transformer architecture. The network intakes a sequence of image frames, processes spatial features through a frozen ViT, and reduces dimensionality via a projector. Separately, time-aligned eye gaze and physiological data are projected linearly to match the projector’s output dimension, enabling fusion through element-wise addition. This is followed by temporal feature extraction through a sequence model. One shared linear layer is followed by separate prediction heads for each task. Task losses are aggregated in a weighted multi-task loss function.

1. **Vision Encoder:** Video frames are individually passed through a pre-trained ViT with frozen weights (Wu et al., 2020).
2. **Projector:** Frame representations are projected to a lower dimension using a Multi-Layer Perceptron (MLP) with one hidden layer.
3. **Fusion Module:** Time-aligned eye gaze and physiological data are projected linearly to match the Projector’s output dimension and then element-wise added to projector outputs.
4. **Sequence Encoder:** Encoder-only transformer model composed of eight causal self-attention (Vaswani et al., 2017) and MLP blocks process the entire sequence of projector outputs, i.e. tokens. Following Touvron et al. (n.d.), we use RMS normalisation (Zhang and Sennrich, 2019) and SwiGLU activations (Shazeer, 2020).
5. **Prediction Heads:** After sequential processing, the final tokens contain all necessary context for prediction. Following a single shared linear layer, each target is predicted separately.

Our modular approach offers distinct advantages over conventional video-native transformer architectures, such as ViViT (Arnab et al., 2021) and TimeS-former (Bertasius et al., 2021). Unlike these models, which intertwine spatial and temporal feature extraction, our design leverages a pre-trained, frozen ViT, focusing computational resources on learning temporal features. We also combine contextual time-series signals (for example, eye gaze data) with

Table 7.2: Overview of model components and their respective trainable parameters. The Vision Encoder (ViT) remains frozen to leverage pre-trained efficiencies, while other components are fully trainable and optimized during training.

Model Component	Trainable Parameters	Frozen
1) Vision Encoder	0	Yes
2) Projector	114,946	No
3) Fusion Module	384	No
4) Sequence Encoder	1,049,628	No
5) Prediction Heads	18,189	No
Total	1,184,043	-

image representations before feeding them into our sequential model, which results in more effective multi-modal temporal feature extraction. Additionally, our design is flexible, allowing for the easy replacement of vision and sequence encoders to meet various application needs. Finally, to improve training stability, we adopt the σ Reparam method for all learnable linear layers, per Zhai et al. (2023).

For a detailed overview of the trainable parameters in each model component, see Table 7.2. For information on training hyperparameters, refer to Table B.2 in the Appendix.

7.5.2 Training Methodology

To train a robust model for trust prediction, we leverage multi-task learning to integrate auxiliary perceptual tasks. This approach enriches feature representations and enhances generalization. The following section introduces the weighted multi-task loss function used to balance these tasks effectively.

7.5.2.1 Data Preprocessing and Training Setup

Prior to training, we time-aligned all data modalities, and time-series observations were normalised using dataset-wide statistics. We transformed the continuous trust measurements from Trusty (values 0-1024) into three discrete classes (low, medium, high) for both technical and practical reasons. First, as suggested by the difficulties encountered with continuous regression in our initial modeling attempts (Section 7.2), classification into discrete trust levels is more robust to noise and small fluctuations in the continuous signal that may not reflect meaningful changes in trust state. Second, discrete classification simplifies the learning problem compared to continuous regression, particularly important given our limited dataset size and the complexity of trust prediction. Third, from an application perspective, robot behavior adaptation likely requires only broad trust level distinctions rather than fine-

grained numerical values. Each input sequence comprised 200 video frames and their corresponding fusion data, totaling around six seconds in duration.

To enhance model robustness, we applied data augmentation during training, using transformations detailed in Appendix Table B.1. Where applicable, these transformations were also applied to corresponding eye gaze coordinates to maintain consistency across modalities.

We performed training on an Nvidia RTX 4090 GPU, Ryzen Threadripper PRO 5965WX, and 128GB RAM for efficient multi-modal processing.

7.5.2.2 Weighted Multi-Task Loss Function

Effective trust prediction requires handling both unstructured video data and sparse trust labels (high, medium, low), which provide limited feedback. To address this, we hypothesized that training the model on auxiliary perceptual tasks would guide it to extract contextually relevant features. These features, such as environmental context or gaze dynamics, are important for understanding trust.

Multi-task learning also brings additional benefits, including reduced overfitting and faster convergence, by sharing knowledge across related tasks [Crawshaw \(2020\)](#). For instance, predicting gaze elevation or azimuth potentially enriches the model’s understanding of user attention, a key determinant of trust in collaborative systems.

We employed a weighted multi-task loss function to balance the contributions of individual tasks. The total loss L is computed as:

$$L = \sum_i w_i L_i \quad (7.1)$$

where L_i represents the loss for task i , and w_i is its corresponding weight. A naive approach to weighting could disproportionately emphasize certain tasks, especially if their loss scales differ. To mitigate this, we normalized the weights as follows:

$$w_i^* = \frac{w_i}{\sum_i w_i} \cdot \text{scale} \quad (7.2)$$

The hyperparameter scale ensures that the total weight sum remains constant, decoupling the task contributions from the number of tasks.

Task weights w_i were empirically determined through preliminary experiments to optimize trust prediction accuracy while maintaining balanced performance across auxiliary tasks. A complete list of auxiliary tasks and their contributions can be found in Table 7.3.

Table 7.3: Experimental results comparing seven model configurations (M1-M7). Models M1-M5 use video-only input with incrementally added auxiliary tasks, improving trust accuracy by 4% (M5 vs M1). Adding gaze input (M6: V+G, all tasks) yields the best performance, boosting trust accuracy by 10% over M5 and more than halving gaze prediction MAE. Adding physiological data (M7: V+G+P) offers nuanced results. Comparisons also show the Transformer architecture outperforms an [LSTM](#) variant (M6 vs M6 [LSTM](#), +12.2% trust accuracy) and supplementary data significantly enhances performance (M6 vs M6 w/o SD, +10.5% trust accuracy).

Model	Inputs	Classification		Regression Tasks (Mean Absolute Error)								
		Accuracy (%)		<i>Lower is better</i>								
		Trust	WC Perf.	X Pos.	Y Pos.	Dist.	Orient.	Vel.	Roll	Pitch	Azim.	Elev.
				(m)	(m)	(m)	(deg)	(m/s)	(deg)	(deg)	(deg)	(deg)
M1	V	50.7	-	-	-	-	-	-	-	-	-	-
M2	V	51.5	86.5	-	-	-	-	-	-	-	-	-
M3	V	49.9	87.9	0.55	0.60	0.09	8.23	0.09	-	-	-	-
M4	V	51.2	81.0	0.46	0.54	0.09	8.64	0.10	3.38	2.79	-	-
M5	V	54.7	84.3	0.49	0.56	0.08	6.42	0.10	3.96	3.21	10.44	10.25
M6	V + G	64.3	88.3	0.50	0.54	0.09	7.71	0.09	3.31	2.55	3.34	3.08
M7	V + G + P	57.3	84.6	0.51	0.59	0.10	6.40	0.10	3.52	3.08	3.89	3.15
M6 (LSTM)	V + G	52.1	76.9	0.62	0.68	0.18	7.51	0.13	3.92	3.68	4.23	3.85
M6 (w/o SD)	V + G	53.8	75.1	0.63	0.69	0.29	9.79	0.12	4.34	3.34	4.34	4.00

V = Video, G = Gaze, P = Physiological Data ([EDA](#) and [HR](#)), WC Perf. = Wheelchair Performance, SD = Supplementary Data

7.6 EXPERIMENTAL RESULTS

We conducted a series of tests to quantify the contribution of A) auxiliary perceptual targets, B) fusion inputs, and C) model architecture and supplementary data to the accuracy of the model’s trust predictions. Table [7.3](#) summarizes the validation results for seven distinct model configurations (M1-M7), detailing the specific inputs used and targets predicted by each.

7.6.1 Multi-Task Contribution

We begin with M1, a baseline model using only video input (V) and trained solely on the trust classification task. To quantify the contribution of auxiliary perceptual targets, we incrementally added task groups while keeping the input modality as video-only: M2 added wheelchair performance classification; M3 added wheelchair state regression tasks (position, distance, orientation, velocity); M4 added head pose regression (roll, pitch); and M5 added gaze regression tasks (azimuth, elevation), thus predicting all available targets. We assessed if adding these task groups improved trust prediction

accuracy. Trust accuracy increased by 4% when all targets were included (M5 vs M1), particularly with the addition of gaze-related targets (azimuth, elevation) in M5. While incorporating these auxiliary tasks improved trust prediction, the performance of these video-only models (M1-M5) remained modest, suggesting a weak but learnable signal in video alone.

7.6.2 Multi-Modal Contribution

We further investigated the contribution of additional input modalities beyond only video. Integrating eye gaze XY coordinates as an input alongside video (Model M6, Inputs: V+G), while training on all tasks, increased trust prediction accuracy by 10% compared to the best video-only model (M6 vs M5). This addition also more than halved the mean-absolute-error for gaze-related perceptual targets, showcasing the fusion module’s efficacy. Adding physiological data (HR and EDA) to video and gaze inputs (Model M7, Inputs: V+G+P) yielded nuanced results: while trust classification accuracy decreased slightly compared to the video and gaze model (M6), it still outperformed video-only models (M1-M5), suggesting a complex contribution that does not linearly enhance performance in all aspects. Our findings suggest that video transformers can predict user’s trust levels (high, medium, low) with about 64% accuracy. This also supports our hypothesis that transformers effectively process rich, high-dimensional temporal data and perform multi-modal fusion, capturing complex trust relationships in HRI. Gaze data, both as input and auxiliary target, boosted trust prediction accuracy, while physiological data’s role requires more research.

7.6.3 Model Architecture and Data Impact

To further validate our architectural choices and data strategy, we conducted two key comparisons using the best-performing configuration, M6 (Video + Gaze inputs, all tasks). First, we compared our standard Transformer-based M6 architecture against a variant using the same inputs and tasks but replacing the Transformer sequence encoder with an LSTM (denoted M6 (LSTM) in Table 7.3). The standard Transformer-based M6 outperformed its LSTM counterpart by 12.2% and 11.4% in trust and wheelchair performance prediction, respectively. This performance difference suggests the transformer’s ability to capture complex temporal relationships in multi-modal data.

Second, we evaluated the impact of the supplementary dataset by comparing our standard M6 model (trained on both participant and supplementary

data) with an identical model trained without the supplementary data (M6 w/o SD). Incorporating supplementary data yielded improvements across all metrics, with notable increases of 10.5% in trust prediction, even though supplementary samples do not have trust labels. This aligns with broader scalability results in transformer research, as performance generally improves with compute (more data or more parameters) (Hoffmann et al., 2022). These findings support our choice of a transformer architecture and the use of supplementary data for enhancing model robustness and performance.

7.6.4 Comprehensive Performance Evaluation

To provide a holistic view of the M6 model’s performance, its predictive accuracy across key metrics is visualized in time-series plots in Figure 7.3, while the confusion matrices for trust and wheelchair performance classification are presented in Figure 7.4. These visualizations contextualize the numerical results discussed above, illustrating how auxiliary perceptual targets, fusion inputs, and architecture enhancements contribute to real-time trust predictions and classification performance.

The time-series plots in Figure 7.3 demonstrate the model’s ability to track complex temporal patterns in wheelchair telemetry (e.g., position, velocity, orientation) and environmental variables (e.g., pitch, elevation). Solid lines represent ground-truth values, while dashed lines show the M6 model’s predictions, highlighting its capacity to approximate dynamic interactions accurately.

Additionally, the confusion matrices presented in Figure 7.4 underscore the model’s classification accuracy. For trust, diagonal dominance reflects consistent prediction of high, medium, and low trust states, with minor misclassifications between adjacent levels (e.g., medium as high). Similarly, high precision and recall for wheelchair performance classification validate the multi-task learning approach, which enables robust feature extraction and integration across modalities. These results visually confirm the quantitative improvements achieved through our model design.

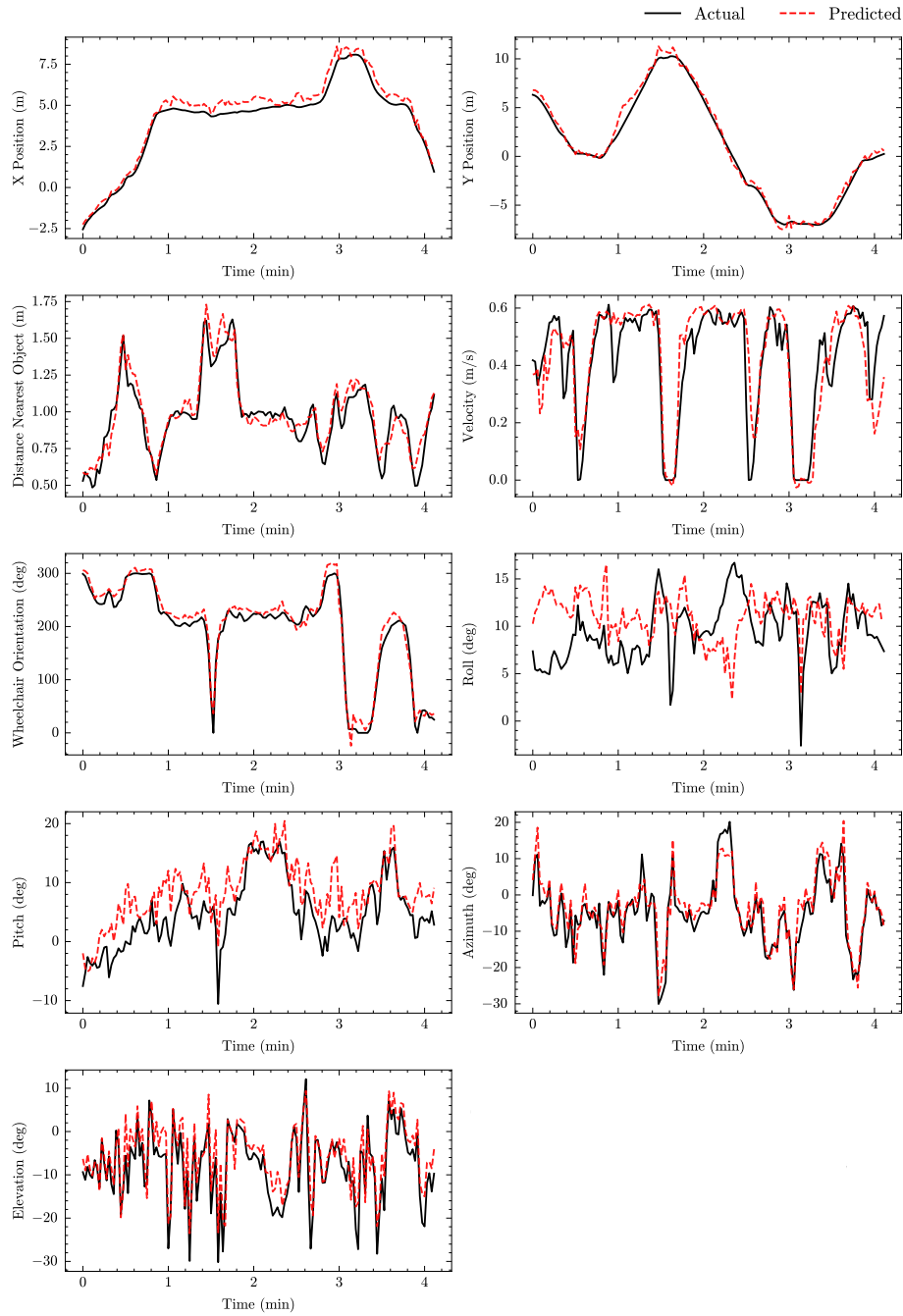


Figure 7.3: Visualisation of M6's performance on validation set. Time-series plots with actual values (solid lines) vs predictions (dashed lines) demonstrate the model's ability to track complex patterns. Confusion matrices for trust and performance provide insights into its classification accuracy.

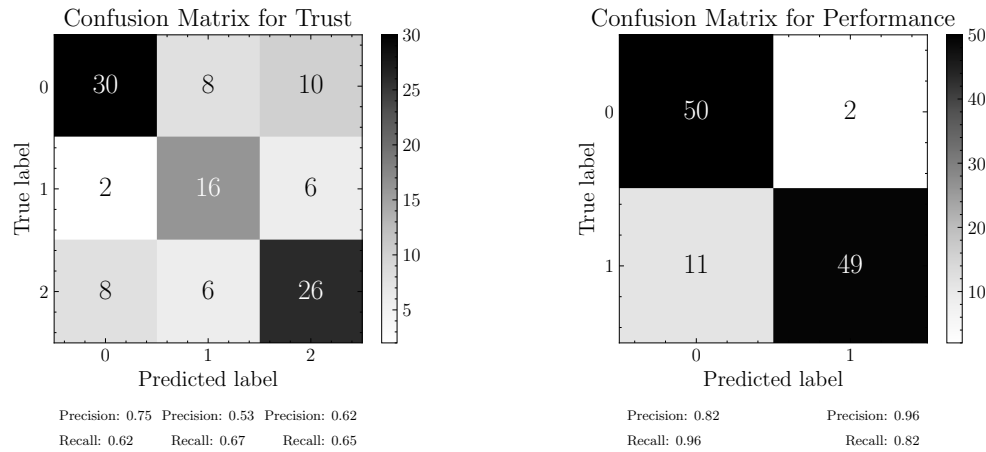


Figure 7.4: Confusion matrices for the M6 model on the validation set. The left matrix shows performance for the three-level trust classification (Classes 0: Low, 1: Medium, 2: High), and the right matrix shows performance for the binary wheelchair performance classification (Classes 0: Bad, 1: Good). Precision and recall scores for each class are provided below each matrix.

7.7 DISCUSSION

The results suggest that trust prediction can benefit significantly from integrating multi-modal data and incorporating auxiliary training tasks. By leveraging diverse data modalities such as gaze patterns, physiological signals, and environmental context, we were able to uncover latent patterns that may not be as accessible with single-modality inputs. These findings align with the idea, explored in earlier chapters (Chapters 5 and 6), that trust is a dynamic, multifaceted construct shaped by a variety of contextual factors. The initial modeling attempts (Section 7.2) further reinforce this, showing the inadequacy of simpler models relying solely on non-visual data for capturing these complexities.

The inclusion of auxiliary perceptual targets in the multi-task learning framework also demonstrated potential benefits. This approach seems to enhance trust prediction accuracy by encouraging the model to capture richer environmental features. These observations resonate with prior research that highlights multi-task learning as a strategy to improve generalization and robustness in machine learning models.

However, the integration of physiological signals yielded mixed results. While HR and EDA contributed some improvement, their overall impact was less pronounced than that of gaze data. This suggests a nuanced relationship between physiological signals and trust. One possibility is that their signal is effectively captured, or even superseded by, visual and gaze cues within

the transformer architecture. Further investigation could help clarify how these signals interact with other modalities and whether they provide more indirect cues to trust dynamics.

Although the results are encouraging, there is room for improvement. The overall prediction accuracy (64.3% for the best-performing model) suggests that while the approach is feasible, further work is needed to enhance its effectiveness. Larger and more diverse datasets, as well as the exploration of alternative model architectures or fusion techniques, could contribute to improved performance in future iterations.

7.8 COMPARISON WITH EXISTING TRUST MODELS

Our multi-modal transformer approach differs from traditional computational trust models in several key aspects. While existing approaches like RoboTrust (Section 2.4.3) or ARMAV (Section 2.4.4) rely on discrete updates and predefined parameters, our implementation learns trust patterns directly from continuous multi-modal interaction data.

Most computational models focus on single data streams or simple performance metrics. Time-series models like ARMAV assume linear relationships between trust and performance factors, while algebraic models reduce complex interactions to binary success/failure indicators. Our transformer architecture processes rich, high-dimensional inputs including visual information, physiological signals, and behavioral patterns without requiring manual feature engineering or simplification.

Traditional models typically serve evaluative purposes, assessing trust levels after interaction completion. DBN-based approaches like OPTIMo (Section 2.4.7) use predefined state transitions and observation models that require manual specification. In contrast, our approach enables real-time trust assessment during active interaction, learning complex temporal dependencies and non-linear relationships directly from empirical data.

The multi-modal integration in our implementation addresses a key limitation in existing models, which often overlook contextual information available during physical human-robot interaction. By processing synchronized video, gaze, and physiological data streams, our model captures the multifaceted nature of trust that single-stream approaches miss.

7.9 CONCLUSION

This chapter presented a multi-modal transformer architecture that successfully predicts user trust by combining video, gaze tracking, and physiological data streams collected during human-robot interaction. Our approach addresses the limitations identified with simpler modeling techniques (Section 7.2) by leveraging the representational power of transformers and the richness of multi-modal context. Our results achieved 64% accuracy in predicting discretized trust levels (low, medium, high), with eye-gaze data emerging as a particularly valuable signal for prediction. The multi-task learning framework further enhanced performance by guiding the model to learn relevant contextual features, such as robot state and environmental variables.

These findings extend the challenges discussed in Chapter 3, particularly the need for richer data sources to capture trust in real-world, dynamic settings. Building on the insights from Chapter 5, our results suggest trust is a dynamic process shaped by various contextual and user-driven factors.

In the next chapter, we introduce the Assistive Mobile Interactive Grasping Agent (AMIGA) platform and present Trust-ACT, our framework for trust-aware imitation learning. Trust-ACT incorporates trust signals into the robot's decision-making process, aligning its actions with human trust metrics.

MULTI-MODAL TRUST-AWARE IMITATION LEARNING FOR ROBOTIC MANIPULATION WITH AMIGA

8.1 INTRODUCTION

Earlier in this work, we examined how trust emerges, evolves, and can be measured in Human-Robot Interaction ([HRI](#)), focusing initially on navigation tasks. We introduced devices and methodologies, such as the Trusty tool, to capture continuous trust signals. Building on this foundation, we now shift our attention from the navigation scenario to a manipulation task, as outlined in the Thesis Roadmap (Figure [1.2](#)), and investigate how we can use trust as a controllable dimension in robotic policy generation.

The central problem explored here is: How can subjective human trust signals, captured continuously during expert demonstrations, be effectively integrated into an imitation learning framework to enhance robotic manipulation skills? Specifically, we investigate if these trust annotations can be used to: (1) enable a robot policy to learn and generate distinct execution styles that reflect the varying levels of trust associated with the demonstrations, and (2) guide the robot during execution to select actions that lead to more reliable or preferred outcomes, as implicitly encoded in the trust signal.

To ground this investigation in a realistic setting, we first introduce [AMIGA](#) (Section [8.2](#)), a versatile mobile manipulation platform developed for [HRI](#) research. [AMIGA](#), integrating a UR10e arm with a powered wheelchair base and advanced sensors, provides the necessary infrastructure for collecting synchronized manipulation demonstrations (via the GELLO teleoperation interface) and continuous trust annotations (via the Trusty device).

Leveraging Assistive Mobile Interactive Grasping Agent ([AMIGA](#))’s capabilities, we propose Trust-ACT (Section [8.3](#)), a method integrating trust signals into the generative Action Chunking with Transformers ([ACT](#)) framework ([Zhao et al., 2023](#)). Trust-ACT conditions its Conditional Variational Autoencoder ([CVAE](#))-based policy on trust annotations to learn distinct exe-

cution styles associated with those annotations. During inference, it utilizes a learned trust prediction model to evaluate multiple trajectories sampled from the policy, selecting the most promising one according to the learned trust criteria. This approach demonstrably improves task success rates (e.g., in block stacking) while enabling both trust-conditioned stylistic control and enhanced performance through trust-guided selection.

8.2 AMIGA MOBILE MANIPULATION PLATFORM

8.2.1 Introduction, Motivation, and Goals

AMIGA, illustrated in Figure 8.1, represents a novel approach to mobile manipulation platforms that addresses a critical challenge in robotics research: cost accessibility. While commercial mobile manipulators often exceed \$100,000, **AMIGA** demonstrates how design choices can dramatically reduce costs while maintaining essential capabilities for **HRI** research.

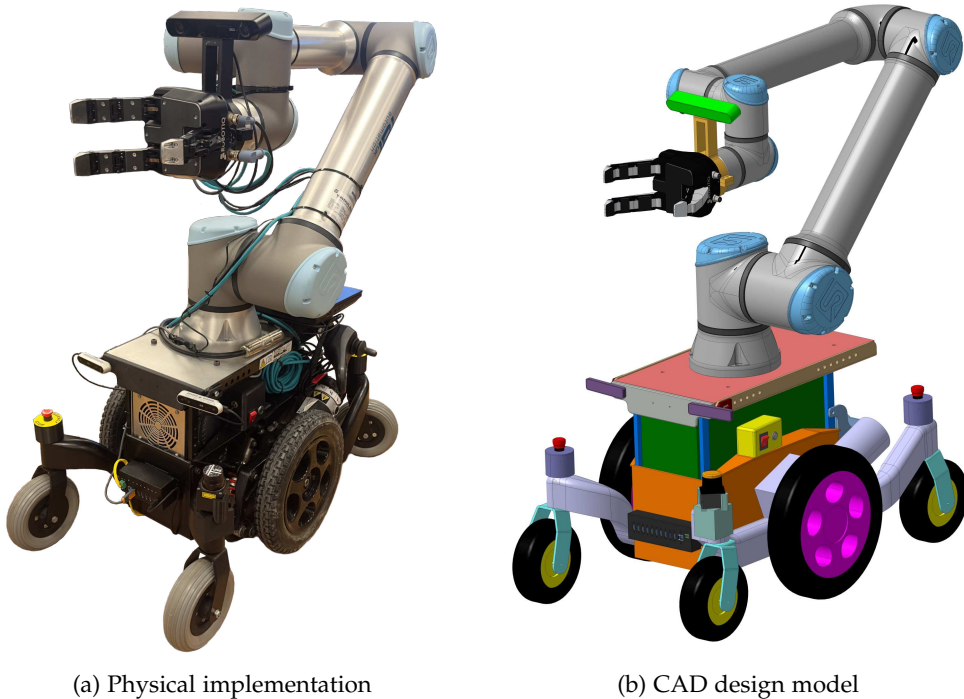


Figure 8.1: **AMIGA**: (a) The physical platform integrating a UR10e collaborative robot arm with a powered wheelchair base. (b) Detailed CAD model developed for multiple purposes: manufacturing specifications, generating the robot's URDF model for Robot Operating System (ROS) integration, and enabling testing in simulation environments like Gazebo.

At the core of AMIGA's design is a complete rethinking of mobile robotics platforms by repurposing a powered wheelchair as its mobile base. This approach leverages the wheelchair's inherent advantages: its robust mechanical design can support heavy payloads, its built-in power system provides reliable operation, and its proven mobility system enables precise navigation - all at a fraction of the cost of conventional robotic mobile platforms. The wheelchair base is integrated with a Universal Robots UR10e collaborative robot arm and equipped with a Robotiq 3-Finger Adaptive Robot Gripper for versatile object manipulation. For comprehensive environmental perception, AMIGA incorporates multiple sensor modalities: a wrist-mounted ZED 2i stereo camera provides detailed views for manipulation tasks, while front and back-mounted Intel RealSense D415 cameras combined with LIDAR sensors and an Inertial Measurement Unit (IMU) enable robust navigation and obstacle detection.

The development of AMIGA was driven by several key requirements. The platform needed sufficient payload capacity to support not only the UR10e arm but also end-effectors, sensors, and computing hardware. Modularity was essential to facilitate easy integration or replacement of components as research needs evolve. Cost-effectiveness remained essential, driving our decision to utilize a second-hand wheelchair base. As a platform intended for HRI studies, safety and reliability were non-negotiable. Finally, the software architecture needed to support multiple control paradigms, from direct teleoperation to autonomous navigation, enabling diverse research applications including trust experiments.

8.2.2 Design Methodology and System Configuration

To systematically address these requirements, we used a comprehensive morphological analysis approach, a creative problem-solving method that breaks down complex systems into their basic functional components and explores different solutions for each. As illustrated in Figure 8.1, this structured method decomposed AMIGA's design into fifteen core subfunctions, from base selection and arm configuration to end-effector choice, mounting solutions, power systems, computing architecture, and sensor integration. For each subfunction, we identified and evaluated multiple potential solutions, creating a systematic decision space for the overall system design.

The analysis revealed three potential system configurations, each represented as a complete path through the solution space in Figure 8.1. Our implemented solution (highlighted as the "red path") combines a wheelchair

base with a UR10e arm in a single-arm configuration, optimally balancing cost-effectiveness and functionality. We also explored two alternative configurations: a Husky mobile base with dual UR5 arms that offered superior outdoor mobility capabilities but at significantly higher cost, and a Ridgeback base with a single UR5 arm that provided enhanced autonomous navigation features but suffered from limited cost-effectiveness and a larger footprint.

In addition to the functional comparisons, a high-level cost analysis (Table 8.2) highlights a significant benefit of our wheelchair-based approach. Specifically, the use of a wheelchair base (Solution 1) resulted in a 90-96% reduction in base platform costs compared to commercial alternatives (Solutions 2 and 3). This substantial cost saving allowed for increased investment in high-quality sensors and computing resources, while maintaining a total system cost considerably below those of commercial options. Although approximate, these figures strongly support using a wheelchair platform to keep system costs manageable.

8.2.3 *Software and Control Architecture*

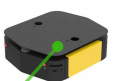
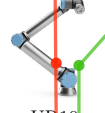
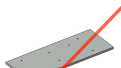
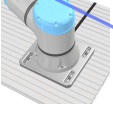
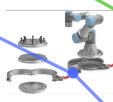
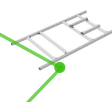
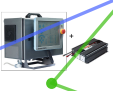
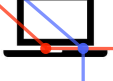
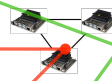
AMIGA's software and control framework is designed as a flexible, distributed architecture, capable of handling various tasks such as navigation, perception, and manipulation. Figure 8.2 provides a block diagram of the overall system, illustrating how data and control flow among the core hardware elements - the powered wheelchair base, the UR10e robotic arm, the Robotiq 3-finger gripper, and the sensor suite - and the computing units responsible for higher-level decision-making and control.

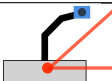
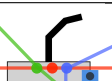
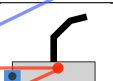
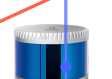
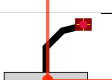
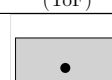
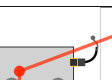
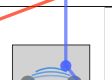
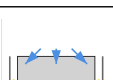
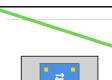
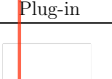
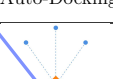
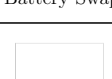
The design philosophy behind AMIGA's software architecture centers on balancing computational demands and ensuring modularity for future expansions. The wheelchair base provides both mobility and power, while two dedicated NVIDIA Jetson Xavier units distribute the processing load: one Jetson manages navigation tasks such as mapping, localization, and path planning, and the other Jetson focuses on perception-driven manipulation, including object recognition and grasp planning.

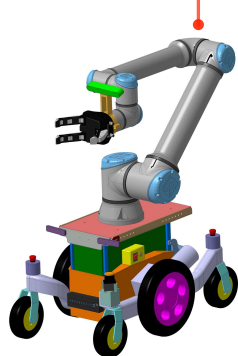
A Raspberry Pi handles low-level commands to the wheelchair's R-Net controller by sending CAN bus messages via the can2RNET¹ library, allowing control of velocity and drive parameters while also reading feedback on battery status and operational states. A laptop adds additional computational resources, for example running advanced machine learning models, and can act as the ROS master in configurations where ROS is utilized.

¹ <https://github.com/redragonx/can2RNET.git>

Table 8.1: Morphological analysis matrix decomposing AMIGA’s design into 15 core subfunctions, from base selection to sensing systems. Each row represents a subfunction with multiple solution alternatives, and colored paths (red, blue, green) show three configurations. The red path highlights our implemented wheelchair-based solution, while blue and green paths represent alternative configurations using Husky and Ridgeback bases respectively. 3D renderings of complete configurations are shown at the bottom.

Morphological Analysis					
Product: AMIGA			Designed by: Nico Lingg		
Subfunction	Solution 1	Solution 2	Solution 3	Solution 4	Solution 5
Base	 Wheelchair	 Ridgeback	 Agile Ranger	 Husky	 Unmotorized
Arm	 UR10	 UR5	 Kinova Gen 3	 Franka Panda	 KUKA
Arm Configuration	 Single Arm	 Dual Arm			
End-Effector	 Robotiq 2-Finger	 Robotiq 3-Finger	 Soft Gripper 4-Finger	 Allegro Hand	 Vacuum Gripper
Mounting	 Custom	 Universal Mount	 Quick Release	 Adjustable Frame	
Power Source	 Wall Power (Tethered)	 Internal Batteries (From Base)	 External Additional Battery	 Power Station (Battery + Inverter)	
Power Control	 Compact DC Control Box	 AC Control Box + Inverter	 AC Control Box		
Computing	 External PC	 Onboard Laptop	 Cloud-Based	 Distributed	

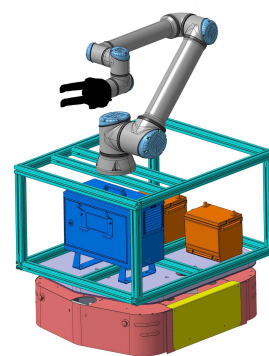
Transmission	 Ethernet	 WiFi	 Bluetooth	 Radio	
Perception Systems	 Mono Camera	 Stereo/RGB-D Camera	 Event Camera	 360° Camera	 Thermal Camera
Perception Systems Config.	 Wrist	 Front	 Back	 Side Left	 Side Right
Distance Sensing	 2D LIDAR	 3D LIDAR	 Ultrasonic	 Radar	 Time-of-Flight (ToF)
Distance Sensing Config.	 Wrist	 Front	 Back	 Side Left	 Side Right
Charging	 Manual Plug-in	 Wireless Charging	 Auto-Docking	 Battery Swap	 Contact Plates
Other Sensors	 IMU	 Wheel Enc.	 GPS/GNSS	 Magnetometer	 Force/Torque



Solution 1



Solution 2



Solution 3

Table 8.2: Cost comparison between wheelchair-based [AMIGA](#) (Solution 1) and alternative mobile manipulation configurations. For simplicity, we list only major components (base, arm, end-effector) separately, with all remaining hardware costs consolidated under additional components. The wheelchair base cost reflects actual purchase price; all other costs are estimates.

Component	Solution 1	Solution 2	Solution 3
Base	\$1,804*	\$18,625	\$45,000
Arm	\$35,000	\$60,000	\$35,000
End-Effector	\$20,000	\$10,000	\$20,000
Additional Components	\$10,000	\$15,000	\$12,000
Total	\$66,804	\$103,625	\$112,000

* Actual cost; all other figures are estimates.

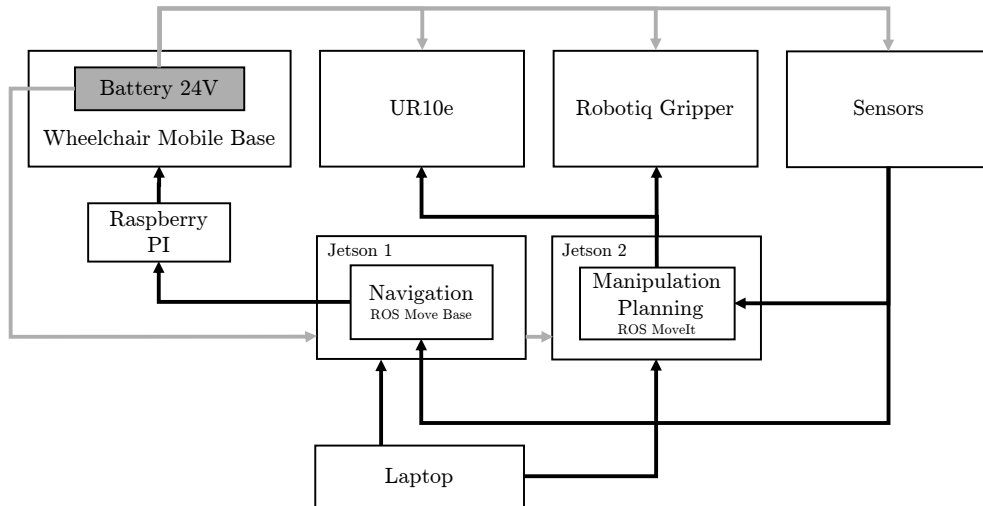


Figure 8.2: Block diagram of [AMIGA](#)'s software and control architecture. It consists of five subsystems: the wheelchair base (powered by a 24V battery), UR10e robot arm, Robotiq 3-finger gripper, the sensor suite, and the compute units (Jetson Xavier modules, laptop, and Raspberry Pi). Navigation tasks are handled by one dedicated processing unit, while manipulation tasks run on another. Additionally, a Raspberry Pi manages low-level CAN bus communication with the wheelchair's R-Net electronics.

A portable router integrates all computing units and sensors into a single network, enabling reliable communication and data sharing. This design allows us to implement new features, modules, or custom software as research needs evolve. The platform supports both [ROS](#) with standard open-source libraries and custom-built control software, making it versatile for many applications.

8.2.4 Control and User Interaction Examples

AMIGA supports multiple options for user interaction, facilitating diverse research scenarios in human-robot collaboration and trust. We have implemented and tested three representative interaction methods.

8.2.4.1 ROS-based with MoveIt

The primary control interface leverages ROS's MoveIt framework for motion planning and execution. Through RViz visualization, operators can monitor the robot's state while an integrated perception pipeline detects and localizes objects in the workspace. The system enables automated grasp planning and trajectory generation - after object detection, MoveIt calculates collision-free paths which the operator can verify before execution.

Figure 8.3 demonstrates this approach.



Figure 8.3: Demonstration of the ROS-based control interface combining object detection, motion planning in MoveIt, and real-time visualization in RViz. The split view shows both the physical robot and its digital planning representation (RViz) ².

² Full video demonstration available at: <https://youtu.be/yE70focMpC0>

8.2.4.2 Vision-Based Arm Control using Apple Vision Pro

This experimental control interface uses Apple Vision Pro's hand tracking to translate user hand gestures into robot motions. After an initial calibration, tracked hand movements are mapped to the robot's end-effector poses, which are then resolved using inverse kinematics to execute the desired movements.

This gesture-based control allows operators to interact with the robot using hand movements. While it provides an intuitive way to guide the robot, it does require the operator to stay aware of their hand position to avoid unintended collisions. This method can be particularly useful in environments where users need to control the robot without physical contact, such as when wearing protective equipment. The system offers flexibility, but users must maintain focus on their gestures to ensure the robot's movements remain accurate and safe.

Figure 8.4 demonstrates this motion control approach.

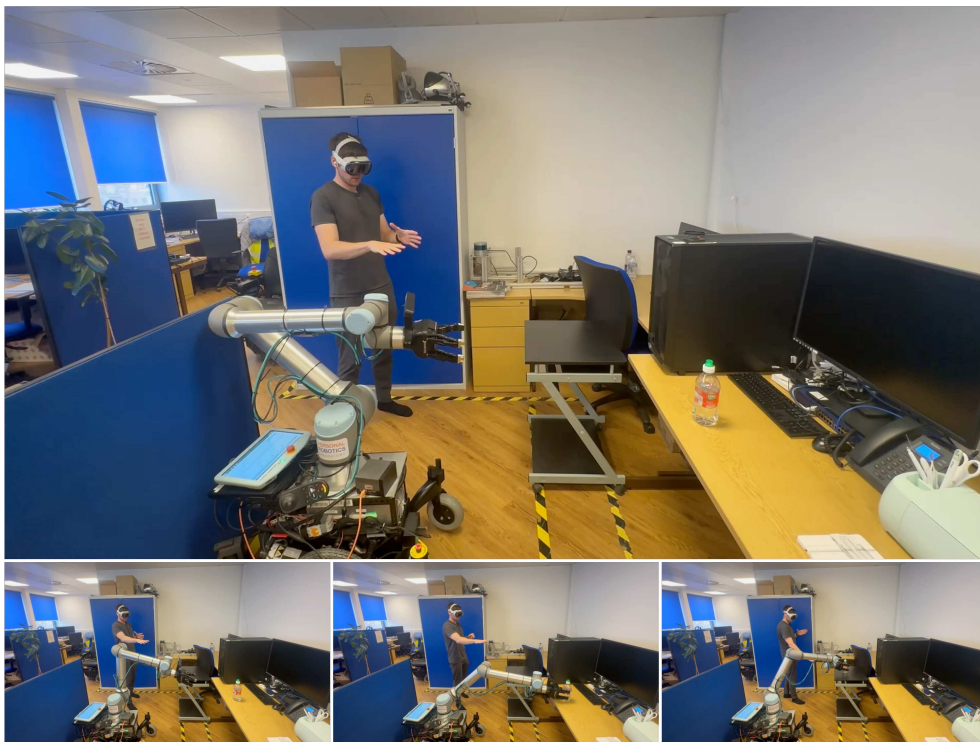


Figure 8.4: Experimental interface using Apple Vision Pro's hand tracking for direct robot teleoperation. The system translates tracked hand positions into robot motions through inverse kinematics ³.

³ Full video demonstration available at: <https://youtu.be/tZJmzDnSWLE>

8.2.4.3 GELLO-based Direct Control

For intuitive manipulation control, we integrated the GELLO framework (Wu et al., 2023), which uses a scaled, kinematically-equivalent physical replica of the UR10e arm. The miniature model is equipped with low-cost motors that function as joint encoders, directly mapping its configuration to the real robot and delivering natural, precise control. With this kinematic mapping, we can easily understand the robot’s constraints and see a direct link between our inputs and the resulting movements.

The GELLO interface stands out for three key characteristics: its cost-effectiveness using readily available components, high precision through direct joint-angle mapping, and intuitive control that requires minimal training for new operators. These features make it an ideal choice for our manipulation experiments, allowing us to focus on trust-related research questions without the interface itself becoming a barrier to natural interaction.

Figure 8.5 demonstrates this approach.

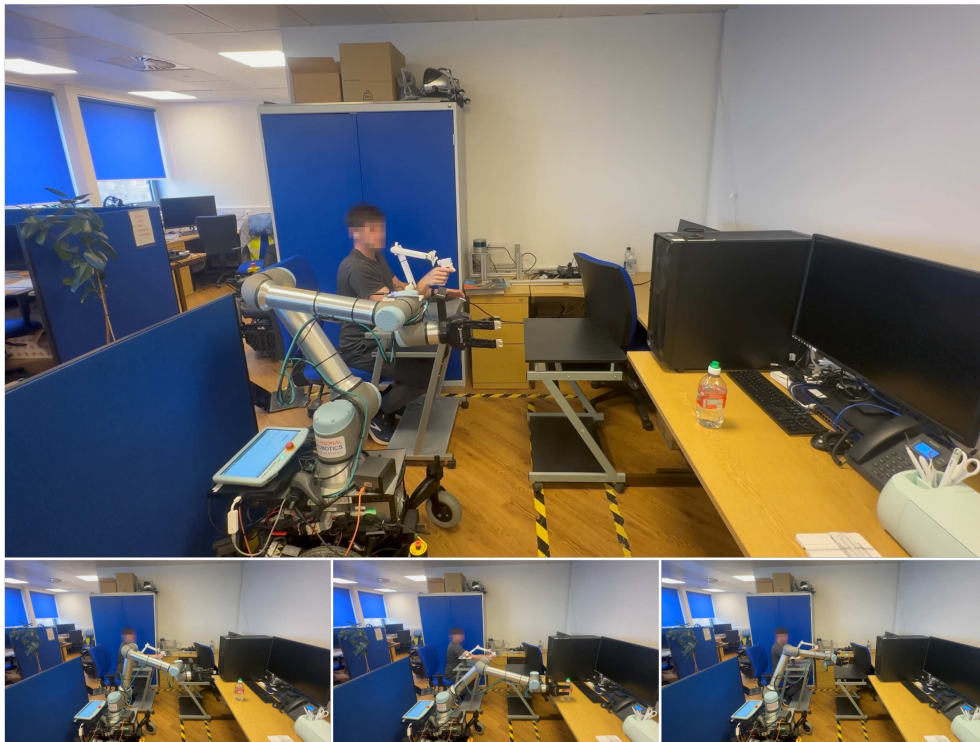


Figure 8.5: The GELLO control interface uses a scaled physical model of the UR10e with joint encoders to enable direct kinematic control of the robot arm, providing intuitive manipulation capabilities while maintaining visual feedback through the planning interface ⁴.

⁴ Full video demonstration available at: <https://youtu.be/K0QMngVUwtw>

8.3 TRUST-ACT: INTEGRATING TRUST IN IMITATION LEARNING FOR ROBOTIC POLICY LEARNING

Building on the capabilities of the [AMIGA](#) platform, this section explores how we can integrate trust into robotic manipulation tasks. Rather than just measuring trust, we focus on using trust annotations to influence imitation learning. We introduce Trust-ACT⁵, a novel approach that leverages human trust signals to condition robotic policy generation. Trust-ACT builds upon the [ACT](#) framework ([Zhao et al., 2023](#)) by conditioning a [CVAE](#)-based policy on trust signals, enabling the robot to generate distinct behaviors corresponding to high, medium, or low trust. In this work, our contributions are threefold: (1) we demonstrate that incorporating trust as a conditioning variable enables real-time modulation of the robot’s motion style; (2) we mitigate the posterior collapse observed in the vanilla [ACT](#) framework through a partial masking strategy (in conjunction with KL annealing) to preserve a meaningful latent space; and (3) we implement a best-of-n sampling strategy, using a trust prediction model to select the most reliable trajectory, which improves task performance.

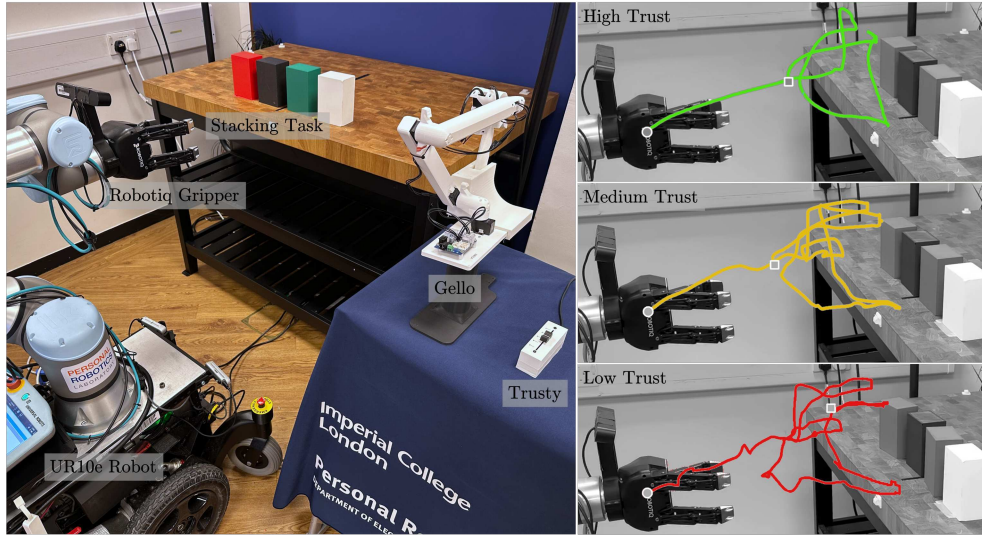


Figure 8.6: Experimental setup and trust-conditioned trajectory visualization: (left) The experimental testbed showing the UR10e robot with Robotiq gripper for block stacking tasks, alongside the Gello and Trusty systems; (right) Generated trajectories under varying trust conditions, demonstrating how path smoothness and directness correlate with trust levels: high trust (green) exhibits direct, confident movements, medium trust (yellow) shows moderate variation, and low trust (red) shows more hesitant, indirect motions ⁶.

⁵ Code available at: https://github.com/NicoLingg/trust_act.git

⁶ Full video demonstration available at: <https://youtu.be/xxqvBuDca4Q>

Figure 8.6 illustrates our experimental setup and highlights the trust-conditioned trajectories produced by Trust-ACT. The left panel shows the testbed featuring the [AMIGA](#) robot with a Robotiq gripper, integrated with the Gello and Trusty systems, while the right panel visualizes trajectories generated under varying trust conditions: high trust (green) exhibits direct, confident movements, medium trust (yellow) shows moderate variation, and low trust (red) displays more hesitant, jerky, and indirect motions.

8.3.1 *Motivation*

A key challenge in robotic manipulation is enabling the robot to adapt its behavior in a manner that aligns with human expectations and trust. Traditionally, trust is evaluated after the fact, serving more as a performance metric. However, if we integrate trust signals directly into the learning process, the robot can learn to generate distinct behavioral distributions that reflect varying levels of trust. In Trust-ACT, by conditioning a [CVAE](#)-based policy on trust annotations, the model learns separate distributions for high, medium, and low trust behaviors. This approach allows the robot to autonomously adjust its motion style, for instance, generating one set of trajectories when the behavior is intended to be trusted and another when it is not. In our work, we opted to map high trust to fast, fluid, and confident movements, while low trust corresponds to slower, jerky, more hesitant, or error-prone trajectories. It is important to note that alternative mappings are conceivable. For example, one might design a system where a drop in trust triggers the robot to deliberately slow down or increase trajectory smoothness to rebuild user confidence. However, in our current framework, the data collection was configured so that high trust is associated with rapid and confident actions, and low trust with less reliable and more erratic motions. Consequently, the model learns to distinguish these behavior patterns directly from the human demonstrations.

8.3.2 *Task and Data Collection*

To gather the necessary data for training Trust-ACT, we designed an experimental task that involves sequentially stacking blocks arranged in a defined workspace. The task setup includes four blocks: red, black, green, and white. The robot is required to perform two stacking operations: placing the red block onto the black block and the white block onto the green block. These operations demand precise grasping, accurate positioning, and stable stacking, making it an ideal scenario for observing how varying trust levels can affect task execution. The task’s sequential nature highlights the influence of trust



Figure 8.7: Block stacking demonstration: The UR10e robot performs a sequential manipulation task: (left) placing the red block onto the black block, (middle) retracting to reposition, and (right) approaching to grasp the white outer block for stacking onto the green block.

on execution styles. Success is measured through binary task completion and execution time, which reflect variations in trajectory smoothness, confidence, and directness under different trust conditions. Figure 8.7 illustrates the task sequence, emphasizing the grasping, retracting, and stacking phases.

To collect data for this task, we employed a two-person protocol. One individual, the teleoperator, used the GELLO interface (Wu et al., 2023) to generate human demonstrations, while a second individual, the observer, continuously rated trust using the Trusty device. This division ensured that the teleoperator could focus on performing specific behaviors, fast and confident or slow, jerky and hesitant, while the observer independently provided real-time trust assessments. This approach helped reduce bias and ensured that the trust ratings reflected genuine external perceptions of the robot’s performance.

To induce clearly differentiated trust ranges in the demonstrations, we designed three main motion styles. High-trust demonstrations, with trust values between 0.66 and 1.0, were executed with fast, fluid movements and minimal hesitation. Medium-trust demonstrations (0.33–0.66) included slower, more cautious maneuvers, while low-trust demonstrations (0.0–0.33) exhibited hesitant, jerky, or error-prone movements. The teleoperator modulated speed, path directness, and pauses to match these trust levels, prompting the observer to maintain trust ratings within the corresponding numerical ranges. In addition to these uniform high-, medium-, and low-trust runs, we also captured mixed-behavior episodes in which the teleoperator changed styles mid-episode. This approach yielded a more varied temporal distribution of trust values within a single trial, adding diversity to the dataset. In total, we recorded 120 episodes: 30 for each trust condition (high, medium, and low), plus 30 mixed-behavior episodes.

The dataset was processed following the preprocessing methodology outlined in the ACT framework (Zhao et al., 2023), with models trained on a NVIDIA RTX 4090 GPU. All episodes were synchronized, and both camera feeds and control signals were recorded at 30 Hz.

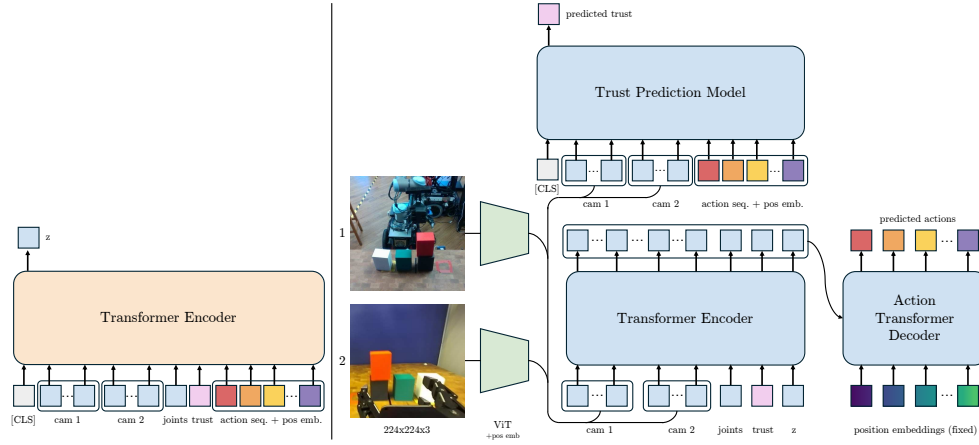


Figure 8.8: Overview of the proposed trust-aware imitation learning architecture. The policy model (bottom) is based on a CVAE framework, where a latent variable z is inferred from demonstrated actions and states (excluding images), and a transformer-based decoder predicts future actions conditioned on the latent, visual inputs from multiple cameras (ViT-encoded), robot states, and an embedded trust label. The trust prediction model (top right) processes the same camera views and actual actions to output a continuous trust score for trajectories. At inference time, multiple candidate action trajectories can be generated by sampling from the latent space and evaluated by the trust prediction model.

8.3.3 Trust-ACT: Trust-Conditioned Policy Architecture

Our policy model, Trust-ACT, builds upon the ACT (Zhao et al., 2023) framework but introduces key modifications to integrate trust as an explicit conditioning signal and evaluation metric. The Trust-ACT framework comprises two core interacting model components:

1. A **Trust-Conditioned Policy Model** (π_θ): Based on a CVAE, this model’s primary role is to generate sequences of robot actions ($\hat{\mathbf{a}}_{t:t+k}$) conditioned on the current robot state ($\mathbf{s}_t$), camera images ($\mathbf{c}_t$), a latent variable ($\mathbf{z}$), and crucially, a desired trust level (r_t).
2. A **Trust Prediction Model** (ν_η): This model’s role is to evaluate a given trajectory (composed of states $\mathbf{s}_t$, images $\mathbf{c}_t$, and actions $\mathbf{a}_{t:t+k}$) and output a predicted trust score ($\hat{r}_t$).

Figure 8.8 illustrates the overall architecture, highlighting the data flow between these components. Distinct from the original ACT framework, Trust-ACT incorporates the desired trust level as an additional conditioning input. Trust is represented as a categorical variable by binning the continuous trust scores (provided by a human observer via the Trusty device during

demonstration, see Section 8.3.2) into discrete classes (e.g., low, medium, high). Each trust class r_t is mapped to a learnable embedding vector. This trust embedding is then concatenated with the visual features (from the ViT-encoded camera images c_t) and the robot state s_t , providing the decoder with explicit information about the desired behavioral style associated with that trust level.

Formally, the encoder $q_\phi(\mathbf{z}|\mathbf{a}_{t:t+k}, \mathbf{s}_t, \mathbf{r}_t)$ infers a latent variable $\mathbf{z}$ from the demonstrated actions, state, and the corresponding trust class. The policy decoder $\pi_\theta(\hat{\mathbf{a}}_{t:t+k}|\mathbf{z}, \mathbf{c}_t, \mathbf{s}_t, \mathbf{r}_t)$ then predicts the action sequence conditioned on the latent variable $\mathbf{z}$, the visual inputs $\mathbf{c}_t$, the robot state $\mathbf{s}_t$, and the embedded desired trust label $\mathbf{r}_t$. We use a pre-trained ViT backbone (Dasari et al., 2023) to embed camera images. The model is trained using a reconstruction loss $\mathcal{L}_{\text{rec}}$ (MSE between predicted and demonstrated actions) and a Kullback-Leibler (KL) divergence loss $\mathcal{L}_{\text{KL}}$ between the learned latent distribution $q_\phi(\mathbf{z}|\cdot)$ and a prior $\mathcal{N}(0, \mathbf{I})$. Additionally, we incorporate a cosine learning rate scheduler during training. The total loss is given by:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}}, \quad (8.1)$$

where β is a weighting term for the KL loss, adjusted via KL annealing (discussed in Section 8.3.3.1) to help preserve the latent space structure.

8.3.3.1 Preventing Posterior Collapse

A common challenge in conditional generative models such as our CVAE-based policy is posterior collapse, where the decoder ignores the latent variable $\mathbf{z}$ because the rich conditional inputs (images, states, trust) provide sufficient information for deterministic reconstruction. In our preliminary experiments with the vanilla ACT framework, this issue rendered the latent space ineffective for generating diverse behaviors.

To address this, we introduce a partial masking scheme during training. Specifically, for a fraction $p\%$ of the samples in each training batch, the conditional inputs ($\mathbf{c}_t$, $\mathbf{s}_t$, and $\mathbf{r}_t$) are masked (set to zero). This forces the decoder to rely more heavily on the latent variable $\mathbf{z}$ for reconstruction, thereby preserving a diverse and meaningful latent space essential for our sampling strategy. In addition, KL annealing (Fu et al., 2019) is employed, gradually increasing the KL weight β during training to further mitigate posterior collapse. Empirically, we found that setting $p = 20\%$ effectively balances latent space preservation and stable training. The full training procedure with the partial masking strategy is detailed in Algorithm 1.

Algorithm 1: Trust-ACT with Partial Masking

Require: Demonstration dataset $\mathcal{D}$, chunk size k , KL weight β , masking probability p

- 1: Initialize encoder $q_\phi(\mathbf{z}|\mathbf{a}_{t:t+k}, \mathbf{s}_t, \mathbf{r}_t)$
- 2: Initialize decoder $\pi_\theta(\hat{\mathbf{a}}_{t:t+k}|\mathbf{z}, \mathbf{c}_t, \mathbf{s}_t, \mathbf{r}_t)$
- 3: **for** $n = 1, 2, \dots$ **do**
- 4: Sample $\mathbf{a}_{t:t+k}, \mathbf{r}_t, \mathbf{s}_t, \mathbf{c}_t$ from $\mathcal{D}$
- 5: Sample $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{a}_{t:t+k}, \mathbf{s}_t, \mathbf{r}_t)$
- 6: Draw $m \sim \mathcal{U}(0, 1)$. If $m < p$, then set $\mathbf{c}_t = \mathbf{o}, \mathbf{s}_t = \mathbf{o}, \mathbf{r}_t = \mathbf{o}$
- 7: Predict $\hat{\mathbf{a}}_{t:t+k} \sim \pi_\theta(\hat{\mathbf{a}}_{t:t+k}|\mathbf{z}, \mathbf{c}_t, \mathbf{s}_t, \mathbf{r}_t)$
- 8: $\mathcal{L}_{\text{rec}} = \text{MSE}(\hat{\mathbf{a}}_{t:t+k}, \mathbf{a}_{t:t+k})$
- 9: $\mathcal{L}_{\text{KL}} = \text{D}_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{a}_{t:t+k}, \mathbf{s}_t, \mathbf{r}_t) \parallel \mathcal{N}(\mathbf{o}, \mathbf{I}))$
- 10: Update ϕ, θ with Adam using $\mathcal{L} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}}$
- 11: **end for**

8.3.3.2 *Trust Prediction and Best-of- n Sampling*

Well-trained generative policies like the CVAE used in Trust-ACT produce varied outputs for the same input condition due to sampling from the latent space $\mathbf{z}$. While this allows for diverse behaviors, it also means that not all generated trajectories will be equally effective or reliable. A single suboptimal sample could lead to task failure. To mitigate this and enhance performance, Trust-ACT employs a best-of- n sampling strategy at inference time, guided by the independently trained Trust Prediction Model (v_η).

The rationale for using the trust predictor v_η as the selection criterion stems directly from our data collection methodology (Section 8.3.2). The policy π_θ learns to generate behaviors matching discrete trust classes (Low, Medium, High). However, the trust predictor v_η is trained on the continuous trust scores $[0, 1]$ recorded via Trusty, which captured performance nuances even within those discrete bins. For instance, smoother, faster, more successful demonstrations within the ‘High’ class were rated closer to 1.0 by the human observer.

Therefore, v_η learns to map observable trajectory features (actions $\mathbf{a}_{t:t+k}$, state $\mathbf{s}_t$, vision $\mathbf{c}_t$) to a continuous score reflecting these performance-related characteristics valued by the human. A higher predicted score $\hat{r}_t$ from v_η indicates that a generated trajectory more closely resembles the successful, efficient, and stylistically preferred demonstrations from the upper range of its corresponding trust class.

The best-of- n process works as follows:

1. For the current state $(\mathbf{s}_t, \mathbf{c}_t)$ and desired trust class $(\mathbf{r}_t)$, sample N latent codes $\mathbf{z}^{(i)} \sim \mathcal{N}(0, \mathbf{I})$.
2. Generate N candidate action trajectories $\hat{\mathbf{a}}_{t:t+k}^{(i)}$ using the policy decoder $\pi_\theta(\cdot | \mathbf{z}^{(i)}, \mathbf{c}_t, \mathbf{s}_t, \mathbf{r}_t)$.
3. Evaluate each candidate trajectory $\hat{\mathbf{a}}^{(i)}$ using the trust prediction model v_η to get a predicted continuous trust score $\hat{r}_t^{(i)} = v_\eta(\hat{\mathbf{a}}_{t:t+k}^{(i)} | \mathbf{s}_t, \mathbf{c}_t)$.
4. Select the trajectory $\hat{\mathbf{a}}^*$ with the highest predicted trust score $\hat{r}_t^{(i)}$ for execution:

$$\hat{\mathbf{a}}_{t:t+k}^* = \arg \max_{\hat{\mathbf{a}}_{t:t+k}^{(i)}} \hat{r}_t^{(i)}. \quad (8.2)$$

In essence, the predicted continuous trust score acts as a learned proxy for performance quality and stylistic adherence within the target class. This selection strategy leverages the fine-grained information embedded in the continuous trust ratings from demonstrations to filter samples and choose the execution predicted to be most performant and reliable, aligning with the implicit human preferences captured during training. As shown in our results (Section 8.3.4), this improves task success rates over single-sample execution.

8.3.4 Experiment Results

In this section, we present the results of the Trust-ACT evaluation on the block stacking task, focusing on trajectory characteristics, execution times, and success rates across varying trust conditions and sampling strategies.

8.3.4.1 Trust Prediction Performance

We demonstrate the effectiveness of our trust prediction model in Figure 8.9, comparing measured and predicted trust values throughout task execution. The predicted trust successfully tracks the overall trends in measured trust, particularly during stable periods, with smoother transitions between states. The model’s ability to identify distinct trust regimes and capture fine-grained trust variations within these regimes enables us to effectively select trajectories in our best-of- n sampling strategy.

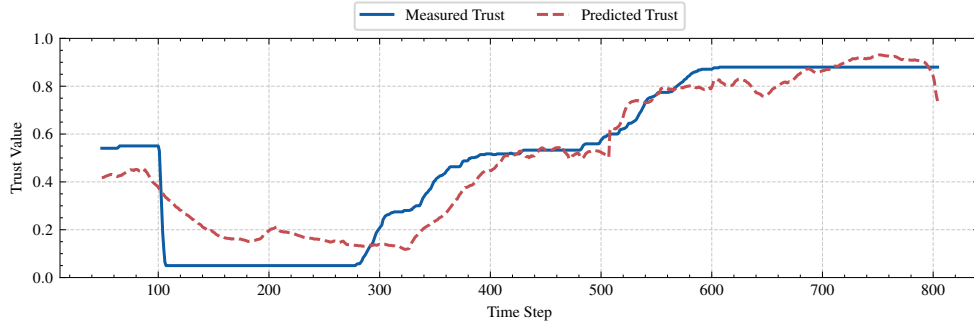


Figure 8.9: Evaluation of the trust prediction model on a held-out validation episode, comparing measured trust (solid blue line) and predicted trust (dashed red line). The model captures three distinct trust levels: low trust (0.0–0.3), moderate trust (0.3–0.6), and high trust (0.6–1.0). The model aligns well with observed trust values across different trust states. During inference, the trust prediction model uses predicted actions to evaluate and select the best trajectory based on trust scores.

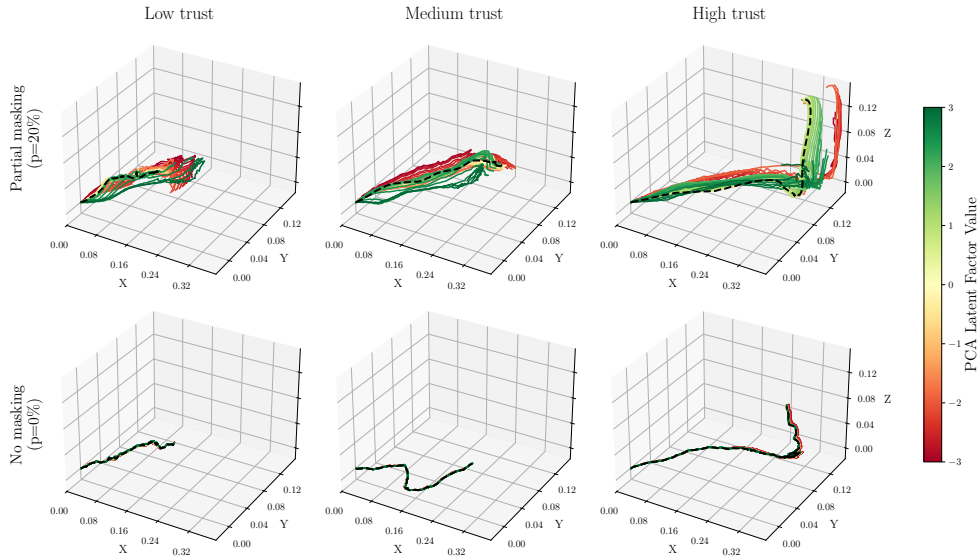


Figure 8.10: Visualization of latent space trajectories across trust levels, showing predicted task sequences of 100 time steps. We first perform PCA on the encoder’s latent means over the dataset to identify principal directions of variation. Trajectories are then generated by traversing the latent space along the first principal component from near the origin. Colors indicate the position along this traversal from start (red) to end (green). The top row shows trajectories with partial masking ($p=20\%$), demonstrating diverse behavior patterns, while the bottom row ($p=0\%$) shows collapsed behavior without masking. The masking strategy effectively preserves meaningful latent space structure.

Table 8.3: Performance comparison of single vs repeated sampling (RS) strategies across robotic trust levels (N=30 trials/condition). For each trust level: success rates shown with and without RS, while execution times (mean $\pm$ SD) and failure modes [grasps (G), stacks (S), drops (D)] are reported for the RS condition only.

Trust	Success Rate		w. Time	Failure Modes		
	w. RS	w/o. RS	Mean $\pm$ SD (s)	G	S	D
High	21/30	19/30	24.1 $\pm$ 5.6	7	2	0
Medium	20/30	18/30	39.5 $\pm$ 7.7	8	1	1
Low	20/30	17/30	51.4 $\pm$ 8.2	7	2	1

8.3.4.2 Trajectory Analysis

In Figure 8.10, we visualize the impact of our partial masking scheme on trajectory generation across trust levels. Each trajectory shows the predicted motion for the first 100 time steps of the task, demonstrating how different regions of the latent space correspond to distinct movement patterns. With masking ($p=20\%$, top row), the model generates diverse trajectories with distinct characteristics: high-trust trajectories exhibit decisive vertical movements and direct paths, while low-trust trajectories show more cautious, exploratory behavior. The medium-trust condition demonstrates intermediate characteristics, balancing between confidence and caution. Without masking ($p=0\%$, bottom row), the trajectories collapse to nearly identical paths within each trust level, demonstrating the effectiveness of our approach in maintaining a meaningful and diverse latent space, which is a prerequisite for the best-of-n sampling strategy to explore varied options.

8.3.4.3 Task Performance Analysis

Our results reveal a clear correlation between the trust level the policy was conditioned on and the resulting task execution times, as shown in Table 8.3. High-trust trajectories exhibited efficient completion times averaging 24.1 $\pm$ 5.6s, while medium and low-trust conditions resulted in progressively longer execution times of 39.5 $\pm$ 7.7s and 51.4 $\pm$ 8.2s respectively. This substantial variation (more than 2x difference between high and low trust) demonstrates that the trust conditioning effectively modulates the robot’s behavior, with high-trust settings leading to more confident and efficient movements, consistent with the styles learned from demonstrations.

8.3.4.4 Success Rates and Sampling Strategies

We evaluated the effectiveness of the best-of-n sampling approach, where the trust prediction model selects the most promising trajectory among N samples. This strategy leverages the predictor’s ability to identify trajectories resembling the high-performance demonstrations within each trust class, as discussed in Section 8.3.3.2. The results, shown in Table 8.3, demonstrate the benefit of this approach across all trust levels. Repeated sampling (RS, using best-of-n) consistently outperformed single-sample execution (w/o. RS) in terms of task success. In high-trust scenarios, we achieved peak performance with a 70% success rate (21/30) using RS, compared to 63% (19/30) without RS, while maintaining the fastest execution times. Both medium and low trust conditions demonstrated similar improvements, reaching 67% success rates (20/30) with RS, compared to their non-RS counterparts of 60% and 57% respectively. Analysis of failure modes under RS conditions revealed that grasp failures remained the most prevalent challenge across all trust levels, followed by stacking failures, while drop failures were comparatively rare. The similar distribution of failure modes across trust levels demonstrates that our trust-based selection approach primarily improves reliability by choosing better executions of the learned style, rather than fundamentally changing the types of errors encountered. It successfully modulates execution style while maintaining consistent baseline manipulation capabilities, regardless of the trust condition.

8.4 DISCUSSION & CONCLUSIONS

In this chapter, we introduced Trust-ACT, a trust-aware imitation learning framework that integrates human trust signals directly into the robot’s policy generation. Our motivation was to move beyond using trust as a purely evaluative metric and instead leverage it as a conditioning variable that can shape the robot’s motion style in real-time. We addressed three core challenges: (1) demonstrating that trust conditioning can yield distinctly different motion behaviors, (2) maintaining a meaningful latent space in a CVAE-based policy despite rich conditional inputs, and (3) leveraging trust predictions to select trajectories that optimize task success.

Our experimental results on a real-world block-stacking task showed clear differentiation between high-trust (24.1s average completion), medium-trust (39.5s), and low-trust (51.4 s) trajectories, confirming that trust conditioning enables the robot to produce distinct motion styles. We further demonstrated

how our partial masking scheme mitigates posterior collapse in the latent space, ensuring that repeated sampling can uncover truly diverse candidate trajectories. By pairing this sampling approach with a trust prediction model, we achieved up to a 70% success rate in high-trust scenarios and observed consistent improvements in medium- and low-trust conditions as well.

While our results demonstrate that trust-conditioned policies can effectively modulate behavior in a block-stacking task, our evaluation is limited to this single scenario. Future work should apply Trust-ACT to a wider variety of robotic tasks to confirm its generality and robustness in different settings. Additionally, our current mapping pairs high trust with fast, fluid movements and low trust with slower, more hesitant, jerky actions. Other mappings are possible, for instance, the robot could slow down and smooth its motions when trust drops, actively working to regain user confidence. Future studies should explore these alternatives to refine trust-conditioned behaviors in real time.

The AMIGA platform enabled data collection and validation for Trust-ACT by providing the necessary infrastructure for integrating the UR10e arm, sensor suite, and teleoperation interfaces. This integration, combined with the Trusty device for trust annotation and GELLO for demonstration collection, allowed us to gather the data required to evaluate trust-conditioned policies in our block stacking tasks.

Having demonstrated Trust-ACT’s capabilities for integrating trust into robotic policies, we now turn to Chapter 9 to synthesize the broader contributions of this thesis and explore future research directions. The following chapter will reflect on how our various contributions collectively advance trust in human-robot interaction while discussing limitations and promising avenues for future work.

CONCLUSIONS AND FUTURE WORK

9.1 SUMMARY OF CONTRIBUTIONS

In this thesis, we investigated the role of trust in Human-Robot Interaction (HRI), particularly in the context of assistive robotics, where effective collaboration between humans and robots is essential for achieving safe and reliable outcomes. Our central objective was to address gaps in the existing literature by developing and evaluating new methods to measure and predict trust in real-world human-robot scenarios. Through experimental platforms, novel measurement tools, and machine learning approaches, we advanced methods for measuring and modeling trust while demonstrating how trust signals can be integrated into robotic learning to improve HRI.

Our first contribution is an extensive review of empirical research on trust in HRI. By analyzing empirical studies, we identified patterns in how trust is studied and revealed opportunities where experimental approaches involving direct physical interaction could complement existing work. While prior research has provided valuable insights, many studies relied on passive observation or controlled environments where participants watched videos or filled out surveys about robots rather than physically interacting with them. Furthermore, across diverse experimental conditions and robot platforms, performance consistently emerged as the dominant factor influencing trust. This review established the need for new experimental methods and tools, forming the basis for the work that followed.

Our second contribution centers on experimental platforms designed to study trust through hands-on human-robot collaboration. The autonomous wheelchair platform immersed participants in a real-world navigation task where they actively depended on the robot's decisions for safety and efficiency. Unlike observing videos or simulations, this setup created a tangible experience where participants gathered rich, multi-modal feedback from the robot's behavior in a real-world environment. By experiencing the wheel-

chair’s navigation decisions firsthand, participants developed a nuanced understanding of the system’s capabilities through direct physical interaction, reacting to its movements, errors, and successes in real time. The experimental design carefully balanced standardized conditions across trials with the natural variability and challenges of an office environment. Additionally, we designed and built AMIGA, a mobile manipulation platform that integrates a robotic arm with a wheelchair base at a fraction of the cost of commercial alternatives. Together, these platforms enabled studying trust through authentic physical interactions between humans and robots.

One key improvement in our work is the introduction of Trusty, a novel handheld device for continuous trust measurement. Trusty enables users to provide real-time trust feedback, offering a richer and more detailed view of trust than traditional discrete measurement approaches. Experimental validation demonstrated that Trusty correlates strongly with established trust metrics while providing a more detailed view of trust dynamics throughout the interaction. Importantly, Trusty also provides a practical and efficient way to collect high-quality labeled trust data, filling a critical gap in the field. Trusty offers a new way to capture continuous trust in real-world settings, addressing a gap in previous methods and offering valuable data for developing and training trust models.

Building on the data collected through these platforms, we trained a multi-modal transformer architecture for trust prediction. The model integrates diverse and complementary data streams — including video, eye gaze tracking, physiological signals, and wheelchair telemetry — to effectively capture the multifaceted nature of trust. Trust can be inferred in various ways: visual cues indicative of the robot’s performance, attention patterns indicated by eye gaze, and physiological responses to the robot’s behavior. The architecture’s ability to learn directly from these diverse data streams reduces the need for manual feature engineering while capturing subtle patterns in the data. To further enhance its capabilities, the model employs auxiliary perceptual tasks, enabling it to extract contextual features that reflect both the robot’s state and the user’s environment. This approach not only improves predictive accuracy but also demonstrates the importance of treating trust as a multi-dimensional phenomenon that relies on diverse sources of information. The inclusion of eye-gaze data, in particular, proved invaluable, highlighting how attention and focus contribute significantly to understanding trust dynamics.

Finally, we developed Trust-ACT, which integrates trust measurements into imitation learning. The approach extends the [ACT](#) framework by conditioning a Conditional Variational Autoencoder (CVAE)-based policy on trust labels

obtained through Trusty. To maintain diversity in the generated behaviors, we introduced a partial masking scheme during training that prevents posterior collapse in the latent space. We simultaneously train a trust prediction model, which processes visual input and proposed actions to evaluate trajectories. At inference time, multiple trajectories are sampled and evaluated using the trust prediction model, enabling selection of higher-performing trajectories based on predicted trust scores. This was validated using the Assistive Mobile Interactive Grasping Agent ([AMIGA](#)) platform and a block-stacking manipulation task.

In summary, we contribute to the field of trust in [HRI](#) by providing a practical, data-driven approach to studying, measuring, predicting, and integrating trust into robotic systems. By developing experimental platforms, introducing tools for continuous trust measurement, and creating trust-aware computational models, we address key challenges in understanding and applying trust in robotics.

9.2 LIMITATIONS

While we have demonstrated several advances, there are notable limitations:

- **Sample Size & Generalizability:** Our real-world experiments involved relatively small participant cohorts. Larger and more diverse populations would enhance statistical power and applicability across demographics, cultures, and user experiences.
- **Task-Specific Focus:** Our studies primarily investigated wheelchair-based navigation and block-stacking with a mobile manipulator. Trust behaviors and requirements in other assistive domains (e.g., meal assistance, home healthcare) may differ significantly.
- **Laboratory-Like Settings:** While our environments were more realistic than pure simulation, they still involved semi-controlled spaces (offices, labs). True in-the-wild deployment in highly dynamic settings could produce different trust outcomes or sensor interferences.
- **Model Evaluation Scope:** While our initial results are promising, future work would benefit from additional baseline comparisons and ablation studies. Expanding the evaluation to include classical methods and investigating the contribution of different architectural components could provide valuable insights. This expanded validation would require additional computational resources and demonstration data.

9.3 FUTURE DIRECTIONS

Based on these findings and limitations, we outline several promising avenues for future research:

- **Trust-Based Metric Design** The Trusty device could be leveraged to design and validate trust-aligned performance metrics across different robotic domains. By correlating continuous trust measurements with various objective metrics, researchers can identify which quantitative measures best reflect user preferences and experiences. This approach could be particularly valuable in domains where standard performance metrics may not fully capture the quality of human-robot interaction.
- **Trust Prediction in Navigation** The multi-modal video transformer architecture developed for trust prediction could be extended to autonomous and semi-autonomous navigation tasks. One promising direction is to investigate the relationship between eye-gaze patterns and trust in driving scenarios, potentially leading to the development of objective metrics based on attention dynamics. This could provide real-time assessment of user trust during autonomous vehicle operation.
- **Scaling Trust-ACT** Current implementations of Trust-ACT are limited in both task diversity and dataset size. Future work should focus on expanding the approach to a broader range of manipulation tasks and collecting larger demonstration datasets. This scaling would help validate the generalizability of trust-conditioned policies and potentially reveal task-specific patterns in trust-performance relationships.
- **Interactive Learning and Preference-Based Methods:** Adding interactive or preference-based approaches (e.g., humans provide feedback on preferred trajectory styles) could further refine trust-conditioned policies and align them with user values.

9.4 CONCLUDING THOUGHTS

Throughout this thesis, we have explored various dimensions of trust in [HRI](#), particularly in the context of assistive robotics. Our work has advanced methods for measuring and predicting trust, from traditional questionnaires to real-time, continuous tracking with tools like Trusty. These efforts were integrated into machine learning models to incorporate trust measurements into robotic processes.

However, our findings highlight a crucial insight: while detailed models of trust are valuable for enhancing collaboration, reliable performance remains the foundation for trust-building. Our experiments demonstrated that consistent and dependable robot behavior was key to gaining trust, while poor performance quickly undermined it. This confirms what prior research has shown: dependable performance is the essential first step in establishing trust in robots, particularly in assistive environments.

Trust, ultimately, is not "designed" — it is earned through repeated demonstrations of competence. It appears that a substantial portion of the effort in creating trustworthy robots, perhaps a large initial focus, should be on ensuring the robot performs its intended tasks consistently and effectively. Only after robots have shown consistent reliability can trust models effectively improve collaboration. Future research should focus on creating robust, reliable systems that work as expected and clearly communicate their actions. This approach will help develop robots that users can trust and rely on as helpful partners in everyday life.



HUMAN-ROBOT INTERACTION TRUST QUESTIONNAIRES

This appendix presents the TOAST, TiA, and TPS questionnaires that we used to assess key trust dimensions in [HRI](#), as discussed in Chapters 5 and 6.

To investigate the dimension of trust in [HRI](#), we conducted experiments in a real-world setting. In these experiments, adult participants interacted with robots within our research lab. In accordance with established methodologies in [HRI](#) and trust research, we evaluated the findings using questionnaires and verbal self-reports from the participants. These tools were designed to measure key factors, including performance-based trust and inclination towards trust, among other variables. During the experiment, the participants filled out these questionnaires on a computer using Microsoft Forms.

TRUST IN AUTOMATED SYSTEMS TEST QUESTIONNAIRE

The Trust in Automated Systems Test ([TOAST](#)) scale, developed by [Wojton et al. \(2020\)](#), is designed to measure two key dimensions of trust in automated systems: understanding the system and system performance. The scale underwent rigorous testing for model fit and was validated against existing trust metrics. It consists of nine survey questions that are sorted into categories based on these two dimensions, as shown below in the table [A.1](#):

Table A.1: Trust in Automated Systems Test Questionnaire by [Wojton et al. \(2020\)](#)

Item	Scale
I understand what the system should do.	System Understanding
I understand the limitations of the system.	System Understanding
I understand the capabilities of the system.	System Understanding
I understand how the system executes tasks.	System Understanding
The system helps me achieve my goals.	System Performance
The system performs consistently.	System Performance
The system performs the way it should.	System Performance
I am rarely surprised by how the system responds.	System Performance
I feel comfortable relying on the information provided by the system.	System Performance

TRUST IN AUTOMATION QUESTIONNAIRE

The Trust in Automation (TIA) scale, introduced by Körber (2019), encompasses five distinct subscales: Trust, Reliability or Competence, Understanding, Intention of Developers and Familiarity. These subscales are made up of two to four items each, totalling 19 items. While previous research has often concentrated on the aspects of reliability and trust, each subscale can also be examined separately. Additionally, an aggregated score from all 19 items can serve as a comprehensive measure of trust. Table A.2 categorizes the items under their respective subscales.

Table A.2: Trust in Automation Questionnaire by Körber (2019)

Item	Scale
The system is capable of interpreting situations correctly.	Reliability/Competence
The system works reliably.	Reliability/Competence
A system malfunction is likely.*	Reliability/Competence
The system is capable of taking over complicated tasks.	Reliability/Competence
The system might make sporadic errors.*	Reliability/Competence
I am confident about the system's capabilities.	Reliability/Competence
The system state was always clear to me.	Understanding/Predictability
The system reacts unpredictably.*	Understanding/Predictability
I was able to understand why things happened.	Understanding/Predictability
It's difficult to identify what the system will do next.*	Understanding/Predictability
I already know similar systems.	Familiarity
I have already used similar systems.	Familiarity
The developers are trustworthy.	Intention of Developers
The developers take my well-being seriously.	Intention of Developers
One should be careful with unfamiliar automated systems.*	Propensity to Trust
I rather trust a system than I mistrust it.	Propensity to Trust
Automated systems generally work well.	Propensity to Trust
I trust the system.	Trust in Automation
I can rely on the system.	Trust in Automation

* indicates inverse item

TRUST PERCEPTION SCALE QUESTIONNAIRE

The Trust Perception Scale (TPS) by Schaefer (2016) is designed for robot interactions, capturing trust levels both before and after user engagement. Unlike traditional scales that rely on self-reported attitudes, the TPS asks participants to estimate the likelihood of a robot meeting specific criteria. The scale consists of 40 items and also includes a 14-item subset for more targeted research. We utilized the 14-item subscale, as shown in A.3.

Table A.3: Trust Perception Scale by [Schaefer \(2016\)](#)

	What % of the time will this robot be...										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1. Dependable	0	0	0	0	0	0	0	0	0	0	0
2. Reliable	0	0	0	0	0	0	0	0	0	0	0
3. Unresponsive	0	0	0	0	0	0	0	0	0	0	0
4. Predictable	0	0	0	0	0	0	0	0	0	0	0

	What % of the time will this robot ...										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
5. Act consistently	0	0	0	0	0	0	0	0	0	0	0
6. Function successfully	0	0	0	0	0	0	0	0	0	0	0
7. Malfunction	0	0	0	0	0	0	0	0	0	0	0
8. Provide feedback	0	0	0	0	0	0	0	0	0	0	0
9. Have errors	0	0	0	0	0	0	0	0	0	0	0
10. Meet the needs of the mission/task	0	0	0	0	0	0	0	0	0	0	0
11. Provide appropriate information	0	0	0	0	0	0	0	0	0	0	0
12. Communicate with people	0	0	0	0	0	0	0	0	0	0	0
13. Perform exactly as instructed	0	0	0	0	0	0	0	0	0	0	0
14. Follow directions	0	0	0	0	0	0	0	0	0	0	0

VERBAL-SELF TRUST REPORT

We used a self-reported verbal trust measure in our study to continuously assess the participants' level of trust. Participants were asked the specific question:

How much did you trust the robot to navigate you through this section?

Responses were captured on a 1 to 7 Likert scale. This approach is consistent with established methodologies in the domain of trust research. Custom self-report scales, often featuring a single item on a Likert or sliding scale, are commonly used for measuring trust as they can be adapted to various experimental settings [Kohn et al. \(2021\)](#).

TRAINING HYPERPARAMETERS FOR MULTI-MODAL TRANSFORMER

This appendix lists the training and augmentation hyperparameters for the multi-modal transformer that we trained for predicting trust in Chapter 7.

AUGMENTATIONS

Table B.1: Training data augmentation parameters. This table lists the probability of application and specific parameters for each augmentation used to enhance the robustness of the model against variations in input data.

Augmentation	Probability	Parameters
CropAndPad	0.2	Percent: 0.05
Blur	0.2	Limit: 4
Perspective	0.2	Scale: [0.05, 0.1]
Brightness	0.2	Limit: [-0.2, 0.2]
Sharpen	0.2	Alpha: [0.2, 0.5], Lightness: [0.5, 1.0]
Affine	0.2	Translate: [-0.1, 0.1]
Emboss	0.2	Alpha: [0.2, 0.5], Strength: [0.2, 0.7]
HueSaturationValue	0.2	Hue Shift Limit: 20, Sat Shift Limit: 30, Val Shift Limit: 20

TRAINING HYPER-PARAMETERS

Table B.2: Model Training Hyperparameters

HYPERPARAMETER	VALUE
GENERAL	
Batch Size	12
Training Epochs	35
Sequence Length	200
Stride	100
Downsample Factor	2
OPTIMIZER	
Type	AdamW
Initial Learning Rate	0.001
Weight Decay	0.01
Betas	(0.9, 0.999)
Eps	1e-08

Table B.2: Model Training Hyperparameters (continued)

HYPERPARAMETER	VALUE
LEARNING RATE SCHEDULER	
Scheduler Type	CosineAnnealingLR
T max	35
Eta min	0.00001
Warmup Steps	500
VISION MODEL	
Model	ViT
Pretrained	TRUE
Image Size Input	224
Number of Channels	3
Patch Size	16
Hidden Size	768
Hidden Layers	12
Number Attention Heads	12
Intermediate Size	3072
Hidden Activation	Gelu
Layer Norm Eps	1.00E-12
Initializer Range	0.02
Encoder Stride	16
Hidden Dropout Prob	0.0
Attention Dropout Prob	0.0
PROJECTOR	
Model	MLP
Number Of Layers	2
Hidden Size	256
RMS Norm Eps	1.00E-06
Hidden Activation	Relu
SEQUENCE MODEL	
Number Of Blocks	4
Intermediate Size	512
Hidden Size	128
Number Attention Heads	8
Number Key Value Heads	8
Max Position Embeddings	200
Hidden Activation	Silu
Initializer Range	0.02
RMS Norm Eps	1.00E-06
Attention Dropout	0.0
TASK HEAD	
Model	MLP
Number Of Layers	1
Hidden Size	256
RMS Norm Eps	1.00E-06
FUSION MODULE	
Model	MLP
Number Of Layers	1
Hidden Size	256
RMS Norm Eps	1.00E-06



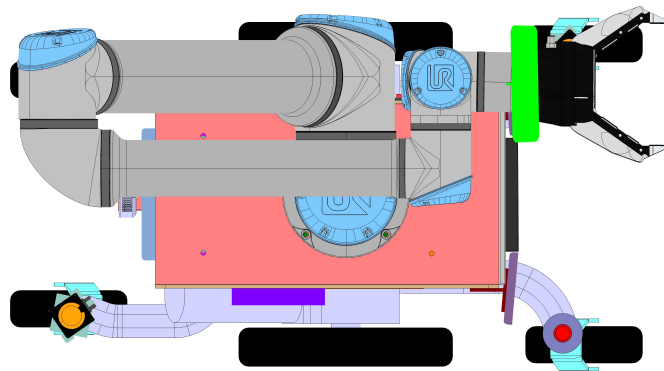
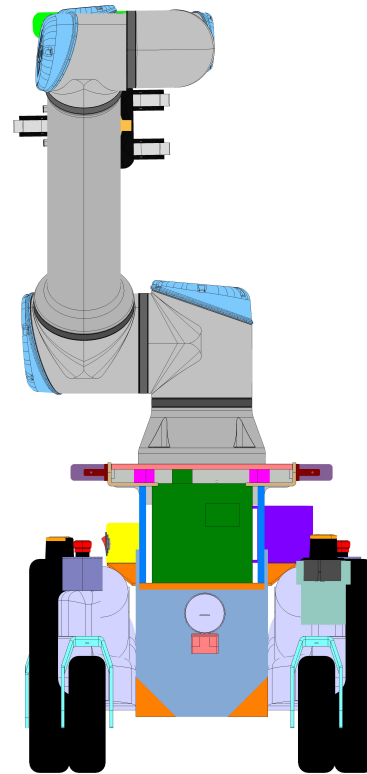
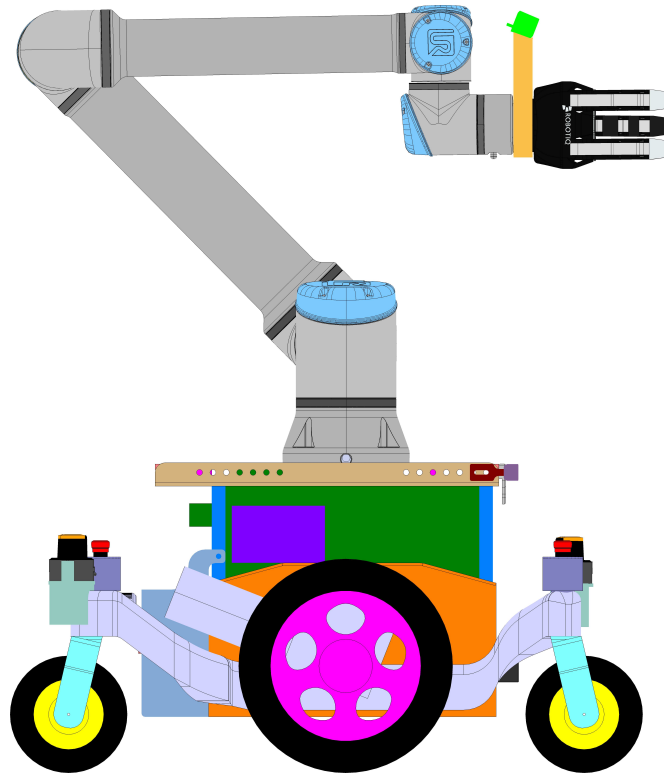
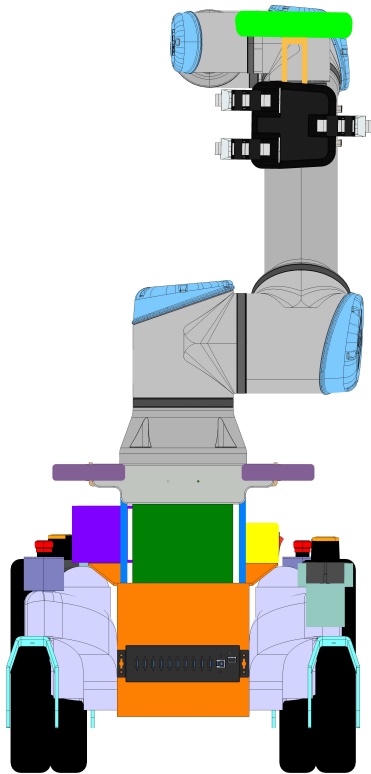
AMIGA SYSTEM DRAWINGS

In this appendix, we present the high-level technical drawings of the [AMIGA](#) mobile manipulation platform. These drawings directly support the system design and integration described in Chapter 8 (see Section 8.2.3 for details on the software and control architecture). In particular, we used the assembly drawing (original A1 format) and the exploded view (original A2 format) to create the URDF model for simulation, and these guided our implementation decisions in the main text.

The following pages include:

- An assembly drawing showing front, side, and top views of [AMIGA](#).
- An exploded reference view that details the system integration and major components.

We have scaled the drawings to A4 size while maintaining aspect ratios.



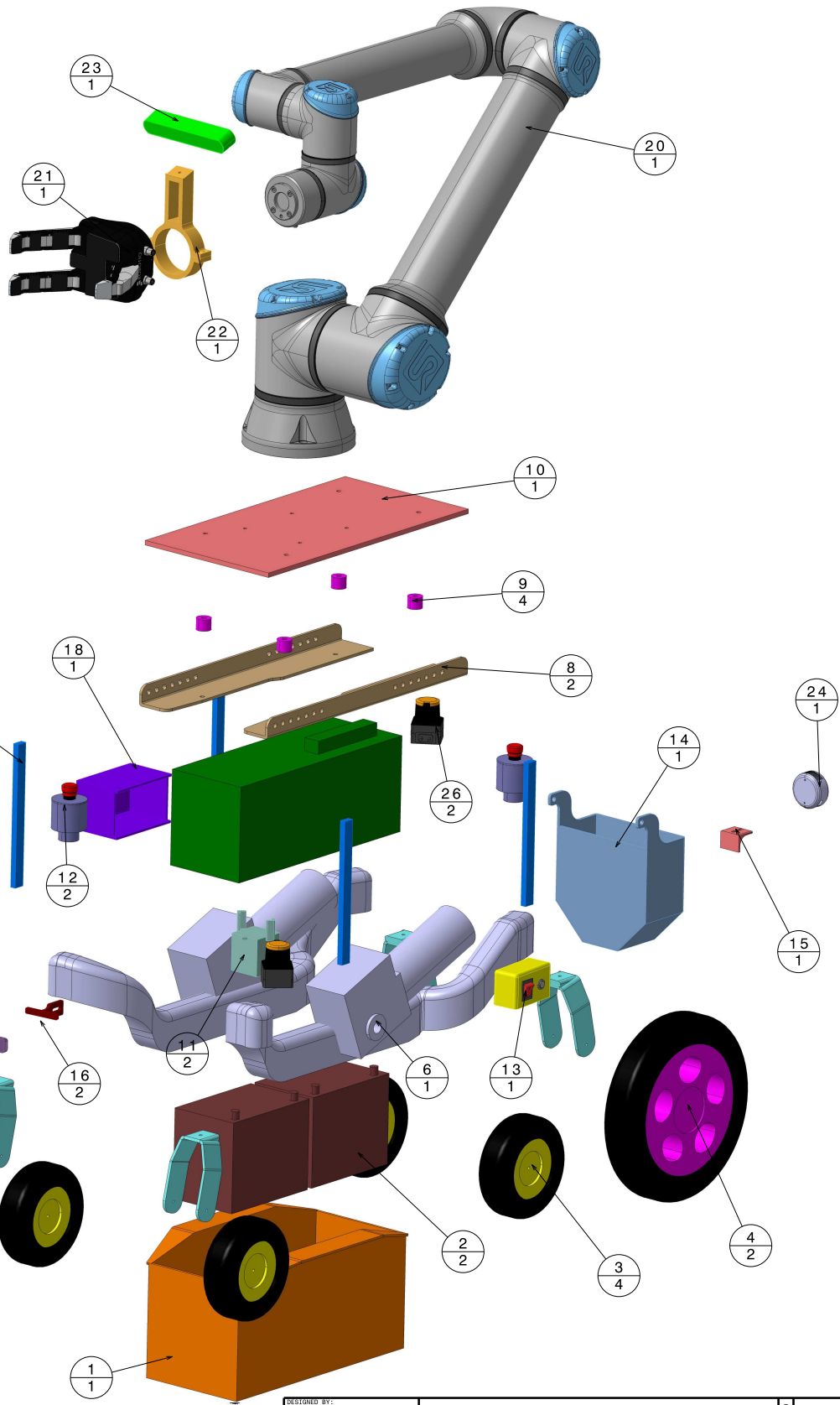
Assembly drawing without dimensions

Do not use for manufacturing

DESIGNED BY: Nico Lingg		Assembly Drawing		I	—
DATE: 2024-12-02				H	—
CHECKED BY:				G	—
DATE:				F	—
SIZE: A1		 AMIGA		E	—
SCALE: 1:4				D	—
REVISION: XXX		DRAWING NUMBER: XXX		C	—
VARIATION NUMBER:		SHEET: 1/1		B	—
A		B		A	—

This drawing is our property. It can't be reproduced or communicated without our written agreement.

Pos	Name	QTY
1	Battery Enclosure	1
2	12 V Battery	2
3	Caster Wheels	4
4	Power Drive Wheels	2
5	Castors	4
6	Frame inc. Motor	1
7	Extention Bar	4
8	Mounting Bar	2
9	Standoffs	4
10	Robot Mounting Plate	1
11	Lidar Mount	2
12	Emergency Button Enclosure	2
13	Robot On/Off Enclosure	1
14	Storage	1
15	Camera Mount Back	1
16	Camera Mount Front	2
17	Front Panel	1
18	Circuit Breaker Enclosure	1
19	Powered USB Hub	1
20	UR10e Robot Arm	1
21	Robotiq 3 Finger Gripper	1
22	Camera Mount Wrist	1
23	ZED Camera	1
24	Intel Realsense L515	1
25	Intel Realsense D415	2
26	Hokuyo Lidar	2



DESIGNED BY:	Nico Lingg	Exploded View Drawing		I	-
DATE:	2024-12-03			H	-
CHECKED BY:	XXX	AMIGA		G	-
DATE:	XXX			F	-
SIZE	A2	1:5		E	-
SCALE	1:5			D	-
WEIGHT (kg)	XXX	XXX		C	-
DRAWING NUMBER	XXX			B	-
1 / 1				A	-

FOR REFERENCE ONLY - NOT FOR MANUFACTURING

TRAINING HYPERPARAMETERS FOR TRUST-ACT

In this chapter, we detail the hyperparameters we used in training the Trust-ACT model described in chapter 8, including both image augmentation strategies and model architecture configurations.

IMAGE AUGMENTATION CONFIGURATION

We applied the following augmentations to enhance the robustness of the model’s visual processing capabilities. These augmentations were carefully selected to maintain the semantic validity of the robot’s visual inputs while introducing meaningful variations.

Table D.1: Image Augmentation Parameters and Their Configuration Values

Category	Parameter	Value	Probability
Global Settings	Interpolation	bilinear	-
	Image Size	[224, 224]	-
	Normalization Mean	[0.485, 0.456, 0.406]	-
	Normalization Std	[0.229, 0.224, 0.225]	-
Camera Augmentations	RandomBrightnessContrast	Brightness: [0.7, 1.3] Contrast: [0.7, 1.3]	0.5
	GaussianBlur	Blur Limit: [3, 7] Sigma Limit: [0.1, 2.0]	0.4
	Affine	Scale: [0.9, 1.1] Translate: [0.0, 0.1] Rotate: [-10, 10]	0.4
	HueSaturationValue	Hue Shift: 0.1 Saturation Shift: 0.1 Value Shift: 0.1	0.3
	Rotate	[-10, 10]	0.3

MODEL ARCHITECTURE AND TRAINING CONFIGURATION

The model architecture and training parameters were chosen to optimize for both performance and computational efficiency. The configuration reflects a balance between model capacity and training stability.

Table D.2: Model Architecture and Training Configuration

HYPERPARAMETER	VALUE
ARCHITECTURE	
Backbone Model	ViT
Hidden Dimension	256
Latent Dimension	8
Feedforward Dimension	2048
Number of Encoder Layers	4
Number of Decoder Layers	4
Number of Attention Heads	8
Dropout Rate	0.1
Number of Action Steps	100
Mask p	0.2
TRAINING CONFIGURATION	
Optimizer	AdamW
Learning Rate (Transformer)	3e-4
Learning Rate (Backbone)	3e-4
Weight Decay	1e-4
Batch Size (Train)	64
Batch Size (Val)	64
Number of Epochs	500
Train Split	0.8
LEARNING RATE SCHEDULE	
Scheduler Type	warmup + cosine
Warmup Steps	1750
Backbone Min LR Ratio	0.01
Transformer Min LR Ratio	0.1
Backbone Warmup Start Ratio	0.001
Transformer Warmup Start Ratio	0.01
KL SCHEDULE	
KL Scheduler Type	cyclic
Min KL Loss Weight	0.001
Max KL Loss Weight	0.07
Epochs per Cycle	50

IPAD CONTROLLER INTERFACE PROTOTYPE

In this appendix, we present an early prototype iPad application that we designed to enhance [HRI](#) for the ARTA wheelchair platform. While this interface demonstrated potential paths for improving user engagement and situational awareness, we later superseded this work with the use of mobile eye-tracking glasses (see Chapters 4 and 7), which provided more comprehensive gaze data collection capabilities that proved valuable for trust assessment as described in the main chapters.

The prototype interface included two primary modes. The basic control mode, as shown in figure [E.1](#), provided a virtual joystick for manual wheelchair control and incorporated real-time gaze tracking through ARKit. While this offered initial insights into user attention patterns, the subsequent use of dedicated eye-tracking glasses enabled more precise and reliable gaze data collection that better supported our trust research goals.

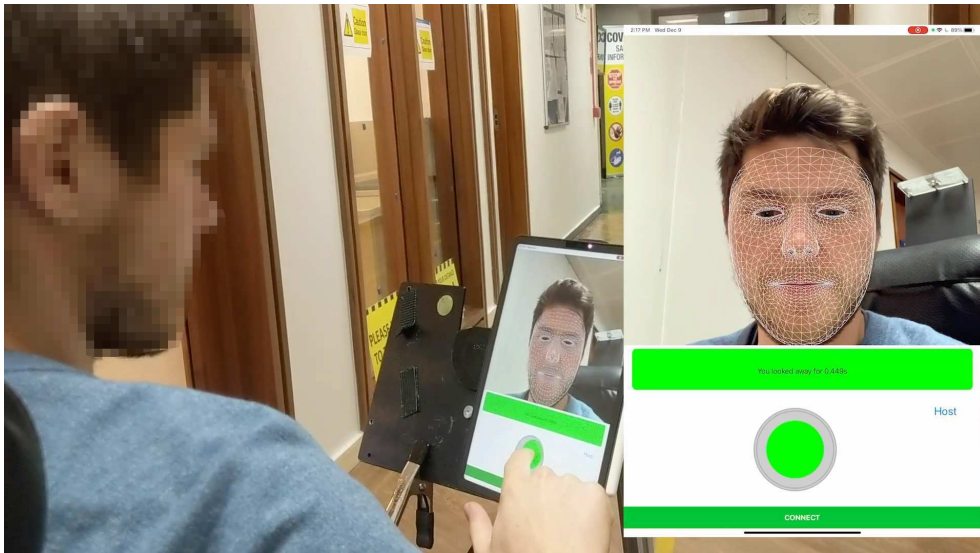


Figure E.1: Basic control mode with face tracking and virtual joystick control.



Figure E.2: Enhanced navigation mode with environmental mapping and object detection.

The enhanced navigation mode (figure E.2) offers an augmented view by combining live video from the wheelchair’s camera with object detection and mapping. The iPad screen displays a map of the environment, showing detected objects and the wheelchair’s position. This mode is intended to give users a more comprehensive understanding of their surroundings and the robot’s “view” of the environment, which could support more informed navigation decisions.¹

Built with Swift, UIKit, and ARKit, the application communicates with the Robot Operating System (ROS) ecosystem through ROSSwift [Gustafsson \(2020\)](#), a Swift library that enables WebSocket-based interaction with the ROS master node. Object detection (YOLOv5), mapping, and control of the robot are handled directly on the robot using the ROS navigation stack, with relevant data streamed to the iPad app for display and interaction.

The implemented features demonstrate the potential of mobile interfaces to enhance human-robot interaction through intuitive control and real-time environmental feedback. While the ARKit-based gaze tracking explored here was eventually replaced by dedicated eye-tracking glasses in our main studies, this early interface work contributed to our understanding of how visual feedback and attention monitoring could support trust development in human-robot interactions.

¹ Video demonstrations of the interface are available on YouTube: [Basic Control Mode](#), [Enhanced Navigation Mode](#), and [App Screen Recording](#).

REPOSITORY AND VIDEO RESOURCES

GITHUB REPOSITORIES

The following repositories contain the implementation code for key components of this thesis:

Table F.1: GitHub repositories associated with key thesis components. Each repository corresponds to a key contribution described in specific thesis chapters, providing code, documentation, and research artifacts.

Repository	Description
trust-in-hri-review-data	Dataset and classification frameworks used in Chapter 3 for analyzing 689 HRI trust studies (2015-2024). Includes data processing scripts and documentation of the review methodology.
trusty_trust_recorder	Hardware design files and ROS package for the Trusty continuous trust measurement device introduced in Chapter 6. Contains Arduino firmware, ROS nodes for data acquisition, and assembly instructions.
multi-modal-video-transformer	Implementation of the multi-modal transformer architecture from Chapter 7 for trust prediction. Features transformer models for fusing video, gaze, and physiological data streams.
trust_act	Trust-ACT framework developed in Chapter 8 for trust-aware imitation learning. Includes CVAE policy model, trust predictor, and sampling infrastructure for the AMIGA platform.

Each repository provides documented implementations of key thesis components, ranging from data analysis tools to deep learning models. The code includes instructions and examples to support future research and extensions.

VIDEO DEMONSTRATIONS

The following videos highlight demonstrations, setups, and technologies developed in this thesis:

Table F.2: Video demonstrations showcasing experimental platforms, research results, and interface prototypes described in this thesis.

Topic	Description
Wheelchair Performance Comparison	A side-by-side comparison of the wheelchair navigation behavior under good and bad performance profiles during the first checkpoint segment. Highlights the influence of reliability on trust.
Autonomous Navigation System	A complete demonstration of the autonomous wheelchair system navigating the experimental office environment, showcasing real-world obstacles and key decision points.
MoveIt Control on AMIGA	Demonstrates the integration of ROS and MoveIt for collision-free motion planning and manipulation tasks on the AMIGA manipulator platform.
Apple Vision Pro Interface	An experimental control interface using Apple Vision Pro's hand tracking capabilities for direct robot teleoperation, emphasizing intuitive and flexible manipulation control.
GELLO-Based Teleoperation	A demonstration of the GELLO-based control interface, allowing precise, scaled physical manipulation of AMIGA using a kinematically equivalent scaled robot arm.
Trust-ACT Robot Trajectories	Visualization of robot trajectories generated under different trust levels (low, medium, high), illustrating variations in movement confidence and caution during stacking tasks.
<i>Prototype Interface Demonstrations (Appendix E)</i>	
iPad Basic Control	Demonstration of the iPad interface's basic control mode with face tracking and virtual joystick.
iPad Enhanced Navigation	Showcase of the enhanced navigation mode featuring environmental mapping and object detection integration.
iPad Interface Overview	Complete screen recording of the iPad application demonstrating all features and modes.

BIBLIOGRAPHY

- Abd, M. A., Gonzalez, I., Nojournian, M. and Engeberg, E. D. (2017), Trust, satisfaction and frustration measurements during human-robot interaction, in 'Proceedings of the 30th Florida Conference on Recent Advances in RoboticsMay', pp. 11–12. (Cited on page 65.)
- Adamik, M., Dudzinska, K., Herskind, A. J. and Rehm, M. (2021), The difference between trust measurement and behavior: Investigating the effect of personalizing a robot's appearance on trust in hri, in '2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)', IEEE, pp. 880–885. doi:10.1109/ro-man50785.2021.9515487 (Cited on page 65.)
- Ahmad, M. I., Bernotat, J., Lohan, K. and Eyssel, F. (2019), 'Trust and cognitive load during human-robot interaction', *arXiv* . doi:10.48550/arxiv.1909.05160 (Cited on page 65.)
- Ajenaghughrure, I. B., Sousa, S. C., Kosunen, I. J. and Lamas, D. (2019), Predictive model to assess user trust: a psycho-physiological approach, in 'Proceedings of the 10th Indian Conference on Human-Computer Interaction - IndiaHCI '19', ACM, pp. 1–10. doi:10.1145/3364183.3364195 (Cited on page 40.)
- Akash, K., Hu, W.-L., Jain, N. and Reid, T. (2018), 'A classification model for sensing human trust in machines using eeg and gsr', *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8(4), 1–20. doi:10.1145/3132743 (Cited on page 65.)
- Akash, K., Jain, N. and Misu, T. (2020), Toward adaptive trust calibration for level 2 driving automation, in 'Proceedings of the 2020 international conference on multimodal interaction', ACM, pp. 538–547. doi:10.1145/3382507.3418885 (Cited on page 65.)
- Albayram, Y., Jensen, T., Khan, M. M. H., Fahim, M. A. A., Buck, R. and Coman, E. (2020), Investigating the effects of (empty) promises on human-automation interaction and trust repair, in 'Proceedings of the 8th international conference on human-agent interaction', ACM, pp. 6–14. doi:10.1145/3406499.3415064 (Cited on page 64.)

- Alhaji, B., Büttner, S., Sanjay Kumar, S. and Prilla, M. (2024), 'Trust dynamics in human interaction with an industrial robot', *Behaviour & Information Technology* pp. 1–23. doi:10.1080/0144929X.2024.2316284 (Cited on page 65.)
- Aliasghari, P., Ghafurian, M., Nehaniv, C. L. and Dautenhahn, K. (2023), 'How do we perceive our trainee robots? exploring the impact of robot errors and appearance when performing domestic physical tasks on teachers' trust and evaluations', *ACM Transactions on Human-Robot Interaction* 12(3), 1–41. doi:10.1145/3582516 (Cited on page 65.)
- Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. and Schmid, C. (2021), Vivit: A video vision transformer, in 'Proceedings of the IEEE/CVF international conference on computer vision', IEEE. doi:10.1109/iccv48922.2021.00676 (Cited on page 109.)
- Ayoub, J., Yang, X. J. and Zhou, F. (2021), 'Modeling dispositional and initial learned trust in automated vehicles with predictability and explainability', *Transportation research part F: traffic psychology and behaviour* 77, 102–116. doi:10.1016/j.trf.2020.12.015 (Cited on page 64.)
- Azevedo-Sa, H., Zhao, H., Esterwood, C., Yang, X. J., Tilbury, D. M. and Robert Jr, L. P. (2021), 'How internal and external risks affect the relationships between trust and driver behavior in automated driving systems', *Transportation research part C: emerging technologies* 123, 102973. doi:10.1016/j.trc.2021.102973 (Cited on page 65.)
- Babel, F., Kraus, J. and Baumann, M. (2022), 'Findings from a qualitative field study with an autonomous robot in public: exploration of user reactions and conflicts', *International Journal of Social Robotics* 14(7), 1625–1655. doi:10.1007/s12369-022-00894-x (Cited on page 65.)
- Ben Ajenaghughrure, I., Sousa, S. and Lamas, D. (2021), Psychophysiological Modelling of Trust in Technology: Comparative Analysis of Psychophysiological Signals, in 'Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications', SCITEPRESS - Science and Technology Publications, pp. 161–173. doi:10.5220/0010237701610173 (Cited on pages 39 and 40.)
- Bernotat, J., Eyssel, F. and Sachse, J. (2021), 'The (fe) male robot: how robot body shape impacts first impressions and trust towards robots', *International Journal of Social Robotics* 13(3), 477–489. doi:10.1007/s12369-019-00562-7 (Cited on page 65.)

- Bertasius, G., Wang, H. and Torresani, L. (2021), Is space-time attention all you need for video understanding?, in 'International Conference on Machine Learning', arXiv, pp. 813–824. doi:10.48550/arxiv.2102.05095 (Cited on pages 52, 54, and 109.)
- Betella, A. and Verschure, P. F. (2016), 'The affective slider: A digital self-assessment scale for the measurement of human emotions', *PloS one* 11(2), e0148037. doi:10.1371/journal.pone.0148037 (Cited on page 51.)
- Biermann, H., Brauner, P. and Ziefle, M. (2020), 'How context and design shape human-robot trust and attributions', *Paladyn, Journal of Behavioral Robotics* 12(1), 74–86. doi:10.1515/pjbr-2021-0008 (Cited on page 65.)
- Biros, D. P., Daly, M. and Gunsch, G. (2004), 'The Influence of Task Load and Automation Trust on Deception Detection', *Group Decision and Negotiation* 13(2), 173–189. doi:10.1023/B:GRUP.0000021840.85686.57 (Cited on pages 33 and 34.)
- Bridgwater, T., Giuliani, M., van Maris, A., Baker, G., Winfield, A. and Pipe, T. (2020), Examining profiles for robotic risk assessment: Does a robot's approach to risk affect user trust?, in 'Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction', ACM, pp. 23–31. doi:10.1145/3319502.3374804 (Cited on page 65.)
- Broekens, J. and Brinkman, W.-P. (2009), Affectbutton: Towards a standard for dynamic affective user feedback, in '3rd international conference on affective computing and intelligent interaction and workshops', IEEE, pp. 1–8. doi:10.1109/acii.2009.5349347 (Cited on page 51.)
- Calvo-Barajas, N., Perugia, G. and Castellano, G. (2020), The effects of robot's facial expressions on children's first impressions of trustworthiness, in '2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)', IEEE, pp. 165–171. doi:10.1109/ro-man47096.2020.9223456 (Cited on page 65.)
- Campagna, G., Chrysostomou, D. and Rehm, M. (2024a), Analysis of facial features for trust evaluation in industrial human-robot collaboration*, in 'IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)', IEEE, pp. 1–6. doi:10.1109/ARS060199.2024.10557748 (Cited on page 51.)
- Campagna, G., Chrysostomou, D. and Rehm, M. (2024b), Investigating electrodermal activity for trust assessment in industrial human-robot collabor-

- ation, in '21st International Conference on Ubiquitous Robots (UR)', IEEE, pp. 880–885. doi:10.1109/UR61395.2024.10597471 (Cited on page 51.)
- Campagna, G., Dadgostar, M., Chrysostomou, D. and Rehm, M. (2024), A data-driven approach utilizing body motion data for trust evaluation in industrial human-robot collaboration, in '33rd IEEE International Conference on Robot and Human Interactive Communication, IEEE RO-MAN 2024', IEEE, pp. 1984–1990. doi:10.1109/ro-man60168.2024.10731437 (Cited on page 64.)
- Campanozzi, L. L., Guglielmelli, E., Cella, E., Ghilardi, G., Michilli, M., Molina, A., Ciccozzi, M. and Tambone, V. (2019), 'Building trust in social robotics: a pilot survey', *IEEE Technology and Society Magazine* 38(4), 45–54. doi:10.1109/mts.2019.2948440 (Cited on page 64.)
- Carlson, T. and Demiris, Y. (2012), 'Collaborative Control for a Robotic Wheelchair: Evaluation of Performance, Attention, and Workload', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42(3), 876–888. doi:10.1109/TSMCB.2011.2181833 (Cited on page 76.)
- Chauhan, H., Jang, Y. and Jeong, I. (2024), 'Predicting human trust in human-robot collaborations using machine learning and psychophysiological responses', *Advanced Engineering Informatics* 62, 102720. doi:10.1016/j.aei.2024.102720 (Cited on pages 39 and 65.)
- Che, Y., Okamura, A. M. and Sadigh, D. (2020), 'Efficient and trustworthy social navigation via explicit and implicit robot–human communication', *IEEE Transactions on Robotics* 36(3), 692–707. doi:10.1109/tro.2020.2964824 (Cited on page 65.)
- Chen, M., Nikolaidis, S., Soh, H., Hsu, D. and Srinivasa, S. (2018), Planning with trust for human-robot collaboration, in 'Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction', ACM, pp. 307–315. doi:10.1145/3171221.3171264 (Cited on page 48.)
- Chi, O. H., Chi, C. G., Gursoy, D. and Nunkoo, R. (2023), 'Customers' acceptance of artificially intelligent service robots: The influence of trust and culture', *International Journal of Information Management* 70, 102623. doi:10.1016/j.ijinfomgt.2023.102623 (Cited on page 65.)
- Chi, V. B. and Malle, B. F. (2023), People dynamically update trust when interactively teaching robots, in 'Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction', ACM, pp. 554–564. doi:10.1145/3568162.3576962 (Cited on page 64.)

- Chien, S.-Y., Lewis, M., Semnani-Azad, Z. and Sycara, K. (2014), 'An empirical model of cultural factors on Trust in Automation', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **58**(1), 859–863. doi:10.1177/1541931214581181 (Cited on page 38.)
- Cogurcu, Y. E., Douthwaite, J. A. and Maddock, S. (2023), 'A comparative study of safety zone visualisations for virtual and physical robot arms using augmented reality', *Computers* **12**(4), 75. doi:10.3390/computers12040075 (Cited on page 64.)
- Conway, D., Dick, I., Li, Z., Wang, Y. and Chen, F. (2013), The Effect of Stress on Cognitive Load Measurement, in 'Human-Computer Interaction – INTERACT 2013', Vol. 8120, Springer, pp. 659–666. doi:10.1007/978-3-642-40498-6_58 (Cited on pages 39 and 40.)
- Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S. and Paiva, A. (2018), Exploring the impact of fault justification in human-robot trust, in 'Proceedings of the 17th international conference on autonomous agents and multiagent systems', International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, pp. 507–513. doi:10.5555/3237383.3237459 (Cited on page 65.)
- Crawshaw, M. (2020), 'Multi-task learning with deep neural networks: A survey', *arXiv*. doi:10.48550/arxiv.2009.09796 (Cited on page 111.)
- Cucciniello, I., Sangiovanni, S., Maggi, G. and Rossi, S. (2021), Validation of robot interactive behaviors through users emotional perception and their effects on trust, in '2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)', IEEE, pp. 197–202. doi:10.1109/ro-man50785.2021.9515352 (Cited on page 65.)
- Da Silva, H. P., Fred, A. and Martins, R. (2014), 'Biosignals for everyone', *IEEE Pervasive Computing* **13**(4), 64–71. doi:10.1109/mprv.2014.61 (Cited on page 81.)
- Dasari, S., Srirama, M. K., Jain, U. and Gupta, A. (2023), An unbiased look at datasets for visuo-motor pre-training, in 'Conference on Robot Learning', arXiv, pp. 1183–1198. doi:10.48550/arxiv.2310.09289 (Cited on page 133.)
- de Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R. and McDonald, C. G. (2018), 'Learning From the Slips of Others: Neural Correlates of Trust in Automated Agents', *Frontiers in Human Neuroscience* **12**, 309. doi:10.3389/fnhum.2018.00309 (Cited on pages 39 and 40.)

- De Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., Parasuraman, R. and Krueger, F. (2017), 'A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents', *Human Factors* **59**, 116–133. doi:10.1177/0018720816687205 (Cited on page 39.)
- de Vries, P., Midden, C. and Bouwhuis, D. (2003), 'The effects of errors on system trust, self-confidence, and the allocation of control in route planning', *International Journal of Human-Computer Studies* **58**(6), 719–735. doi:10.1016/S1071-5819(03)00039-9 (Cited on pages 33 and 56.)
- Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A. and Yanco, H. (2013), Impact of robot failures and feedback on real-time trust, in 'ACM/IEEE International Conference on Human-Robot Interaction (HRI)', IEEE/ACM, pp. 251–258. doi:10.1109/HRI.2013.6483596 (Cited on page 38.)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020), 'An image is worth 16x16 words: Transformers for image recognition at scale', *arXiv*. doi:10.48550/arxiv.2010.11929 (Cited on pages 52 and 54.)
- Dyck, L., Beierling, H., Helmert, R. and Vollmer, A.-L. (2023), Technical transparency for robot navigation through ar visualizations, in 'Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction', ACM, pp. 720–724. doi:10.1145/3568294.3580181 (Cited on page 64.)
- Elias, A. and Trigo, M. J. G. (2024), Unveiling trust dynamics with a mobile service robot: Exploring various interaction styles for an agricultural task, in '2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)', IEEE, pp. 638–645. doi:10.1109/ro-man60168.2024.10731301 (Cited on page 64.)
- Elkins, A. C. and Derrick, D. C. (2013), 'The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents', *Group decision and negotiation* **22**(5), 897–913. doi:10.1007/s10726-012-9339-x (Cited on page 39.)
- Emaminejad, N., Kath, L. and Akhavian, R. (2024), 'Assessing trust in construction ai-powered collaborative robots using structural equation modeling', *Journal of Computing in Civil Engineering* **38**(3), 04024011. doi:10.1061/JCCEE5.CPENG-5660 (Cited on page 65.)

- Esterwood, C. and Robert, L. P. (2023), 'The theory of mind and human-robot trust repair', *Scientific Reports* **13**(1), 9877. doi:10.1038/s41598-023-37032-0 (Cited on page 64.)
- Floyd, M., Drinkwater, M. and Aha, D. (2015), Trust-guided behavior adaptation using case-based reasoning, in 'Twenty-Fourth International Joint Conference on Artificial Intelligence', AAAI Press, pp. 4261–4267. doi:10.5555/2832747.2832851 (Cited on page 46.)
- Fooladi Mahani, M., Jiang, L. and Wang, Y. (2021), 'A bayesian trust inference model for human-multi-robot teams', *International Journal of Social Robotics* **13**(8), 1951–1965. doi:10.1007/s12369-020-00705-1 (Cited on page 50.)
- Fox, D., Burgard, W., Dellaert, F. and Thrun, S. (1999), 'Monte carlo localization: Efficient position estimation for mobile robots', **1999**, 343–349. doi:10.1109/ROBOT.1999.772544 (Cited on page 76.)
- Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A. and Carin, L. (2019), 'Cyclical annealing schedule: A simple approach to mitigating kl vanishing', *arXiv*. doi:10.48550/arxiv.1903.10145 (Cited on page 133.)
- Fuentes, C., Herskovic, V., Rodríguez, I., Gereia, C., Marques, M. and Rosset, P. O. (2017), 'A systematic literature review about technologies for self-reporting emotional information', *Journal of Ambient Intelligence and Humanized Computing* **8**, 593–606. doi:10.1007/s12652-016-0430-z (Cited on page 51.)
- García-Corretjer, M., Ros, R., Mallol, R. and Miralles, D. (2023), 'Empathy as an engaging strategy in social robotics: a pilot study', *User Modeling and User-Adapted Interaction* **33**(2), 221–259. doi:10.1007/s11257-022-09322-1 (Cited on page 64.)
- Giorgi, I., Tiroto, F. A., Hagen, O., Aider, F., Gianni, M., Palomino, M. and Masala, G. L. (2022), 'Friendly but faulty: A pilot study on the perceived trust of older adults in a social robot', *IEEE Access* **10**, 92084–92096. doi:10.1109/access.2022.3202942 (Cited on page 64.)
- Goubard, C. and Demiris, Y. (2024), Learning self-confidence from semantic action embeddings for improved trust in human-robot interaction, in '2024 IEEE International Conference on Robotics and Automation (ICRA)', IEEE, pp. 7859–7866. doi:10.1109/icra57147.2024.10611445 (Cited on page 64.)
- Graf, L., Torkar, M., Stüchelmaier, E., Sichler, R., Malafosse, P., Fischer, K. and Palinko, O. (2022), Perceived trustworthiness of an interactive robotic sys-

- tem, in 'ACM/IEEE International Conference on Human-Robot Interaction (HRI)', IEEE / ACM, pp. 773–777. doi:10.1109/hri53351.2022.9889667 (Cited on page 65.)
- Gustafsson, T. (2020), 'Rosswift: A pure swift implementation of ros', <https://github.com/tgu/RosSwift>. Accessed: 2024-11. (Cited on page 156.)
- Guznov, S., Lyons, J., Pfahler, M., Heironimus, A., Woolley, M., Friedman, J. and Neimeier, A. (2020), 'Robot transparency and team orientation effects on human–robot teaming', *International Journal of Human–Computer Interaction* 36(7), 650–660. doi:10.1080/10447318.2019.1676519 (Cited on page 64.)
- Hald, K., Rehm, M. and Moeslund, T. B. (2021), Human-robot trust assessment using top-down visual tracking after robot task execution mistakes, in '2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)', IEEE, pp. 892–898. doi:10.1109/ro-man50785.2021.9515501 (Cited on page 64.)
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J. and Parasuraman, R. (2011), 'A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction', *Human Factors* 53(5), 517–527. doi:10.1177/0018720811417254 (Cited on pages 35, 36, 41, 42, 43, 56, 74, and 93.)
- Hinz, A., Michalski, D., Schwarz, R. and Herzberg, P. Y. (2007), 'The acquiescence effect in responding to a questionnaire', *GMS Psycho-Social Medicine* 4. (Cited on page 38.)
- Hoff, K. A. and Bashir, M. (2015), 'Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust', *Human Factors* 57(3), 407–434. doi:10.1177/0018720814547570 (Cited on pages 41, 43, and 56.)
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A. et al. (2022), 'Training compute-optimal large language models', *arXiv*. doi:10.48550/arxiv.2203.15556 (Cited on page 114.)
- Hopko, S. K., Mehta, R. K. and Pagilla, P. R. (2023), 'Physiological and perceptual consequences of trust in collaborative robots: An empirical investigation of human and robot factors', *Applied Ergonomics* 106, 103863. doi:10.1016/j.apergo.2022.103863 (Cited on page 64.)
- Hopko, S. K., Zhang, Y., Yadav, A., Pagilla, P. R. and Mehta, R. K. (2024), 'Behavior relationships of trust in shared space human–robot collaboration',

- ACM Transactions on Human-Robot Interaction* **13**(1), 1–23. doi:10.1145/3632149 (Cited on page 64.)
- Hulcelle, M., Varni, G., Rollet, N. and Clavel, C. (2021), Turin: A coding system for trust in human robot interaction, in ‘2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)’, IEEE, pp. 1–8. doi:10.1109/acii52823.2021.9597448 (Cited on page 64.)
- Ioannidis, J. P. (2005), ‘Why most published research findings are false’, *PLoS medicine* **2**(8), e124. doi:10.1371/journal.pmed.0020124 (Cited on page 21.)
- Itoh, M. and Tanaka, K. (2000), Mathematical modeling of trust in automation: Trust, distrust, and mistrust, in ‘Proceedings of the human factors and ergonomics society annual meeting’, Vol. 44, SAGE Publications, pp. 9–12. doi:10.1177/154193120004400103 (Cited on page 46.)
- Jian, J.-Y., Bisantz, A. M. and Drury, C. G. (2000), ‘Foundations for an Empirically Determined Scale of Trust in Automated Systems’, *International Journal of Cognitive Ergonomics* **4**(1), 53–71. doi:10.1207/S15327566IJCE0401_04 (Cited on page 38.)
- Kaniarasu, P., Steinfeld, A., Desai, M. and Yanco, H. (2012), Potential measures for detecting trust changes, in ‘7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)’, ACM, pp. 241–242. doi:10.1145/2157689.2157775 (Cited on pages 39 and 40.)
- Kessler, T. T., Larios, C., Walker, T., Yerdon, V. and Hancock, P. A. (2017), A Comparison of Trust Measures in Human–Robot Interaction Scenarios, in ‘Advances in Human Factors in Robots and Unmanned Systems’, Vol. 499, Springer, pp. 353–364. doi:10.1007/978-3-319-41959-6_29 (Cited on page 39.)
- Khalid, H. M., Shiung, L. W., Nooralishahi, P., Rasool, Z., Helander, M. G., Kiong, L. C. and Ai-vyrn, C. (2016), Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction, in ‘Proceedings of the human factors and ergonomics society annual meeting’, Vol. 60, pp. 697–701. doi:10.1177/154193121360116 (Cited on page 39.)
- Khavas, Z. R., Ahmadzadeh, S. R. and Robinette, P. (2020), Modeling Trust in Human-Robot Interaction: A Survey, in ‘Social Robotics’, Vol. 12483, Springer International Publishing, pp. 529–541. doi:10.1007/978-3-030-62056-1_44 (Cited on pages 35 and 36.)

- Khavas, Z. R., Perkins, R., Ahmadzadeh, S. R. and Robinette, P. (2021), 'Moral-trust violation vs performance-trust violation by a robot: Which hurts more?', *arXiv*. doi:10.48550/arxiv.2110.04418 (Cited on page 65.)
- Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C. and Shaw, T. H. (2021), 'Measurement of trust in automation: A narrative review and reference guide', *Frontiers in psychology* **12**, 604977. doi:10.3389/fpsyg.2021.604977 (Cited on page 147.)
- Kraus, J., Miller, L., Klumpp, M., Babel, F., Scholz, D., Merger, J. and Baumann, M. (2024), 'On the role of beliefs and trust for the intention to use service robots: an integrated trustworthiness beliefs model for robot acceptance', *International Journal of Social Robotics* **16**(6), 1223–1246. doi:10.1007/s12369-022-00952-4 (Cited on page 64.)
- Kraus, M., Wagner, N., Untereiner, N. and Minker, W. (2022), Including social expectations for trustworthy proactive human-robot dialogue, in 'Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization', ACM, pp. 23–33. doi:10.1145/3503252.3531294 (Cited on page 64.)
- Körber, M. (2019), Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation, in 'Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)', Vol. 823, Springer, pp. 13–30. doi:10.1007/978-3-319-96074-6_2 (Cited on pages 17, 38, 80, 88, and 146.)
- Lagomarsino, M., Lorenzini, M., Constable, M. D., De Momi, E., Becchio, C. and Ajoudani, A. (2023), 'Maximising efficiency of human-robot handovers through reinforcement learning', *IEEE Robotics and Automation Letters* **8**(8), 4378–4385. doi:10.1109/lra.2023.3280752 (Cited on page 65.)
- Lee, J. D. and Moray, N. (1994), 'Trust, self-confidence, and operators' adaptation to automation', *International Journal of Human-Computer Studies* **40**(1), 153–184. doi:10.1006/ijhc.1994.1007 (Cited on page 38.)
- Lee, J. D. and See, K. A. (2004), 'Trust in Automation: Designing for Appropriate Reliance', *Human Factors* **46**(1), 50–80. doi:10.1518/hfes.46.1.50_30392 (Cited on pages 33 and 87.)
- Lee, J. and Moray, N. (1992), 'Trust, control strategies and allocation of function in human-machine systems', *Ergonomics* **35**(10), 1243–1270. doi:10.1080/00140139208967392 (Cited on pages 46 and 47.)

- Lee, K. M., Krishna, A., Zaidi, Z., Paleja, R., Chen, L., Hedlund-Botti, E., Schrum, M. and Gombolay, M. (2023), The effect of robot skill level and communication in rapid, proximate human-robot collaboration, *in* 'Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction', ACM, pp. 261–270. doi:[10.1145/3568162.3577002](https://doi.org/10.1145/3568162.3577002) (Cited on page 65.)
- Leichtmann, B., Meneweger, T., Busch, C., Reiterer, B., Meyer, K., Rammer, D., Haring, R. and Mara, M. (2024), 'Teaming with a robot in mixed reality: Dynamics of trust, self-efficacy, and mental models affected by information richness', *International Journal of Human-Computer Interaction* pp. 1–18. doi:[10.1080/10447318.2024.2331878](https://doi.org/10.1080/10447318.2024.2331878) (Cited on page 65.)
- Lingg, N. and Demiris, Y. (2023a), Beyond self-report: A continuous trust measurement device for hri, *in* '2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)', IEEE, pp. 2220–2225. doi:[10.1109/ro-man57019.2023.10309660](https://doi.org/10.1109/ro-man57019.2023.10309660) (Cited on page 65.)
- Lingg, N. and Demiris, Y. (2023b), Building trust in assistive robotics: Insights from a real-world mobile navigation experiment, *in* 'Proceedings of the First International Symposium on Trustworthy Autonomous Systems', ACM, pp. 1–7. doi:[10.1145/3597512.3597519](https://doi.org/10.1145/3597512.3597519) (Cited on page 64.)
- Lingg, N. and Demiris, Y. (2024), Trust prediction in assistive robotics using multi-modal video transformers, *in* '2024 16th International Conference on Social Robotics (ICSR)'. (Cited on page 64.)
- Loizaga, E., Bastida, L., Sillaurren, S., Moya, A. and Toledo, N. (2024), 'Modelling and measuring trust in human-robot collaboration', *Applied Sciences* **14**(5), 1919. doi:[10.3390/app14051919](https://doi.org/10.3390/app14051919) (Cited on page 64.)
- Luo, R., Huang, C., Peng, Y., Song, B. and Liu, R. (2021), Repairing human trust by promptly correcting robot mistakes with an attention transfer model, *in* 'IEEE 17th International Conference on Automation Science and Engineering (CASE)', IEEE, pp. 1928–1933. doi:[10.1109/case49439.2021.9551502](https://doi.org/10.1109/case49439.2021.9551502) (Cited on page 64.)
- Lyons, J. B., Vo, T., Wynne, K. T., Mahoney, S., Nam, C. S. and Galimore, D. (2021), 'Trusting autonomous security robots: The role of reliability and stated social intent', *Human factors* **63**(4), 603–618. doi:[10.1177/001872082090162](https://doi.org/10.1177/001872082090162) (Cited on page 64.)

- Madhavan, P. and Wiegmann, D. A. (2007), 'Similarities and differences between human-human and human-automation trust: an integrative review', *Theoretical Issues in Ergonomics Science* **8**(4), 277–301. doi:10.1080/14639220500337708 (Cited on pages 33 and 34.)
- Madsen, M. and Gregor, S. (2000), Measuring human-computer trust, in 'Proceedings of the 11 th Australasian Conference on Information Systems', pp. 6–8. doi:10.1006/ijhc.1994.1007 (Cited on page 38.)
- Maehigashi, A., Tsumura, T. and Yamada, S. (2024), 'Impacts of robot beep timings on trust dynamics in human-robot interaction', *International Journal of Social Robotics* **16**, 2205–2222. doi:10.1007/s12369-024-01181-7 (Cited on page 39.)
- Maehigashi, A. and Yamada, S. (2023), 'Modeling trust and reliance with wait time in a human-robot interaction', *arXiv*. doi:10.48550/arxiv.2302.08064 (Cited on page 39.)
- Malle, B. F. and Ullman, D. (2021), A multidimensional conception and measure of human-robot trust, in 'Trust in human-robot interaction', Elsevier, pp. 3–25. doi:10.1016/b978-0-12-819472-0.00001-0 (Cited on page 65.)
- Mangalindan, D. H., Rovira, E. and Srivastava, V. (2023), On trust-aware assistance-seeking in human-supervised autonomy, in 'American Control Conference (ACC)', IEEE, pp. 3901–3906. doi:10.23919/acc55779.2023.10156103 (Cited on page 47.)
- Menolotto, M., Komaris, D.-S., O'Sullivan, P. and O'Flynn, B. (2024), Assessing trust in collaborative robotics with different human-robot interfaces, in 'IEEE International Instrumentation and Measurement Technology Conference (I2MTC)', IEEE, pp. 1–6. doi:10.1109/i2mtc60896.2024.10561211 (Cited on page 65.)
- Merritt, S. M. (2011), 'Affective Processes in Human-Automation Interactions', *Human Factors* **53**(4), 356–370. doi:10.1177/0018720811411912 (Cited on page 38.)
- Mikulski, D. G., Lewis, F. L., Gu, E. Y. and Hudas, G. R. (2012), Trust method for multi-agent consensus, in 'Unmanned systems technology xiv', Vol. 8387, SPIE, pp. 146–159. doi:10.1117/12.918927 (Cited on page 45.)
- Miller, L., Kraus, J., Babel, F. and Baumann, M. (2021), 'More than a feeling—interrelation of trust layers in human-robot interaction and the role

- of user dispositions and state anxiety', *Frontiers in psychology* **12**, 592711. doi:[10.3389/fpsyg.2021.592711](https://doi.org/10.3389/fpsyg.2021.592711) (Cited on page [65](#).)
- Mota, R. C. R., Rea, D. J., Le Tran, A., Young, J. E., Sharlin, E. and Sousa, M. C. (2016), Playing the 'trust game' with robots: Social strategies and experiences, in '2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)', IEEE, pp. 519–524. doi:[10.1109/roman.2016.7745167](https://doi.org/10.1109/roman.2016.7745167) (Cited on page [64](#).)
- Muir, B. M. (1994), 'Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems', *Ergonomics* **37**(11), 1905–1922. doi:[10.1080/00140139408964957](https://doi.org/10.1080/00140139408964957) (Cited on pages [33](#), [41](#), and [87](#).)
- Nenna, F., Zanardi, D., Maria Orlando, E., Mingardi, M., Buodo, G. and Gamberini, L. (2024), Addressing trust and negative attitudes toward robots in human-robot collaborative scenarios: Insights from the industrial work setting, in 'Proceedings of the 17th International Conference on Pervasive Technologies Related to Assistive Environments', ACM, pp. 259–264. doi:[10.1145/3652037.3663905](https://doi.org/10.1145/3652037.3663905) (Cited on page [65](#).)
- Nikolaidis, S., Zhu, Y. X., Hsu, D. and Srinivasa, S. (2017), Human-robot mutual adaptation in shared autonomy, in 'Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction', ACM, pp. 294–302. doi:[10.1145/2909824.3020252](https://doi.org/10.1145/2909824.3020252) (Cited on page [64](#).)
- Okamura, K. and Yamada, S. (2020), 'Adaptive trust calibration for human-ai collaboration', *Plos one* **15**(2), e0229132. doi:[10.1371/journal.pone.0229132](https://doi.org/10.1371/journal.pone.0229132) (Cited on page [64](#).)
- Palmer, S., Richards, D., Shelton-Rayner, G., Izzetoglu, K. and Inch, D. (2020), Assessing variable levels of delegated control—a novel measure of trust, in 'HCI International 2020—Late Breaking Papers: Cognition, Learning and Games: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22', Vol. 12425, Springer Science+Business Media, pp. 202–215. doi:[10.1007/978-3-030-60128-7_16](https://doi.org/10.1007/978-3-030-60128-7_16) (Cited on page [64](#).)
- Parron, J., Nguyen, T. T. and Wang, W. (2023), Development of a multimodal trust database in human-robot collaborative contexts, in 'IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)', IEEE, pp. 0601–0605. doi:[10.1109/uemcon59035.2023.10316014](https://doi.org/10.1109/uemcon59035.2023.10316014) (Cited on page [65](#).)

- Pinney, J., Carroll, F. and Newbury, P. (2022), 'Human-robot interaction: the impact of robotic aesthetics on anticipated human trust', *PeerJ Computer Science* 8, e837. doi:10.7717/peerj-cs.837 (Cited on page 65.)
- Pompe, B. L., Velner, E. and Truong, K. P. (2022), The robot that showed remorse: Repairing trust with a genuine apology, in '31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)', IEEE, pp. 260–265. doi:10.1109/ro-man53752.2022.9900860 (Cited on page 64.)
- Rabby, M. K. M., Khan, M. A., Karimoddini, A. and Jiang, S. X. (2020), Modeling of trust within a human-robot collaboration framework, in 'IEEE International Conference on Systems, Man, and Cybernetics (SMC)', IEEE, pp. 4267–4272. doi:10.1109/smc42975.2020.9283228 (Cited on page 46.)
- Rehm, M., Pontikis, I. and Hald, K. (2024), Automatic trust estimation from movement data in industrial human-robot collaboration based on deep learning, in 'IEEE International Conference on Robotics and Automation (ICRA)', IEEE, pp. 11245–11251. doi:10.1109/ICRA57147.2024.10610822 (Cited on page 51.)
- Roesmann, C., Feiten, W., Woesch, T., Hoffmann, F. and Bertram, T. (2012), Trajectory modification considering dynamic constraints of autonomous robots, in '7th German Conference on Robotics', VDE-Verlag, pp. 1–6. (Cited on page 77.)
- Roesmann, C., Hoffmann, F. and Bertram, T. (2017), 'Integrated online trajectory planning and optimization in distinctive topologies', *Robotics and Autonomous Systems* 88, 142–153. doi:10.1016/j.robot.2016.11.007 (Cited on page 77.)
- Romeo, M., Torre, I., Le Maguer, S., Cangelosi, A. and Leite, I. (2023), Putting robots in context: Challenging the influence of voice and empathic behaviour on trust, in '2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)', IEEE, pp. 2045–2050. doi:10.1109/ro-man57019.2023.10309631 (Cited on page 64.)
- Sadrifaridpour, B., Saeidi, H., Burke, J., Madathil, K. and Wang, Y. (2016), 'Modeling and control of trust in human-robot collaborative manufacturing', *Robust intelligence and trust in autonomous systems* pp. 115–141. doi:10.1007/978-1-4899-7668-0_7 (Cited on page 47.)
- Schaefer, K. E. (2016), Measuring trust in human robot interactions: Development of the "trust perception scale-hri", in 'Robust intelligence

- and trust in autonomous systems', Springer, pp. 191–218. doi:10.1007/978-1-4899-7668-0_10 (Cited on pages 17, 38, 80, 88, 146, and 147.)
- Seet, M., Harvy, J., Bose, R., Dragomir, A., Bezerianos, A. and Thakor, N. (2020), 'Differential impact of autonomous vehicle malfunctions on human trust', *IEEE transactions on intelligent transportation systems* 23(1), 548–557. doi:10.1109/tits.2020.3013278 (Cited on page 64.)
- Shazeer, N. (2020), 'Glu variants improve transformer', *arXiv* . doi:10.48550/arxiv.2002.05202 (Cited on page 109.)
- Shidujaman, M., Mi, H. and Jamal, L. (2020), " i trust you more": A behavioral greeting gesture study on social robots for recommendation tasks, in '2020 International Conference on Image Processing and Robotics (ICIP)', IEEE, pp. 1–5. doi:10.1109/icip48927.2020.9367364 (Cited on page 64.)
- Soh, H., Xie, Y., Chen, M. and Hsu, D. (2020), 'Multi-task trust transfer for human–robot interaction', *The International Journal of Robotics Research* 39(2-3), 233–249. doi:10.1177/0278364919866905 (Cited on page 49.)
- Song, Y. and Luximon, Y. (2024), 'When trustworthiness meets face: Facial design for social robots', *Sensors* 24(13), 4215. doi:10.3390/s24134215 (Cited on page 65.)
- Sun, X., Li, J., Tang, P., Zhou, S., Peng, X., Li, H. N. and Wang, Q. (2020), 'Exploring personalised autonomous vehicles to influence user trust', *Cognitive Computation* 12(6), 1170–1186. doi:10.1007/s12559-020-09757-x (Cited on page 65.)
- Tonsen, M., Baumann, C. K. and Dierkes, K. (2020), 'A high-level description and performance evaluation of pupil invisible', *arXiv* . doi:10.48550/arxiv.2009.00508 (Cited on page 81.)
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (n.d.), 'Llama: open and efficient foundation language models, 2023', URL <https://arxiv.org/abs/2302.13971> . (Cited on page 109.)
- van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. and Haselager, P. (2014), 'Do robot performance and behavioral style affect human trust? a multi-method approach', *International journal of social robotics* pp. 519–531. doi:10.1007/s12369-014-0231-5 (Cited on page 93.)

- Van Lange, P. A., Finkenauer, C., Popma, A. and van Vugt, M. (2011), 'Electrodes as social glue: Measuring heart rate promotes giving in the trust game', *International Journal of Psychophysiology* **80**(3), 246–250. doi:10.1016/j.ijpsycho.2011.03.007 (Cited on pages 39 and 40.)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**, 5998–6008. (Cited on pages 52, 53, and 109.)
- Wald, S., Puthuveetil, K. and Erickson, Z. (2024), Do mistakes matter? comparing trust responses of different age groups to errors made by physically assistive robots, in '33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)', IEEE, pp. 373–380. doi:10.1109/ro-man60168.2024.10731271 (Cited on page 64.)
- Walker, F., Dey, D., Martens, M., Pfleging, B., Eggen, B. and Terken, J. (2019), Feeling-of-safety slider: Measuring pedestrian willingness to cross roads in field interactions with vehicles, in 'Conference on Human Factors in Computing Systems', ACM, pp. 1–6. (Cited on page 51.)
- Walker, F., Wang, J., Martens, M. H. and Verwey, W. B. (2019), 'Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles', *Transportation Research Part F: Traffic Psychology and Behaviour* **64**, 401–412. doi:10.1016/j.trf.2019.05.021 (Cited on pages 39 and 40.)
- Weerawardhana, S., Akintunde, M. E., Masters, P. et al. (2024), More than trust: Compliance in instantaneous human-robot interactions, in '33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)', IEEE, pp. 1556–1563. doi:10.1109/ro-man60168.2024.10731378 (Cited on page 39.)
- Wojton, H. M., Porter, D., T. Lane, S., Bieber, C. and Madhavan, P. (2020), 'Initial validation of the trust of automated systems test (toast)', *The Journal of Social Psychology* **160**(6), 735–750. doi:10.1080/00224545.2020.1749020 (Cited on pages 17, 38, 80, 88, and 145.)
- Wu, B., Hu, B. and Lin, H. (2017), Toward efficient manufacturing systems: A trust based human robot collaboration, in 'American Control Conference (ACC)', IEEE, pp. 1536–1541. doi:10.23919/acc.2017.7963171 (Cited on page 48.)

- Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K. and Vajda, P. (2020), ‘Visual transformers: Token-based image representation and processing for computer vision’, *arXiv* . doi: [10.48550/arxiv.2006.03677](https://doi.org/10.48550/arxiv.2006.03677) (Cited on page [109](#).)
- Wu, P., Shentu, Y., Yi, Z., Lin, X. and Abbeel, P. (2023), ‘Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators’, *arXiv* pp. 12156–12163. doi:[10.48550/arxiv.2309.13037](https://doi.org/10.48550/arxiv.2309.13037) (Cited on pages [128](#) and [131](#).)
- Xu, A. and Dudek, G. (2015), OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations, in ‘Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction’, ACM, pp. 221–228. doi:[10.1145/2696454.2696492](https://doi.org/10.1145/2696454.2696492) (Cited on pages [39](#) and [49](#).)
- Xu, A. and Dudek, G. (2016), Towards modeling real-time trust in asymmetric human–robot collaborations, in ‘Robotics Research: The 16th International Symposium ISRR’, Springer, pp. 113–129. doi:[10.1007/978-3-319-28872-7_7](https://doi.org/10.1007/978-3-319-28872-7_7) (Cited on page [45](#).)
- Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J. and Susskind, J. M. (2023), Stabilizing transformer training by preventing attention entropy collapse, in ‘International Conference on Machine Learning’, arXiv, pp. 40770–40803. doi:[10.48550/arxiv.2303.06296](https://doi.org/10.48550/arxiv.2303.06296) (Cited on page [110](#).)
- Zhang, B. and Sennrich, R. (2019), ‘Root mean square layer normalization’, *Advances in Neural Information Processing Systems* **32**, 12360–12371. doi: [10.5167/uzh-177483](https://doi.org/10.5167/uzh-177483) (Cited on page [109](#).)
- Zhang, Y., Yadav, A., Hopko, S. K. and Mehta, R. K. (2024), 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’, ACM, pp. 1188–1193. doi:[10.1145/3610978.3640649](https://doi.org/10.1145/3610978.3640649) (Cited on page [65](#).)
- Zhao, T. Z., Kumar, V., Levine, S. and Finn, C. (2023), ‘Learning fine-grained bimanual manipulation with low-cost hardware’, *Robotics: Science and Systems Foundation* . doi:[10.15607/rss.2023.xix.016](https://doi.org/10.15607/rss.2023.xix.016) (Cited on pages [119](#), [129](#), [131](#), and [132](#).)