



Dynamic Trust Modeling for Human–Robot Interaction

A Structured Mathematical Model with End-to-End Identification and Validation

Pedro Silva | Master Thesis in Aerospace Engineering

Dynamic Trust Modeling for Human–Robot Interaction

A Structured Mathematical Model with End-to-End Identification and Validation

by

Pedro Silva

in partial fulfillment of the requirements for the degree of

Master of Science
in Aerospace Engineering

at the Delft University of Technology,
to be defended publicly on Monday 2 March 2026 at 12:45.

Thesis committee

Chair: Dr. ir. Coen de Visser
Supervisor: Dr. Anahita Jamshidnejad
Examiner: Dr. Geethu Joseph

Project Duration: February 2025 - March 2026

Cover image inspired by Michelangelo's The Creation of Adam, designed by Pedro Silva, 2026.

An electronic version of this thesis is available at <https://repository.tudelft.nl/>.

Remember: the rules, like streets, can only take you to known places.
- Ocean Vuong, in "On Earth We're Briefly Gorgeous"

Acknowledgments

This thesis is the culmination of my time as a student at TU Delft. This study meant more than simply receiving a Master's Degree in Aerospace Engineering. When I decided to leave Portugal four years ago, I knew an exciting adventure awaited in the cold and rainy streets of this small town in the Netherlands, and what a ride it has been. I have since lived in both the Netherlands and Germany and have been able to meet people from all around the world. This has been a journey of many ups and downs, but the challenging moments make this achievement feel all the more rewarding.

Every significant accomplishment is, like trust, the product of many contributing factors worth acknowledging. I would like to sincerely thank my supervisor, Dr. Anahita Jamshidnejad, for her guidance and support over the past year. From the early stages of defining the topic to the completion of this thesis, her knowledge and expertise were instrumental in shaping this work.

I am also deeply grateful to my family. To my mom, Catarina, and my dad, Rui, thank you for all your support and for allowing me to explore life in a different country. I know it has not been easy, but I am forever grateful for this opportunity. I would not be who I am today without the support of my grandparents. I would especially like to express my sincere gratitude to my grandfather Adolfo, who passed away while I was working on this project. Thank you for the countless car rides to Jobra and for the coffee candies always present in your pockets, which you never refused to share. Your absence will be deeply felt. Finally, a big thank you to all my extended family. Even though we do not see each other frequently, I always look forward to seeing you whenever I visit home.

Moreover, special thanks are due to all the friends I made along the way. There are too many to name, but please allow me to mention a few in particular. To Jai, who was already eager to have breakfast together the morning after I arrived at Lisztstraat, it was a pleasure to share number 99 with you. To Haohua, thank you for being my first true friend in Delft, for the endless trips to the cinema, and for the late-night talks. I hope you are doing well. To David, thank you for not giving up and actually wanting to meet after so long. I would probably be homeless or back to Portugal by now if it were not for you. It is a pleasure to be your friend, and I wish you the best of luck with your PhD. To Atbin, thank you for inadvertently pushing me to learn more about myself. I will forever cherish those ephemeral summer months of endless possibility. Additionally, I must thank all my Aeroamigos, without whom this journey would certainly not have been as fun. Vamos todos beber um café a Civil para festejar.

Furthermore, I would like to acknowledge all the people I had the privilege of working with at Airbus Defence and Space in (even colder) Friedrichshafen during my internship, particularly my supervisors Ralf Schubert and Simon Delchambre. To all the friends I made outside of work, thank you for making the German winter a little warmer.

Finally, I cannot help but reflect on the journey that led me to this moment. In these (almost) four years, a lot has happened. The list is long and includes moving five times, spending a birthday having a wisdom tooth removed in the middle of Germany, running away from my conspiracy-theorist landlady, and many a heartbreak. At the same time, these have been years filled with friendships, laughter, and experiences that I will always look back on fondly. All these experiences, good and bad, allowed me to grow, to learn more about myself, and to build a life far beyond what I had imagined when I first set foot in Delft. Through it all, I never lost myself, and for that I am truly proud. It really has been a hell of an adventure.

Now, onto the next one!

*Pedro Silva
Delft, March 2026*

Abstract

Trust plays a central role in human–robot interaction, as it shapes how humans rely on and collaborate with autonomous systems under uncertainty. Existing approaches often treat trust as a static value or rely on highly data-driven models that lack interpretability and require extensive amounts of context-specific datasets. This limits their applicability in simulation and control. Human trust is modeled as a latent, time-varying internal state and is formulated as an end-to-end identification problem grounded in empirical interaction data. A modular, deterministic dynamic model integrates dispositional, situational, and learned components within a discrete-time framework, capturing both stable predispositions and interaction-driven adaptation. A human-subject experimental paradigm including 35 participants elicits informative trust dynamics and enables parameter identification and evaluation in terms of fit, generalization, and robustness. The resulting model remains competitive against validated baselines in reproducing observed trust trajectories while maintaining interpretability, modularity, and practical identifiability. Closed-loop coupling with probabilistic decision models examines the relationship between inferred trust and observable behavior, showing that trust is behaviorally relevant but insufficient on its own to explain human actions. The resulting framework enables executable, empirically grounded modeling of trust dynamics in human–robot interaction and supports the development of trust-aware simulation environments and human-centered robotic systems.

Keywords: Human-robot interaction, Trust dynamics, Latent state modeling, System identification, Parameter estimation, Discrete-time systems, Trust-aware systems, Autonomous systems, Behavioral modeling.

Nomenclature

List of Abbreviations

AIC	Akaike Information Criterion	OLS	Ordinary Least Squares
AUC	Area Under the Curve	RMSE	Root Mean Square Error
BIC	Bayesian Information Criterion	ROC	Receiver Operating Characteristic
HAI	Human–Automation Interaction	SaR	Search and Rescue
HRI	Human–Robot Interaction	WLS	Weighted Least Squares
MAE	Mean Absolute Error	wMAE	Weighted Mean Absolute Error
MAP	Maximum A Posteriori	wRMSE	Weighted Root Mean Square Error
NLL	Negative Log-Likelihood		

List of Symbols

τ	Overall Trust Vvalue	E_k	Interaction Event Indicator
τ^{disp}	Dispositional Trust	k	Discrete Time Step Index
τ^{sit}	Situational Trust	n_k	Consecutive Outcome Index
τ^{rep}	Reputation	r_k	Contextual Tisk Input at Time Step k
τ^{exp}	Personal Experience	t_i	Measurement Time Index
τ^{lat}	Latent Trust	β	Self-confidence
$\hat{\tau}(t; \theta)$	Predicted Trust Trajectory	$\sigma(\cdot)$	Logistic (Sigmoid) Function
θ	Global model parameter vector		

Model Parameters

α^{sit}	Situational Trust Weighting Factor	ϕ^{succ}	Trust Increase After First Success
λ^{sit}	Situational Trust Decay Factor	ϕ^{fail}	Trust Decrease After First Failure
λ^{rep}	Reputation Decay Factor	κ^{lat}	Latent Trust Growth Rate
a^{succ}	Success Curve Lower Asymptote	λ^{lat}	Latent Trust Decaying Rate

Contents

Acknowledgments	iv
Abstract	v
Nomenclature	vi
1 Introduction	1
2 Literature Review	3
2.1 Scope and Research Trajectory	3
2.2 Introduction to Human-Technology Interaction	3
2.3 Trust in Human-Technology Systems	4
2.4 Factors Influencing Human-Robot Trust	6
2.5 Modeling Trust Dynamics	7
2.6 Application Domain: Search and Rescue	15
2.7 Integrating Trust Dynamics and Agent-Based SaR Simulation	16
3 Problem Formulation and Contributions	18
3.1 Identification of Research Gap	18
3.2 Research Objective	19
3.3 Contributions	19
3.4 Research Questions	20
4 Theoretical Background for Model Identification and Validation	22
4.1 Trust Modeling as a Practical Inverse Identification Problem	22
4.2 Least-Squares Identification under Model-Data Mismatch	23
4.3 Model Validation, Generalization, and Robustness	23
4.4 Latent States, Observable Behavior, and Closed-Loop Evaluation	24
5 Proposed Trust Dynamics Model	25
5.1 Design Requirements for Dynamic Trust Modeling	25
5.2 Modular Architecture and Trust State Definition	26
5.3 Dispositional Trust	28
5.4 Situational Trust	29
5.5 Learned Trust I: Reputation	31
5.6 Learned Trust II: Personal Experience	32
5.7 Latent Trust	34
5.8 Discrete-Time Specification of the Trust Dynamics Model	36
5.9 Integrated Trust Dynamics Model	39
5.10 Synthesis and Implications for Research Question 1	42
6 Experimental Setup	44
6.1 Description of the Experimental Task	44
6.2 Experimental Design Considerations	47
6.3 Experimental Conditions and Trial Set Design	49
6.4 Mapping Trust Dynamics Model Components to Experimental Setup	52
6.5 Data Validation and Inclusion Criteria	59
6.6 Trust Measurements Calibration	59
6.7 Participant Recruitment, Ethics, and Data Handling	67
6.8 Synthesis and Implications for Research Question 2	68
7 Identification of the Proposed Trust Dynamics Model	69
7.1 Overview of the Identification Pipeline and Research Question	69

7.2	Dataset Overview and Validation	71
7.3	Parameter Identification Problem Formulation for Trust Dynamics	79
7.4	Summary of Experimental Data Used for Identification	80
7.5	Model Validation and Evaluation Metrics	81
7.6	Robustness and Sensitivity Analysis Framework	83
7.7	Baseline Models and Comparative Evaluation Protocol	85
7.8	Summary of Identification Methodology	90
7.9	Identification Results	92
7.10	Discussion of Trust Model Identification Results	104
7.11	Synthesis and Implications for Research Question 3	111
8	Trust-Aware Behavioral Model	113
8.1	Purpose and Scope of Trust-Aware Behavioral Modeling	113
8.2	Candidate Trust-to-Action Mappings	114
8.3	Behavioral Dataset Construction and Causal Alignment	115
8.4	Behavioral Model Estimation and Probabilistic Evaluation	116
8.5	Individual Differences in Reliance Mechanisms	117
8.6	Closed-Loop Generative Evaluation of Interaction Patterns	118
8.7	Trust Trajectory Sanity Checks under Closed-Loop Coupling	121
8.8	Summary of Behavioral Model Methodology	123
8.9	Behavioral Model Results	125
8.10	Discussion of Behavioral Modeling Results	129
8.11	Synthesis and Implications for Research Question 4	132
9	Integrated Discussion and System-Level Implications	134
9.1	Reframing the End-to-End Modeling Objective	134
9.2	Trust Dynamics Identification as a Core Contribution	135
9.3	Behavioral Coupling as a Diagnostic Stress Test	135
9.4	System-Level Lessons for Trust-Aware HRI Modeling	136
9.5	Positioning Within the Broader Trust Modeling Landscape	136
10	Conclusions and Future Recommendations	137
10.1	Problem Context and Thesis Positioning	137
10.2	Summary of Core Contributions	138
10.3	Answers to the Research Questions	138
10.4	Interpretation of Key Findings	139
10.5	Limitations of the Present Work	140
10.6	Implications for Trust-Aware Human-Robot Interaction Systems	141
10.7	Future Research and Recommendations	141
A	Full Derivations of the Proposed Trust Dynamics Model	144
A.1	Dispositional Trust	144
A.2	Situational Trust	144
A.3	Learned Trust Components	144
A.4	Latent Trust	152
A.5	Integrated Trust Dynamics Model	162
B	Online Experiment Implementation and Data Acquisition	164
B.1	Purpose of the Experimental Software	164
B.2	Scope of the Implemented Experiment Platform	164
B.3	Mapping Between Experimental Procedure and Software Modules	165
B.4	Participant Timeline and Door-Trial Structure	165
B.5	Condition Sets and Reproducible Assignment	165
B.6	Data Logging, Durability, and Withdrawal Support	166
B.7	Opening Statement and Participant Consent	166
B.8	Review Stimulus Pool	167
B.9	Experiment Walkthrough	168
B.10	Documentation and Repository Access	171

C	Software Implementation and Reproducibility	172
C.1	Purpose of the Software Framework	172
C.2	Scope of the Implemented Codebase	172
C.3	End-to-End Pipeline Implemented in Software	173
C.4	Reproducibility and Run Management	173
C.5	Role of Software in the Research Contribution	173
C.6	Reuse and Research Impact	174
C.7	Documentation and Repository Access	174
	References	175

List of Figures

5.1	Overview of the proposed modular trust dynamics architecture.	27
5.2	Expected behavior of dispositional trust.	29
5.3	Expected behavior of situational trust.	30
5.4	Expected behavior of reputation.	32
5.5	Illustration of asymmetric trust adaptation.	34
6.1	High-level timeline of the experimental protocol.	45
7.1	End-to-end identification and evaluation pipeline for the proposed trust dynamics model.	70
7.2	Age-range distribution of the validated analytic sample.	74
7.3	Distribution of validated participants across door-trial sequence sets.	74
7.4	Distribution of expected review scores prior to task interaction.	75
7.5	Expected versus response review scores across validated participants.	75
7.6	Distribution of pre-interaction trust questionnaire scores.	75
7.7	First intermediate trust questionnaire scores distribution.	77
7.8	Second intermediate trust questionnaire scores distribution.	77
7.9	Post-interaction trust questionnaire scores distribution.	77
7.10	Balance of participant binned pre-interaction trust scores across experimental sets.	78
7.11	Balance of participant expected review scores across experimental sets.	78
7.12	Distribution of participant-level weighted RMSE on the training set.	94
7.13	Distribution of participant-level weighted RMSE on the validation set.	94
7.14	Residual distributions and error metrics stratified by measurement type on the validation set.	95
7.15	Pairwise validation wRMSE response surface for ϕ^{fail} and ϕ^{succ}	98
7.16	Best-performing validation set participant (noticeably low wRMSE): simulated trust trajectories.	100
7.17	Typical validation set participant (near-median wRMSE): simulated trust trajectories.	101
7.18	Worst-performing validation set participant (highest wRMSE): simulated trust trajectories.	102
7.19	Additional representative validation set participant: simulated trust trajectories.	103
7.20	Additional representative validation set participant: proposed model versus best baseline.	103
8.1	BIC values by validation set participant for the three proposed behavioral models.	126
8.2	Distribution of split-to-split SD of k for the thresholded mapping model.	127
8.3	Distribution of split-to-split SD of k for the flexible mapping model.	127
8.4	Pooled transition structure on the validation split.	128
8.5	Pooled trust trajectory error over normalized time.	129
A.1	Unit exponential decay curve.	145
A.2	Sampled exponential decay.	146
A.3	Role of sampling interval.	147
A.4	Role of a	148
A.5	Role of $x_{1/2}$	149
A.6	Role of sampling interval.	150
A.7	Relationship between HRI successes and failures curves.	152
A.8	First transformation of the failure curve for latent trust modeling.	153
A.9	Second transformation of the failure curve for latent trust modeling.	154
A.10	Full transformation of the failure curve for latent trust modeling.	156
A.11	First transformation of the success curve for latent trust modeling.	157
A.12	Second transformation of the success curve for latent trust modeling.	159
A.13	Full transformation of the success curve for latent trust modeling.	160

B.1	Reputation exposure stage.	168
B.2	Demonstration (training) phase.	169
B.3	Drone scan phase.	169
B.4	Recommendation and decision stage.	170
B.5	Outcome feedback — empty door.	170
B.6	Outcome feedback — victim found.	171

List of Tables

5.1	Classification of model parameters by identification role.	39
6.1	Trial configuration for Set A.	49
6.2	Trial configuration for Set B.	49
6.3	Trial configuration for Set C.	50
6.4	Trial configuration for Set D.	50
6.5	Trial configuration for Set E.	50
6.6	Trial configuration for Set F.	51
6.7	Trial configuration for Set G.	51
6.8	Trial configuration for Set H.	51
6.9	Mapping between discrete simulation risk levels and visual/temporal parameters.	55
6.10	Summary of trust model quantities and their mapping to experimental signals.	58
7.1	Raw event counts by event type in the collected log.	71
7.2	Aggregated exclusion breakdown across the discarded participant sessions.	73
7.3	Gender distribution of the validated analytic sample.	74
7.4	Participant emergency decision choice.	74
7.5	Baseline parameters calibrated on the training set of participants.	89
7.6	Numerical values of the selected parameters following the WLS optimization solution.	93
7.7	Overall identification performance of the selected trust dynamics model.	93
7.8	Overall identification performance stratified by measurement-type on validation participant set.	95
7.9	Validation performance of candidate parameter estimates from different identification methods.	96
7.10	Pairwise Euclidean distances between candidate parameter vectors.	96
7.11	One-at-a-time sensitivity rankings based on median absolute validation wRMSE change.	97
7.12	Morris-like screening summaries (μ^* , σ) for all model parameters.	97
7.13	Overall baseline comparison on the validation set.	98
7.14	Validation wRMSE by measurement kind.	99
7.15	Overall improvement of the proposed model relative to baselines on the validation set.	99
8.1	Behavioral dataset composition and summary statistics per split (door-trial level).	125
8.2	Pooled probabilistic performance on validation set participants.	125
8.3	Discrimination (ranking) performance on validation set participants (ROC-AUC).	126
8.4	Per-participant behavioral fits on validation set.	127
8.5	Pooled closed-loop interaction signatures on the validation split.	128
B.1	Mapping between the experimental procedure and implemented software components.	165
B.2	Very positive reviews.	167
B.3	Slightly positive reviews.	167
B.4	Mixed reviews.	167
B.5	Slightly negative reviews.	167
B.6	Very negative reviews.	168
C.1	Mapping between methodological stages and implemented software modules.	173

1

Introduction

As autonomous and semi-autonomous robotic systems increasingly move from controlled laboratory environments into real-world operational contexts, they are expected to interact more closely with human operators, teammates, and decision-makers. In many of these settings, robots are no longer passive tools but active participants that provide information, recommendations, or guidance under uncertainty. The effectiveness and safety of such interactions depend not only on the technical capabilities of the robotic system, but also on how humans perceive, interpret, and rely on its actions. Among the psychological and cognitive factors that shape this interaction, trust has emerged as a central determinant of whether and how humans choose to engage with robotic assistance (J. D. Lee and See, 2004).

Trust in human-robot interaction is inherently dynamic. Human trust evolves over time as a function of experience, system performance, contextual risk, and individual predispositions, often changing rapidly after salient interaction events such as failures or recoveries. These temporal dynamics are particularly consequential in settings where humans must repeatedly decide whether to rely on a robot's recommendations or intervene manually, and where inappropriate trust calibration can lead to misuse, disuse, or unsafe reliance (Hoff and Bashir, 2015). Understanding how trust forms, adapts, and stabilizes during interaction is therefore essential for designing robotic systems that can be integrated responsibly into human decision-making processes.

While the importance of trust is widely acknowledged, modeling trust in a way that is both scientifically grounded and operationally useful remains challenging. Trust is not directly observable; it is a latent internal state that must be inferred from subjective reports, behavioral indicators, or physiological signals, all of which are noisy and context-dependent. As a result, many existing studies focus either on identifying factors that influence trust or on predicting trust measurements at specific moments in time, without explicitly addressing how trust evolves as a dynamic process or how such dynamics can be empirically identified from realistic interaction data. This gap becomes particularly apparent when trust models are intended for use in simulation, control, or decision-support pipelines, where models must be executable, interpretable, and amenable to validation.

This thesis adopts the viewpoint that progress in trust-aware human-robot interaction requires treating trust modeling as an end-to-end identification problem. Rather than viewing trust solely as a descriptive psychological construct or a static predictor of behavior, trust is modeled here as a latent, time-varying internal state whose evolution can be formally specified, empirically estimated, and stress-tested against observed data. From this perspective, the central challenge is not only to propose a plausible mathematical structure for trust dynamics, but also to demonstrate that such a model can be identified from human-subject experiments, generalized beyond the data used for estimation, and meaningfully related to observable interaction outcomes.

The work is motivated by high-stakes human-robot collaboration scenarios, such as search and rescue, where uncertainty, time pressure, and risk naturally amplify the consequences of trust miscalibration. These scenarios provide a concrete and intuitive context in which trust dynamics are expected to be strongly excited and observable. However, the thesis does not aim to deliver a deployable system for any specific application domain. Instead, such contexts are used to guide experimental design and to ground modeling choices, while the primary contribution lies in the development and evaluation of a reusable modeling and identification framework that can be adapted to other human-robot interaction settings.

A key design choice throughout the thesis is the explicit separation between trust modeling and behavioral decision modeling. Trust is treated as an internal state that influences, but does not uniquely determine, human actions. This distinction allows the proposed trust model to be evaluated on its own merits, while also enabling an exploratory investigation of how inferred trust states can be coupled with simple behavioral models to assess their explanatory limits. Rather than assuming a fixed mapping from trust to behavior, the thesis uses this coupling as a diagnostic tool to reveal where trust-based explanations are sufficient and where additional cognitive or contextual factors are required.

The remainder of this thesis is structured as follows. Chapter 2 reviews the literature on trust in automation and human-robot interaction, surveying established definitions, influencing factors, dynamic trust models, and measurement challenges, and identifying gaps that motivate the present work. Chapter 3 formalizes these gaps into a set of research objectives, contributions, and research questions, framing trust modeling as an end-to-end identification and validation problem. Chapter 4 introduces the theoretical and methodological background required for the modeling and identification approach adopted in the thesis. Chapter 5 presents the proposed dynamic mathematical model of human trust, including its structure, assumptions, and design rationale. Chapter 6 describes the human-subject experiment and data collection protocol designed to elicit informative trust dynamics. Chapter 7 develops and applies the parameter identification and validation pipeline, assessing model fit, generalization, and parameter robustness. Chapter 8 explores the coupling of the identified trust model with a behavioral decision model in a closed-loop interaction setting, using this coupling to analyze behavioral relevance and limitations. Chapter 9 provides an integrated discussion of the results and examines their system-level implications for trust-aware human-robot interaction and modeling. Finally, chapter 10 synthesizes the overall findings, reflects on limitations, and outlines directions for future research.

2

Literature Review

This chapter provides a comprehensive review of the literature that motivates and constrains the development of an identifiable, trust-aware human-robot interaction model, with Search and Rescue (SaR) serving as the primary motivating deployment context. It first clarifies the scope and research trajectory that led from an initial goal of trust-aware SaR simulation integration to the need for formulating and empirically validating a structured trust dynamics model. The chapter then introduces foundational concepts of trust in human–technology interaction and summarizes established definitions and dimensions of trust in automation and human-robot interaction (HRI), including key factors known to influence trust. Building on this foundation, it surveys computational approaches for modeling trust dynamics over time, highlighting their typical assumptions, strengths, and limitations, as well as the associated challenges of measuring trust reliably in HRI. The chapter subsequently reviews SaR as a high-risk application domain and discusses the specific HRI challenges that arise in emergency contexts. Finally, it examines representative works at the intersection of trust dynamics and agent-based SaR simulation, using them to expose a gap in the current literature that motivates the research objectives and questions addressed in this thesis.

2.1. Scope and Research Trajectory

The original objective motivating this research was the implementation of a human trust dynamics model within a simulated Search and Rescue environment, with the aim of enabling trust-aware human-robot interaction in complex, high-risk missions. However, a preliminary review of the existing literature revealed that currently available trust models are either highly context-specific, largely probabilistic and descriptive in nature, or insufficiently structured to support explainable, modular, and generalizable integration into simulation or control pipelines. As a result, the focus of this work shifted progressively backward along the implementation pipeline: from simulation integration, to trust model formulation, to empirical validation, and ultimately to the design of a dedicated experimental setup for data collection. This literature review therefore reflects this discovery process, in which SaR simulation remains the motivating deployment context, while the formulation and validation of a dynamic trust model constitutes the primary contribution of this thesis.

2.2. Introduction to Human-Technology Interaction

The rapid advancements in automation, artificial intelligence (AI), and robotics are fundamentally reshaping various aspects of human life, from daily routines to highly critical professional domains such as search and rescue missions, healthcare, and military operations (Murphy, 2004; Unhelkar et al., 2014; Rantanen et al., 2017). As these technological systems become increasingly sophisticated, their roles are evolving from mere tools to collaborative partners, or even autonomous agents. This evolution necessitates a deeper understanding of the human-technology relationship, particularly the concept of trust, which is widely recognized as a cornerstone for effective and synergistic human-machine teaming (J. D. Lee and See, 2004).

In contexts where humans and automated systems or robots collaborate, the level of trust an operator places in the technology is paramount (Lewis et al., 2018). Inappropriate trust, whether it be overtrust (misuse or over-reliance) or undertrust (disuse or under-reliance), can lead to detrimental outcomes, including compromised safety, reduced performance, and even rejection of potentially beneficial systems (Du and Yang, 2019). For instance, overtrust in an autonomous system might lead an operator to become complacent and

fail to intervene when necessary, while undertrust might cause an operator to manually perform tasks that the automation could handle more efficiently, thereby increasing workload and decreasing overall system effectiveness (Ullrich et al., 2021). This highlights the critical need for appropriate trust calibration, where a human's trust level accurately matches the system's actual capabilities and reliability.

While the literature on human-automation interaction (HAI) has extensively explored trust (J. D. Lee and See, 2004), human-robot interaction (HRI) presents unique challenges and opportunities that warrant specific investigation (Hancock et al., 2011). Unlike many conventional automated systems, robots are often mobile, can exhibit human-like or animal-like forms, and are designed to effect action at a distance. Moreover, the increasing sophistication of robots, including their ability to communicate through speech, can lead people to attribute intentionality and impute motivation to them, further complicating trust dynamics (Schaefer et al., 2016). Therefore, understanding and modeling human trust in robots is crucial for designing, evaluating, and training approaches that promote appropriate trust and reliance, especially as robots become integral members of human teams in dynamic and uncertain environments. This literature review will systematically explore the conceptual foundations of trust, the varied factors influencing HRI trust, existing approaches to modeling trust dynamics, and the current state of research in high-risk, human-robot collaborative contexts like search and rescue, ultimately identifying critical knowledge gaps that inform the objectives of this thesis.

2.3. Trust in Human-Technology Systems

Trust is a fundamental aspect of human interaction, serving as an "adhesive" that connects people across various levels, from interpersonal relationships to communities, organizations, governments, nations, cultures, and societies (Hoff and Bashir, 2015). The significance of trust extends beyond the interpersonal domain to define how people interact with technology (Sapienza et al., 2022). With the increasing prevalence of automated technologies, such as flight management systems and GPS route planners, the concept of trust in automation has become a substantial area of research (Hoff and Bashir, 2015).

Automation is used for various functions, including information acquisition and analysis, decision-making, action execution, and monitoring. When human-automation teams operate optimally, labor system efficiency can drastically improve, but introducing automation into critical systems can also create new pathways for error, sometimes with severe consequences (Hoff and Bashir, 2015). Trust is crucial in guiding human reliance on automation (S. M. Merritt et al., 2013). However, the recent surge in research related to trust and reliance has produced many confusing and seemingly conflicting findings, highlighting the need for integrated reviews (Schaefer et al., 2016).

Trust itself is a rich and multi-faceted construct, studied across many disciplines, including philosophy (Plato, 1955; Baier, 1986), computer science (Sabater and Sierra, 2005), neuroscience, psychology, and economics (Krueger and Meyer-Lindenberg, 2019). Each discipline may rely on its own definitions and findings, often lacking agreement and generalization (J. D. Lee and See, 2004). Despite this, at its core, trust involves a belief in the competence and reliability of another entity (Salem et al., 2015). In the context of human-automation interaction, J. D. Lee and See (2004) define trust as the "attitude that an agent will help achieve an individual's goals in situations characterized by uncertainty and vulnerability". This definition is widely used in human-autonomy interaction research (S. M. Merritt et al., 2013; Lu and Sarter, 2019; Guo and Yang, 2021). Their integrated review elucidated the role of trust and associated feelings in guiding human behavior toward automation, stimulated numerous studies on factors influencing trust variability, and encapsulated key elements crucial for understanding trust in technological systems:

- **Expectation:** trust involves an anticipation or subjective probability that certain future events or behaviors of the trustee will occur favorably. Rotter (1967) additionally defined trust as "an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon".
- **Uncertainty:** trust becomes relevant in situations where outcomes are not fully predictable or guaranteed. If an outcome were certain, trust would not be necessary.
- **Vulnerability:** the trustor puts something at stake (e.g., safety, performance, resources) whose outcome depends on the actions of the trustee. This willingness to be vulnerable is a core aspect of trust.

While trust is primarily conceptualized as a belief or attitude, it also manifests as behavioral trust, which refers to the actions resulting from that subjective expectation, such as relying on or complying with the automation. This interplay between belief and behavior is dynamic; for instance, if a system is not trusted, it is unlikely to be used, which in turn limits the operator's ability to gather information about its capabilities, thus

hindering the growth of trust. The foundation of trust can also be understood through various dimensions or attributional abstractions. Muir (1994) described trust in terms of predictability, dependability, and faith, forming a long-standing foundation for empirical studies in automation. J. D. Lee and See (2004) further updated this taxonomy to highlight three key dimensions:

- **Performance:** reflects the agent's competence, ability, and reliability in achieving goals. This is often considered the most influential dimension in human-automation and human-robot trust.
- **Process:** refers to the underlying mechanisms, integrity, persistence, and predictability of the agent's operation. Transparency, i.e. understanding how the system works, is often linked to this dimension.
- **Purpose:** encompasses loyalty, benevolence, and value congruence, reflecting the agent's perceived intentions to act in the trustor's best interest.

A fundamental challenge in extrapolating human-human trust concepts to human-automation or human-robot trust is that automation often lacks intentionality. While loyalty and benevolence are central to interpersonal trust, automation is designed with a purpose and thus embodies the intentionality of its designers. However, as robots become more sophisticated and exhibit human-like characteristics (e.g., speech communication, anthropomorphic features), people may attribute intentionality and impute motivation to them, making intention-based trust an "unignorable component" in human-robot teaming.

The intrinsic link between trust and risk is a recurring theme across the literature (Lewis et al., 2018; Ullrich et al., 2021; Chen et al., 2020). Trust is not merely a component of decision-making in risky situations; it serves as a mechanism for reducing social complexity and managing uncertainty and vulnerability in such contexts. As Das and Teng (2004) argue, subjective trust and perceived risk are often mirror images of each other, where higher perceived risk correlates with lower subjective trust, and vice versa. Furthermore, behavioral trust can be viewed as risk-taking, as the decision to rely on an agent in a high-stakes situation inherently involves accepting potential negative consequences (Das and Teng, 2004).

An individual's risk propensity, i.e., their general tendency to be either risk-seeking or risk-averse, also significantly influences their trust decisions. This dispositional characteristic can lead to variations in how individuals assess risk and subsequently decide to trust an automated system or robot. For instance, a risk-averse individual may exhibit lower initial trust or be less willing to engage in high-risk reliance, regardless of the system's capabilities. This complex interplay between the nature of the risk, the perceived trustworthiness of the agent, and the trustor's individual characteristics underscores the multifaceted nature of trust in human-technology interaction.

Much of the existing research on factors guiding human-automation interaction centers around trust, as it often determines the willingness of human operators to rely on automation (Hoff and Bashir, 2015). Various studies have utilized different automated systems and experimental paradigms to identify factors impacting operators' trust. To synthesize recent findings and address the diversity of variables and methodologies, systematic reviews (Hoff and Bashir, 2015) and meta-analyses (Hancock et al., 2011; Schaefer et al., 2016) have been conducted. Hoff and Bashir (2015) conducted a systematic review of empirical research on trust in automation from 2002 to 2013, identifying 101 papers with 127 studies. Their analysis led to a three-layered trust model encompassing dispositional trust, situational trust, and learned trust. This model provides a framework for conceptualizing the variability of trust and can guide future research, training interventions, and design procedures for appropriate trust.

2.3.1. Trust in Human-Robot Interaction

Human-Robot Interaction (HRI) represents a specific domain within human-automation interaction, where trust is particularly critical (J. D. Lee and See, 2004; Hutter and Siegwart, 2018; Sapienza et al., 2022). While trust in automation has been extensively studied (Hancock et al., 2011), robots differ from typical automated machines or computer interfaces in key aspects such as mobility and a greater degree of embodiment (Salem et al., 2015). These differences can lead to a potentially richer interaction and introduce different dimensions of risks and safety concerns, suggesting that human trust may vary for robots compared to other forms of automation or even Human-Computer Interaction (Hancock et al., 2011). Given the growing need for synergistic human-machine teaming (Schaefer et al., 2016; Mahani et al., 2021; Sapienza et al., 2022), understanding trust in HRI is increasingly important.

Early reviews of trust in HRI were primarily qualitative or descriptive. To provide a fundamental empirical foundation, quantitative reviews like meta-analyses have been performed. Hancock et al. (2011) conducted a meta-analysis of factors affecting trust in HRI. Their objective was to evaluate and quantify the effects of

human, robot, and environmental factors on perceived trust in HRI. They collected a total of 29 empirical studies published between 1996 and 2010, of which 10 met the selection criteria for correlational analysis and 11 for experimental analysis. These studies provided 69 correlational and 47 experimental effect sizes. Their findings produced quantitative estimates of human, robot, and environmental factors influencing HRI trust. Their analysis led to the three-factor model of human-robot trust, emphasizing the triad of human, robot, and environmental factors and their associated antecedents of trust (Hancock et al., 2011).

2.4. Factors Influencing Human-Robot Trust

The meta-analysis by Hancock et al. (2011), initially constrained to HRI, and later expanded to all automation (Schaefer et al., 2016), solidified the understanding that factors influencing trust development in HRI situations can be dynamically influenced by the robotic system itself, the surrounding operational environment, and the nature and characteristics of the respective human team members (Hancock et al., 2011). These factors are often categorized into robot-related, human-related, and environment-related factors (Schaefer et al., 2016; Guo and Yang, 2021).

2.4.1. Robot-Related Factors

The Hancock et al. (2011) meta-analysis indicated that robot-related factors are the most important element of trust. These factors, such as robot performance and attributes, can often be directly manipulated by designers within technological constraints, allowing for some predictability in trust development in human-robot partnerships. Subsequent meta-analyses on trust in automation, which included HRI, also highlighted robot features and capabilities as significant (Schaefer et al., 2016).

Performance-based Factors

Reliability and validity of an automated system's functions are important antecedents of trust (Lu and Sarter, 2019). Reliability refers to the consistency of functions, and validity refers to performing the intended task (Hoff and Bashir, 2015). Predictability (consistency with operator expectations) and dependability (frequency of breakdowns) are also crucial, with operators trusting systems more when they are predictable and dependable (Schaefer et al., 2016). In fact, a robot's consistent and reliable performance is the most impactful factor in building and maintaining trust.

Empirical studies have consistently shown that higher trust is associated with higher reliability (Mahani et al., 2021). Poor robot performance negatively affects operator trust (Salem et al., 2015; Robinette et al., 2016), and these changes can be measured over time. Even a single failure by a robot can sharply reduce a person's trust in it for future interactions.

System faults, a specific form of reliability issue, significantly influence trust. The magnitude and timing of errors matter; larger faults negatively affect trust more severely and take longer to recover. Research suggests that early reliability drops can have a profound, sometimes masked, impact on trust (Desai et al., 2013). Feedback mechanisms that convey robot confidence can also interact with trust and control allocation strategies.

Attribute-based Factors

The inherent characteristics of the robot, beyond its task performance, also play a role in trust development. Namely, robot attributes such as type, size, proximity, and behavior affect trust (Hancock et al., 2011). Design features like appearance, ease of use, communication style, transparency, and level of control are important considerations (Hoff and Bashir, 2015).

The physical form, human-like features, or even gestures of a robot can also influence perceived trust. Appearance-based anthropomorphism, while perhaps more specific to embodied agents like robots, is a more recent research area. Increasing the humanness of an agent may increase trust resilience, but caution is advised as different perceptions of reasoning and motivation can lead to incorrect trust calibration (Culley and Madhavan, 2013). In fact, people may attribute intentionality and impute motivation to robots that appear more human-like, which can lead to incorrect trust calibration if they are misled about the robot's true capabilities.

A robot's communication style (e.g., speech, non-verbal cues, explanations of its actions) and the transparency of its underlying mechanisms can also affect trust. Transparency, which allows humans to understand how an autonomous system works and predict its behavior, is believed to influence trust.

The degree of control the automation is designed to exercise can impact trust, although there may not be a simple direct relationship (Rani et al., 2000; S. M. Merritt et al., 2013). Appropriate levels of automation

(Parasuraman et al., 2000), matched to task requirements and human capabilities, are crucial for fostering trust (Davis, 2019).

2.4.2. Human-Related Factors

The Hancock et al. (2011) HRI meta-analysis found limited evidence for human-related factors, suggesting a strong need for future experimental efforts in this area. However, the broader meta-analysis on trust in automation found a larger empirical basis for the influence of stable characteristics of the individual on trust development (Schaefer et al., 2016). Factors pertaining to the individual characteristics of the human interacting with the robot include traits and states, as well as cognitive and emotive dimensions.

An individual's general disposition or tendency to trust others, including technology, is an important starting point for trust development (Davis, 2019). This propensity to trust, also referred to as dispositional trust, is often considered a relatively stable trait (Hoff and Bashir, 2015).

A user's self-confidence in their own ability to perform a task can inversely correlate with trust in automation; when self-confidence is low, trust in automation may increase, leading to greater use (Schaefer et al., 2016). Similarly, an operator's expertise or familiarity with a task or system can influence their trust levels.

Cognitive aspects like attentional capacity, workload and situation awareness, and emotive factors like mood and stress influence trust development. For instance, positive emotions can lead to higher initial trust. High cognitive workload can also moderate the effects of implicit attitudes on trust.

Past interactions with robots, both successful and unsuccessful, heavily influence dynamically learned trust (Ullrich et al., 2021). This prior experience builds a "generalised expectancy" that shapes future trust perceptions.

2.4.3. Environment-Related Factors

Environmental factors were found to be moderately influential on trust development in the HRI meta-analysis (Hancock et al., 2011). These factors describe the operational setting and task characteristics that mediate trust development, and include task complexity, workload, system accuracy (Desai et al., 2013), team collaboration characteristics, and tasking factors (Hancock et al., 2011).

The nature of the task (e.g., cognitive, control, perceptual), its complexity, and the multi-tasking requirements influence trust. For example, high-risk, time-critical situations, such as emergency evacuations, uniquely engage human cognitive processes and fight-or-flight responses, impacting trust in difficult-to-predict ways (Kinsey et al., 2019; Kinatader and Warren, 2021). The perceived risk, including the type and amount of potential loss, directly influences a person's decision to trust a robot.

The automation literature highlights that risk and uncertainty are of essential importance in trust development, deriving from both internal (e.g., data corruption) and external sources (e.g., physical terrain) (Schaefer et al., 2016). Task and function-specific aspects of automation also play an important role. Characteristics of the human-robot team, including shared mental models and the presence of other agents, can affect trust. Environment-related factors have been more frequently studied in HRI, potentially due to efforts in teaming and social interaction (Schaefer et al., 2016).

2.5. Modeling Trust Dynamics

Understanding the factors that influence human-robot trust is crucial for developing appropriate trust models that can predict and manage human-robot interaction effectively. However, the interplay among these factors is complex and often non-linear, posing significant challenges for comprehensive modeling.

The majority of early literature on human-automation trust adopted a "snapshot" view, typically evaluating trust at a single point in time, through questionnaires usually administered at the end of an experiment (Campagna and Rehm, 2025). However, this static approach poses a significant limitation, as it fails to acknowledge that trust is a dynamic variable that can strengthen or decay over time and repeated interactions (Yang et al., 2023). Trust can be dynamically influenced by the factors discussed previously, and the process by which trust develops over time requires further investigation (Hancock et al., 2011). Understanding how trust evolves during a human's interaction with a system, particularly how it is impacted by events like failures, is crucial.

Modeling trust dynamics usually involves capturing how a human agent updates their trust based on interacting with a robotic agent and observing its performance over time. Recent research efforts have increasingly focused on developing computational models capable of capturing these real-time trust dynamics (Guo and Yang, 2021).

The seminal work by Muir (1994), which conceptualized trust in terms of predictability, dependability, and faith, provided an early taxonomy that served as a foundation for empirical studies. J. D. Lee and See (2004) further refined this by highlighting performance, process, and purpose as key dimensions for understanding human-automation trust.

2.5.1. Three-Layered Trust Model Framework

Hoff and Bashir (2015) presented a systematic review of empirical research on factors influencing trust in automation. Their comprehensive analysis, which included 101 papers and 127 human-subjects experiments, culminated in a three-layered trust model designed to conceptualize the variability of trust in automation across different systems and human operators. The fundamental premise is that appropriate levels of trust are crucial to reduce misuse and disuse of automated systems. The three proposed layers of trust proposed are dispositional, situational and learned trust (Hoff and Bashir, 2015).

Dispositional Trust

This layer represents an individual's inherent tendency to trust automation. It is considered a relatively stable characteristic over time, unlike situational or learned trust (S. Merritt and Ilgen, 2008). Factors influencing dispositional trust are rooted in biological and environmental aspects of the individual. While Hoff and Bashir (2015) noted a scarcity of research in this area at the time of their review, they identified potential influences such as culture, age, gender, and personality traits. Other sources reinforce the idea of a "propensity to trust" as a stable, long-term individual characteristic (Davis, 2019). Research suggests that positive implicit and explicit attitudes towards automation can have an additive effect on trust.

From a modeling perspective, these findings imply that any trust dynamics formulation must include a persistent latent component reflecting individual predispositions, which cannot be fully overridden by short-term interaction outcomes or isolated events.

Situational Trust

This layer describes trust that is dependent on the specific context of an interaction with an automated system. It is highly variable and can change significantly based on the environment and the operator's mental state. Factors are generally divided into external environmental factors and internal variations in the operator's mental state.

The former encompasses elements such as perceived risk, workload, and the presence of distractions. For example, studies suggest that under high-risk conditions, operators might reduce reliance on complex automation but increase reliance on simple automation (Hoff and Bashir, 2015). Conversely, high workloads may compel operators to use automation (Wetzel et al., 2004). The perception of risk is a critical aspect, with some viewing subjective trust and perceived risk as mirror images (Das and Teng, 2004). The willingness to take risks is often seen as a common characteristic across trust situations, and the possibility of failure is necessary for trust to be present (Hardin, 2006).

The latter includes factors like self-confidence, expertise, mood, and attentional capacity. For instance, operators with lower self-confidence in their manual ability may be more likely to trust a robot. Similarly, a user's expertise or familiarity with a subject can influence trust. Mood can also play a role, with individuals in a pleasant mood being more prone to initially trust in robots. Hoff and Bashir (2015) also highlighted that trust has a greater influence on reliance when operators have a better ability to evaluate the automation's performance relative to their own unaided performance.

This body of work suggests that trust dynamics models must explicitly account for exogenous, time-varying situational factors, enabling trust to respond to contextual changes independently of accumulated interaction history.

Learned Trust

This layer is formed and modified based on past experiences relevant to a specific automated system. It is closely intertwined with situational trust, but specifically relates to experiences with the automated system itself. Hoff and Bashir (2015) further divided this layer into initial and dynamic learned trust.

Initial learned trust refers to any preexisting knowledge about the system before direct interaction, such as its reputation (Ullrich et al., 2021) or brand awareness. This type of trust can be influenced by information from others, like "gossip" about false alarm rates.

Dynamic learned trust evolves over time through direct interaction and experience with the automated system, and is influenced by key factors such as design features, performance and preexisting attitudes and expectations.

Consequently, trust dynamics models are required to support history-dependent evolution, allowing trust to be shaped cumulatively by interaction outcomes rather than being determined solely by instantaneous observations.

Hoff and Bashir (2015) emphasized five major design features: appearance, ease of use, communication style, transparency, and level of control. Designing transparent systems that provide accurate, useful feedback about their reliability and operation is crucial for appropriate trust. Ease of use and even "automation personality" conveyed through communication style and politeness can influence trust. The perceived aesthetics of an interface can also induce trust.

Trust is highly dependent on the real-time performance of the automation. Reliability, validity, predictability, and dependability of the system's functions are critical antecedents. The effect of automation faults on trust is dynamic, with an acute decline after failures and gradual recovery over time. Studies indicate that errors, especially early ones, can significantly impact trust and recovery rates. The distinction between false alarms and misses in automation errors can also differentially affect trust depending on the context. Trust tends to stabilize over repeated interactions with the same autonomous agent Yang et al., 2022.

These can be condensed into three central characteristics of trust dynamics in human-robot interactions validated by Yang et al. (2022), namely:

1. trust is hard to gain, i.e., individual failed interactions have a substantial impact on human-robot trust loss;
2. trust is easy to lose, i.e., individual successful interactions have a marginal positive impact on human-trust formation;
3. trust is bounded, i.e., it stabilizes after repeated human-robot interactions.

Both conscious (explicit) and unconscious (implicit) attitudes can alter the trust formation process and subsequent usage decisions. People may initially have a "positivity bias" or assume machines are perfect, leading to high initial trust based on "faith" that rapidly dissolves after errors (Hoff and Bashir, 2015).

Hoff and Bashir (2015) emphasized that trust guides, but does not completely determine, reliance on automation. Reliance and decision-making are influenced by many factors beyond trust, including situational constraints such as workload and time pressure, as well as properties of the automation and the task (Hoff and Bashir, 2015; M. Y. Zhang and Yang, 2017). There is a closed-loop dynamic between trust and reliance: if a system is not trusted, it will not be used, which limits the opportunity to gather information about its capabilities and hinders trust growth. The strength of the relationship between trust and reliance is enhanced when operators have a greater ability to evaluate the automation's performance and possess more decisional freedom (Hoff and Bashir, 2015).

This three-layered framework provides a valuable lens for conceptualizing the variability of trust and can guide the design of appropriate trust measurements. More recent work by Alzahrani and Ahmad (2026) and Ahmad et al. (2023) has focused on mathematically emulating this three-layered framework to estimate human trust in robots in real-time, even validating the model in repeated HRI user studies. Their work aims to design robotic systems that can adapt to optimize human trust using these validated model metrics.

2.5.2. From Trust Dynamics to Decision-Making: The Trust-Behavior Gap

A recurring theme in the trust-in-automation and HRI literature is that trust is best understood as a latent psychological attitude that guides interaction with automation, rather than a direct behavioral policy. In particular, J. D. Lee and See (2004) emphasize that trust influences reliance while not completely determining it, implying that observable behavior cannot be treated as a one-to-one proxy for the underlying trust state. This conceptual separation is echoed in integrated evidence reviews, which discuss how reliance decisions may vary substantially even under similar subjective trust levels due to other contextual and cognitive influences (Hoff and Bashir, 2015). For computational modeling, this distinction is important: a trust dynamics model may be valid as a predictor of trust trajectories while still being insufficient, on its own, to reproduce human decision-making behavior in closed loop.

A central reason is that reliance and broader decision-making in HRI are inherently multifactorial. Even in paradigms where trust correlates with the tendency to comply with or depend on automation, behavior is also shaped by situational constraints and operator state variables that do not reduce to trust. For example, reviews of trust in automation highlight the influence of factors such as workload, time pressure, and attentional limitations on how operators use automation, including cases where these factors affect reliance without necessarily reflecting a corresponding change in subjective trust (Hoff and Bashir, 2015). Empirical findings

in dual-task contexts further suggest that workload can moderate the relationship between subjective trust and dependence-related behavior, such that reliance patterns may change even when reported trust does not exhibit a comparable shift (M. Y. Zhang and Yang, 2017). These findings motivate a cautious interpretation of trust-behavior couplings: while trust is behaviorally relevant, its explanatory power is necessarily limited when other latent decision variables are unobserved.

This gap is also directly connected to the long-standing objective of appropriate reliance and trust calibration. Inappropriate trust can manifest as misuse or disuse, with safety and performance consequences that motivate modeling approaches capable of reasoning about both trust formation and its behavioral implications (J. D. Lee and See, 2004; Hancock et al., 2011; Schaefer et al., 2016). At the same time, the literature on behavioral and agent-based decision modeling suggests that credible models of human action in complex environments often require additional decision structure beyond a single trust variable. In cognitive modeling traditions, the Belief-Desire-Intention (BDI) framework represents behavior as emerging from interacting beliefs, goals, and intentions (Bratman, 1987). Building on this idea, modular social-agent architectures incorporate affective and social mechanisms such as emotions, norms, and social relations alongside cognition to generate richer and more explainable decision-making (Bourgais et al., 2020), and evacuation-oriented agent models similarly incorporate socio-cultural and affective dynamics to reproduce group behavior under stress (van der Wal et al., 2017). These perspectives support the interpretation that trust is best viewed as one informative component of a broader decision-making process in HRI, rather than as a standalone substitute for a behavioral model.

In this thesis, this literature motivates a deliberate separation between (i) identifying and validating a structured trust dynamics model as a latent-state predictor, and (ii) treating trust-behavior coupling as an exploratory diagnostic layer to assess where trust-informed decision models succeed or break down under realistic empirical constraints. This framing is consistent with the view that trust guides reliance while not fully determining it (J. D. Lee and See, 2004; Hoff and Bashir, 2015), and it ensures that limitations in behavior reproduction are interpreted as informative evidence about model scope rather than as a failure of trust modeling itself.

2.5.3. Existing Trust Dynamics Models

Early quantitative models for trust in automation included regression-based models, neural network models, and time-series models (Y. Wang et al., 2024). While regression models established quantitative relationships between trust and influencing factors, they often lacked the ability to capture the dynamic variance in trust development. Neural networks could discover complex relationships but required large datasets and often lacked interpretability.

Y. Wang et al. (2024) meticulously categorizes computational human-robot trust models into five principal classes, distinguished by their underlying modeling techniques: performance-centric algebraic, time-series, Markov Decision Process (MDP)/Partially Observable MDP (POMDP), Gaussian, and dynamic Bayesian network (DBN)-based trust models. This comprehensive categorization provides a structured overview of the various approaches employed to understand and predict trust dynamics in human-robot interaction (HRI).

A significant contribution of this survey is its emphasis on the utilization of computational human trust models in control and robotics applications. Quantifying human trust can be used to design robot motion planning and control algorithms, predict and optimize the dynamic allocation of autonomous functions, and provide objective metrics for selecting appropriate Levels of Autonomy. This directly addresses the need for models that can inform practical robot design and deployment. The authors highlight that computational models can complement and reduce the reliance on extensive empirical data for initial model development and testing, particularly by enabling numerical simulations to demonstrate overall frameworks where human psychological aspects are simplified or modeled rationally. However, they also acknowledge that most current models are "trust-like" rather than fully comprehensive, often restricting trust to discrete values and neglecting many important influencing factors like risk or true benevolence. The challenge of data acquisition from human subject tests is a common limitation for trust modeling, which computational models and simulations can help to mitigate by allowing for exploration of scenarios that might be impractical or unethical to test empirically.

A brief discussion on the five identified model categories follows. For a thorough review of all these models, please refer to the original source by Y. Wang et al. (2024).

Performance-Centric Algebraic Trust Models

These models are characterized by their deterministic and algebraic nature, fundamentally centering on robot performance or the performance of the human-robot team as the primary determinant of trust. They employ algebraic formulations to compute performance and, subsequently, human-robot trust, based on observed

data.

For instance, Floyd et al. (2015) proposed a performance-centric computational trust model for a simulated wheeled unmanned ground robot collaborating with a human agent in movement and patrol tasks. The model's objective was to enable the robot to adapt its behavior to enhance team trustworthiness, specifically by modifying elements like speed and obstacle padding. Another example involved the development of a trust-based information management system designed to identify anomalous Unmanned Ground Vehicles (UGVs) that require human intervention. This system functioned by decreasing the perceived importance of information originating from untrusted UGVs, prioritizing security even at the expense of efficiency due to potentially slower trust recovery (Li et al., 2019).

The main benefit of these models is their efficient mechanism for evaluating trust based on observable evidence. However, they often encounter limitations, such as conflating trust with trustworthiness, which can lead to inaccuracies. While trustworthiness is an intrinsic, objective property of the robot, trust is a perceived quality influenced by human factors and emotions. Another significant drawback is their tendency to disregard the inherent uncertainty, stochasticity, and dynamic emotional aspects of human nature. Emotions are critical for directing attention and influencing decision-making, cooperation, and social interaction. Furthermore, these models typically do not incorporate human subject tests for parameter learning, and a unified framework or quantitative metrics for consistently modeling robot performance or trust across varied scenarios remains largely absent (Y. Wang et al., 2024).

Time-Series Trust Models

These models are specifically crafted to capture the dynamic progression of trust and the multifaceted causal factors that exert influence on human-robot trust. These factors typically encompass the prior trust value (reflecting trust memory), current and historical robot and human performance, alongside other robot-specific, human-specific, and contextual elements.

J. Lee and Moray (1992) proposed an auto-regressive moving average vector (ARMAV) time series model, which calculated trust at the present moment as a function of previous trust, task performance, and the occurrence of automation failures. This model highlighted key properties of trust dynamics.

- **Dependence on Past Trust:** trust at the current moment is significantly influenced by trust at the previous moment (J. Lee and Moray, 1992).
- **Asymmetric Influence of Success/Failure:** negative experiences with autonomy (failures) typically exert a stronger influence on trust adjustment than positive experiences (successes) (J. Lee and Moray, 1992; Manzey et al., 2012; Yang et al., 2016).
- **Stabilization:** human agents' trust tends to stabilize over repeated interactions with the same autonomous agent (J. Lee and Moray, 1992; Yang et al., 2017; Yang et al., 2022).

More advanced time-series models have been developed to encode human intention into multi-robot motion tasks, requiring massive amounts of human-robot interaction data and often relying on Bayesian inference and Markov Chain Monte Carlo (MCMC) sampling for parameter estimation (Shahriari et al., 2016). These models can explain the influence of environmental attributes on human trust in multi-robot systems, but their reliance on self-reported trust introduces potential bias and simulation fidelity issues.

An illustrative example is the work by Zheng et al. (2023), who extended time-series models to a human multi-robot trust model. This model was utilized to integrate human intention into multi-robot motion tasks within offroad environments. It quantifies the influence of robot attributes, such as traversability and visibility, using a linear state-space equation that also incorporates human dispositional trust. Notably, this model explicitly accounts for the influence of a preceding robot on trust in a succeeding robot during formation control situations.

The primary strengths of time-series models are their ability to effectively capture the dynamic and multi-dimensional aspects of trust, illustrating that changes in trust are a gradual process exhibiting inertia. They are also capable of accommodating data uncertainty and identifying parameters for optimal estimation and forecasting. Conversely, a common limitation is that many time-series models presume linearity, which can oversimplify the complex relationship between trust and its influencing variables. Like algebraic models, they also suffer from a lack of a unified framework for modeling robot performance and trust consistently across diverse application scenarios.

Markov Decision Process (MDP)-Based Trust Models

These models are employed to delineate stochastic decision-making processes in discrete time, encompassing states, actions, and transitions. Their utility is particularly pronounced in optimizing overall system performance through the integration of the trust state with robot behaviors. These are also frequently expanded to Partially Observable MDP (POMDP) problems. In a POMDP framework, human trust is treated as a hidden, partially observable state that can be inferred from human actions. These models can provide systematic descriptions of the transition process for human trust levels and offer optimal policies for human-robot collaboration, especially for task assignment based on satisfying tasks specifications while minimizing costs.

Nam et al. (2020) developed a human robotic swarm trust model for the supervisory control of foraging and search missions. Their analysis across different Levels of Automation (manual, mixed-initiative, and fully autonomous) revealed that human trust was more significantly affected by the swarm's physical characteristics (e.g., heading angles, convex hull area) than by task performance itself (Nam et al., 2017; Nam et al., 2020).

Akash et al. (2020) established a POMDP model designed to capture the dynamic shifts in human trust and workload when collaborating with an intelligent decision-aid for a teleoperated robot. This model simplified trust and workload into binary values (low/high) and assumed independent propagation to streamline the training process (Akash, Polson, et al., 2019; Akash, Reid, and Jain, 2019; Akash et al., 2020).

Zahedi et al. (2023) proposed a Meta-MDP trust model specifically for human supervision tasks. This model introduced the innovative concept of trust transition, based on the disparity between the robot's policy and the human's anticipated policy.

The core advantage of MDP/POMDP models is their capacity to integrate the trust state directly with robot behaviors, which is highly beneficial for optimizing overall system performance. However, they present challenges, as it can be difficult to explicitly reveal the precise relationship between trust and other variables such as workload and intervention, primarily because these factors are often modeled collectively as a single state. Moreover, these models frequently rely on group-based training rather than individualized models, which can limit personalization. They commonly consider discrete action and state spaces, which only provide an approximation of the continuously evolving nature of human trust. Scalability also emerges as a significant hurdle when dealing with larger state and action spaces.

Gaussian-Based Trust Models

These models offer a methodology for representing trust as a continuous random variable or stochastic process, a representation that more accurately reflects the probabilistic and continuous nature of human trust in robots.

A pertinent example comes from Soh et al. (2020), who utilized Gaussian Processes (GP) for trust prediction in both automated driving scenarios and household tasks involving a Fetch robot. This approach incorporates human actions as observations, which are then used to update trust values through approximate Bayesian inference.

Moreover, these models are not limited to Gaussian distributions, as other works use different probability distributions to model trust. For example, Azevedo Sá et al. (2021) proposed a bidirectional trust model, while Guo and Yang (2021) used a Beta distribution-based human-robot trust model. They interpreted the mean of the Beta distribution as the predicted trust, with both parameters of the Beta distribution being updated based on prior trust information. This model captured the non-linear behavior of trust dynamics.

The benefits of Gaussian-based (or in general, probability distribution-based) models include their ability to model trust as a continuous dynamic process. They are also sufficiently flexible to be integrated into other decision-making and learning frameworks, such as POMDP and Bayesian inference. Additionally, these models can leverage human subject data for parameter learning, enabling the creation of either personalized or population-based trust models. A notable limitation, however, is the inherent uncertainty regarding whether human trust genuinely adheres to a Gaussian, Beta, or another specific statistical distribution, which can consequently impact the model's reliability and prediction accuracy. Furthermore, existing Gaussian-based models primarily rely on robot performance as the sole factor influencing trust, often overlooking other crucial elements such as task difficulty or risk (Y. Wang et al., 2024).

Dynamic Bayesian Network (DBN)-Based Trust Models

DBN models are sophisticated graphical models that illustrate causal relationships among various factors influencing trust, hidden trust states, and observed trust feedback over time. They are particularly adept at capturing the dynamic property of trust and its temporal evolution through defined state transition distributions.

Xu and Dudek (2015) introduced OPTIMo, a personalized performance-centric trust model that infers a human operator's degree of trust in an autonomous robot as a probability distribution at each time step. OPTIMo

combines causal reasoning about robot trustworthiness based on task performance with evidence from direct experiences. It uniquely aims to estimate operator trust in near real-time and has shown superior predictive accuracy compared to existing methods. Its formulation adheres to the three properties of trust dynamics mentioned earlier, thus guaranteeing high model explicability and generalizability. Notably, OPTIMO's strength lies in its ability to infer trust from observable robot performance history without continuous querying of the human, addressing a significant operational challenge in time-critical settings.

Guo and Yang (2021) proposed a personalized trust prediction model using Bayesian inference, aiming to model each individual human agent's trust dynamics when interacting with a robotic agent over time. Their underlying premise is that a rational human agent will update their current trust based on the robot's performance history. This model adheres to three major properties of trust dynamics previously mentioned in this section. They use a Beta distribution to model a human agent's temporal trust, which is consistent with the bounded self-reported trust, fits the exploration-exploitation scheme, and adheres to the aforementioned properties. The model does not depend on repeatedly querying the human or collecting physiological information. The model was tested using an existing dataset involving human participants interacting with robotic agents in a simulated surveillance mission. This demonstrated success at accurately predicting trust-induced behaviors, outperforming existing prediction methods. Moreover, the study identified three distinct types of trust dynamics: the Bayesian decision maker, the oscillator, and the disbeliever. The authors note that this prediction model can be used for the design of individualized and adaptive technologies (Guo and Yang, 2021).

Similarly, Mahani et al. (2021) developed a Bayesian trust inference model for human-multi-robot teams using a DBN structure. Their model captures the causal relationship between trust impacting factors, latent trust states, and human-robot interaction evidence. Taking robot performance measures as input and human subjective feedback and intervention data as observations, the model infers hidden trust states. Tested in a simulated human-UAV collaborative search mission, the model could infer degrees of human trust in multiple robots and predict human interventions with relatively high accuracy. They highlight that their DBN trust model is computational, quantitative, dynamic, probabilistic, and personalized, applicable to multi-robot systems and capable of predicting trust for individual robots.

These dynamic trust models often rely on a trust-behavior model, which predicts how a human's trust level influences their actions towards the robot (e.g., compliance, intervention). Previous research has often assumed a "disuse" trust-behavior model (Chen et al., 2020), where a human stops relying on a robot when trust is low and follows their own behavioral policies. However, psychological research suggests that people may display "reverse psychology" behaviors (Guo et al., 2021), doing the exact opposite of what the robot suggests. Examining different trust-behavior models is an area of ongoing research (Guo et al., 2021).

The advantages of DBN models are their effectiveness in capturing the dynamic nature of trust and characterizing its evolution over time. Their graphical representation enhances interpretability, clearly depicting causal relationships. They can also effectively describe non-symmetry in trust within multi-robot systems. However, the accuracy of DBN models is heavily contingent on the accurate selection of initial probability distributions (initial bias), which can be challenging to determine precisely. While theoretically capable of accommodating continuous variables, most existing DBN-based models simplify by assuming discrete levels of trust.

Although probabilistic trust models provide valuable descriptive insights and are well suited for uncertainty representation, they are often formulated in a way that can limit direct parameter interpretability and modular decomposition, particularly when intended for integration into control-oriented or simulation-based frameworks. In such cases, the trust formulation may be tightly coupled to a specific inference or decision architecture, making reuse across tasks and system designs more difficult. The emphasis in this thesis is therefore not on avoiding probabilistic reasoning, but on defining a structured, executable trust dynamics model whose components admit clear interpretation and can be identified and validated from data.

In a broader context, it is acknowledged that while many factors affecting human trust in robots have been identified, current computational models often incorporate only a small subset of these, such as robot performance, false alarm rate, human performance, fault, and task difficulty. This sometimes leads to these models being considered "trust-like" rather than fully comprehensive representations of trust (Y. Wang et al., 2024). The simplification of trust into discrete or binary variables is also a common limitation, largely due to the computational demands of handling larger continuous trust state spaces and their corresponding parameter learning algorithms. This underscores the ongoing need for more generalized models that can integrate a wider array of critical trust-impacting factors, coupled with efficient parameter estimation and learning algorithms capable of managing large state spaces.

2.5.4. Trust Measurement Challenges

Despite the proliferation of trust models, standardized and objective measures for trust remain elusive. Various types of measures have been employed to examine trust in human-automation interaction, including subjective ratings, behavioral data, and psychophysiological measurements (Lu and Sarter, 2019; Robinette et al., 2016; N. Wang et al., 2016; Salem et al., 2015).

Subjective ratings, such as questionnaires administered at the end of an experiment (Desai et al., 2013), are easy to implement but often fail to capture how trust evolves over extended periods of interaction (Guo and Yang, 2021). Repeatedly asking for subjective ratings can capture some temporal evolution but can be disruptive (Lu and Sarter, 2019). Self-report questionnaires are also prone to biases and may not capture real-time trust fluctuations accurately, especially in high-workload or time-critical settings.

Behavioral indicators, such as compliance (Salem et al., 2015; N. Wang et al., 2016), control allocation (e.g., take-over or hand-over behavior) (Xu and Dudek, 2015), or interventions (Gao et al., 2013), infer trust based on observable actions. While trust is considered an attitude, this attitude forms the basis for intentions and mediates behavior (J. D. Lee and See, 2004). However, behavioral indicators tend to be discrete and domain-specific, making comparisons across studies difficult (Lu and Sarter, 2019). Moreover, while objective measures offer promise for real-time assessment, they are often used individually rather than in combination, leading to a less robust representation of experienced trust.

Psychophysiological measures have also been used to infer trust in real-time, initially in interpersonal trust research (Hancock et al., 2011) and more recently for trust in modern technologies (Lu and Sarter, 2019). Operationalizing trust for measurement can involve focusing on perceived "trustworthiness" based on factors like ability, benevolence, and integrity, or through behavioral indicators like compliance (N. Wang et al., 2016).

The absence of clear empirical grounding and standardized metrics for trust in automation has even led to criticisms of its construct validity. Therefore, a significant challenge remains in developing objective, reliable, and valid measures that can capture the continuous and complex dynamics of human trust in robots.

2.5.5. Measuring Trust in Human–Robot Interaction

To overcome the methodological limitations of prior work and produce a robust instrument for quantifying dynamic human–robot trust, Schaefer (2013) developed a comprehensive measurement framework. Their goal was to design a reliable and valid scale capable of capturing temporal variations in an individual's trust toward a robot, thereby enabling both empirical assessment and computational modeling of trust dynamics.

The development of the scale was conceptually guided by three central components: (1) a theoretical grounding in the application of trust within human–robot interaction (HRI), (2) a structured classification of the perceptual dimensions of robots, and (3) the systematic psychometric construction of the measurement instrument itself. Building upon the Triadic Model of Trust, Schaefer (2013) ensured that the scale incorporated items representing measurable factors associated with the Human (e.g., personality traits, cognitive state, prior experience), the Robot (e.g., reliability, transparency, communication), and the Environment/Task (e.g., risk level, team composition, situational stress). This triadic framing reflects the interdependent nature of trust as a relational and context-sensitive construct, which aligns with the multidimensional structure assumed in the present work's trust dynamics model.

The initial item pool consisted of 172 statements encompassing a broad range of antecedents and manifestations of trust. Through iterative refinement, including internal review, paired-sample *t*-tests, exploratory and confirmatory factor analyses, and expert content validation by Subject Matter Experts (SMEs), the pool was distilled to a psychometrically sound 40-item scale. Each item is rated on a continuous scale from 0% to 100%, allowing the computation of an overall percentage trust score with high granularity. The resulting instrument was designed specifically for HRI contexts, capturing domain-specific attributes such as perceived reliability, robot affect, and environmental constraints that are absent from traditional automation trust measures.

Validation studies confirmed both the reliability and the sensitivity of the 40-item scale to temporal changes in trust (Schaefer, 2013). In particular, Schaefer's fifth validation study evaluated trust evolution across three time points in a simulated task involving manipulation of robot reliability (initially 100% reliable, followed by 25%). The results demonstrated significant mean differences across time ($\mu_{T1} < \mu_{T2} > \mu_{T3}$), evidencing that the scale effectively captured dynamic trust variations induced by performance history. This dynamic responsiveness substantiated the scale's suitability for modeling trust as a continuous, state-dependent process rather than a static attitude.

Recognizing the practical constraints of real-time data collection in interactive or time-critical environments, Schaefer and SMEs identified a condensed 14-item subscale that maintains a strong positive correlation with the full instrument. Although the 40-item version provides a finer level of measurement granularity, the reduced

subscale has been validated as an efficient proxy for capturing rapid fluctuations in trust during continuous human–robot interaction. This makes it particularly well-suited for integration into short experimental blocks or real-time simulations, such as those proposed in the current study.

In the context of this thesis, the trust scales proposed by Schaefer (2013) serve as the essential psychometric foundation for the empirical validation of the proposed rule-based mathematical model of trust dynamics. Specifically, the full 40-item instrument will be employed in pre and post-interaction phases to capture dispositional and aggregate changes in perceived trust, while the 14-item subscale will be administered at intermediate points to monitor within-interaction dynamics. When combined with the behavioral measures collected during the experimental trials, such as reliance decisions and response times, these subjective trust scores will provide the quantitative grounding required for parameter identification and validation of the trust evolution equations introduced in chapter 5. In this way, Schaefer’s validated methodology bridges the gap between theoretical modeling and empirical observation, enabling the development of a computationally interpretable and psychologically grounded model of human–robot trust.

2.6. Application Domain: Search and Rescue

Search and Rescue (SaR) missions are critical operations in response to catastrophic incidents like urban fires. These missions require prompt response due to life endangerment and the risk of propagation. SaR in general is challenging because the number of trapped victims may be unknown and their behavior during evacuation is prone to variations influenced by individual cognition and social interactions (Kinatader, 2012; Kinatader et al., 2014; Kinatader and Warren, 2016; Kinatader et al., 2018; Kinatader and Warren, 2021; Y. Zhang et al., 2025). A common decisive factor for success is the speed at which rescue workers can conduct missions.

Robotics has emerged as a valuable resource for urban SaR scenarios. Robots can offer assistance by taking advantage of their expendability in dangerous environments (Murphy, 2004). Examples include search and rescue operations, coverage of areas, and providing guidance to evacuees. Research has explored the potential advantages of human-robot cooperation in urban fire SaR, such as overcoming communication issues and improving situational awareness. Modeling the behavior of human rescuers is considered vital before integrating robotic agents into the team.

In this thesis, SaR scenarios are therefore not treated as a validation domain for specific mission outcomes, but rather as a motivating deployment context that exposes the limitations of existing trust models and highlights the need for a formally defined and empirically validated trust dynamics formulation.

2.6.1. HRI Challenges in High-Risk Environments

SaR missions, particularly emergency evacuations, present unique challenges for HRI due to their high-risk nature (Couceiro et al., 2021), that inherently involve high levels of uncertainty, vulnerability, and significant consequences. In such situations, evacuees may place their lives at risk depending on the actions of robots providing guidance (Robinette et al., 2016). High-risk situations may engage fight-or-flight responses and other cognitive faculties that impact a person’s trust in difficult to predict ways (Robinette et al., 2016). Research has begun to investigate human-robot trust specifically in comparatively high-risk situations like simulated emergency evacuations (N. Wang et al., 2016).

A key question in this context is whether people will trust and follow the directions of a robot in an emergency, and under what conditions they might stop following it (Freedy et al., 2007; N. Wang et al., 2016; J. D. Lee and See, 2004; Yang et al., 2023). Unlike lower-risk scenarios or tasks where humans can take control from an autonomous system (Desai et al., 2013), emergency evacuation often requires humans to decide whether to follow the robot’s guidance or not, representing a different dimension of reliance and trust (Robinette et al., 2016). A potential concern is overtrust of robots in emergency evacuation scenarios (J. D. Lee and See, 2004; Robinette et al., 2016; Lewis et al., 2018).

Robinette et al. (2016) further delve into the complexities of human-robot trust in high-risk emergency contexts. They explore the phenomenon of overtrust, revealing that people may continue to follow a faulty robot even when it makes mistakes in emergency evacuation scenarios. This overtrust can be dangerous, as participants in their real-world experiments showed compliance with a robot guide even after its reliability was ranked lower due to errors. The authors hypothesize that factors like adrenaline-based cognitive processing in emergencies might lead people to focus on obvious evacuation routes, potentially ignoring other options or the robot’s flaws. This work underscores the critical need for robots to recognize when people trust them too much and act to dissuade individuals from placing themselves at unwarranted risk (Robinette et al., 2016).

In a subsequent study, Robinette et al. (2017) complement this by confirming that robot performance is a critical factor influencing trust, with a significant drop in self-reported trust (53%) observed when the robot performed poorly. The study found that even a single failure strongly impacts a person's trust, and the manner of failure (e.g., inefficient vs. complete failure) did not significantly alter this reduction. Importantly, in low-risk situations (monetary incentives), people might act as if they trust the robot after a failure, even if they self-report little trust, suggesting that other motivations (like fun or time) can override trust in such contexts. However, in higher-risk scenarios (simulated emergency), participants' self-reports aligned with their decision to use the robot. This research illuminates the methodological challenges of investigating trust in high-risk scenarios, as ethical guidelines prevent truly putting participants in danger, which can impact the realism of perceived risk and thus the validity of trust assessments. These findings underscore the need for models that can predict how and why a robot's behavior impacts trust, influencing how emergency guidance robots should be designed and communicate failure states.

Understanding human behavior in dangerous situations, including evacuation behavior and social influence, is crucial for designing effective robotic assistance (Kinatder and Warren, 2016; Kinatder and Warren, 2021). Concepts such as risk perception (the personalization of risk related to an event like a fire emergency) are relevant, influenced by emotions and cognitive biases (Kinsey et al., 2019). Risk perception is differentiated from other concepts like situation awareness (Kinatder et al., 2014) and is studied within various theoretical frameworks relevant to fire evacuation (Kinatder, 2012; Kinatder et al., 2014).

These characteristics make SaR-inspired tasks particularly suitable for studying trust calibration, because they naturally couple perceived risk, outcome uncertainty, and reliance decisions, i.e., the key ingredients required to excite and observe trust dynamics over time.

2.7. Integrating Trust Dynamics and Agent-Based SaR Simulation

Simulating human behavior during evacuation, especially in multi-actor settings, has often utilized separate frameworks for individual and group levels, covering either macroscopic or microscopic perspectives (Cristiani et al., 2014). Integrating approaches is seen as beneficial (Wilmes and Jamshidnejad, 2025). Agent-based modeling environments, like NetLogo, are suitable for simulating complex systems with many interacting agents and have been used for various applications, including traffic simulation and social systems. These platforms provide a means to model individual agent behaviors and their collective emergent outcomes.

One notable work that bridges the concepts of agent-based simulation, human behavior modeling, HRI, and trust in a SaR context is the GT-BDI model proposed by Wilmes and Jamshidnejad (2025). Their research focuses on simulating the behavior of trapped victims during an evacuation process to gain insights for SaR missions and to analyze the influence of robots. The GT-BDI model is novel in that it integrates game theory and the Belief-Desire-Intention (BDI) framework to incorporate both the interactions of trapped victims (game theory, group level) and their cognitive processes (BDI, individual level). This combined approach aims to provide a more precise prediction of human actions during an evacuation by considering different intentions and cooperation levels.

Crucially for this thesis, the GT-BDI model incorporates human-robot interactions by proposing a dynamic equation for modeling the evolution of the trust of victims in robots. This equation considers factors like the robot's failure class (e.g., class 1 for minor faults, class 3 for major faults) and trust recovery rate after interacting with a reliable robot. The paper analyzes the impact of this trust on human-robot interactions and, consequently, on the outcomes of the evacuation process. The model was validated using extensive simulations in NetLogo against benchmark models. Using the validated model, the paper assessed the effectiveness of evacuation robots, finding that their presence reduces evacuation time. The simulated scenarios varied factors like the number and types of faulty robots, providing insights into performance and dynamic trust evolution.

The GT-BDI model's ability to assess the influence of robots with varying conditions (e.g., perfectly functioning, partially faulty) within evacuation processes, and its validation against benchmark models, demonstrates its utility as a reliable behavioral model for planning and analyzing SaR missions. This approach of developing and validating complex behavioral models in simulation, like the GT-BDI model in NetLogo, provides a robust method for exploring trust dynamics without the immediate need for extensive new empirical data from real-world high-risk situations. Such models are vital for making distinctions in human behavior, understanding trust evolution, and designing robots that can appropriately foster trust.

While trust is explicitly acknowledged in this work as a relevant factor for human-robot interaction in SaR scenarios, the trust component itself remains largely underspecified and is not formulated as an identifiable or

executable dynamic model. Notably, the authors explicitly state that alternative modeling approaches, such as non-linear difference equations, should be considered, thereby directly highlighting the absence of a suitable trust dynamics formulation within the proposed framework (Wilmes and Jamshidnejad, 2025).

Another relevant work by Couceiro et al. (2021), focuses specifically on urban fire SaR. This paper highlights the critical role of robot assistance in urban SaR scenarios, such as urban fires in basement garages, by leveraging their expendability. Couceiro et al. (2021) present a study that models key features of a firefighting mission by developing a behavioral model of human firefighting teams, based on interviews with a Portuguese Fire Department. Their work then studies the addition of robotic teams in cooperation with human firefighters to overcome communication issues and improve situation awareness. The results highlight the importance of human-robot partnership for effective response and mitigating life endangerment. While this work models team behavior and evaluates the impact of robots, the specific dynamics of human trust in the robots are not directly addressed, although trust is implicitly related to cooperation and reliance.

A key aspect of this research is its methodology, which involves implementing the models in a distributed agent-based simulation environment, Multi-Robot Simulator (MRSim), drawing evidence from existing literature. This simulation-based approach allows for intensive simulations under different configurations to draw conclusions based on quantitative data. This is particularly important because, as the authors note, while numerous robots are proposed for real-world rescue missions, only a few proposed robots have actually participated in a rescue or been allowed to operate on site after a disaster for testing purposes; all of these have been teleoperated (Couceiro et al., 2021). The reliance on simulation addresses the practical limitations and safety concerns of deploying autonomous robots in real-world, high-risk scenarios, enabling comprehensive testing and analysis of human-robot cooperation without direct exposure to danger or the need for extensive new empirical data from human subjects in hazardous environments. The results underscore the importance of such human-robot partnerships for a more effective response and mitigation of life endangerment for both first responders and victims. The ability to model rescuers' behavior prior to extending the team with robotic agents is vital for understanding and optimizing mixed teams.

Despite the high level of technical detail and mission realism achieved by the proposed SaR simulation environment, the framework does not incorporate any explicit representation of human trust or trust dynamics in human-robot interaction. As a consequence, while the simulation provides a rich testbed for autonomous coordination and mission planning, it is unable to capture trust-mediated human decision-making, underscoring the absence of a validated trust dynamics model suitable for integration into such environments.

Taken together, these works illustrate a clear gap in the current literature: existing SaR simulation frameworks either acknowledge the importance of trust without providing an operational model, or achieve high-fidelity mission simulation while omitting trust entirely, thereby motivating the need for a generalizable and identifiable trust dynamics model.

Problem Formulation and Contributions

This chapter establishes the conceptual and methodological foundation of the thesis. It first identifies the key gaps in existing research on trust modeling in human-robot interaction, motivating the need for a unified, empirically grounded approach to dynamic trust modeling. Building on this analysis, the chapter articulates the overall research objective, clarifies the scope and intended contributions of the work, and situates the thesis within a broader modeling and identification pipeline. The chapter concludes by formalizing the research questions that structure the remainder of the thesis and guide the progression from model definition to empirical validation and exploratory system-level evaluation.

3.1. Identification of Research Gap

Human trust has long been recognized as a critical factor in effective human-robot interaction (HRI), particularly in scenarios characterized by uncertainty, time pressure, and asymmetric information (J. D. Lee and See, 2004; Hancock et al., 2011; Schaefer et al., 2016). Extensive research has investigated the psychological foundations of trust, identified factors influencing trust formation and erosion, and proposed computational models to capture trust dynamics over time (J. D. Lee and See, 2004; Hoff and Bashir, 2015; Y. Wang et al., 2024). In parallel, robots are increasingly envisioned as collaborative partners in safety-critical and decision-support roles, motivating the need for principled models that can represent and reason about human trust (Schaefer et al., 2016; Sapienza et al., 2022; Robinette et al., 2016).

Despite this progress, the existing body of work remains fragmented in terms of modeling assumptions, empirical grounding, and validation practices (Hoff and Bashir, 2015; Schaefer et al., 2016; Y. Wang et al., 2024). While numerous trust models have been proposed, there is limited consensus on how such models should be structured, how they can be empirically identified from realistic data, and how their relevance can be evaluated beyond fitting trust measurements (Y. Wang et al., 2024; Guo and Yang, 2021; Xu and Dudek, 2015). As a result, it remains difficult to compare models, assess their generalizability, or reuse them across application domains (Y. Wang et al., 2024).

A first key gap lies in the lack of identification-ready trust models (Y. Wang et al., 2024). Many existing computational trust models are either strongly grounded in psychological theory but difficult to parameterize and validate from empirical data, or computationally simple but limited in interpretability and structural expressiveness. Few models are explicitly designed with identifiability, boundedness, and execution in mind, making it challenging to assess whether their parameters can be meaningfully inferred from realistic human-robot interaction data (Hoff and Bashir, 2015).

A second gap concerns the absence of purpose-designed empirical pipelines for trust model identification (Schaefer, 2013; Guo and Yang, 2021). Trust data is inherently noisy, heterogeneous, and costly to collect, and many existing datasets are designed for hypothesis testing rather than dynamic model estimation (Lu and Sarter, 2019; Schaefer, 2013; N. Wang et al., 2016). Experimental designs, measurement instruments, and validation procedures are often insufficiently aligned with the requirements of identifying latent trust dynamics, limiting the extent to which proposed models can be rigorously evaluated or compared (Guo and Yang, 2021; Hoff and Bashir, 2015).

A third gap arises from the limited end-to-end validation of trust models beyond trust trajectories themselves (J. D. Lee and See, 2004; Hoff and Bashir, 2015; Y. Wang et al., 2024). While many studies focus

on predicting or explaining trust measurements, it is frequently assumed rather than tested that inferred trust states meaningfully influence human behavior (J. D. Lee and See, 2004; Chen et al., 2020). The relationship between latent trust and observable actions such as reliance or intervention remains underexplored, and attempts to close the trust–behavior feedback loop are rare (Hoff and Bashir, 2015; M. Y. Zhang and Yang, 2017). Without such stress-testing, it is unclear where trust models succeed, where they break down, and what their limitations imply for downstream human-robot interaction modeling (Y. Wang et al., 2024).

These gaps point not only to the need for improved trust models, but more fundamentally to the absence of a unified, end-to-end framework for defining, identifying, validating, and stress-testing dynamic trust models under realistic empirical constraints. Addressing this structural bottleneck is essential for advancing trust modeling in HRI in a cumulative and reusable manner (Hoff and Bashir, 2015; Y. Wang et al., 2024).

3.2. Research Objective

The primary objective of this thesis is to develop and empirically evaluate an end-to-end modeling and identification pipeline for dynamic human trust in human-robot interaction. Rather than treating trust modeling, data collection, and validation as isolated tasks, this work aims to explicitly link these stages into a coherent methodological framework that supports principled model construction and empirical assessment.

To achieve this objective, the thesis pursues four sequential goals. First, a modular, dynamic mathematical model of human trust is formally defined, grounded in the three-layered trust framework that distinguishes between dispositional, situational, and learned trust components (Hoff and Bashir, 2015). The model is designed to be interpretable, bounded, and executable, while maintaining sufficient structure to support parameter identification.

Second, a controlled human-subject experiment is designed with the explicit purpose of eliciting trust dynamics that are informative for model identification. Particular attention is given to experimental task design, trust measurement strategies, and data validation criteria, enabling the collection of empirical data suitable for estimating latent trust dynamics in the presence of noise and inter-individual variability.

Third, a principled identification and validation pipeline is developed and applied to assess whether the proposed trust model can be empirically identified from the collected data. This includes evaluating generalization to unseen participants, assessing parameter identifiability and robustness, and analyzing the extent to which observed trust trajectories can be reproduced, including comparison with representative baseline trust models from the literature (Xu and Dudek, 2015; Wilmes and Jamshidnejad, 2025) to contextualize empirical performance.

Finally, as an exploratory and diagnostic step, the identified trust model is coupled with a behavioral decision model to form a closed-loop human-robot interaction pipeline. This coupling is not intended to produce an optimal or definitive predictor of human behavior, but rather to stress-test the behavioral relevance of inferred trust states and to investigate the limitations that emerge when attempting to map latent trust to observable actions.

The experimental task and interaction scenario are inspired by high-stakes human-robot collaboration contexts, such as search and rescue, which provide a motivating setting characterized by uncertainty and risk. However, the thesis does not claim deployment readiness or mission-level performance improvement. Instead, the focus is on developing and evaluating a reusable modeling and identification framework that can be adapted to other HRI domains, trust models, and experimental contexts.

3.3. Contributions

This thesis makes the following contributions to the study of trust in human-robot interaction:

- **Dynamic trust modeling:** a modular mathematical model of human trust dynamics grounded in the three-layered trust framework, explicitly separating trust components, inputs, and parameters, and designed for interpretability, boundedness, and identifiability.
- **Experimental and measurement framework:** a purpose-designed human-subject experiment and trust measurement strategy aimed at identifying dynamic trust models, including principled data validation and inclusion criteria aligned with model estimation requirements.
- **Identification and validation pipeline:** a complete parameter identification and validation pipeline for dynamic trust models, incorporating training–validation separation, robustness analysis, empirical assessment of parameter identifiability and generalization, and comparison with representative baseline trust models from the literature for contextual validation.

- **Exploratory trust–behavior coupling:** an exploratory closed-loop coupling of inferred trust dynamics with a behavioral decision model, used to evaluate the behavioral relevance and limitations of trust-based modeling assumptions.
- **End-to-end methodological framework:** an end-to-end, reusable framework for defining, identifying, validating, and stress-testing trust dynamics in HRI, demonstrated in a search-and-rescue-inspired context but applicable to a broader range of domains.

3.4. Research Questions

The objective of this thesis is to develop and empirically evaluate an end-to-end modeling and identification pipeline for dynamic human trust in human-robot interaction (HRI). Within this pipeline, trust is modeled as a latent, time-varying internal state that evolves through interaction and can be estimated from empirical human data. To achieve this objective, the thesis is structured around four main research questions, each addressing a distinct stage of the modeling, identification, validation, and exploratory evaluation process.

RQ1 How can a dynamic mathematical model of human trust in HRI be formally defined?

- RQ1.a** How can the three-layered trust framework (dispositional, situational, and learned trust) be translated into a mathematically explicit, executable model suitable for simulation and analysis?
- RQ1.b** How can the temporal evolution of trust be modeled as a dynamic process that integrates interaction events, robot performance outcomes, and contextual factors while remaining bounded and interpretable?
- RQ1.c** What states, inputs, and parameters are required to construct a modular trust dynamics model that balances expressive power with identifiability?
- RQ1.d** How can the proposed continuous-time trust dynamics model be discretized in time and made fully explicit, deterministic, and reproducible for use in simulation-based evaluation and parameter identification?

RQ2 How can the parameters of the proposed trust dynamics model be identified from empirical human-robot interaction data?

- RQ2.a** How can a controlled human-subject experiment be designed specifically to elicit trust dynamics that are informative for parameter identification, while remaining consistent with established findings in the human-robot-trust literature?
- RQ2.b** How can noisy, heterogeneous trust measurements (e.g., self-reports and behavioral indicators) be incorporated into a principled estimation framework for identifying latent trust dynamics?

RQ3 To what extent can the proposed trust dynamics model be empirically identified and validated?

- RQ3.a** How can an identification strategy be formulated that enables the estimation of shared trust model parameters while accounting for inter-individual variability and supporting generalization beyond the training data?
- RQ3.b** Can the model reproduce observed trust trajectories across participants when evaluated on held-out validation data?
- RQ3.c** Which model parameters are consistently identifiable from the available data, and which exhibit weak identifiability or high uncertainty?
- RQ3.d** How robust are the identified parameters and model predictions to variations in experimental conditions, measurement subsets, and estimation assumptions?

RQ4 To what extent can the identified trust model be coupled with a behavioral decision model to form a closed-loop HRI pipeline, and what limitations emerge from this coupling?

- RQ4.a** How can the inferred latent trust state be mapped to probabilistic models of human reliance and decision-making behavior?
- RQ4.b** Can the coupled trust–behavior model reproduce aggregate behavioral patterns observed in the experimental data when simulated in closed loop?
- RQ4.c** What limitations emerge when attempting to close the trust–behavior feedback loop, and what do these limitations reveal about the relationship between trust and observable human actions?

The first research question focuses on the formal definition of a dynamic trust model for human-robot interaction. Its purpose is to establish a mathematically explicit representation of trust as a latent, time-varying internal state, grounded in the three-layered trust framework from literature. This question addresses how different trust components can be structured, how they evolve over time in response to interaction events, and what assumptions are required to render the model executable and interpretable. The outcome of this research question is a clearly defined trust dynamics model, which serves as the theoretical foundation for all subsequent stages of the thesis.

The second research question addresses the methodological challenge of identifying the parameters of the proposed trust dynamics model, given the lack of publicly available empirical data. Since trust is not directly observable, this question focuses on how human-robot interaction experiments can be designed to elicit informative trust dynamics and how heterogeneous trust measurements can be incorporated into a principled estimation framework. The goal of this research question is not to assess model performance, but to define a systematic identification and validation pipeline that enables parameter estimation while accounting for measurement noise and inter-individual variability.

The third research question evaluates the extent to which the proposed trust dynamics model can be empirically identified and validated. Building on the identification framework established in the previous question, this research question examines whether the model can reproduce observed trust trajectories on held-out data, which parameters are reliably identifiable, and how robust the results are to changes in estimation assumptions and data subsets. The purpose of this question is to assess the empirical support for the proposed model and to delineate the limits of what can be inferred from the available data.

The fourth research question explores the coupling of the identified trust model with a behavioral decision model in a closed-loop human-robot interaction pipeline. Rather than aiming to optimize or predict task performance, this question is explicitly exploratory in nature and investigates how inferred trust states can be mapped to probabilistic models of human reliance and decision-making. The objective is to examine whether the coupled model can reproduce aggregate behavioral patterns observed in the experimental data and to identify limitations that arise when attempting to close the trust–behavior feedback loop. This analysis provides insight into the relationship between latent trust and observable behavior, while clarifying the scope and limitations of trust-aware behavioral modeling.

By systematically addressing these questions, this thesis aims to contribute to both the theoretical understanding of human-robot trust dynamics and the practical development of more intelligent, trustworthy, and effective robotic systems for critical applications like emergency search and rescue. The focus on a computationally tractable model implemented within simulation environments will provide a scalable and ethical approach to exploring complex trust phenomena in high-risk contexts.

4

Theoretical Background for Model Identification and Validation

This chapter introduces the theoretical concepts that underpin the model identification, validation, and behavioral coupling procedures used throughout this thesis. Instead of providing a broad survey of related theory, the focus is deliberately selective and pragmatic. In particular, this chapter clarifies how trust modeling is treated as a practical inverse identification problem, how least-squares estimation is interpreted under model-data mismatch, how model validation and robustness are assessed in the presence of human variability, and how latent trust states relate to observable behavior under closed-loop evaluation. These foundations establish the interpretive framework needed to assess both successful and limited results without conflating methodological adequacy with behavioral sufficiency.

4.1. Trust Modeling as a Practical Inverse Identification Problem

The estimation of a trust dynamics model from experimental data constitutes an inverse problem (Aster et al., 2013). In the forward direction, a dynamical model specifies how an internal trust state evolves in response to interaction events and contextual factors. In the inverse direction, this thesis seeks to infer the parameters and structure of such a model from observed trust reports collected during human-robot interaction. This distinction is important, as inverse problems are inherently ill-posed: multiple models or parameter configurations may reproduce the observed data to a similar degree, particularly in the presence of noise, limited measurements, and inter-individual variability.

Within this framing, a central concept is identifiability. Structural identifiability concerns whether model parameters are uniquely recoverable in principle, assuming ideal data and noise-free observations. Practical identifiability, by contrast, reflects whether parameters can be reliably constrained given finite, noisy data and realistic experimental conditions. In human-centered studies, practical identifiability is often the dominant limitation, as internal cognitive states such as trust are only indirectly observable and are measured through sparse, subjective reports. Consequently, lack of unique parameter recovery does not necessarily indicate a flawed model, but may instead reflect fundamental limits imposed by the data.

For this reason, the objective of model identification in this thesis is not strict parameter recovery, but predictive and structural adequacy. A model is considered informative if it reproduces salient features of observed trust trajectories across participants, generalizes beyond the training data, and yields interpretable insight into how trust responds to interaction outcomes and contextual influences. Parameter estimates are therefore interpreted as elements of an equivalence class of solutions that induce similar trust dynamics, rather than as uniquely determined psychological constants.

This perspective motivates the emphasis on population-level modeling and robustness analysis in subsequent chapters. Rather than tailoring highly individualized models that risk overfitting, the identification strategy prioritizes shared dynamics that capture consistent patterns across participants, while explicitly evaluating generalization to unseen individuals. Within this inverse-problem framing, variability across parameter estimates and sensitivity to estimation assumptions are expected and informative, provided that predictive behavior and qualitative trust dynamics remain stable.

4.2. Least-Squares Identification under Model-Data Mismatch

The trust model identification problem in this thesis is formulated as a weighted least-squares optimization problem (Strutz, 2010). Let $\hat{\tau}(t; \theta)$ denote the trust trajectory predicted by a parametric trust dynamics model with parameter vector θ , and let $\tau^{\text{obs}}(t_i)$ denote the corresponding reported trust measurements at discrete time points t_i , $i = 1, \dots, N$. Parameter estimation is performed by minimizing the discrepancy between predicted and observed trust values over the available measurements. In its basic form, this corresponds to the ordinary least-squares (OLS) objective

$$J(\theta) = \sum_{i=1}^N (\tau^{\text{obs}}(t_i) - \hat{\tau}(t_i; \theta))^2. \quad (4.1)$$

This formulation provides a simple and interpretable criterion for aligning model predictions with observed trust trajectories.

In practice, the estimation problem is complicated by the presence of heterogeneous uncertainty and structural mismatch between the model and the data. Trust reports are subjective, sparsely sampled, and influenced by factors not explicitly modeled, while the trust dynamics model is an abstraction of a richer cognitive process. To account for these considerations, a weighted least-squares (WLS) objective is employed:

$$J_{\text{WLS}}(\theta) = \sum_{i=1}^N w_i (\tau^{\text{obs}}(t_i) - \hat{\tau}(t_i; \theta))^2, \quad (4.2)$$

where $w_i \geq 0$ are weights that modulate the relative influence of different measurements or temporal regions.

Crucially, least-squares estimation in this context is not treated as a statistically optimal estimator under an assumed noise distribution. The dominant source of error is not assumed to be measurement noise alone, but rather model–data mismatch, arising from unmodeled cognitive factors, inter-individual variability, and simplifications inherent in the trust dynamics formulation. As a result, the residuals $\tau^{\text{obs}}(t_i) - \hat{\tau}(t_i; \theta)$ are not expected to satisfy the assumptions required for maximum-likelihood or minimum-variance interpretations. Least-squares is therefore used as a consistency-seeking criterion that penalizes systematic deviations without asserting probabilistic optimality.

This interpretation aligns with the inverse-problem perspective introduced in Section 4.1. Because multiple parameter vectors θ may yield similar values of $J_{\text{WLS}}(\theta)$ under model–data mismatch, estimator optimality is less informative than solution stability. Subsequent analyses therefore emphasize robustness to solver choice, initialization, and data subsets, as well as consistency of predictive behavior, rather than reliance on a single nominal optimum. Within this framework, least-squares estimation serves as a transparent and defensible tool for identifying structurally adequate trust dynamics models under realistic experimental constraints.

4.3. Model Validation, Generalization, and Robustness

Model validation in the context of human-robot trust differs fundamentally from validation in purely physical systems (Barlas, 1996). Human trust dynamics exhibit substantial inter-individual variability, are influenced by latent cognitive factors, and are only indirectly observed through subjective reports. As a result, validation cannot reasonably be defined as exact reproduction of every individual trajectory or as minimization of prediction error for each participant. Instead, validation must assess whether a model captures consistent structure in the data and generalizes beyond the specific individuals and interaction instances used for identification.

In this thesis, validation is therefore framed in terms of out-of-sample generalization across participants. Model parameters are identified using a training participant set and evaluated on a disjoint validation set, ensuring that reported performance reflects population-level regularities rather than within-participant overfitting. This design explicitly prioritizes generalization to unseen individuals, which is essential for trust models intended to support analysis and simulation in broader HRI contexts. Under this framing, degradation in performance on some participants is expected and informative, as it delineates the limits of shared-parameter modeling rather than indicating failure of the underlying trust dynamics formulation.

Because trust dynamics models are simplified representations of complex cognitive processes, validation also requires attention to robustness. Robustness refers to the stability of model behavior and conclusions under reasonable variations in estimation assumptions, solver choice, initialization, and data subsets. A model whose qualitative predictions and generalization performance remain consistent across such variations is considered more informative than a model that achieves marginally better fit under narrowly tuned conditions.

Consequently, robustness analyses play a central role alongside predictive metrics, and variability in parameter estimates is interpreted in light of its impact on predicted trust trajectories rather than as an end in itself.

Finally, validation is treated as a multi-level concept. Quantitative error metrics provide a necessary but incomplete view of model adequacy, particularly when data are noisy and heterogeneous. Qualitative diagnostics, such as trajectory comparisons and representative-case analyses, complement numerical scores by revealing whether the model produces coherent and interpretable trust evolution patterns. Together, these validation criteria support a disciplined interpretation: a trust dynamics model is considered successful if it generalizes across participants, reproduces salient temporal structure, and remains robust under realistic perturbations, even if it does not achieve uniformly low prediction error at the individual level.

4.4. Latent States, Observable Behavior, and Closed-Loop Evaluation

Trust, as modeled in this thesis, is a latent internal state that evolves over time in response to interaction outcomes and contextual factors. While such latent states are commonly hypothesized to influence human behavior, they are not directly observable and do not uniquely determine action (Borsboom et al., 2003). Observable reliance decisions, such as whether a human follows or overrides a system recommendation, are the result of multiple interacting cognitive, situational, and strategic factors, of which trust is only one component. Consequently, the existence of a well-defined and predictive latent trust state does not imply that behavior can be accurately or exhaustively predicted from trust alone.

This distinction has important implications for behavioral modeling. Mapping an inferred latent state to observable actions is not a validation of the latent state itself, but a test of whether that state is sufficient to explain behavior under a given set of modeling assumptions. Weak predictive performance or poor calibration of trust-conditioned behavioral models therefore does not invalidate the underlying trust dynamics model. Instead, it indicates that additional latent variables, decision mechanisms, or contextual dependencies are required to account for observed actions. In this sense, behavioral modeling serves as a diagnostic probe that clarifies the explanatory limits of trust, rather than as a definitive test of trust model correctness.

Closed-loop evaluation further amplifies this distinction. When behavioral decisions sampled from a trust-conditioned model are fed back into the trust dynamics, the resulting interaction trajectories reflect the combined assumptions of both the trust model and the behavioral coupling. Discrepancies between simulated and observed interaction patterns may arise from limitations in the behavioral mapping, the coupling assumption, or their interaction, even when the trust dynamics are well identified in open loop. Such discrepancies are therefore informative: they reveal which trust-to-action assumptions are too strong, which forms of feedback destabilize interaction dynamics, and which additional behavioral structure may be required for realistic loop closure.

Within this framework, closed-loop simulations are not intended as predictive tools for human behavior, but as structured stress tests of modeling assumptions. Their value lies in exposing failure modes and boundary conditions of trust-aware pipelines, thereby guiding principled interpretation and future refinement. Trust estimation and trust-conditioned behavior modeling are thus treated as conceptually distinct stages, with downstream coupling results interpreted conditionally and without retroactively undermining the validity of the inferred trust dynamics.

Proposed Trust Dynamics Model

This chapter presents the formal definition of a dynamic mathematical model of human trust in human-robot interaction (HRI) and explicitly tackles the first research question. The model is designed to represent trust as a latent, time-varying internal state that evolves through interaction, contextual exposure, and accumulated experience, and that can be executed within discrete-time simulations and empirically identified from human-subject data. Drawing inspiration from the three-layered trust framework proposed by Hoff and Bashir (2015), the model explicitly separates dispositional, situational, and learned trust mechanisms while integrating them into a unified, interpretable state representation.

Rather than targeting a specific deployed robotic system, the proposed model is intended as a reusable and identification-ready trust dynamics formulation that can serve as the core of an end-to-end modeling pipeline. The design emphasizes modularity, boundedness, and interpretability, ensuring structural compatibility with empirical parameter identification and subsequent validation, without presupposing their outcome. A search-and-rescue-inspired decision-support scenario is used throughout the thesis as a motivating application context characterized by uncertainty and risk, but the modeling principles developed here are not restricted to that domain.

The chapter first introduces key design requirements used in the development of the proposed trust dynamics model. It then motivates the modular architecture of the proposed model, with explicit reference to each of its components. The chapter further describes each of the proposed trust modules in detail. Both continuous-time formulations and discrete-time implementations are reported. Finally, it proposes the final integrated trust dynamics model. The chapter concludes by explicitly addressing its implications for the first research question.

5.1. Design Requirements for Dynamic Trust Modeling

The objective of this chapter is to formally define a trust dynamics model that is suitable for empirical identification and integration within a broader human-robot interaction modeling pipeline. Achieving this objective requires treating trust not as a static or post-hoc assessment, but as an internal state that evolves over time in response to interaction outcomes, contextual conditions, and prior beliefs. In contrast to snapshot-based trust measurements or purely descriptive models, a dynamic formulation is necessary to capture temporal inertia, trust calibration effects, and adaptation across repeated interactions.

Accordingly, the proposed model is designed to balance psychological interpretability with mathematical structure and computational executability. While probabilistic approaches to trust modeling have been explored in prior work, this thesis adopts a deterministic formulation that prioritizes transparency of assumptions, parameter identifiability, and suitability for discrete-time simulation. This choice aligns with performance-centric algebraic trust models identified in the literature, while retaining the ability to represent dynamic temporal evolution. Based on the literature findings of chapter 2, the resulting design requirements (R1 through R6) define the structural properties that the trust dynamics model must satisfy in order to serve as a reliable foundation for subsequent identification, validation, and exploratory closed-loop coupling.

- R1 Boundedness and interpretability.** Trust must remain bounded within a normalized range to preserve psychological interpretability and enable comparison across participants, conditions, and simulation runs.
- R2 Modular decomposition aligned with trust theory.** The model must explicitly separate dispositional, situational, and learned trust components, consistent with the three-layer trust framework identified in the literature, while allowing these components to interact through a unified state representation.
- R3 Event-driven updates with continuous between-event evolution.** Trust must respond to discrete human-robot interaction outcomes (e.g., successes and failures), while also evolving continuously between interactions to preserve temporal memory and long-term dynamics.
- R4 Asymmetric learning and diminishing returns.** The model must capture the empirically observed asymmetry between trust loss and trust gain, as well as saturation effects whereby repeated successes/failures yield diminishing increases/drops in trust.
- R5 Baseline attraction and long-term stabilization.** In the absence of new evidence, trust should gradually relax toward an individual-specific baseline level, reflecting dispositional tendencies and preventing permanent over- or under-calibration following isolated events.
- R6 Contextual and risk-dependent modulation.** Trust must be continuously influenced by situational factors such as perceived operational risk, even when no explicit interaction occurs, reflecting the close relationship between trust, uncertainty, and vulnerability emphasized in prior human–automation studies.

Together, these requirements rule out purely interaction-triggered updates and static trust representations, and motivate a continuously evaluated, modular trust architecture. It is important to emphasize that the design requirements defined here pertain exclusively to the formal definition of a trust dynamics model. At this stage, no claims are made regarding empirical validity, parameter identifiability, or predictive performance. Instead, the requirements are formulated to ensure that the resulting model is interpretable, executable, and structurally compatible with subsequent identification, validation, and exploratory behavioral coupling. The assessment of whether these objectives are met is explicitly deferred to later chapters.

The following section introduces the proposed model structure that satisfies these requirements by decomposing trust into interpretable components and defining a deterministic update mechanism suitable for identification and closed-loop simulation.

5.2. Modular Architecture and Trust State Definition

To satisfy the design requirements established in section 5.1, the proposed trust dynamics model adopts a modular architecture in which distinct psychological mechanisms are represented as separate, interpretable components. This structure enables trust to evolve dynamically while preserving theoretical clarity, computational tractability, and extensibility to future modeling efforts.

The overall trust score τ is not treated as a monolithic dynamical state. Instead, it is reconstructed from multiple conceptually distinct components that correspond directly to the three-layer trust framework identified in the literature. This separation ensures that short-term reactions, long-term internalized beliefs, contextual modulation, and prior knowledge are not conflated within a single update mechanism.

The role of this architectural definition is to make explicit the structure, assumptions, and update semantics of the trust dynamics model; empirical identification, validation, and behavioral relevance are addressed in subsequent chapters.

Trust Components

The proposed model decomposes trust into the following five components:

- **Dispositional Trust** (τ^{disp}): a static, individual-specific baseline representing the operator's inherent tendency to trust automation. This component does not evolve during the task and serves as the long-term attractor for trust dynamics.
- **Situational Trust** (τ^{sit}): an instantaneous, context-dependent contribution that captures the effect of environmental risk and operational conditions. Situational trust is evaluated as a function of observed risk and does not exhibit internal memory.
- **Reputation** (τ^{rep}): an initial learned trust component representing pre-existing beliefs or information about the system acquired prior to direct interaction. Reputation evolves continuously over time through gradual decay as direct evidence accumulates.

- **Personal Experience** (τ^{exp}): an event-driven learned trust component capturing the immediate impact of successes and failures during human-robot interaction. This component produces discrete trust adjustments with asymmetric gain and loss behavior.
- **Latent Trust** (τ^{lat}): a slowly varying internal trust state that integrates the effects of accumulated experience and relaxes toward dispositional trust between interaction events. Latent trust encodes temporal inertia and long-term stabilization.

Trust Reconstruction

The instantaneous trust score τ is reconstructed as an additive combination of the dynamic components:

$$\tau = \tau^{\text{lat}} + \tau^{\text{rep}} + \alpha^{\text{sit}} \tau^{\text{sit}}(r), \quad (5.1)$$

where α^{sit} weights the influence of situational trust relative to internalized belief components. Importantly, the personal experience term τ^{exp} does not contribute directly to τ ; instead, it modifies the latent trust state, ensuring that instantaneous outcomes affect long-term belief rather than transient contextual perception.

This formulation ensures that each trust component evolves according to its own mechanism, preventing unintended cross-effects between psychological processes and preserving interpretability.

State Variables and Inputs

Within this architecture, only a subset of components constitutes the system state:

- **Dynamic state variables:** Latent trust (τ^{lat}) and reputation (τ^{rep}), which evolve continuously over time.
- **Static parameters:** Dispositional trust (τ^{disp}) and model calibration parameters, fixed per individual or experimental condition.
- **Time-varying inputs:** Discrete interaction outcomes (success or failure), the index of consecutive interactions, observed environmental risk, and elapsed time.

The trust dynamics are therefore governed by a hybrid update structure: discrete event-driven updates occur during human-robot interactions, while continuous evolution governs trust behavior between interactions.

Model Overview

Figure 5.1 provides a conceptual overview of the proposed modular trust dynamics architecture and its main components. The figure organizes trust into four interacting elements: dispositional trust, situational trust, learned trust, and the latent trust state. Each element is impacted by relevant factors, shown in yellow. It follows from the three-layered trust framework proposed by Hoff and Bashir (2015).

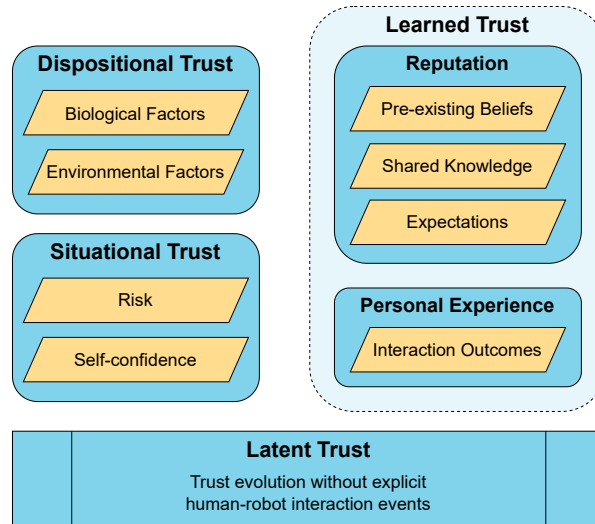


Figure 5.1: Overview of the proposed modular trust dynamics architecture. Dispositional trust provides a stable baseline, situational trust captures instantaneous contextual effects, reputation decays over time, personal experience produces discrete learning impulses, and latent trust evolves during periods without explicit HRI events.

The following sections detail the mathematical formulation of each trust component individually, before presenting the integrated trust dynamics and update rules.

5.3. Dispositional Trust

Dispositional Trust (τ^{disp}) represents an individual's inherent and stable tendency to trust automation, independent of the system's immediate performance or operational context. Within the proposed trust dynamics model, dispositional trust serves two primary roles: it defines the long-term baseline around which trust stabilizes, and it acts as the attractor toward which latent trust evolves during periods without direct interaction.

Consistent with findings in the trust literature, dispositional trust is treated as a slowly varying personal characteristic that remains constant over the duration of a single task or mission (Hoff and Bashir, 2015; Davis, 2019).

Inputs, State, and Output

Dispositional trust is a static component of the model:

- **Inputs:** individual biological predisposition and environmental conditioning,
- **State:** none (time-invariant during simulation),
- **Output:** $\tau^{\text{disp}} \in [0, 1]$.

Once defined, τ^{disp} does not evolve dynamically and is not directly affected by interaction outcomes or contextual risk.

Design Constraints

The formulation of dispositional trust satisfies the following constraints:

- boundedness within the normalized trust interval $[0, 1]$,
- interpretability as an individual-specific baseline,
- invariance over the task duration,
- suitability as an asymptotic reference for latent trust dynamics.

Proposed Functional Form

Dispositional trust is modeled as a weighted linear combination of biological and environmental factors. While the distinction between inherent personal traits and contextual or experiential influences is well established in the trust literature, the decomposition adopted here serves solely as a modeling abstraction and does not impose additional structure on the dispositional trust component itself.

The central modeling assumption is that dispositional trust constitutes a largely constant, individual-specific baseline. The subdivision of this baseline into biological and environmental contributions is introduced to provide an interpretable parameterization within an agent-based framework, rather than to assert a complete or exclusive psychological account of how dispositional trust is formed.

In this formulation, biological factors are modeled as an agent-specific baseline propensity to trust, while environmental factors are modeled as a shared offset reflecting common training or preparatory experience. This choice enables controlled introduction of inter-agent variability while preserving a consistent reference level across agents, without restricting the range or role of dispositional trust in the overall model.

$$\tau^{\text{disp}} = \alpha^{\text{disp}} \delta^{\text{bio}} + (1 - \alpha^{\text{disp}}) \delta^{\text{env}}, \quad \alpha^{\text{disp}}, \delta^{\text{bio}}, \delta^{\text{env}} \in [0, 1]. \quad (5.2)$$

Under this parameterization, δ^{bio} governs inter-agent variability in dispositional trust, while δ^{env} acts as a uniform shift applied to all agents. Importantly, this representation preserves the interpretation of τ^{disp} as a free, bounded baseline parameter, and does not constrain its influence as the long-term attractor of latent trust dynamics.

Expected Behavior

Figure 5.2 illustrates the resulting dispositional trust distribution across different agents for representative parameter values. Variations in δ^{bio} shift the baseline trust level between individuals, while δ^{env} acts as a uniform offset reflecting common training or background knowledge.

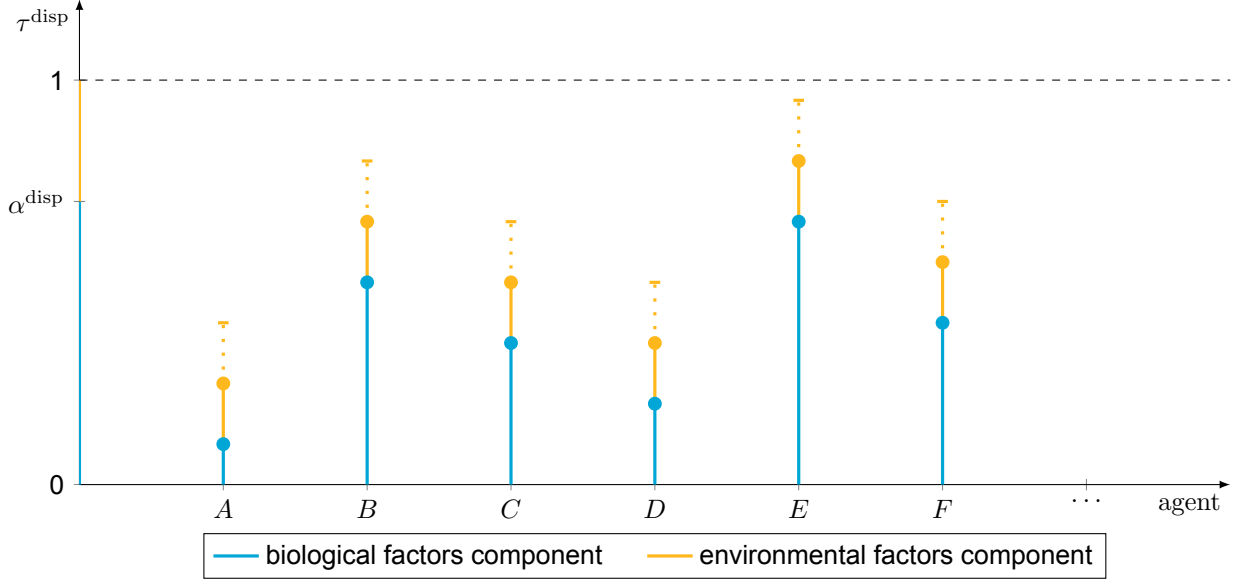


Figure 5.2: Dispositional trust (τ^{disp}) values generated by different combinations of biological and environmental factors for a representative value of the weight α^{disp} .

Calibration and Identification Note

Although the general formulation above supports simulation-based generation of dispositional trust values, in the empirical identification pipeline used in this thesis, τ^{disp} is obtained directly from pre-interaction questionnaire data and treated as a fixed, participant-specific parameter. It is therefore not estimated during model identification (Chapter 6), but instead defines the reference level toward which latent trust dynamics converge.

5.4. Situational Trust

Situational Trust (τ^{sit}) captures the instantaneous modulation of trust induced by the current operational context. Unlike learned or latent trust components, situational trust does not encode memory of past interactions; instead, it reflects how contextual factors such as perceived risk influence the operator's momentary willingness to rely on the system.

In the proposed model, situational trust provides a mechanism through which trust can respond continuously to changes in the environment, even in the absence of explicit human-robot interaction events.

Inputs, State, and Output

Situational trust is defined as a memoryless function:

- **Input:** a scalar risk variable r representing the current operational hazard,
- **State:** none (no temporal accumulation or internal dynamics),
- **Output:** $\tau^{\text{sit}}(r) \in [-1, 1]$.

In the Search and Rescue application considered in this thesis, task complexity and workload are treated as constant, leaving risk as the dominant situational factor.

Design Constraints

The situational trust formulation satisfies the following constraints:

- instantaneous evaluation at every time step,
- bounded output suitable for weighted combination with other trust components,
- independence from latent and reputational dynamics,
- interpretable interaction with stable individual characteristics.

Role of Self-Confidence

Situational trust incorporates a self-confidence parameter β , representing the operator's confidence in their own abilities relative to reliance on automation (de Vries et al., 2003). In this model, β serves as a structural parameter that modulates how contextual risk interacts with the individual's baseline trust level.

Importantly, β is introduced as a modeling abstraction that enables different situational responses to risk, rather than as a comprehensive psychological measure. Its role is twofold:

1. it determines the qualitative form of the situational trust response relative to dispositional trust,
2. it is later reused as a behavioral decision threshold in the trust-to-action mapping defined in chapter 8.

Proposed Functional Form

Situational trust is defined as a function of risk using a dual exponential formulation:

$$\tau^{\text{sit}}(r) = \begin{cases} 2e^{-\lambda^{\text{sit}} r} - 1, & \text{if } \tau^{\text{disp}} \leq \beta, \\ 1 - 2e^{-\lambda^{\text{sit}} r}, & \text{if } \tau^{\text{disp}} > \beta, \end{cases} \quad (5.3)$$

where λ^{sit} controls the sensitivity of situational trust to changes in risk.

This conditional structure allows situational trust to either amplify or attenuate reliance as a function of risk. When the operator's dispositional trust exceeds their self-confidence ($\tau^{\text{disp}} > \beta$), increasing risk raises situational trust, representing a shift toward reliance on automation under perceived vulnerability. Conversely, when self-confidence exceeds baseline trust ($\tau^{\text{disp}} \leq \beta$), increasing risk reduces situational trust, reflecting cautious disengagement in hazardous contexts.

Expected Behavior

Figure 5.3 illustrates the resulting situational trust curves for representative parameter values. Higher values of λ^{sit} produce a more rapid transition toward the asymptotic trust response as risk increases, while the conditional structure ensures that situational trust interacts meaningfully with stable individual traits.

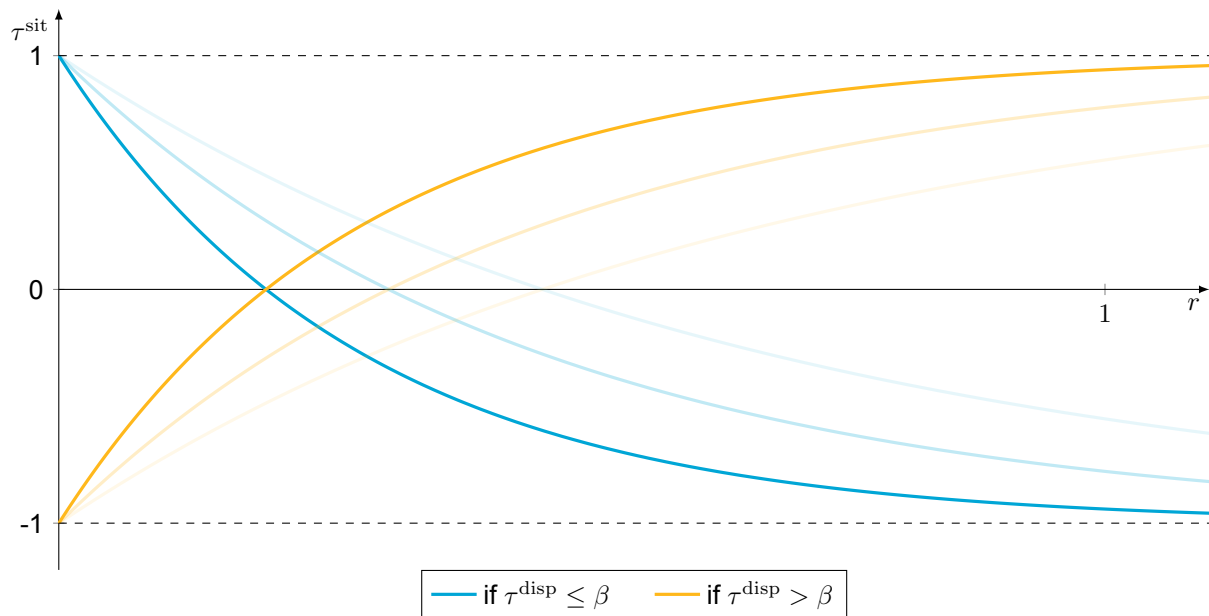


Figure 5.3: Situational trust (τ^{sit}) as a function of risk (r) for representative values of the decay rate λ^{sit} . Curves with lower opacity represent lower values of the decay rate.

Modeling Note

Situational trust contributes directly to the reconstructed total trust score, but does not modify latent trust or reputation. Consequently, situational effects do not accumulate over time and disappear as soon as contextual conditions change, ensuring a clear separation between contextual modulation and learned trust dynamics.

5.5. Learned Trust I: Reputation

Reputation (τ^{rep}) represents learned trust acquired prior to direct interaction with the system. It captures pre-existing beliefs, shared knowledge, or expectations formed through training, instruction, or third-party information, before any personal experience is accumulated. Following the findings of the literature review, reputation effects are assumed to be relevant prior to any direct personal experience, but rapidly decaying once interaction with robots accumulates (Ullrich et al., 2021).

Within the proposed trust dynamics model, reputation provides an initial learned trust contribution that gradually diminishes as direct interaction evidence becomes available.

Inputs, State, and Output

Reputation is modeled as a continuously evolving state:

- **Input:** elapsed time t ,
- **State:** a single scalar trust value that decays monotonically over time,
- **Output:** $\tau^{\text{rep}}(t) \in [-1, 1]$.

Reputation is initialized before the start of direct interaction and evolves independently of interaction outcomes.

Design Constraints

The reputation component is designed to satisfy the following constraints:

- boundedness within a symmetric interval to allow both positive and negative prior beliefs,
- monotonic decay in the absence of reinforcement,
- independence from personal experience and latent trust dynamics,
- interpretability as pre-interaction knowledge rather than adaptive learning.

Proposed Functional Form

Reputation is modeled using an exponential decay function:

$$\tau^{\text{rep}}(t) = \tau_0^{\text{rep}} e^{-\lambda^{\text{rep}} t}, \quad \tau^{\text{rep}} \in]-1, 1[, \quad (5.4)$$

where τ_0^{rep} denotes the initial reputation value at the start of the task.

To reflect differences in prior training or familiarity, the decay rate λ^{rep} is parameterized as

$$\lambda^{\text{rep}} = \Lambda^{\text{rep}} \delta^{\text{training}}, \quad (5.5)$$

where $\Lambda^{\text{rep}} > 0$ denotes the maximum decay rate and $\delta^{\text{training}} \in [0, 1]$ represents a normalized training level.

As with the decomposition adopted for dispositional trust, this parameterization constitutes a modeling abstraction rather than a claim about the precise psychological mechanisms governing reputation decay. Its purpose is to provide an interpretable and controllable means of introducing systematic variation in how quickly agents discount pre-existing beliefs, while preserving reputation as a distinct, well-defined trust component.

Expected Behavior

Figure 5.4 illustrates the characteristic exponential decay of the reputation component over time. Regardless of its initial value, reputation gradually approaches zero as interaction progresses, ensuring that long-term trust is dominated by accumulated experience and latent belief rather than pre-task assumptions.

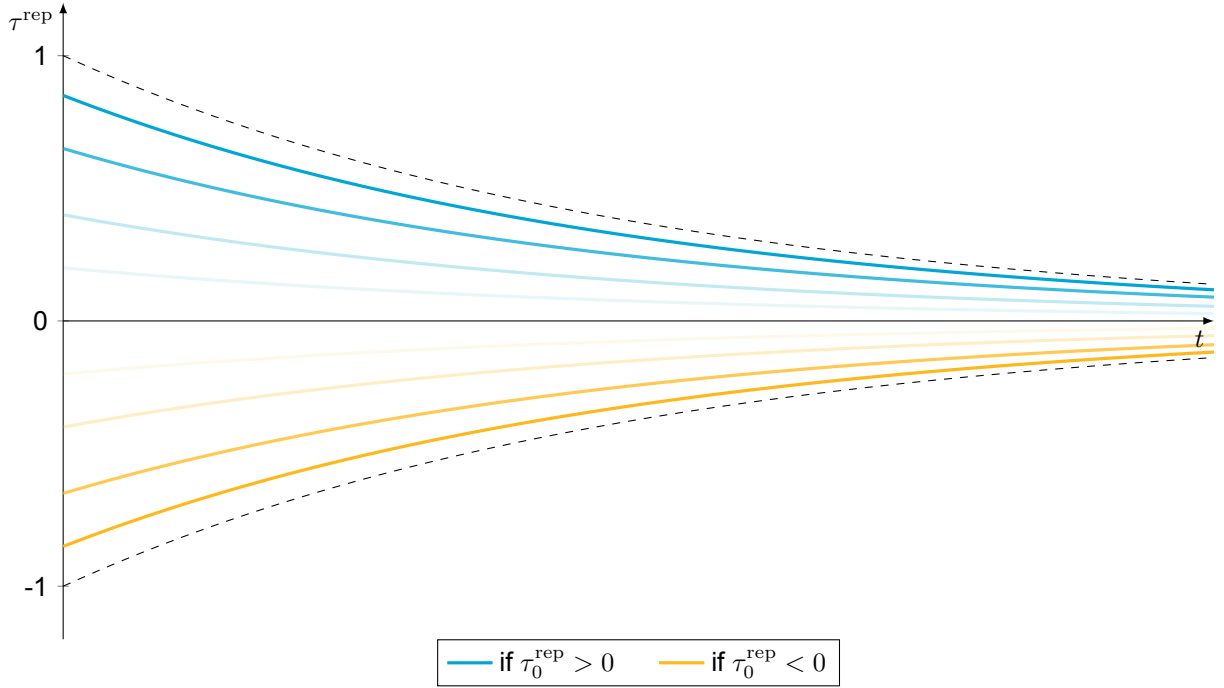


Figure 5.4: Exponential decay of the reputation component (τ^{rep}) over time (t) for representative values of the initial reputation τ_0^{rep} . All curves share the same representative value of the decay rate λ^{rep} . Curves with lower opacity represent lower absolute values of initial reputation.

Modeling Note

In the proposed formulation, reputation evolves continuously with time and is not directly modified by interaction outcomes. An alternative modeling choice would have been to define reputation decay as a function of the number of human-robot interaction events, using an equivalent exponential formulation indexed by interaction count rather than elapsed time.

While such an approach could offer increased robustness to variations in interaction timing and may be advantageous in contexts where interaction frequency is irregular or task-dependent, it would also introduce a tighter coupling between reputation and personal experience. In particular, interaction events would then contribute simultaneously to multiple trust components, reducing the conceptual independence between pre-interaction knowledge and experience-based learning.

The time-dependent, interaction-independent formulation is therefore adopted to preserve a clear separation between trust components and to maintain modularity of the overall model. The event-based alternative remains a viable extension and is identified as a potential direction for future refinement, particularly in multi-context or highly asynchronous interaction scenarios.

5.6. Learned Trust II: Personal Experience

Personal Experience captures the acute, discontinuous trust adjustments that arise from direct human-robot interaction outcomes. Unlike reputation (pre-interaction knowledge) and latent trust (slow belief dynamics), this component represents instantaneous learning impulses driven by observed successes and failures.

A central premise, consistent with empirical findings in the trust literature, is that trust is typically easier to lose than to gain, and that repeated outcomes exhibit diminishing marginal impact (Yang et al., 2022). To encode these effects, the magnitude of the experience-induced trust change is modeled as a function of the number of consecutive outcomes, denoted by $n \in \mathbb{Z}^+$.

Inputs, State, and Output

Personal experience is implemented as an event-driven stateless mapping:

- **Inputs:** outcome type (success or failure) and consecutive outcome index n ,
- **State:** none. The dependence on past outcomes is represented only through the externally computed summary variable n , while long-term temporal integration is handled by the latent trust state,
- **Output:** $\tau^{\text{exp}}(n) \in [-1, 1]$.

In other words, the module introduces no additional dynamical state beyond what is already present in τ^{lat} ; it only maps the current event and the current value of n to an impulse magnitude.

Design Constraints

The personal experience module is designed to satisfy:

- **asymmetry:** failures induce sharper initial drops than the gains induced by successes,
- **diminishing returns:** successive identical outcomes have decreasing incremental effect,
- **boundedness and stabilization:** the induced adjustments should not grow unbounded with repeated outcomes,
- **event locality:** experience affects trust only at interaction events, not continuously over time.

Modeling Failures

To model trust loss under repeated failures, the experience impulse is defined using an exponentially decaying sequence whose magnitude diminishes with the number of consecutive failures. The first failure is constrained to produce a fixed drop of magnitude $\phi^{\text{fail}} \in [0, 1]$, ensuring interpretability and direct control over initial sensitivity.

Let $\lambda^{\text{fail}} > 0$ denote the failure decay rate. The experience impulse associated with the n^{th} consecutive failure is defined as:

$$\tau^{\text{exp, fail}}(n) = e^{-\lambda^{\text{fail}} n} (1 - e^{\lambda^{\text{fail}}}), \quad n \in \mathbb{Z}^+, \quad (5.6)$$

which yields negative increments and decreasing magnitude as n increases.

The parameter λ^{fail} is constrained by fixing the first-failure trust drop:

$$\tau^{\text{exp, fail}}(1) = -\phi^{\text{fail}} \Rightarrow \lambda^{\text{fail}} = -\ln(1 - \phi^{\text{fail}}). \quad (5.7)$$

Modeling Successes

To model trust gain under repeated successes, a bounded logistic formulation is used to capture slow initial growth (“hard to gain”) and saturation toward a maximum level (“stabilization”). The induced trust increment is defined as the difference between consecutive samples of a shifted logistic curve with upper bound 1 and negative lower asymptote $-a$, where $a \in]0, 1[$ shapes the initial reluctance to increase trust.

The experience impulse associated with the n^{th} consecutive success is defined as:

$$\tau^{\text{exp, suc}}(n) = (1 + a) \left(\frac{1}{1 + e^{-\kappa^{\text{succ}}(n - x_{1/2})}} - \frac{1}{1 + e^{-\kappa^{\text{succ}}(n - 1 - x_{1/2})}} \right). \quad (5.8)$$

As with failures, the first-success increment is constrained to a fixed, interpretable value $\phi^{\text{succ}} \in]0, 1[$

$$y^{\text{succ}}(1) = \phi^{\text{succ}}, \quad (5.9)$$

which defines $(\kappa^{\text{succ}}, x_{1/2})$ as functions of (a, ϕ^{succ}) :

$$\kappa^{\text{succ}} = -\ln \left(\frac{1 - \phi^{\text{succ}}}{\phi^{\text{succ}} + a} \right) - \ln(a). \quad (5.10)$$

$$x_{1/2} = \frac{\ln(a)}{\ln \left(\frac{1 - \phi^{\text{succ}}}{\phi^{\text{succ}} + a} \right) + \ln(a)}. \quad (5.11)$$

This constraint ensures that the initial trust gain is controlled directly, while the logistic structure ensures diminishing returns and bounded growth.

Relating Gain and Loss Dynamics

The asymmetry between trust gain and trust loss can be enforced either by directly relating the first-step magnitudes (e.g., ϕ^{fail} larger than ϕ^{succ}), or by comparing the relative speeds at which the success and failure curves approach their asymptotes. A convenient summary metric is the gain/loss ratio χ , which compares characteristic timescales of convergence in an envelope γ (derivation in Appendix A).

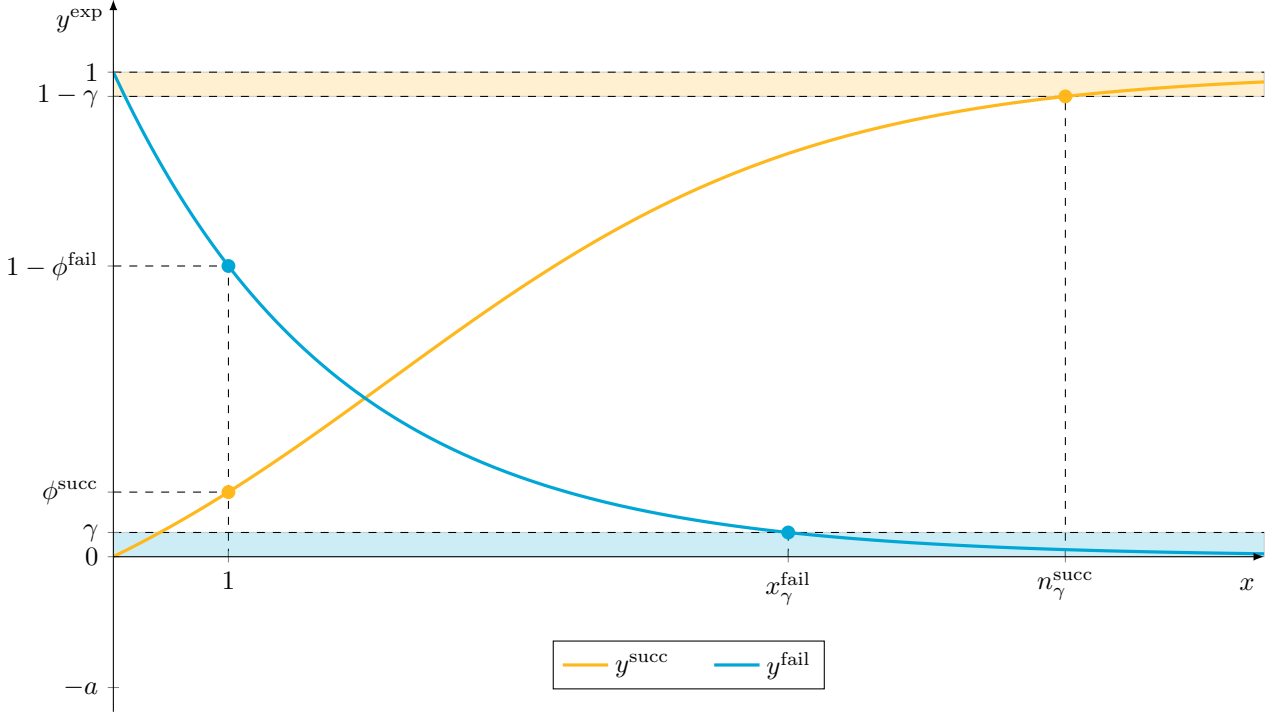


Figure 5.5: Illustration of asymmetric trust adaptation: trust loss under failures is sharper than trust gain under successes, with both exhibiting diminishing marginal effects.

Modeling Note

Personal experience contributes only through instantaneous impulses at HRI events. In the integrated architecture (section 5.2), these impulses update the latent trust state rather than directly contributing to the reconstructed total trust score. This design enforces a clear separation between short-term outcome-driven learning signals and the slowly evolving internalized belief dynamics represented by τ^{lat} .

5.7. Latent Trust

Latent Trust (τ^{lat}) represents the slowly varying, internalized trust state that evolves during periods without explicit human-robot interaction (HRI) events. Its fundamental role is to provide temporal continuity and inertia: acute experience-driven impulses (as defined in section 5.6) do not directly define long-term trust, but instead perturb an internal state that gradually relaxes toward a stable baseline.

In this model, the stable attractor is the agent-specific Dispositional Trust (τ^{disp}). Latent dynamics therefore enforce the empirical property that trust tends to stabilize over time, preventing it from remaining indefinitely inflated or depressed following isolated outcomes.

A non-interaction episode is defined as the interval that begins immediately after τ^{lat} is perturbed by a personal experience impulse at an HRI event, and ends at the next HRI event. During this interval, no new discrete outcome evidence is introduced, and latent trust evolves autonomously under the relaxation dynamics defined below.

Inputs, State, and Output

Latent trust is the model's primary dynamical belief state:

- **Inputs:** elapsed time t (or Δt in simulation) and the baseline τ^{disp} ; latent trust is also perturbed by experience impulses at HRI events,
- **State:** $\tau^{\text{lat}}(t) \in [0, 1]$,
- **Output:** an updated latent trust value that will later be combined with other modules to reconstruct total trust.

Design Constraints

The latent dynamics are designed to satisfy:

- **baseline attraction:** $\tau^{\text{lat}}(t)$ relaxes toward τ^{disp} in the absence of new evidence,
- **asymmetric relaxation:** consistent with the empirical finding that trust is generally easier to lose than to gain, recovery from low trust need not mirror decay from high trust,
- **approximately consistent absolute relaxation:** relaxation should not depend excessively on the initial deviation magnitude (i.e., not purely relative-rate behavior),
- **stability:** convergence toward τ^{disp} without oscillation,
- **validity near the attractor:** a narrow deadband around τ^{disp} avoids the undefined region induced by the rate-transformations.

Proposed Functional Form

Latent trust evolves piecewise depending on whether it starts above or below the dispositional baseline at the beginning of a non-interaction episode. Let τ_0^{lat} denote the latent trust value at the start time t_0 of such an episode.

Case A: Trust Above Dispositional Trust (Decay). If $\tau_0^{\text{lat}} > \tau^{\text{disp}} + \varepsilon$, latent trust decays exponentially toward τ^{disp} :

$$\tau^{\text{lat}}(t) = \tau^{\text{disp}} + (\tau_0^{\text{lat}} - \tau^{\text{disp}}) \exp(-\lambda_{\downarrow}^{\text{lat}}(t - t_0)), \quad t \geq t_0, \quad (5.12)$$

where $\lambda_{\downarrow}^{\text{lat}}$ is an episode-specific decay rate.

A key modeling requirement is that relaxation should exhibit approximately consistent absolute decay behavior across different values of τ_0^{lat} (rather than purely relative decay governed by amplitude). To achieve this, $\lambda_{\downarrow}^{\text{lat}}$ is defined as a transformed rate:

$$\lambda_{\downarrow}^{\text{lat}} = f(\tau_0^{\text{lat}}, \tau^{\text{disp}}; \gamma, \varepsilon) \lambda^{\text{lat}}, \quad (5.13)$$

where λ^{lat} is a global latent decay parameter and $f(\cdot)$ is a scaling function that enforces the desired absolute relaxation characteristics. The closed-form expression and derivations for $f(\cdot)$ is provided in Appendix A.

Case B: Trust Below Dispositional Trust (Recovery). If $\tau_0^{\text{lat}} < \tau^{\text{disp}} - \varepsilon$, latent trust recovers toward τ^{disp} via a bounded logistic formulation:

$$\tau^{\text{lat}}(t) = \tau_0^{\text{lat}} - \tau^{\text{offset}} + \frac{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}{1 + \exp(-\kappa_{\uparrow}^{\text{lat}}(t - t_{1/2} - t_0))}, \quad t \geq t_0, \quad (5.14)$$

where $\tau^{\text{offset}} > 0$ defines a negative lower asymptote relative to τ_0^{lat} and $\kappa_{\uparrow}^{\text{lat}}$ is an episode-specific growth rate.

As in the decay case, the recovery dynamics are designed to exhibit approximately consistent absolute growth across different initial trust levels. This is implemented by defining:

$$\kappa_{\uparrow}^{\text{lat}} = g(\tau_0^{\text{lat}}, \tau^{\text{disp}}; \gamma, \varepsilon, \tau^{\text{offset}}) \kappa^{\text{lat}}, \quad (5.15)$$

where κ^{lat} is a global latent growth parameter and $g(\cdot)$ is a scaling function derived to preserve the intended absolute relaxation behavior. The closed-form expression and derivations for $g(\cdot)$ (and the corresponding midpoint definition) is provided in Appendix A.

Case C: Deadband Around Dispositional Trust. If $|\tau_0^{\text{lat}} - \tau^{\text{disp}}| \leq \varepsilon$, latent trust is held constant:

$$\tau^{\text{lat}}(t) = \tau_0^{\text{lat}}, \quad \text{if } |\tau_0^{\text{lat}} - \tau^{\text{disp}}| \leq \varepsilon. \quad (5.16)$$

This deadband is not introduced for simulation convenience, but to ensure mathematical consistency of the rate-transformations used to define the episode-specific parameters $\lambda_{\downarrow}^{\text{lat}}$ and $\kappa_{\uparrow}^{\text{lat}}$ from the global rates λ^{lat} and κ^{lat} . Near τ^{disp} , these transformations become undefined due to the normalization terms involving the initial deviation. The deadband therefore excludes this narrow region from the piecewise dynamics, and since it is centered at the attractor τ^{disp} , the approximation $\tau^{\text{lat}}(t) \approx \tau_0^{\text{lat}}$ is appropriate.

For this to remain conceptually valid, ε is chosen to be small, so that the deadband represents a very narrow neighborhood around τ^{disp} .

Expected Behavior

The latent component is intended to evolve slowly relative to event-driven experience impulses, so that it stabilizes and regularizes trust trajectories between interactions rather than dominating short-term trust dynamics.

The resulting dynamics ensure that latent trust converges toward the agent's stable baseline τ^{disp} between interactions. When trust is temporarily elevated, it relaxes downward; when it is depressed, it recovers upward in a bounded manner. The scaling functions $f(\cdot)$ and $g(\cdot)$ ensure that relaxation is not dominated by purely relative-rate effects, improving interpretability and consistency across heterogeneous initial trust levels.

Modeling Note

Latent trust is the only component in the architecture that carries long-term memory as a dynamical state. In the integrated model, personal experience contributes through instantaneous impulses that perturb τ^{lat} , while reputation and situational trust remain conceptually separate and do not alter the latent belief dynamics directly.

Finally, while the latent dynamics are defined here in continuous time, their discrete-time evaluation for simulation is treated explicitly in a dedicated discretization section later in this chapter, to preserve a clear separation between model definition and implementation details.

5.8. Discrete-Time Specification of the Trust Dynamics Model

The trust dynamics model proposed in this chapter is formulated primarily in continuous time in order to preserve its psychological interpretability, including gradual belief relaxation, context-driven modulation, and event-triggered learning. However, the objective of this thesis is not only to define a conceptual trust model, but to establish an end-to-end modeling pipeline that can be executed, identified from data, and embedded within closed-loop human-robot interaction simulations.

For this reason, the continuous-time trust processes must be transformed into an explicit, deterministic discrete-time update law. This transformation is not treated as a numerical approximation detail, but as a necessary modeling step that defines how trust evolves at each simulation step, how interaction events are handled, and how model parameters can be identified from time-indexed human data.

This section therefore provides a complete executable specification of the proposed trust dynamics model. All time-varying trust components are expressed as discrete-time update equations evaluated at a fixed sampling interval Δt , and the full rule-based update logic governing event-driven learning and inter-event evolution is made explicit. The resulting formulation constitutes the definitive implementation-level description of the model and directly supports simulation-based validation, parameter identification, and subsequent behavioral coupling.

5.8.1. Discrete-Time Realization of Trust Components

We denote discrete simulation time by the index $k \in \mathbb{Z}_{\geq 0}$, corresponding to physical time $t_k = t_0 + k\Delta t$. At each time step, the model updates (or recomputes) the trust components according to their intended temporal role:

- **Continuously evolving components:** reputation $\tau^{\text{rep}}(t)$ and latent trust $\tau^{\text{lat}}(t)$ are advanced over the interval $[t_k, t_{k+1}]$ using their discrete-time realizations derived from the underlying continuous-time dynamics.
- **Instantaneous context modulation:** situational trust $\tau^{\text{sit}}(r)$ is evaluated algebraically from the current risk input r_k at each step, without internal state.

- **Event-triggered learning:** personal experience $\tau^{\text{exp}}(n)$ is applied only when an HRI event occurs, producing an instantaneous increment/decrement that perturbs the latent trust state and starts a new non-interaction episode.

This discrete-time realization does not change the conceptual model; it provides an implementation layer that makes the resulting trust dynamics executable, reproducible, and compatible with simulation-based evaluation.

5.8.2. Dispositional Trust

Dispositional Trust represents an individual's inherent and stable tendency to trust automation. By design, this component is time-invariant and does not evolve during the simulation. It serves as a fixed baseline toward which latent trust converges and against which situational and experiential effects are interpreted.

Since Dispositional Trust is constant, no discrete-time update equation is required. Instead, it is initialized once at the start of the simulation and remains unchanged for all subsequent time steps:

$$\tau^{\text{disp}}[k] = \tau^{\text{disp}}, \quad \forall k \geq 0. \quad (5.17)$$

Treating Dispositional Trust as constant ensures that all long-term trust evolution arises from interaction-driven learning and contextual modulation, rather than from drift in the individual's inherent trust propensity.

5.8.3. Situational Trust

Situational trust is an algebraic, time-invariant function of the instantaneous risk measure r and therefore does not require discretization. At each simulation step, it is evaluated directly as

$$\tau^{\text{sit}}[k] = \tau^{\text{sit}}(r_k), \quad (5.18)$$

using the functional form defined in section 5.4.

5.8.4. Reputation Trust

Reputation trust evolves continuously over time according to an exponential decay process and is independent of interaction outcomes and latent trust dynamics. Its discrete-time realization is given by

$$\begin{cases} \tau^{\text{rep}}[0] = \tau_0^{\text{rep}}, \\ \tau^{\text{rep}}[k+1] = \tau^{\text{rep}}[k] e^{-\lambda^{\text{rep}} \Delta t}, \end{cases} \quad (5.19)$$

where λ^{rep} is the reputation decay rate.

5.8.5. Personal Experience

Personal experience represents instantaneous, event-driven trust adjustments resulting from HRI successes or failures. These updates are not time-integrated and therefore admit no intrinsic discrete-time dynamics. Instead, at time step k , the appropriate trust increment

$$\tau^{\text{exp}}(n) \quad (5.20)$$

is computed based on the outcome type (success or failure) and the consecutive outcome index n , and is applied instantaneously to the latent trust component when an HRI event occurs.

5.8.6. Latent Trust

Latent trust governs the slow inter-interaction evolution of trust and is the only component whose discrete-time realization depends on the current trust regime relative to the dispositional trust level τ^{disp} .

Latent Trust Above Dispositional Trust

When a latent episode starts with $\tau_0^{\text{lat}} > \tau^{\text{disp}} + \varepsilon$, the latent trust decays exponentially toward τ^{disp} . The discrete-time update is

$$\begin{cases} \tau_{\text{above}}^{\text{lat}}[0] = \tau_0^{\text{lat}}, \\ \tau_{\text{above}}^{\text{lat}}[k+1] = \tau^{\text{disp}} + (\tau_{\text{above}}^{\text{lat}}[k] - \tau^{\text{disp}}) e^{-f \lambda^{\text{lat}} \Delta t}, \end{cases} \quad (5.21)$$

where the scaling function $f(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^\tau)$ is defined as

$$f = \frac{\ln\left(\frac{\gamma}{1-\tau^{\text{disp}}}\right) \ln\left(\frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}}\right)}{\ln(\tau^{\text{disp}}) \ln\left(\frac{1-\tau_0^{\text{lat}}}{1-\tau^{\text{disp}}}\right)}. \quad (5.22)$$

Latent Trust Below Dispositional Trust

When a latent episode starts with $\tau_0^{\text{lat}} < \tau^{\text{disp}} - \varepsilon$, trust recovers toward τ^{disp} following a logistic growth process. Introducing the auxiliary state $\sigma[k]$, the discrete-time formulation is

$$\begin{cases} \sigma[0] = \frac{\tau^{\text{offset}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}, \\ \sigma[k+1] = \sigma[k] (1 + g\kappa^{\text{lat}} (1 - \sigma[k]) \Delta t), \\ \tau_{\text{below}}^{\text{lat}}[k+1] = \tau_0^{\text{lat}} - \tau^{\text{offset}} + (\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}) \sigma[k+1], \end{cases} \quad (5.23)$$

where the scaling function $g(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^\tau)$ is given by

$$\begin{aligned} g &= \frac{\ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right) - \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right)}{\ln(\tau^{\text{offset}}) + \ln\left(\frac{1-\tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right)} \\ &\times \frac{\ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\varepsilon}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)}{\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)}. \end{aligned} \quad (5.24)$$

The constants $\gamma > 0$ and $\varepsilon > 0$ are small, user-defined tolerances introduced by the serial transformations used to enforce comparable absolute growth and decay rates in the latent trust dynamics across different initial trust levels. Specifically, γ defines how close the transformed latent dynamics are required to approach their asymptotic value when constraining the rate parameters, while ε defines a narrow admissible neighborhood around the dispositional trust level τ^{disp} in which latent trust is treated as constant.

These constants are not psychological parameters and are not estimated from data. Instead, they serve a structural role, ensuring mathematical validity and numerical robustness of the transformed dynamics. In practice, both γ and ε are chosen to be small relative to the trust scale, resulting in a very narrow deadband around τ^{disp} and preserving the intended interpretation of latent trust as a slow, secondary process. Please refer to the Appendix for a more in-depth discussion on the role of these transformation constraints.

Deadband Around Dispositional Trust

If the latent episode starts within the deadband

$$|\tau_0^{\text{lat}} - \tau^{\text{disp}}| \leq \varepsilon, \quad (5.25)$$

the latent trust is held constant during the non-interaction interval,

$$\tau^{\text{lat}}[k+1] = \tau^{\text{lat}}[k], \quad (5.26)$$

reflecting both the mathematical constraints of the transformation-based formulation and the intended insensitivity to negligible deviations around the dispositional trust level.

Together, the discrete-time equations presented in this section fully define the implementable dynamics of all time-varying trust components and serve as the direct foundation for the integrated trust dynamics model introduced next.

5.8.7. Event Semantics and Consecutive-Outcome Index

At each simulation step k , the model receives (i) the contextual input r_k , and (ii) an event label E_k indicating whether a human-robot interaction (HRI) outcome occurred during the interval $[t_k, t_{k+1}]$. We encode

$$E_k \in \{\emptyset, \text{succ}, \text{fail}\}, \quad (5.27)$$

where $E_k = \emptyset$ denotes no HRI event during the time step.

The Personal Experience impulse depends on the number of consecutive identical outcomes. This dependence is captured by the consecutive-outcome index $n_k \in \mathbb{Z}^+$, which is updated only at event steps:

$$n_k = \begin{cases} 1, & E_k \in \{\text{succ}, \text{fail}\} \wedge (E_{k^-} = \emptyset \vee E_k \neq E_{k^-}), \\ n_{k^-} + 1, & E_k \in \{\text{succ}, \text{fail}\} \wedge E_k = E_{k^-}, \end{cases} \quad (5.28)$$

where k^- denotes the index of the most recent time step strictly before k such that $E_{k^-} \in \{\text{succ}, \text{fail}\}$. Intuitively, n_k resets to 1 whenever the outcome type changes (success $\leftrightarrow$ failure), and increments otherwise. When $E_k = \emptyset$, n_k is not evaluated and no Personal Experience impulse is applied.

This definition ensures that the Personal Experience module remains stateless, and that the only memory of outcome history it uses is the externally computed summary variable n_k .

5.9. Integrated Trust Dynamics Model

Building on the continuous-time trust processes defined in the previous sections and their discrete-time realizations described in subsection 5.8.1, the complete trust dynamics model is assembled into a unified update system. The model defines how multiple trust components, each governed by distinct mechanisms and time scales, combine to produce a single, interpretable trust estimate suitable for closed-loop human–robot interaction (HRI) simulation.

5.9.1. Parameter Roles and Identification Pathway

For interpretability and empirical reproducibility, it is useful to distinguish (i) parameters obtained directly from measurements or experimental design, (ii) structural tolerances introduced for mathematical validity, and (iii) parameters that must be fitted from observed trust trajectories. Table 5.1 provides this classification for the proposed model.

Table 5.1: Classification of model parameters by identification role.

Category	Examples	Obtained from	Purpose
Measured / participant- or design-defined	$\tau^{\text{disp}}; r_k; \Delta t;$ event labels E_k	Pre-task questionnaire; environment/simulation; experiment logging	Fix baseline trust and provide time-varying inputs required to run the model.
Fitted dynamics parameters	$\alpha^{\text{sit}}, \lambda^{\text{sit}}, \lambda^{\text{rep}}, \phi^{\text{fail}},$ $\phi^{\text{succ}}, a^{\text{succ}}, \lambda^{\text{lat}}, \kappa^{\text{lat}}$	Estimated from time series of trust (and/or reliance) using the iden- tification procedure in chapter 7	Control the magnitude and time- scale of learned trust effects: reputational decay, experience impulses, and latent relaxation.
Derived parameters	$\lambda^{\text{fail}}; \text{success-curve}$ $\text{shift/rate } (x_{1/2}, \kappa^{\text{succ}});$ scaling functions $f(\cdot),$ $g(\cdot)$	Computed deterministi- cally from fitted param- eters and episode initial conditions	Enforce interpretable first-step constraints and produce episode- specific effective rates without introducing additional free parameters.
Fixed structural tolerances	γ, ε	Chosen by the modeler (small values)	Ensure mathematical validity and numerical robustness of the latent- rate transformations; not interpreted as psychological parameters.

This classification clarifies that only a subset of parameters must be estimated from behavioral data, while others are either directly measured or introduced strictly for structural consistency of the mathematical formulation. More specifically, it identifies a minimum set of 8 parameters that require estimation to define the behavior of the trust dynamics model. For notation consistency throughout the remainder of this thesis report, we compactly define this set of parameters as a global parameter vector

$$\theta = [\lambda^{\text{rep}} \quad \alpha^{\text{sit}} \quad \lambda^{\text{sit}} \quad \phi^{\text{fail}} \quad \phi^{\text{succ}} \quad a^{\text{succ}} \quad \lambda^{\text{lat}} \quad \kappa^{\text{lat}}]^{\top}.$$

5.9.2. Overall Trust Decomposition

At any discrete time step k , the overall trust score $\tau[k]$ is defined as the additive composition

$$\tau[k] = \tau^{\text{lat}}[k] + \tau^{\text{rep}}[k] + \alpha^{\text{sit}} \tau^{\text{sit}}(r_k), \quad (5.29)$$

where:

- τ^{lat} is the latent trust component, representing the internalised trust state that evolves slowly between interactions and relaxes toward the individual's dispositional trust level.
- τ^{rep} is the component, capturing pre-existing beliefs that decay continuously over time.
- $\tau^{\text{sit}}(r_k)$ is the situational trust component, instantaneously computed from the current risk level r_k .
- α^{sit} is a weighting factor regulating the influence of situational trust on the total score.

Crucially, trust is not treated as a single dynamical variable. Instead, each component evolves according to its own update mechanism, and the overall trust score is reconstructed algebraically at each time step. This modular structure prevents unintended cross-effects between trust processes operating at different time scales.

5.9.3. Rule-Based Trust Update Structure

The integrated model operates using two mutually exclusive update rules, depending on whether a human–robot interaction occurs during the current time step.

Rule 1: HRI Event Occurs

When an HRI event (success or failure) occurs at time step k , the model performs an instantaneous trust update driven by personal experience. Before this event-driven update, all continuously evolving components are first advanced over the elapsed time interval Δt :

- τ^{lat} evolves according to its latent dynamics,
- τ^{rep} decays according to the reputation model,
- τ^{sit} is recomputed using the observed risk value.

Following this continuous update, the personal experience contribution $\tau^{\text{exp}}(n)$ —corresponding to the n^{th} consecutive success or failure—is applied exclusively to the latent trust component:

$$\tau^{\text{lat}}[k] \leftarrow \tau^{\text{lat}}[k] + \tau^{\text{exp}}(n). \quad (5.30)$$

This design choice ensures that interaction outcomes modify the internalised trust state, rather than directly altering reputation or situational assessments.

Immediately after this update, a new latent trust episode begins. The updated latent trust value becomes the episode's initial condition, and the episode-specific relaxation rate is computed based on whether trust lies above or below the dispositional baseline.

Rule 2: No HRI Event Occurs

If no HRI event occurs during the interval $[k, k+1]$, the system is considered to be in a non-interaction episode. During this period:

- τ^{lat} evolves autonomously toward the dispositional trust level,
- τ^{rep} continues its time-based decay,
- τ^{sit} is recomputed using the observed risk value.

The latent dynamics follow asymmetric relaxation laws:

- When $\tau_0^{\text{lat}} > \tau^{\text{disp}}$, trust decays exponentially toward τ^{disp} .
- When $\tau_0^{\text{lat}} < \tau^{\text{disp}}$, trust recovers logistically toward τ^{disp} .

This asymmetry reflects established findings that trust is typically easier to lose than to regain. The latent component is intentionally designed as a slow stabilizing process, ensuring that it does not dominate short-term trust dynamics driven by interaction outcomes.

To ensure mathematical validity of the transformed decay and growth rates, a narrow deadband region is defined around τ^{disp} . If the latent episode starts within $|\tau_0^{\text{lat}} - \tau^{\text{disp}}| \leq \varepsilon$, the latent trust value is held constant for the duration of the episode. The parameter ε is chosen to be small, ensuring that this region does not affect meaningful trust evolution.

5.9.4. Trust Reconstruction and Boundedness

At every time step, the overall trust score is reconstructed using (A.59). Due to the additive structure of the model, transient excursions outside the nominal interval $[0, 1]$ may occur, particularly during strong situational influences or immediately following extreme sequences of HRI outcomes.

For a properly calibrated model, such excursions are expected to be rare. The internal dynamics of the individual trust components are explicitly designed to remain bounded and to promote convergence toward stable regimes. Consequently, the combined trust value typically remains within $[0, 1]$ during nominal operation.

Nevertheless, when excursions do occur, they can be interpreted meaningfully. Values exceeding the upper bound correspond to states of overtrust, while values below the lower bound correspond to undertrust, both of which are well-established concepts in the trust literature. The interval $[0, 1]$ therefore represents calibrated trust, whereas values outside this range indicate miscalibration rather than model failure.

To preserve interpretability and ensure compatibility with downstream decision-making or simulation logic, the final trust estimate is clipped to the closed interval $[0, 1]$:

$$\tau[k] = \begin{cases} 0, & \tau^{\text{combined}} < 0, \\ \tau^{\text{combined}}, & 0 \leq \tau^{\text{combined}} \leq 1, \\ 1, & \tau^{\text{combined}} > 1, \end{cases} \quad (5.31)$$

where τ^{combined} denotes the unconstrained sum of trust components.

Importantly, this bounding operation is applied only to the reconstructed trust output and does not interfere with the internal evolution of the latent, reputational, or experiential trust dynamics. As such, clipping serves as an output-level normalization mechanism rather than a corrective intervention in the trust dynamics themselves.

5.9.5. Algorithmic Realization of the Integrated Trust Update

For reproducibility, Algorithm 1 summarizes in pseudocode the step-by-step execution of the integrated trust dynamics model at discrete time step k .

Algorithm 1 Integrated Trust Dynamics Update (single step $k \rightarrow k + 1$)

Require: Current states $\tau^{\text{lat}}[k]$, $\tau^{\text{rep}}[k]$; latent-episode regime and parameters; inputs r_k , $E_k \in \{\emptyset, \text{succ}, \text{fail}\}$, Δt ; fitted parameters θ ; structural tolerances γ , ε .

Ensure: Updated states $\tau^{\text{lat}}[k + 1]$, $\tau^{\text{rep}}[k + 1]$, and reconstructed trust $\tau[k + 1]$.

```

1: (A) Contextual evaluation: compute  $\tau^{\text{sit}}[k] \leftarrow \tau^{\text{sit}}(r_k)$ .
2: (B) Reputation update:  $\tau^{\text{rep}}[k + 1] \leftarrow \tau^{\text{rep}}[k] \exp(-\lambda^{\text{rep}} \Delta t)$ .
3: if  $E_k = \emptyset$  then
4:   (C) Latent evolution (episode dynamics):
5:   if  $|\tau_0^{\text{lat}} - \tau^{\text{disp}}| \leq \varepsilon$  then
6:      $\tau^{\text{lat}}[k + 1] \leftarrow \tau^{\text{lat}}[k]$ .
7:   else if  $\tau_0^{\text{lat}} > \tau^{\text{disp}} + \varepsilon$  then
8:      $\tau^{\text{lat}}[k + 1] \leftarrow \tau^{\text{disp}} + (\tau^{\text{lat}}[k] - \tau^{\text{disp}}) \exp(-f \lambda^{\text{lat}} \Delta t)$ .
9:   else
10:    update auxiliary  $\sigma[k] \rightarrow \sigma[k + 1]$  via  $\sigma[k + 1] \leftarrow \sigma[k] (1 + g \kappa^{\text{lat}} (1 - \sigma[k]) \Delta t)$ .
11:     $\tau^{\text{lat}}[k + 1] \leftarrow \tau_0^{\text{lat}} - \tau^{\text{offset}} + (\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}) \sigma[k + 1]$ .
12:  end if
13: else
14:   (D) Latent evolution over  $\Delta t$ : advance  $\tau^{\text{lat}}[k] \rightarrow \hat{\tau}^{\text{lat}}$  using the current episode rule as (C) (with  $\hat{\tau}^{\text{lat}}$  in place of  $\tau^{\text{lat}}[k + 1]$ ).
15:   (E) Compute consecutive index: compute  $n_k$  via (5.28).
16:   (F) Experience impulse: compute  $\tau^{\text{exp}}(n_k)$  using the success/failure mapping corresponding to  $n_k$ .
17:   (G) Apply impulse to latent state:  $\tau^{\text{lat}}[k + 1] \leftarrow \hat{\tau}^{\text{lat}} + \tau^{\text{exp}}(n_k)$ .
18:   (H) Start new latent episode: set  $\tau_0^{\text{lat}} \leftarrow \tau^{\text{lat}}[k + 1]$  and determine episode regime:
19:   if  $|\tau_0^{\text{lat}} - \tau^{\text{disp}}| \leq \varepsilon$  then
20:     set deadband regime.
21:   else if  $\tau_0^{\text{lat}} > \tau^{\text{disp}} + \varepsilon$  then
22:     compute  $f$  (and thus the effective rate  $f \lambda^{\text{lat}}$ ) for the new episode.
23:   else
24:     compute  $g$  (and thus the effective rate  $g \kappa^{\text{lat}}$ ) for the new episode, initialize  $\sigma[0]$ .
25:   end if
26: end if
27: (I) Trust reconstruction:  $\tau^{\text{combined}} \leftarrow \tau^{\text{lat}}[k + 1] + \tau^{\text{rep}}[k + 1] + \alpha^{\text{sit}} \tau^{\text{sit}}[k]$ .
28: (J) Output clipping:  $\tau[k + 1] \leftarrow \text{Clip}_{0,1}(\tau^{\text{combined}})$ .

```

5.10. Synthesis and Implications for Research Question 1

This section synthesizes the contributions of chapter 5 and explicitly addresses the first research question posed in this thesis, which concerns the formal definition of a dynamic mathematical model of human trust in human-robot interaction.

Research Question 1: *How can a dynamic mathematical model of human trust in HRI be formally defined?*

- RQ1.a** How can the three-layered trust framework (dispositional, situational, and learned trust) be translated into a mathematically explicit, executable model suitable for simulation and analysis?
- RQ1.b** How can the temporal evolution of trust be modeled as a dynamic process that integrates interaction events, robot performance outcomes, and contextual factors while remaining bounded and interpretable?
- RQ1.c** What states, inputs, and parameters are required to construct a modular trust dynamics model that balances expressive power with identifiability?
- RQ1.d** How can the proposed continuous-time trust dynamics model be discretized in time and made fully explicit, deterministic, and reproducible for use in simulation-based evaluation and parameter identification?

Translation of trust theory into an executable model (RQ1.a). This chapter demonstrates how the three-layered trust framework can be translated into a mathematically explicit and executable trust dynamics model by decomposing trust into conceptually distinct but interacting components. Dispositional trust, situational trust, and learned trust are not treated as abstract psychological constructs, but are instead mapped to well-defined mathematical objects with clear roles, inputs, and update mechanisms. Learned trust is further decomposed into reputation, personal experience, and latent trust, enabling the separation of pre-interaction beliefs, event-driven learning impulses, and slow internalized belief dynamics. This structured decomposition preserves theoretical fidelity while yielding a model that can be simulated, analyzed, and extended within computational human-robot interaction frameworks.

Dynamic integration of events, outcomes, and context (RQ1.b). Trust evolution is modeled as a hybrid dynamic process that integrates discrete interaction events, continuous-time belief relaxation, and instantaneous contextual modulation. Interaction outcomes generate asymmetric, diminishing learning impulses that perturb an internal latent trust state, while latent trust evolves continuously between events toward an individual-specific baseline. In parallel, situational trust provides instantaneous modulation based on contextual risk, allowing trust to respond even in the absence of direct interaction. Boundedness, stability, and interpretability are enforced through explicit saturation mechanisms, baseline attraction, and carefully structured nonlinear dynamics, ensuring that the resulting trust trajectories remain meaningful across time scales and interaction regimes.

Model states, inputs, and identifiable parameters (RQ1.c). The proposed model balances expressive power and identifiability by restricting internal dynamical state to a small, interpretable set of variables while treating other influences as measured inputs or static parameters. Latent trust and reputation constitute the only continuously evolving internal states, while dispositional trust is fixed per participant and situational trust is computed algebraically from observed risk. Interaction outcomes enter the model through a stateless personal experience module driven by a compact summary variable (the consecutive-outcome index). This design yields a clear separation between measured quantities, fitted parameters, derived parameters, and structural tolerances, enabling principled parameter identification without overparameterization or loss of interpretability.

Discrete-time realization and executability (RQ1.d). To support simulation-based evaluation and empirical parameter identification, the continuous-time trust dynamics are discretized into a fully explicit, deterministic, and reproducible discrete-time update system. Each trust component admits a well-defined discrete-time realization aligned with its conceptual role: reputation and latent trust evolve via closed-form updates, situational trust is evaluated instantaneously, and personal experience is applied as an event-triggered impulse. The resulting rule-based update structure specifies unambiguous execution semantics at each time step and is summarized algorithmically to ensure reproducibility. This discrete-time specification is a core outcome of the modeling process that enables end-to-end integration of trust dynamics into simulation, identification, and closed-loop HRI pipelines.

Together, these results provide a complete and operational answer to Research Question 1. The proposed model establishes a formally defined, modular, and executable representation of trust dynamics that is explicitly designed for empirical identification and simulation-based analysis. This foundation directly enables the identification and validation procedures developed in the subsequent chapters.

6

Experimental Setup

This chapter describes the experimental and methodological foundation used to identify the parameters of the proposed trust dynamics model. It addresses the second research question by specifying how empirical human-robot interaction data are generated, measured, and prepared to support principled identification of latent trust dynamics. Positioned between the formal model definition in chapter 5 and the identification and validation procedures in chapter 7, this chapter establishes the necessary bridge between theoretical trust modeling and data-driven parameter estimation.

The chapter first introduces the experimental task and interaction scenario, detailing how the task is structured, implemented, and presented to participants. It then motivates the key experimental design choices, including interaction mechanics, interface elements, and trust measurement strategies, with explicit reference to their role in enabling model identification. The chapter further describes the resulting dataset, preprocessing steps, and validation criteria, yielding an identification-ready dataset that serves as the sole empirical basis for the trust model identification and validation performed in the following chapter. Finally, it documents participant recruitment, inclusion criteria, and ethical considerations, including consent procedures and data handling safeguards. The chapter concludes by explicitly addressing its implications for the second research question.

6.1. Description of the Experimental Task

Participants take part in a simulated urban search-and-rescue task in which they assume the role of a human rescuer assisted by an autonomous drone. The task is implemented as an interactive web-based experiment and consists of a sequence of time-critical decision trials in which participants decide whether to follow or ignore the drone's recommendations. In the remainder of this section, each of the components that make up the full experimental task are described.

6.1.1. Experiment Timeline

The experiment follows a fixed high-level structure for all participants. After providing informed consent, participants complete a short demographics questionnaire (age range and gender) and are introduced to the scenario and the simulated autonomous drone they will interact with. They then complete a pre-task trust assessment consisting of a 40-item trust questionnaire, a single-item trust probe, and a self-confidence question probing their default reliance preference in emergency situations.

Prior to the main task, participants are also exposed to a reputation manipulation in which they are shown a set of short reviews describing other simulated rescuers' experiences with the drone, followed by a reputation probe. Participants then receive task instructions and observe a single non-interactive demonstration door trial illustrating the task mechanics without requiring any decisions or revealing outcomes.

The main task consists of three consecutive blocks of 20 door trials each, yielding a total of 60 trials. During the main task, single-item trust probes are interleaved between trials at predefined points, and two 14-item trust questionnaires are administered at block midpoints. A post-task 40-item trust assessment is administered after completion of the final block.

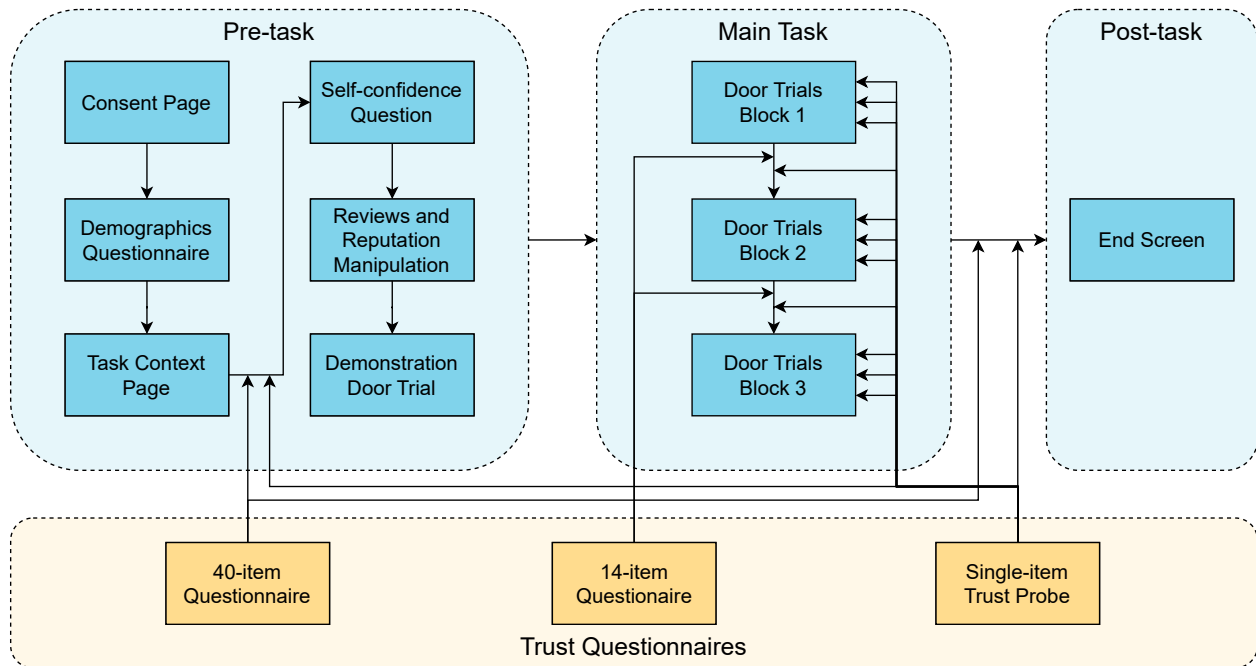


Figure 6.1: High-level timeline of the experimental protocol, showing consent and introduction, pre-task trust assessments, reputation manipulation, demonstration door trial, three blocks of door trials with interleaved trust probes, and post-task trust assessment.

6.1.2. Multi-item Trust Questionnaires Overview

Participants are probed using three types of trust measurement instruments. All instruments are derived from, or inspired by, validated human-robot trust measurement frameworks, but differ in length, reliability and temporal resolution.

40-item Trust Questionnaire

The 40-item human-robot Trust Scale developed by Schaefer Schaefer, 2013 serves as the primary reference instrument in this study. As mentioned in chapter 2, this scale was specifically designed to assess trust in human-robot interaction contexts and was derived through a multi-study psychometric validation process involving item pool construction, expert review, factor analysis, and repeated experimental validation. The finalized instrument yields a continuous trust score on a 0–100% scale and has been shown to be sensitive to both pre-interaction expectations and post-interaction trust changes while maintaining strong internal consistency and construct validity (Schaefer, 2013).

In the present experiment, the 40-item questionnaire is administered only twice per participant: once prior to any interaction with the drone (pre-interaction) and once after completion of all task blocks (post-interaction).

14-item Trust Questionnaire

The 14-item questionnaire is a reduced version of the 40-item scale, also proposed by Schaefer Schaefer, 2013. It retains a subset of items that capture the core dimensions of trust while substantially reducing response burden. As a strict subset of the 40-item instrument, it is conceptually aligned with the same trust construct but exhibits higher measurement noise due to the reduced number of items.

In the experimental design, the 14-item questionnaire is administered at intermediate points between task blocks.

Single-Item Trust Probes

To capture fine-grained temporal dynamics of trust during task execution, single-item trust probes are employed. Participants are periodically asked:

“Right now, how much do you trust this drone?”

and respond on a discretized scale from 0 to 100 in steps of 10.

6.1.3. Self-confidence Question Overview

After the 40-item pre-interaction questionnaire is administered, a single binary probe is shown to participants with the following question:

“In an emergency situation, whose judgement do you generally trust more?”

Participants select one of two options: *My own judgement* or *A rescue robot’s judgement*.

6.1.4. Reviews and Reputation Probe Overview

Before the main task takes place, participants are shown a set of five short textual reviews describing other simulated rescuers’ prior experiences with the drone. The reviews are written as human testimonials and designed to convey a controlled overall tone ranging from strongly negative to strongly positive.

Immediately after reading the reviews, participants responded to the probe question:

“How did these reviews change your opinion of the drone?”

using one of five discrete options: *Much worse*, *Worse*, *No change*, *Better*, or *Much better*.

6.1.5. Demonstration Trial Overview

The final section before the first door trial consists of a single demonstration trial where participants are shown the drone entering and analyzing a search area. This demonstration is visually identical to the subsequent door trials. Crucially, during this demonstration the user makes no choice and no outcome of the drone recommendation is shown.

6.1.6. Single Door Trial Overview

In each trial, participants are presented with a scene containing two closed doors, behind exactly one of which a victim is located. Due to the simulated spread of fire and smoke, there is sufficient time to open only one door before proceeding to the next search area. The overarching objective of the task is to locate as many victims as possible over the course of the mission. This information is clearly stated to participants prior to the 60 door trials take place.

An autonomous drone enters the scene at the start of each trial, visually scans both doors, and then positions itself above the door it recommends to open. Participants must decide whether to follow the drone’s recommendation or ignore it and select the opposite door. Decisions are made under time pressure, with a visible countdown indicating the remaining environmental integrity. If no response is provided before the timer expires, the drone’s recommendation is followed automatically.

After a decision is made, the selected door opens, either revealing a victim behind it, or an empty area. Additional visual feedback in the form of text overlaid on screen is provided indicating whether a victim has been found.

Each door trial therefore follows a fixed sequence of phases:

1. **Drone entry and positioning**, during which the drone enters the scene and stabilizes at mid-height.
2. **Scanning phase**, in which a visual sweep of both doors conveys sensor-based information gathering.
3. **Recommendation phase**, where the drone moves to hover above the door it recommends.
4. **Decision phase**, during which a decision interface appears and participants choose whether to follow or ignore the recommendation under a decreasing time budget.
5. **Outcome phase**, in which the chosen door opens and success or failure feedback is displayed on screen.
6. **Transition phase**, consisting of a brief blackout separating consecutive trials.

This structure ensures consistent temporal dynamics across trials while allowing systematic manipulation of drone reliability and environmental risk.

6.1.7. Experimental Manipulations

Two independent between-participant manipulations define the experimental conditions, namely reviews and trial sets.

Before beginning the task, participants are assigned to one of seven review conditions. Each review condition determines the mixture of review tones presented to the participant, drawn from a fixed pool of

predefined peer-style reviews describing prior experiences with the drone. This manipulation is designed to induce different prior expectations regarding the drone's reliability before direct interaction.

Additionally, each participant is assigned one of eight predefined condition sets. A condition set fully specifies the sequence of sixty door trials, including victim location, drone recommendation, the resulting correctness of the drone on each trial, environmental risk level, and placement of trust probes between trials.

Each set encodes a distinct temporal pattern of drone reliability and environmental risk across the three blocks. Risk is conveyed at each trial through multiple simulated visual cues that jointly influence the perceived situational urgency and hazard of the environment. Within each block, the ratio and ordering of successful and unsuccessful drone recommendations, as well as the variation in environmental risk, follow a predefined structure intended to reproduce trust dynamics reported in prior literature, such as trust calibration, degradation following failures, and recovery after renewed reliability. Drone accuracy is deterministic and entirely defined by the assigned condition set.

6.1.8. Recorded Measures

At the trial level, the experiment records participants' door choices, reaction times, whether the drone's recommendation is followed, trial outcomes, and timeout events. Trust is assessed using both the multi-item questionnaires and repeated single-item probes administered before, during, and after task execution. Together, these measures provide the empirical basis for analyzing the dynamics of trust and reliance under varying conditions of drone reliability and environmental risk.

6.2. Experimental Design Considerations

The experimental protocol presented in this chapter is not intended solely for descriptive analysis of trust-related behavior. Instead, it is purpose-designed to support the identification and validation of the proposed dynamic trust model. This requires repeated interaction events, controlled outcome sequences, independent manipulation of contextual risk, and time-resolved trust measurements. The design choices described are thus directly motivated by the requirements of trust model identification rather than by hypothesis testing alone.

Identifiability and Persistent Excitation

A primary design requirement is that the collected data provide repeated and attributable changes in trust over time. Because trust is modeled as a latent internal state, the experiment must repeatedly drive this state away from equilibrium to avoid degenerate trajectories (e.g., trust saturating early and remaining flat). This motivates the use of structured interaction regimes that induce formation, degradation, recovery, and stabilization phases within a finite number of trials per participant. In particular, the experiment was constructed to include salient events such as reliability regime shifts and clustered failures/successes, which are expected to induce measurable trust changes and thereby improve parameter identifiability.

Separation of Causal Influences

A second core requirement is to separate, as much as possible, the principal factors that influence trust in the proposed model. The protocol is structured to disentangle: (i) prior expectation induced by reputation cues presented before interaction, (ii) situational context induced by exogenous risk cues at decision time, and (iii) experience-driven learning induced by the observed correctness of the robot's recommendations. This separation is implemented through both experimental orthogonalization and temporal ordering: reputation information is shown before any direct interaction begins, risk cues are displayed during each door trial to modulate situational uncertainty independently of correctness, and objective outcomes are revealed only after the participant commits to a decision. This design supports attributing observed trust changes to identifiable drivers, rather than to confounded mixtures of perception and performance.

Task Abstraction as a Methodological and Ethical Constraint

The two-door decision task is a deliberate abstraction of urban search-and-rescue decision-making. Restricting each trial to a binary choice simplifies the decision space while preserving the interaction ingredients required for trust dynamics: a recommendation produced by an autonomous teammate, uncertainty about outcomes, and time pressure under hazard cues. This abstraction also enables a clear operationalization of reliance as whether the participant follows or ignores the drone's recommendation, which is used later as an observation channel for trust model identification. Moreover, realism in hazardous scenarios is ethically constrained: the study cannot place participants under actual physical risk. The task therefore uses controlled visual cues and time pressure to approximate perceived hazard without exposing participants to real danger.

Reduced Data Constraints

A practical constraint of the study is the limited number of participants that can be recruited within the scope of a master's thesis project, combined with the non-negligible duration and cognitive load of the task. A fully stochastic design (e.g., random performance and risk on each trial) would require substantially more data to guarantee that characteristic trust regimes and salient events occur frequently enough to be observed. Accordingly, the experiment adopts a finite set of deliberately structured trial sequences that are designed to "excite" distinct trust dynamics modes while maintaining feasibility under limited data. This choice prioritizes within-subject time-series informativeness over between-condition inference typical of hypothesis-testing studies.

Trust Measurement as an Observability-reactivity Trade-off

Because trust is a latent, time-varying state, its identification requires time-resolved observations rather than only pre/post endpoint assessments. However, trust measurement is not neutral: repeatedly prompting participants to reflect on their trust may impose cognitive load and may also induce measurement reactivity, whereby participants become increasingly aware of their own reported trust trajectory and adjust subsequent judgments and behavior. Since the proposed trust dynamics model does not explicitly account for self-reflective feedback mechanisms, the protocol balances observability against intrusion. This motivates a multi-resolution measurement strategy combining sparse single-item probes with block-level and session-level questionnaires, and motivates deliberate probe placement at informative points in the interaction rather than after every trial.

Engagement, Fatigue Mitigation, and Immersion

To reduce fatigue and maintain attention throughout repeated trials, the interface is designed to minimize monotony without altering the underlying decision structure. Door trials are rendered using a pool of randomly selected visual "skins" (door/background/smoke/fire compositions) with selection constraints to avoid immediate repetition across consecutive trials, such that no two consecutive door trials share the same skin configuration. Short transition text between trials (e.g., indicating the drone moving to the next search area) are included to preserve a sense of spatial continuity, reinforced by consistent motion conventions (the drone always enters from the left and exits to the right of the scene). Fullscreen mode is used to promote immersion and reduce external distractions during the task.

Concrete Design Choices for Reputation Cues and Reliance Behavior

Reputation manipulation is implemented via a pool of multiple reviews rather than a single review. Using multiple reviews improves granularity of the intended tone, reduces the risk that a single miscalibrated review dominates perception, and increases realism without significant added cognitive load to participants. Reviews are presented alongside firefighter figures with varied visual skins and facial/body-expression variants consistent with review tone. Within the door task, the trial interaction was iteratively refined to ensure that the robot's recommendation and action are unambiguous: the drone visually scans doors before suggesting one, and then moves above the suggested door to make its choice explicit. To avoid unintended associations between interface layout and the left/right placement of doors, response buttons were stacked vertically rather than placed side-by-side. For unambiguous assessment of the driving force in participants' behavior, reliance choices were presented as *follow* / *not follow* decisions, rather than asking participants to choose a specific door.

Single-session, Automated Execution and Repeat Participation Safeguards

The experiment is deployed as a fully automated, web-based single-session protocol to ensure consistent execution and to avoid experimenter-induced variability. Because repeated participation would bias the dataset toward participants with prior exposure to the task and robot behavior (learning effects), the protocol includes safeguards to detect immediate re-entry into the study flow (e.g., page refresh/restart) and to prevent logging data from duplicate sessions while preserving participant anonymity.

Table 6.2: Trial configuration for Set B. Each cell shows outcome (correct: ✓, incorrect: ✗), victim/suggestion pair (L: left, R: right), and risk level (low/medium/high/extreme). TP markers indicate positions where a trust probe is administered immediately after that trial.

Set C - Late Betrayal

Displays high reliability throughout most of the session, with a sudden cluster of errors toward the end of the final block. This condition tests the vulnerability of established trust to unexpected failures occurring after long periods of dependable interaction. Table 6.3 collects the trial configuration for Set C.

Table 6.3: Trial configuration for Set C. Each cell shows outcome (correct: ✓, incorrect: ✗), victim/suggestion pair (L: left, R: right), and risk level. TP markers indicate positions where a trust probe is administered immediately after that trial.

Block	Trial Number																				Success rate [%]
1	✓ (L,L) low	✓ (R,R) low	✓ (L,L) low	✓ (R,R) high	✓ (R,R) medium	✓ ^{TP} (R,R) low	✓ (L,L) low	✗ (R,L) low	✓ (L,L) medium	✓ (L,L) low	✗ (R,L) low	✓ ^{TP} (R,R) medium	✓ (L,L) low	✓ (R,R) high	✓ (R,R) low	✗ ^{TP} (R,L) low	✓ ^{TP} (L,L) low	✓ (R,R) medium	✓ (L,L) low	✓ (L,L) low	85
	✓ (L,L) low	✓ (R,R) medium	✓ (L,L) medium	✓ (R,R) low	✓ ^{TP} (R,R) medium	✓ (L,L) low	✗ (L,L) high	✓ (R,R) high	✓ (L,L) extreme	✓ ^{TP} (L,L) medium	✓ (L,L) low	✓ (R,R) medium	✓ (L,L) low	✓ (R,R) low	✓ ^{TP} (R,R) low	✗ (R,R) high	✓ (L,L) low	✓ (R,R) medium	✗ (L,L) medium	✗ (L,R) low	
2	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) medium	✓ ^{TP} (R,R) medium	✓ (R,R) high	✓ (R,R) high	✓ (L,L) high	✗ ^{TP} (R,L) medium	✗ (L,R) extreme	✓ (L,L) medium	✓ (L,L) medium	✓ ^{TP} (R,R) low	✓ (L,L) low	✗ (L,R) medium	✓ (R,R) high	✗ (R,R) medium	✗ ^{TP} (L,L) high	✓ (L,L) high	✓ (R,R) high	✗ (R,L) medium	60
	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) medium	✓ ^{TP} (R,R) medium	✓ (R,R) high	✓ (R,R) high	✓ (L,L) high	✗ ^{TP} (R,L) medium	✗ (L,R) extreme	✓ (L,L) medium	✓ (L,L) medium	✓ ^{TP} (R,R) low	✓ (L,L) low	✗ (L,R) medium	✓ (R,R) high	✗ (R,R) medium	✗ ^{TP} (L,L) high	✓ (L,L) high	✓ (R,R) high	✗ (R,L) medium	

Set D - Volatility and Inconsistency

Alternates between short bursts of correct and incorrect trials across all blocks, producing a highly variable reliability pattern. It is intended as a “trust volatility” condition, where users experience difficulty forming a stable expectation of the drone’s performance. Table 6.4 collects the trial configuration for Set D.

Table 6.4: Trial configuration for Set D. Each cell shows outcome (correct: ✓, incorrect: ✗), victim/suggestion pair (L: left, R: right), and risk level (low/medium/high/extreme). TP markers indicate positions where a trust probe is administered immediately after that trial.

Block	Trial Number																				Success rate [%]
1	✓ (L,L) medium	✗ (R,L) medium	✓ (L,L) medium	✓ (R,R) medium	✓ ^{TP} (R,R) medium	✗ (R,L) medium	✗ (L,R) low	✗ (R,L) medium	✓ ^{TP} (L,L) medium	✓ (L,L) medium	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) medium	✓ (R,R) high	✗ (R,L) medium	✗ ^{TP} (R,L) medium	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) extreme	✓ (L,L) medium	70
	✗ (L,R) low	✗ (R,L) medium	✓ (L,L) high	✓ (R,R) high	✓ ^{TP} (R,R) high	✓ (L,L) high	✓ (R,R) high	✓ (L,L) medium	✓ (L,L) medium	✗ (L,R) low	✗ ^{TP} (R,L) low	✗ (L,R) low	✗ (L,R) high	✓ (R,R) high	✓ (R,R) high	✓ ^{TP} (R,R) high	✓ (L,L) high	✓ (R,R) high	✓ (L,L) high	✓ (L,L) high	
2	✓ (L,L) low	✓ (R,R) medium	✓ ^{TP} (L,L) extreme	✗ (R,L) medium	✓ (R,R) low	✓ ^{TP} (R,R) medium	✓ (L,L) high	✗ (R,L) extreme	✓ (L,L) low	✓ (R,R) medium	✓ ^{TP} (L,L) high	✗ (R,L) extreme	✓ (L,L) low	✗ (L,R) medium	✓ (R,R) high	✓ ^{TP} (R,L) low	✓ (L,L) extreme	✓ (R,R) medium	✓ (L,L) extreme	✗ (L,R) high	65
	✓ (L,L) low	✓ (R,R) medium	✓ ^{TP} (L,L) extreme	✗ (R,L) medium	✓ (R,R) low	✓ ^{TP} (R,R) medium	✓ (L,L) high	✗ (R,L) extreme	✓ (L,L) low	✓ (R,R) medium	✓ ^{TP} (L,L) high	✗ (R,L) extreme	✓ (L,L) low	✗ (L,R) medium	✓ (R,R) high	✓ ^{TP} (R,L) low	✓ (L,L) extreme	✓ (R,R) medium	✓ (L,L) extreme	✗ (L,R) high	

Set E - High Reliability under Varying Risk

Maintains near-perfect success rates across all blocks while systematically manipulating the environmental risk level (from low to extreme). This isolates the influence of situational danger cues on trust formation, independent of reliability itself. Table 6.5 collects the trial configuration for Set E.

Table 6.5: Trial configuration for Set E. Each cell shows outcome (correct: ✓, incorrect: ✗), victim/suggestion pair (L: left, R: right), and risk level (low/medium/high/extreme). TP markers indicate positions where a trust probe is administered immediately after that trial.

Block	Trial Number																				Success rate [%]
1	✓ (L,L) low	✓ (R,R) low	✓ (L,L) low	✓ (R,R) low	✓ ^{TP} (R,R) low	✓ (R,R) medium	✗ (L,L) medium	✓ (R,R) medium	✓ (L,L) medium	✓ ^{TP} (L,L) medium	✗ (R,L) high	✓ (R,R) high	✓ (L,L) high	✓ (R,R) high	✓ ^{TP} (R,R) high	✓ (R,R) extreme	✗ (R,L) extreme	✓ (R,R) extreme	✓ (L,L) extreme	✓ (L,L) extreme	85
	✓ (L,L) extreme	✓ (R,R) extreme	✗ (L,R) extreme	✓ (R,R) extreme	✓ ^{TP} (L,L) extreme	✓ (R,R) high	✓ (L,L) high	✗ (L,R) high	✓ (L,L) high	✓ ^{TP} (L,L) high	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) medium	✓ (R,R) medium	✓ ^{TP} (R,R) medium	✓ (R,R) low	✗ (R,L) low	✓ (R,R) low	✓ (L,L) low	✓ (L,L) low	
2	✓ (L,L) medium	✓ (R,R) low	✓ (L,L) medium	✓ (R,R) high	✓ ^{TP} (R,R) medium	✓ (R,R) medium	✓ (L,L) medium	✗ (L,R) high	✓ (L,L) extreme	✓ ^{TP} (L,L) low	✓ (L,L) medium	✗ (R,L) medium	✓ (L,L) high	✓ (R,R) medium	✓ ^{TP} (R,R) low	✗ (L,R) medium	✓ (R,R) high	✓ (L,L) medium	✓ (L,L) low	✓ (L,L) medium	85
	✓ (L,L) medium	✓ (R,R) low	✓ (L,L) medium	✓ (R,R) high	✓ ^{TP} (R,R) medium	✓ (R,R) medium	✓ (L,L) medium	✗ (L,R) high	✓ (L,L) extreme	✓ ^{TP} (L,L) low	✓ (L,L) medium	✗ (R,L) medium	✓ (L,L) high	✓ (R,R) medium	✓ ^{TP} (R,R) low	✗ (L,R) medium	✓ (R,R) high	✓ (L,L) medium	✓ (L,L) low	✓ (L,L) medium	

Set F - Trust Rebound

Begins with sustained unreliability across the first two blocks and improves markedly in the final one, examining how users recover trust after extended exposure to a poorly performing system. Table 6.6 collects the trial configuration for Set F.

Table 6.6: Trial configuration for Set F. Each cell shows outcome (correct: ✓, incorrect: ✗), victim/suggestion pair (L: left, R: right), and risk level (low/medium/high/extreme). TP markers indicate positions where a trust probe is administered immediately after that trial.

Block	Trial Number																				Success rate [%]
1	✓ (L,L) medium	✓ (R,R) medium	✗ (L,R) high	✓ (R,R) medium	✓ (R,R) extreme	TP	✗ (R,L) low	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) high	✓ (L,L) low	TP	✗ (L,R) medium	✓ (R,R) medium	✗ (L,R) low	✓ (R,R) medium	✗ (R,L) extreme	TP	✗ (R,L) high	✓ (L,L) medium	60
	✗ (R,L) medium	✓ (R,R) medium	✗ (L,R) low	✓ (R,R) medium	✓ (R,R) low	TP	✗ (R,L) medium	✗ (L,R) high	✓ (R,R) medium	✗ (R,L) medium	✓ (L,L) low	TP	✓ (L,L) medium	✗ (L,R) low	✓ (R,R) medium	✗ (R,R) extreme	✓ (R,R) medium	TP	✓ (R,R) high	✗ (R,L) medium	
2	✓ (L,L) low	✓ (R,R) low	✓ (L,L) low	✗ (R,L) low	TP	✓ (R,R) medium	✓ (R,R) low	✗ (R,L) low	✗ (R,L) low	TP	✓ (L,L) extreme	✓ (L,L) extreme	✗ (L,R) medium	✓ (R,R) medium	✓ (L,L) low	✓ (R,R) low	✓ (R,R) high	TP	✓ (R,R) high	✓ (L,L) high	80
	✓ (L,L) low	✓ (R,R) low	✓ (L,L) low	✗ (R,L) low	TP	✓ (R,R) medium	✓ (R,R) low	✗ (R,L) low	✗ (R,L) low	TP	✓ (L,L) extreme	✓ (L,L) extreme	✗ (L,R) medium	✓ (R,R) medium	✓ (L,L) low	✓ (R,R) low	✓ (R,R) high	TP	✓ (R,R) high	✓ (L,L) high	

Set G - Trust Relapse

Exhibits low reliability at first, strong improvement in the middle phase, and another decline near the end. This pattern probes the fragility of recovered trust and how quickly users' confidence collapses after renewed errors. Table 6.7 collects the trial configuration for Set G.

Table 6.7: Trial configuration for Set G. Each cell shows outcome (correct: ✓, incorrect: ✗), victim/suggestion pair (L: left, R: right), and risk level (low/medium/high/extreme). TP markers indicate positions where a trust probe is administered immediately after that trial.

Block	Trial Number																				Success rate [%]
1	✓ (L,L) low	✗ (R,L) low	✗ (L,R) low	✓ (R,R) low	✓ (R,R) medium	TP	✓ (R,R) high	✓ (L,L) low	✗ (L,R) low	✓ (L,L) low	TP	✓ (L,L) extreme	✓ (L,L) low	✗ (L,R) medium	✗ (R,L) low	✗ (R,L) low	✗ (R,L) low	TP	✓ (R,R) low	✓ (L,L) low	60
	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) high	✓ (R,R) medium	✓ (R,R) low	TP	✓ (R,R) medium	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) extreme	TP	✓ (L,L) medium	✗ (L,R) low	✗ (R,L) medium	✓ (R,R) medium	✓ (R,R) medium	✓ (R,R) medium	TP	✓ (R,R) medium	✓ (L,L) medium	
2	✗ (L,R) low	✗ (R,L) low	✗ (L,R) low	✓ (R,R) low	TP	✓ (R,R) medium	✓ (R,L) low	✓ (L,L) low	✗ (L,R) medium	✓ (L,L) low	TP	✓ (L,L) extreme	✓ (L,L) low	✗ (L,R) medium	✗ (R,L) high	✓ (R,R) low	✓ (R,R) low	TP	✓ (R,R) low	✓ (L,L) low	60
	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) high	✓ (R,R) medium	✓ (R,R) low	TP	✓ (R,R) medium	✓ (L,L) medium	✓ (R,R) medium	✓ (L,L) extreme	TP	✓ (L,L) medium	✗ (L,R) low	✗ (R,L) medium	✓ (R,R) medium	✓ (R,R) medium	✓ (R,R) medium	TP	✓ (R,R) medium	✓ (L,L) medium	

Set H - Boundary Conditions (Ceiling–Floor–Ceiling)

Starts with an almost perfect first block, followed by a severe degradation in the second, and partial recovery in the final block. This extreme pattern serves to test the limits of the “trust asymptote” hypothesis, i.e. whether users' trust rebounds symmetrically after reaching near-complete failure. Table 6.8 collects the trial configuration for Set H.

Table 6.8: Trial configuration for Set H. Each cell shows outcome (correct: ✓, incorrect: ✗), victim/suggestion pair (L: left, R: right), and risk level (low/medium/high/extreme). TP markers indicate positions where a trust probe is administered immediately after that trial.

Block	Trial Number																				Success rate [%]
1	✓ (L,L) low	✓ (R,R) low	✓ (L,L) high	✓ (R,R) low	✓ (R,R) medium	TP	✓ (R,R) low	✓ (L,L) extreme	✓ (R,R) high	✓ (L,L) medium	✓ (L,L) low	TP	✓ (L,L) extreme	✓ (R,R) low	✓ (L,L) high	✓ (R,R) low	✗ (R,L) low	TP	✓ (R,R) low	✓ (L,L) low	95
	✗ (L,R) medium	✗ (R,L) low	✗ (L,R) medium	✓ (R,R) medium	TP	✓ (R,R) low	✓ (R,R) medium	✗ (L,R) medium	✗ (R,L) medium	✗ (L,R) high	✓ (L,L) medium	TP	✓ (L,L) extreme	✓ (R,R) low	✗ (L,R) medium	✗ (R,L) medium	✗ (R,L) low	TP	✓ (R,R) high	✓ (L,L) medium	
2	✓ (L,L) medium	✓ (R,R) high	✓ (L,L) low	✓ (R,R) high	TP	✓ (R,R) medium	✗ (R,L) low	✓ (L,L) medium	✓ (R,R) high	✓ (L,L) extreme	✗ (L,R) high	TP	✓ (L,L) medium	✗ (L,R) low	✗ (R,L) medium	✓ (R,R) high	✓ (R,R) low	TP	✓ (R,R) extreme	✓ (L,L) extreme	75
	✓ (L,L) medium	✓ (R,R) high	✓ (L,L) low	✓ (R,R) high	TP	✓ (R,R) medium	✗ (R,L) low	✓ (L,L) medium	✓ (R,R) high	✓ (L,L) extreme	✗ (L,R) high	TP	✓ (L,L) medium	✗ (L,R) low	✗ (R,L) medium	✓ (R,R) high	✓ (R,R) low	TP	✓ (R,R) extreme	✓ (L,L) extreme	

The eight sets collectively span a representative spectrum of reliability and risk trajectories, from highly consistent to erratic, from low to extreme environmental risk, and from early to late failure events. This structured diversity ensures that the resulting behavioral data capture multiple trust dynamics like formation, decay, and stabilization, allowing for a robust validation of the theoretical trust model across distinct experiential contexts.

6.4. Mapping Trust Dynamics Model Components to Experimental Setup

This section describes how the abstract components of the proposed trust dynamics model are instantiated using signals extracted from the experimental setup. The objective is not to redefine the trust model itself, but to specify how participant-level and trial-level quantities required by the model are obtained from the experimental task, questionnaires, and logged interaction events.

The trust model is formulated in terms of latent internal states and driving inputs that are, by design, independent of any specific experimental implementation. To apply the model to the present study, a minimal set of quantities must therefore be inferred or constructed from the experimental observations. These quantities include participant-specific initial conditions and trait-like parameters, as well as trial-specific inputs that modulate trust dynamics over time. Importantly, this mapping concerns only the instantiation of the model for a given experimental context; all global model parameters governing trust dynamics are identified later from behavioral data and are not fixed at this stage.

Throughout this section, we distinguish clearly between: (i) experimental signals directly observed or elicited during the study (e.g., review conditions, questionnaire responses, trial outcomes, and environmental cues), and (ii) model-level quantities required to simulate trust evolution (e.g., initial reputation, dispositional trust, situational inputs, and experience outcomes).

The mappings defined below provide the inputs and initial conditions needed to evaluate the trust dynamics equations for each participant and each trial. Trust measurements collected during the door-trial task (e.g., repeated trust probes and post-interaction questionnaires) are not used to drive the trust model and are therefore not discussed here; they are reserved for independent validation and calibration of model outputs, as described in the subsequent section on trust measurement calibration.

For clarity, the end of this section summarizes all mappings in a compact table linking each trust model component to its corresponding experimental signal and encoding rule.

6.4.1. Mapping of Dispositional Trust

Dispositional Trust represents an individual's stable, context-independent tendency to trust automation or robotic systems. In the proposed trust dynamics model, this component is denoted by τ^{disp} and is assumed to remain constant throughout the interaction. From the perspective of model instantiation, the experimental setup must therefore provide a single scalar value per participant corresponding to this baseline trust level.

Measurement Instrument

Dispositional Trust was measured using Schaefer's 40-item human-robot Trust Scale developed (Schaefer, 2013). In the present experiment, the 40-item scale was administered prior to any task interaction, denoted as the pre-interaction trust measurement ($Q_{\text{pre}}^{(40)}$). This measurement captures participants' baseline trust disposition toward the robotic system before exposure to performance outcomes, situational risk, or experiential learning effects. As such, $Q_{\text{pre}}^{(40)}$ provides a direct empirical estimate of dispositional trust.

Mapping to the Trust Dynamics Model

The trust dynamics model requires a single dispositional trust parameter per participant:

$$\tau^{\text{disp}} \in [0, 1]. \quad (6.1)$$

The experimentally measured pre-interaction trust score $Q_{\text{pre}}^{(40)}$, expressed as a percentage in $[0, 100]$, is mapped into the model domain via linear normalization:

$$\tau_i^{\text{disp}} = \frac{Q_{\text{pre}, i}^{(40)}}{100}, \quad (6.2)$$

where i indexes participants.

This mapping preserves the ordinal and metric properties of the original scale while ensuring compatibility with the bounded trust representation used throughout the model.

Relation to the Model Formulation

In the trust dynamics formulation, Dispositional Trust is expressed as a weighted combination of biological and environmental factors:

$$\tau^{\text{disp}} = \alpha^{\text{disp}} \delta^{\text{bio}} + (1 - \alpha^{\text{disp}}) \delta^{\text{env}}. \quad (6.3)$$

However, these internal parameters are not identified or estimated from the experimental data. They are introduced solely to support agent-based simulation and to provide a mechanistic interpretation of how dispositional trust could arise from individual variability and shared training background. Because the experiment directly measures the full dispositional trust construct via $Q_{\text{pre}}^{(40)}$, no decomposition into δ^{bio} , δ^{env} , or α^{disp} is required.

Accordingly, the experimental setup supplies τ^{disp} directly as an observed parameter, and dispositional trust enters the trust dynamics model as a fixed, participant-specific constant throughout the simulation.

Rationale

This approach offers several advantages:

- it leverages a validated, HRI-specific trust instrument,
- it avoids unnecessary parameter identification or latent inference,
- it preserves interpretability of the dispositional trust component,
- and it cleanly separates experimental measurement from simulation-oriented model structure.

6.4.2. Reputation Assessment and Integration into the Trust Model

In the proposed trust dynamics model, reputation represents prior information about the robotic system acquired before any direct interaction takes place. Conceptually, it captures socially transmitted or externally provided knowledge that influences initial trust but gradually loses relevance as personal experience accumulates. Within the model, reputation contributes an initial condition τ_0^{rep} to the trust state and decays over time according to the dynamics defined in section 5.5.

This subsection specifies how the experimental setup provides an estimate of τ_0^{rep} for each participant, and how this estimate is mapped into the trust model without introducing additional fitted parameters.

Reputation Elicitation via Review-Based Prior

To operationalize reputation experimentally, participants were shown a set of short textual reviews describing prior experiences with the drone before the first door-trial block. Participants were assigned to one of seven predefined review conditions. Each condition corresponds to a review set intended to convey a distinct reputation tone, encoded by a system-assigned expected reputation value

$$v \in \{-2, -1.2, -0.8, 0, +0.8, +1.2, +2\}. \quad (6.4)$$

Positive values indicate favorable prior impressions, negative values indicate unfavorable impressions, and $v = 0$ represents a neutral or mixed review set. The value v is constant throughout the session and represents the experimenter-defined reputation signal supplied to the trust model.

Participants' subjective responses are encoded separately on a five-level scale

$$s \in \{-2, -1, 0, +1, +2\}, \quad (6.5)$$

where positive values correspond to improved perception and negative values to worsened perception. The system-assigned reputation value v and the participant response s intentionally differ in resolution and step size, reflecting their distinct conceptual roles.

Methodological Assessment of Negativity Bias

Negative information is often reported to exert a stronger influence on judgments than equivalently positive information (Kanouse and Hanson, 1987; Baumeister et al., 2001). In the present context, such a negativity bias would manifest if negative review sets produce stronger perceptual shifts than positive review sets of comparable intended magnitude. While the trust model itself does not require this assumption, its presence would motivate alternative mappings between review conditions and model initialization.

To assess whether such asymmetry is present, the following diagnostic quantities are defined. Let v_i and s_i denote the expected reputation value and corresponding participant response for participant i , $i = 1, \dots, N$. Participants are partitioned by review valence:

$$\mathcal{P}_+ = \{i \mid v_i > 0\}, \quad \mathcal{P}_- = \{i \mid v_i < 0\}, \quad \mathcal{P}_0 = \{i \mid v_i = 0\}. \quad (6.6)$$

The mean absolute response magnitude is computed separately for positive and negative review conditions:

$$\mu_+ = \frac{1}{N_+} \sum_{i \in \mathcal{P}_+} |s_i|, \quad \mu_- = \frac{1}{N_-} \sum_{i \in \mathcal{P}_-} |s_i|, \quad (6.7)$$

leading to a raw negativity-bias ratio

$$\eta_{\text{raw}} = \frac{\mu_-}{\mu_+}. \quad (6.8)$$

Because the magnitudes of v_i are not symmetric across all conditions, a complementary normalized influence index is also defined for all $v_i \neq 0$:

$$I_i = \frac{|s_i|}{|v_i|}, \quad (6.9)$$

with group means

$$\bar{I}_+ = \frac{1}{N_+} \sum_{i \in \mathcal{P}_+} I_i, \quad \bar{I}_- = \frac{1}{N_-} \sum_{i \in \mathcal{P}_-} I_i, \quad (6.10)$$

and corresponding normalized bias ratio

$$\eta_{\text{norm}} = \frac{\bar{I}_-}{\bar{I}_+}. \quad (6.11)$$

These quantities are defined here for methodological completeness. Their numerical evaluation and interpretation are reported separately in subsection 7.2.5 and do not introduce additional parameters into the trust model.

Mapping to the Trust Model

In the trust dynamics model, reputation is represented as a bounded component

$$\tau^{\text{rep}} \in [-1, 1], \quad (6.12)$$

with an initial value τ_0^{rep} and exponential decay over time, as defined in section 5.5. To initialize this component, the system-assigned expected reputation value v_i is mapped into the model domain via a linear rescaling:

$$\tau_{0,i}^{\text{rep}} = \frac{v_i}{2}. \quad (6.13)$$

This mapping ensures that the most extreme review conditions correspond to $\tau_{0,i}^{\text{rep}} = \pm 1$, while neutral reviews yield $\tau_{0,i}^{\text{rep}} = 0$. The initialization is fixed and deterministic for each participant.

Participant response values s_i are not incorporated into the reputation initialization or subsequent dynamics. Instead, they are retained solely for diagnostic analysis of the review manipulation. After initialization, the reputation component evolves according to the exponential decay governed by the rate parameter λ^{rep} , which is estimated during model identification and may depend on training as defined in the trust model formulation.

This integration strategy preserves interpretability of the reputation signal, avoids contamination of the trust dynamics by noisy self-reports, and maintains compatibility with the additive structure of the overall trust model.

6.4.3. Mapping of Situational Trust

Situational trust represents the portion of trust that depends on the immediate operational context rather than on prior experience with the automated system. In human–automation interaction, situational trust is particularly sensitive to perceived environmental risk and uncertainty, as these factors shape the operator's assessment of task vulnerability and the automation's ability to cope with adverse conditions.

Within the proposed trust dynamics model, the situational trust component τ^{sit} is determined by three elements: (i) a scalar representation of perceived environmental risk r , (ii) the participant's stable Dispositional Trust τ^{disp} , and (iii) an internal Self-Confidence parameter β that represents a participant's confidence in their own abilities. The role of the experimental setup is therefore to provide interpretable measurements for the situational inputs required by the model, without constraining how situational trust itself is computed.

Perceptual Operationalization of Environmental Risk

The trust dynamics model requires as input a bounded scalar representation of perceived risk,

$$r(t) \in [0, 1], \quad (6.14)$$

which may vary across trials. Importantly, the model does not prescribe how risk must be sensed or computed; it only assumes that a proxy for perceived environmental hazard is available. The experimental setup is thus designed to provide a controlled and interpretable mapping from perceptual conditions to this scalar input.

In the present experiment, situational risk was operationalized through a combination of visual and temporal cues designed to emulate varying levels of environmental danger. Rather than attempting to estimate physical quantities such as absolute distance to fire sources or temperature (which could be relevant candidates in more immersive experimental setups), the simulation relies on perceptual proxies that participants can immediately interpret as indicators of risk.

Because the simulation environment is more naturally described through discrete perceptual states, while the trust model requires a continuous-valued input, four discrete risk levels were defined:

low, medium, high, extreme.

Each level controls a set of visual and temporal parameters that jointly convey the proximity, intensity, and urgency of the surrounding hazard.

Mapping Between Simulation Parameters and Risk Input

Environmental risk is conveyed through the following simulation elements:

- Scene warmth, implemented as a global background color tint (`cool`, `warm`, `hot`), representing atmospheric intensity.
- Fire and smoke encroachment, defined by the inward motion of fire and smoke layers, conveying perceived proximity of danger.
- Camera shake, controlled via oscillation amplitude and period, used to signal turbulence and urgency.
- Integrity countdown time, which decreases with increasing risk to simulate reduced temporal safety margins for decision-making.

Each discrete risk level is associated with a normalized scalar value,

$$r \in [0, 1], \quad (6.15)$$

which serves as the direct input to the situational trust component of the model. The complete parameterization used in the experiment is summarized in Table 6.9.

Table 6.9: Mapping between discrete simulation risk levels and visual/temporal parameters.

Risk level	Scene warmth	Fire encroach. [px]	Smoke encroach. [px]	Shake level	Timeout scale	Risk value r
Low	Cool	60	50	0	1.00	0.2
Medium	Warm	90	80	1	0.85	0.5
High	Hot	110	100	3	0.70	0.8
Extreme	Hot	130	120	5	0.55	1.0

During each trial, the simulation selects one of the predefined risk configurations and applies the corresponding environmental parameters. While the visual and temporal cues influence the participant's perception of environmental danger, only the associated scalar risk value r is recorded and passed to the trust model.

Estimation of Self-confidence under Emergency Conditions

In addition to environmental risk, the situational trust model incorporates a Self-Confidence parameter β , which represents the operator's confidence in their own judgment. Within the trust dynamics formulation, β plays a conditional role: it determines how situational risk modulates trust depending on whether the operator's baseline Dispositional Trust τ^{disp} exceeds their self-confidence or not.

Self-confidence is estimated experimentally using the single binary probe administered prior to task interaction that asks participants to disclose whose judgement they trust more in an emergency situation: their own judgement, or a rescue robot's judgement.

This response provides a direct qualitative comparison between self-confidence and dispositional trust in automation:

- If the participant selects *My own judgement*, self-confidence is inferred to exceed dispositional trust, i.e.,

$$\beta > \tau^{\text{disp}}. \quad (6.16)$$

- If the participant selects *A rescue robot's judgement*, self-confidence is inferred to be lower than dispositional trust, i.e.,

$$\beta \leq \tau^{\text{disp}}. \quad (6.17)$$

This binary comparison is the only information required by the situational trust component to determine which functional regime of $\tau^{\text{sit}}(r)$ applies.

Although the situational trust model itself relies only on the relative comparison between β and τ^{disp} , a numerical estimate of $\beta \in [0, 1]$ is useful for subsequent behavioral modeling. To obtain such an estimate without introducing additional free parameters, a conservative midpoint rule is adopted.

Specifically, β is assigned as the midpoint between the participant's dispositional trust value and the nearest boundary of the unit interval consistent with the qualitative response:

$$\beta = \begin{cases} \frac{\tau^{\text{disp}} + 1}{2}, & \text{if } \beta > \tau^{\text{disp}} \\ \frac{\tau^{\text{disp}} + 0}{2}, & \text{if } \beta \leq \tau^{\text{disp}} \end{cases}. \quad (6.18)$$

This rule preserves the ordinal information provided by the emergency-response probe while avoiding overconfident numerical assumptions.

Relation to the Situational Trust Model

In summary, the experimental setup provides:

- a trial-varying scalar risk input r ,
- a participant-specific qualitative relation between self-confidence and dispositional trust,
- and a participant-specific self-confidence parameter β derived from that qualitative emergency judgment.

Together, these quantities fully specify the inputs required to compute situational trust $\tau^{\text{sit}}(r)$ according to the proposed trust dynamics model, while maintaining a clear separation between experimentally observed signals and model-internal dynamics.

6.4.4. Mapping of Performance Evidence and Personal Experience Signals

The Personal Experience component of the trust dynamics model captures discrete trust updates driven by the outcomes of successive human-robot Interaction (HRI) events. In the proposed formulation, the magnitude of the trust change depends not only on whether the most recent interaction constitutes a success or a failure, but also on the number of consecutive successes or failures experienced up to that point. Consequently, the experimental setup must provide, for each trial, (i) a binary HRI outcome label and (ii) a rule for constructing the corresponding consecutive-count variable n .

Available Experimental Signals and Outcome Feedback per Door Trial

Each door trial produces a participant choice and a drone suggestion, together with ground-truth victim location. Denote, for trial k of participant i :

- $g_{i,k} \in \{L, R\}$: ground-truth door containing the victim,
- $a_{i,k} \in \{L, R\}$: drone suggestion,
- $c_{i,k} \in \{L, R\}$: participant choice.

Crucially, each trial provides unambiguous outcome feedback immediately after the participant's choice: the selected door is opened and is either empty or contains a victim, and a corresponding on-screen SUCCESS or FAILURE card is displayed. Therefore, the participant observes a clear success/failure signal on every trial, and the trust model's Personal Experience updates can be interpreted as being driven by explicit trial outcomes rather than by latent inference.

Defining HRI Success and Failure from Trust-relevant Evidence

The Personal Experience model is driven by whether the interaction provides evidence that the automation was correct or was incorrect with respect to the participant's decision context. In the present task, the participant chooses whether to follow the drone's recommendation; therefore, the trust-relevant outcome must account for both suggestion correctness and the participant's reliance decision.

We define a binary suggestion-correctness indicator

$$\text{corr}_{i,k} = \mathbb{I}\{a_{i,k} = g_{i,k}\} \in \{0, 1\}, \quad (6.19)$$

and a binary reliance indicator

$$\text{rel}_{i,k} = \mathbb{I}\{c_{i,k} = a_{i,k}\} \in \{0, 1\}. \quad (6.20)$$

A door trial constitutes an HRI success (positive personal-experience evidence) if the participant obtains evidence that the drone's recommendation aligns with the correct outcome in the context of the participant's reliance decision. This is operationalized as:

$$y_{i,k} = \text{success} \iff (\text{rel}_{i,k} = 1 \wedge \text{corr}_{i,k} = 1) \vee (\text{rel}_{i,k} = 0 \wedge \text{corr}_{i,k} = 0). \quad (6.21)$$

Equivalently, $y_{i,k} = \text{success}$ when the participant finds the victim, either by correctly following a correct suggestion or by correctly rejecting an incorrect suggestion.

Conversely, a door trial constitutes an HRI failure (negative personal-experience evidence) if the participant obtains evidence that the drone's recommendation leads to an incorrect outcome under reliance, or that rejecting the drone prevents the correct outcome when the drone was right:

$$y_{i,k} = \text{failure} \iff (\text{rel}_{i,k} = 1 \wedge \text{corr}_{i,k} = 0) \vee (\text{rel}_{i,k} = 0 \wedge \text{corr}_{i,k} = 1). \quad (6.22)$$

Equivalently, $y_{i,k} = \text{failure}$ when the participant does not find the victim, either by following an incorrect suggestion or by rejecting a correct one.

This definition ensures that Personal Experience is driven by the trust-relevant evidence available to participants in the task, and it matches the literature findings that both over-reliance and under-reliance can produce negative experience outcomes.

Constructing the Consecutive-count Variable n

The Personal Experience model uses different update laws for consecutive successes and consecutive failures, parameterized by the run length n . For each participant, the experimental trajectory is therefore transformed into consecutive counts:

$$n_{i,k}^{\text{suc}}, \quad n_{i,k}^{\text{fail}}. \quad (6.23)$$

Let $y_{i,k} \in \{\text{success}, \text{failure}\}$ denote the outcome label from Section 6.4.4. The consecutive-count variable is constructed recursively:

$$n_{i,k}^{\text{suc}} = \begin{cases} n_{i,k-1}^{\text{suc}} + 1, & \text{if } y_{i,k} = \text{success} \\ 0, & \text{if } y_{i,k} = \text{failure} \end{cases}, \quad n_{i,k}^{\text{fail}} = \begin{cases} 0, & \text{if } y_{i,k} = \text{success} \\ n_{i,k-1}^{\text{fail}} + 1, & \text{if } y_{i,k} = \text{failure} \end{cases} \quad (6.24)$$

with initial conditions $n_{i,0}^{\text{suc}} = n_{i,0}^{\text{fail}} = 0$. The resulting run-length sequences provide the required integer argument n for the personal-experience update terms $\tau^{\text{exp,suc}}(n)$ and $\tau^{\text{exp,fail}}(n)$ defined in the trust dynamics formulation.

In summary, the experimental setup provides all required inputs for the Personal Experience component through trial-level fields $(g_{i,k}, a_{i,k}, c_{i,k})$, which are deterministically mapped to a binary outcome label $y_{i,k}$ and a consecutive-count variable n . The presence of explicit per-trial feedback (door content reveal and SUCCESS/FAILURE cards) ensures that these outcomes correspond to salient, directly observed events for participants. The subsequent trust update magnitudes are determined by the model-defined success and failure functions (logistic growth for consecutive successes and exponential decay for consecutive failures), and do not require additional experimental measurements beyond the per-trial outcomes defined above.

6.4.5. Summary of Model Component-Experiment Mapping

The preceding subsections have detailed how each component of the proposed trust dynamics model is extracted from the experimental setup. Taken together, these mappings define a clear and minimal contract between experimentally observed signals and the abstract trust model. This subsection consolidates those mappings and explicitly enumerates which quantities are extracted from the experiment, which are fixed model parameters, and which elements of the model do not require experimental input.

A key design principle of the proposed approach is the strict separation between:

- experimentally observed or derived quantities, which are used to initialize or drive the trust dynamics, and
- model-internal parameters, which govern the functional form and temporal evolution of trust components and are identified separately during model fitting or simulation.

Crucially, with the exception of the initial dispositional trust measurement, no self-reported trust probes are used as inputs to the trust dynamics. Instead, trust evolution is driven by controlled experimental manipulations (reputation and risk) and by explicit task outcomes (personal experience), while trust measurements are reserved exclusively for model validation, as described in the subsequent section.

Overview of Extracted Experimental Quantities

Table 6.10 provides a compact overview of the relationship between trust model components and the corresponding experimental signals used to instantiate them. For each component, the table identifies the required model quantity, its experimental source, and its functional role within the trust dynamics formulation.

Table 6.10: Summary of trust model quantities and their mapping to experimental signals.

Trust module	Model quantity	Experimental signal	Role in model
Dispositional Trust	τ^{disp}	Pre-interaction 40-item HRI Trust Scale ($Q_{\text{pre}}^{(40)}$)	Participant-specific, time-invariant baseline trust; used as a stable reference level and as input to situational trust
Reputation	τ_0^{rep}	Assigned review condition ($v \in \{-2, \dots, +2\}$)	Initial trust contribution reflecting prior information; decays exponentially over time
Situational Trust	r	Discrete environmental risk level mapped to scalar value	Trial-varying input modulating situational trust as a function of perceived hazard
	β	Binary emergency-response probe comparing self vs. robot judgment	Determines situational-trust regime relative to τ^{disp}
Personal Experience	$n^{\text{suc}}, n^{\text{fail}}$	Door-trial outcomes (victim found / not found)	Input defining discrete trust updates driven by consecutive HRI successes and failures
Latent Trust	τ^{lat}	No experimental input required	Internal model assumption capturing unobserved trust state; not directly instantiated

Scope and Limitations of Experimental Instantiation

As summarized in Table 6.10, the experimental setup provides only the quantities required to initialize and drive the trust dynamics model at an abstract level. Parameters governing the shape, rate, or asymptotic behavior of the trust components (e.g., decay rates, growth rates, weighting coefficients) are not directly measured and are treated as model parameters to be identified separately.

Equally important, trust measurements collected during the experiment, including repeated trust probes administered throughout the task, are not used as inputs to the trust dynamics. Instead, they serve exclusively as external reference signals against which the model-generated trust trajectories are evaluated. This design choice preserves the explanatory power of the model, avoids circularity between measurement and prediction, and allows trust to be treated as a latent internal state rather than as a directly observed variable.

6.5. Data Validation and Inclusion Criteria

Following raw data normalization and participant-level reconstruction, a systematic validation procedure is applied to determine which participant sessions satisfy all requirements and are therefore suitable for subsequent analysis, ensuring that all retained participants provide complete experimental trajectories.

Validation is performed at the participant level and assesses multiple dimensions of data integrity. A participant session is considered valid only if all of the criteria listed below are satisfied.

- **Non-empty participant timeline:** the reconstructed event timeline must be present and non-empty (guarding against malformed or incomplete session logs).
- **Door-trial task integrity:** door trials must be present, and each door trial must contain non-empty drone suggestion and participant choice fields. Sessions with no door trials, or with any door trial containing missing core fields, are excluded.
- **Trust probe validity and completeness:**
 - (a) Probes must exist (non-empty).
 - (b) The total number of probes must match the expected experimental design, i.e., 13 trust probes per participant.
 - (c) All probe values must be finite and lie within the admissible range $[0, 100]$.
- **Questionnaire completeness and numeric consistency:** the participant record must include all questionnaire instruments required by the protocol:
 - (a) 40-item pre-interaction trust questionnaire and 40-item post-interaction trust questionnaire, each with a finite scalar score and a finite scalar equivalent field (used later for cross-instrument calibration).
 - (b) 14-item intermediate questionnaires, each with a finite scalar score.
- **Probe-to-task alignment:** trust probes with origin `after_door` must point to a door identifier that actually exists in the participant's timeline. This check ensures that each probe can be unambiguously linked to the door trial it is intended to follow, preventing temporal misalignment between events and measurements.

In addition, two soft validation checks are performed to flag anomalies without being used as strict exclusion criteria in the default preprocessing configuration.

- **Demographics sanity:** if demographics are present, their types and fields are checked for consistency; however, missing demographics do not gate participant inclusion, as they are only relevant for participant dataset characterization purposes.
- **Emergency trial response validity:** the emergency trial is checked for a non-empty admissible choice ("self" or "robot"); however, missing emergency trial responses are treated as non-fatal.¹

6.6. Trust Measurements Calibration

The trust dynamics model proposed in this thesis generates a continuous time-varying estimate of simulated trust, which represents the model's internal prediction of a participant's trust state based on dispositional factors, situational inputs, reputation, and accumulated personal experience. In contrast, the experiment provides a collection of observed trust measurements, obtained through multiple self-report instruments administered at different points in time and with varying levels of granularity. Because these measurements differ in reliability, temporal resolution, and measurement burden, they cannot be used interchangeably without adjustment.

This section introduces a calibration and weighting framework that reconciles heterogeneous trust measurements onto a common scale and assigns them relative influence based on their expected uncertainty. It is intentionally detailed, as calibration directly affects parameter identifiability and the validity of model–data comparisons, making it necessary to ensure that identification results reflect trust dynamics rather than instrument artifacts. For convenience, a brief interpretive summary is provided at the end of the section to consolidate the methodological rationale.

¹ Emergency choice is used as an estimation of self-confidence. If missing, a conservative assumption for self-confidence is still possible in downstream analysis.

6.6.1. Purpose and Scope of Calibration

A central methodological challenge arises from the use of different types of observed trust measurements. Even though all measurements aim to capture the same underlying psychological construct, they differ substantially in scale resolution, psychometric reliability, temporal placement, and susceptibility to noise. As a result, raw trust measurements cannot be directly compared to one another nor used interchangeably for model identification without introducing bias or invalid assumptions.

The purpose of the trust measurement calibration procedure is threefold:

1. to place heterogeneous trust measurements onto a common, interpretable metric;
2. to quantify instrument-specific measurement uncertainty in a principled manner;
3. and to enable the joint use of all available trust observations during model identification without over-weighting noisier instruments.

Importantly, calibration is treated here as a methodological operation rather than a model-specific or implementation-driven step. It establishes the statistical foundation that allows multiple trust instruments to be combined consistently when estimating the parameters of the trust dynamics model. All calibration mappings and uncertainty estimates are defined independently of the trust dynamics parameters and are fixed prior to model identification.

Throughout this section, a clear distinction is maintained between:

- **Observed trust:** empirical trust measurements collected from participants using questionnaires and probes;
- **Simulated trust:** the trust trajectory predicted by the trust dynamics model.

The goal of calibration is not to modify the meaning of trust as measured by any instrument, but rather to reconcile these measurements onto a shared scale that supports quantitative comparison with simulated trust.

6.6.2. Trust Measurement Instruments and Their Roles

The experimental protocol employs three types of trust measurement instruments, each serving a distinct methodological role. Beyond differences in psychometric properties, the selection and placement of trust measurements are also influenced by concerns regarding measurement reactivity, as discussed in section 6.2. Explicit trust assessments may themselves influence subsequent judgments and behavior by prompting reflection on one's own trust state. Since the proposed trust dynamics model does not explicitly account for such self-reflective feedback, the experimental protocol balances the need for temporal observability of trust with the risk of perturbing the dynamics it aims to identify.

The 40-item human-robot Trust Scale is administered at the start and end of the experiment. While its length already imposes a non-negligible cognitive load, a more critical concern is that frequent administration during task execution could prompt sustained self-reflection about trust, potentially influencing subsequent trust formation. For this reason, the 40-item questionnaire is deliberately restricted to anchor measurements outside the interactive task flow and is used exclusively as a high-fidelity reference for calibration.

The 14-item questionnaire is administered at intermediate points between task blocks. This placement reflects a compromise: the instrument is short enough to limit participant fatigue and introspective bias, yet sufficiently structured to provide more reliable snapshots of trust evolution than single-item probes. Importantly, the 14-item questionnaire is not administered repeatedly within blocks, thereby limiting its potential to interfere with trial-by-trial trust dynamics.

Single-item probes offer minimal response burden and can be administered repeatedly with limited disruption to task flow. Nevertheless, they are inherently noisier than multi-item questionnaires and more sensitive to transient contextual factors. Moreover, frequent probing still carries a risk of increasing participant awareness of their own trust trajectory. To mitigate this effect, probes are sparsely distributed and phrased in a neutral, momentary manner, avoiding explicit reference to prior responses or expectations of consistency.

As such, single-item probes are not treated as interchangeable with questionnaire-based trust measures. Instead, they are interpreted as low-intrusion, high-variance observations that complement more reliable instruments by increasing temporal resolution.

6.6.3. Common Trust Scale and Notation

Before introducing the cross-instrument calibration procedure, we define the notation used to describe the empirical trust measurements and their relation to the trust dynamics model. This subsection is concerned exclusively with observed trust measurements and their calibration; it does not describe the internal trust state of the proposed model, separately introduced in chapter 5.

Let $i \in \{1, \dots, N\}$ index participants in the experiment. Each participant contributes a sequence of trust measurements collected at predefined time points throughout the experimental protocol. Three types of trust measurement instruments are used, all presented to participants on a nominal $[0, 100]$ scale for usability, but differing substantially in psychometric reliability and measurement noise.

First, the 40-item questionnaire yields high-fidelity trust measurements and is treated as the reference instrument. For participant i , these measurements are denoted by

$$Q_{i,\text{pre}}^{(40)}, \quad Q_{i,\text{post}}^{(40)}, \quad (6.25)$$

corresponding to pre-interaction and post-interaction administration, respectively.

Second, the 14-item questionnaire provides intermediate trust measurements with reduced response burden. These measurements are denoted by

$$Q_{i,\text{mid1}}^{(14)}, \quad Q_{i,\text{mid2}}^{(14)}, \quad (6.26)$$

and are collected between task blocks.

Finally, single-item probes provide low-intrusion, high-frequency trust measurements. For participant i , these are denoted generically as

$$Q_{i,k}^{(1)}, \quad (6.27)$$

where k indexes the time at which the probe is administered. This includes probes collected immediately after anchor questionnaires, at block midpoints, and within task blocks.

Although all instruments are displayed on a $[0, 100]$ range, their outputs are not directly interchangeable. Let $q_{i,t}^{(j)}$ denote an uncalibrated observed trust measurement from instrument $j \in \{1, 14, 40\}$ for participant i at time t , expressed on the unit interval.

$$q_{i,t}^{(j)} = \frac{Q_{i,t}^{(j)}}{100}. \quad (6.28)$$

To enable joint use of all measurements, all trust observations must be calibrated onto a common trust scale that is numerically aligned with the 40-item questionnaire and linearly scaled to the unit interval.

Let $\theta \in \mathbb{R}^8$ denote the parameter vector of the proposed trust dynamics model. Given participant-specific experimental inputs $\mathcal{X}_i$ (e.g., drone recommendations, outcomes, risk conditions), the model produces a simulated trust trajectory

$$\hat{\tau}_i(t; \theta) \in [0, 1], \quad (6.29)$$

which is expressed on the same numerical scale as the calibrated observed trust measurements $\hat{q}_{i,t}^{(j)}$.

Importantly, $\hat{\tau}_i(t; \theta)$ represents the model-predicted trust (hereafter referred to as simulated trust), whereas $\hat{q}_{i,t}^{(j)}$ represents calibrated observed trust. The distinction between observed and simulated trust is maintained throughout the remainder of this chapter to avoid conceptual ambiguity.

The objective of cross-instrument calibration is to establish a statistically principled correspondence between heterogeneous trust measurements and the common trust scale, while explicitly accounting for differences in measurement reliability. This calibrated representation enables all available trust observations to be used jointly during model identification, without conflating instrument noise with trust dynamics.

The calibration procedure is methodological in nature and independent of the specific numerical implementation used to compute it. Its mathematical formulation is detailed in the following subsections.

6.6.4. Calibration Objectives and Assumptions

The experimental protocol yields multiple trust measurements per participant that differ in length, frequency, and psychometric reliability. While these measurements are all intended to reflect the same underlying trust construct, they cannot be treated as interchangeable without adjustment. To avoid over-weighting noisier instruments, the purpose of calibration is therefore twofold: to reconcile heterogeneous instruments onto a common scale, and to account for systematic differences in measurement uncertainty.

The calibration framework is designed to satisfy the following objectives:

1. **Scale alignment.** All observed trust measurements should be expressed on a common numerical scale that is directly comparable to the simulated trust output of the model.
2. **Instrument consistency.** Measurements derived from different instruments should be comparable in expectation, meaning that—after calibration—no instrument is systematically biased relative to another.
3. **Reliability-aware integration.** Instruments with higher measurement noise should contribute less strongly to model identification than more reliable instruments, without being discarded entirely.
4. **Minimal modeling assumptions.** Calibration should avoid introducing latent variables or additional psychological constructs beyond those already defined in the trust dynamics model.

These objectives ensure that the calibration procedure serves as a principled bridge between empirical measurement and model-based trust estimation, rather than an additional modeling layer. To achieve them, the following assumptions are adopted.

Common Underlying Construct

All instruments are assumed to measure the same latent trust construct, differing only in resolution and noise characteristics. This assumption is supported by the fact that the 14-item questionnaire is a strict subset of the 40-item scale, and that the single-item probes are phrased as direct trust judgments.

Instrument-specific Noise

Each measurement instrument is assumed to introduce additive, instrument-specific noise with zero mean and constant variance. The variance may differ across instruments but is assumed to be homogeneous within each instrument type.

Temporal Neutrality of Measurement

The act of measuring trust is assumed not to alter the underlying trust state in a systematic way. While repeated probing may increase self-reflection or awareness, such effects are not explicitly modeled and are treated as part of measurement noise rather than trust dynamics.

Absence of Ground Truth

No instrument is treated as a perfect or ground-truth measurement of trust. Instead, the 40-item questionnaire is used as a reference scale due to its higher reliability, not because it is assumed to be error-free.

6.6.5. Cross-Instrument Calibration

The purpose of calibration is to express heterogeneous trust measurements on a single comparable numerical scale. The 40-item human-robot Trust Scale is selected as the reference scale for calibration. This choice is motivated by its extensive psychometric validation, higher internal consistency, and lower sensitivity to momentary context compared to shorter instruments. Calibration is therefore defined as the construction of a mapping that converts each instrument output into an estimate on the 40-item scale.

All calibrated trust observations are expressed on a scale numerically aligned with the 40-item questionnaire, and raw 40-item measurements are simply scaled down to the unit interval.

We denote a generic calibrated observed trust measurement by

$$\hat{q}_{i,t}^{(j)} \in [0, 1], \quad (6.30)$$

where t indexes the measurement time. This calibrated observed trust can therefore be associated to different measurement types, namely a 40-item questionnaire measurement $Q_{i,t}^{(40)}$, a 14-item questionnaire measurement $Q_{i,t}^{(14)}$, or a single-item probe measurement $Q_{i,t}^{(1)}$.

For each instrument, a mapping to the common trust scale is assumed. We can therefore define this calibrated observed trust as

$$\hat{q}_{i,t}^{(j)} = \frac{\hat{Q}_{i,t}^{(40)}}{100}, \quad (6.31)$$

where $\hat{Q}_{i,t}^{(40)}$ is the calibrated (but not scaled) measurement in the 40-item trust scale. In turn, this calibrated measurement is conditioned on the measurement instrument type so that

$$\widehat{Q}_{i,t}^{(40)} = \begin{cases} Q_{i,t}^{(40)} & , t \in \{\text{pre}, \text{post}\} \\ a_{14}Q_{i,t}^{(14)} + b_{14} & , t \in \{\text{mid1}, \text{mid2}\} , \\ a_1Q_{i,t}^{(1)} + b_1 & , t \in \{1, 2, \dots, 9\} \end{cases} \quad (6.32)$$

where a_{14} , b_{14} , a_1 and b_1 are global calibration coefficients. These calibration mappings (a_{14}, b_{14}) and (a_1, b_1) are estimated using anchor measurements where instruments co-occur in time. Once estimated, these mappings are treated as fixed and are applied uniformly across all participants and time points.

These mappings ensure that, in expectation, calibrated measurements from different instruments are comparable on the common trust scale. Please note that, by definition, $\widehat{q}_{i,t}^{(j)} = q_{i,t}^{(j)}$ for $j = 40$. However, the same equality does not hold for $j = 1$ and $j = 14$.

6.6.6. Instrument-Specific Uncertainty and Reliability Estimation

While linear calibration aligns different trust instruments onto a common numerical scale, it does not address a fundamental limitation: calibrated measurements obtained from different instruments do not carry the same level of uncertainty. Shorter instruments and single-item probes are inherently noisier than longer questionnaires, and this difference must be explicitly accounted for when combining observations for model identification. This subsection formalizes how instrument-specific uncertainty is defined and estimated, and how it is subsequently used to modulate the influence of each observation.

Because the underlying trust state $\tau_i(t)$ is unobservable, absolute measurement variance cannot be identified directly. Instead, instrument reliability is estimated relative to a reference instrument. The 40-item questionnaire is adopted as the reference due to its higher internal consistency and extensive validation in literature.

The relevance of estimating instrument-specific relative variances lies in their use to assign relative weights to calibrated observations. Intuitively, observations from noisier instruments should exert less influence on model identification than observations from more reliable instruments. From a statistical perspective, this follows directly from maximum-likelihood estimation under Gaussian noise assumptions: minimizing a weighted sum of squared residuals with weights

$$w_j = \frac{1}{\sigma_j^2} \quad (6.33)$$

yields an efficient estimator that properly accounts for heterogeneous measurement uncertainty.

Crucially, this approach allows all available trust measurements to contribute to model identification while preventing high-frequency but noisy probes from dominating the estimation process.

It is important to emphasize that instrument reliability and calibration operate purely at the measurement level. They do not modify the trust dynamics model itself, nor do they introduce additional psychological assumptions about how trust evolves.

Instead, they ensure that discrepancies between simulated trust and observed trust are interpreted in light of known measurement uncertainty, enabling principled comparison between simulated trust trajectories and heterogeneous empirical observations.

The next subsection details how these variance estimates are computed for each instrument type and how they are combined into a unified weighting scheme.

6.6.7. Estimation of Instrument Reliability Weights

This subsection describes how instrument-specific relative weights are estimated from experimental data. The objective is to obtain empirically grounded weights that reflect the relative trustworthiness of each measurement instrument when combining observations for model identification.

Conservative Relative WEighting of the 14-item Questionnaire

The 14-item questionnaire is a reduced version of the 40-item trust scale and is administered as a standalone instrument at intermediate time points. Because the 14-item score at anchor points is computed from a strict subset of the same responses used to form the 40-item score, the two quantities do not constitute independent measurements and cannot be used to empirically identify a relative noise level without additional assumptions.

To avoid exploiting this structural dependence and to maintain a conservative calibration strategy, the relative reliability of the 14-item questionnaire is therefore defined a priori based on item count. Under the

standard assumption that averaging a larger number of comparable items reduces measurement variance, the uncertainty of a mean score scales approximately inversely with the number of items. Accordingly, the 14-item questionnaire is assigned a relative weight

$$w_{14} = \frac{14}{40} w_{40}, \quad (6.34)$$

with w_{40} denoting the reference weight attributed to the 40-item questionnaire score measurements. This choice intentionally penalizes the reduced instrument due to its shorter length, providing a transparent and assumption-light weighting rule.

This assumption, coupled with the earlier definition of $w_j = \frac{1}{\sigma_j^2}$, is equivalent to formulating this relation in terms of σ_{14}^2 and σ_{40}^2

$$\sigma_{14}^2 = \frac{40}{14} \sigma_{40}^2. \quad (6.35)$$

An estimation of single-item probe reliability relative to the 14-item questionnaire is required to fully define the triple of relative weights (w_{40} , w_{14} and w_1). This is addressed in the following subsections.

Measurement Error Model

For a measurement instrument j , we can model a calibrated trust observation of participant i at time t as:

$$\hat{q}_{i,t}^{(j)} = \tau_i(t) + \varepsilon_{i,t}^{(j)}, \quad (6.36)$$

where $\tau_i(t)$ is the participant's underlying (unobserved) trust state at time t , and $\varepsilon_{i,t}$ is an additive measurement error term.

The error term is assumed to satisfy:

$$\mathbb{E}[\varepsilon_{i,t}^{(j)}] = 0, \quad \text{Var}(\varepsilon_{i,t}^{(j)}) = \sigma_j^2, \quad (6.37)$$

where once again $j \in \{1, 14, 40\}$ indexes the measurement instrument used (40-item, 14-item, or single-item probe). Thus, all observations from the same instrument share a common error variance, while different instruments may exhibit different levels of noise.

Uncorrelated Cross-Instrument Measurement Errors

To derive relative reliability weights across trust measurement instruments, we adopt the assumption that, conditional on the underlying trust state at a given measurement occasion, the residual measurement errors of the single-item trust probe are approximately uncorrelated with those of the questionnaire-based instruments:

$$\text{Cov}(\varepsilon_{i,t}^{(40)}, \varepsilon_{i,t}^{(1)} | \tau_i(t)) \approx 0, \quad \text{Cov}(\varepsilon_{i,t}^{(14)}, \varepsilon_{i,t}^{(1)} | \tau_i(t)) \approx 0. \quad (6.38)$$

This assumption is intentionally weaker than full statistical independence and is sufficient for the variance relations employed in the calibration procedure.

The justification for this approximation lies in the fundamentally different response processes elicited by the two instrument classes. Questionnaire-based measurements (40-item and 14-item) require participants to evaluate multiple specific facets of human-robot interaction (e.g., reliability, predictability, team coordination) and to integrate these judgments across many items. As a result, their measurement error is dominated by item-level interpretation variability and is substantially reduced through averaging. In contrast, the single-item probe elicits a direct, momentary global assessment of trust, whose variability is dominated by transient contextual influences and response spontaneity rather than item-specific noise.

Because these dominant sources of measurement uncertainty differ qualitatively across instruments, systematic cross-instrument error covariance is not expected to persist once the underlying trust state is conditioned upon. In this sense, any remaining dependence between probe and questionnaire responses is treated as second-order relative to the instrument-specific noise mechanisms that govern their reliability. Modeling cross-instrument errors as approximately uncorrelated therefore provides a parsimonious and defensible first-order approximation that enables relative reliability estimation without introducing additional latent variables or covariance parameters.

Importantly, this assumption is applied only to probe-questionnaire pairs. No such assumption is made between the 40-item and 14-item questionnaires at anchor points, as the latter is computed from a subset of the same item responses and therefore shares structural measurement noise by construction.

Out-of-Sample Estimation and Bias Avoidance

To avoid optimistic bias, and given the expected moderate number of participants, variance estimation is performed using out-of-sample procedures at the participant level. Specifically, calibration mappings between instruments are estimated using all participants except one, and residuals are then computed for the held-out participant. Repeating this process across all participants yields a pooled set of residuals that reflects genuine out-of-sample error, while leveraging the largest possible number of participants for measurement calibration.

This procedure ensures that variance estimates capture the generalizability of each instrument rather than overfitting idiosyncratic participant responses. Importantly, these variance estimates are computed independently of trust dynamics model identification and are therefore not influenced by model structure or parameter choices.

Effective Prediction-error Variance of the Single Trust Probe

For single trust probe measurements, we have two distinct types of available differences. At anchor times ($t \in \{\text{pre}, \text{post}\}$), we can directly compare calibrated single trust probes to 40-item trust scores, while at mid-block times ($t \in \{\text{mid1}, \text{mid2}\}$), we can compare calibrated single trust probes to 14-item trust scores. We can define these differences as measurement residuals so that

$$e_{i,t}^{\text{anchor}} = \hat{q}_{i,t}^{(40)} - \hat{q}_{i,t}^{(1)}, \quad (6.39)$$

$$e_{i,t}^{\text{mid}} = \hat{q}_{i,t}^{(14)} - \hat{q}_{i,t}^{(1)}. \quad (6.40)$$

Because we assume that all instruments are measuring the same unobservable trust τ , these components cancel out in the residual formulations so that

$$e_{i,t}^{\text{anchor}} = \varepsilon_{i,t}^{(40)} - \varepsilon_{i,t}^{(1)}, \quad (6.41)$$

$$e_{i,t}^{\text{mid}} = \varepsilon_{i,t}^{(14)} - \varepsilon_{i,t}^{(1)}. \quad (6.42)$$

We can collect out-of-sample residuals from leave-one-participant-out (LOPO) cross-validation into the following sets:

$$\mathcal{E}_{\text{anchor}} = \{e_{i,\text{pre}}^{\text{anchor}}, e_{i,\text{post}}^{\text{anchor}}\}_{i=1}^N, \quad (6.43)$$

$$\mathcal{E}_{\text{mid}} = \{e_{i,\text{mid1}}^{\text{mid}}, e_{i,\text{mid2}}^{\text{mid}}\}_{i=1}^N \quad (6.44)$$

The variances of the residual differences can therefore be estimated directly from the empirical variances of the LOPO residuals:

$$\widehat{\text{Var}}(e^{\text{anchor}}) = \frac{1}{|\mathcal{E}_{\text{anchor}}| - 1} \sum_{e \in \mathcal{E}_{\text{anchor}}} (e - \bar{e})^2, \quad (6.45)$$

$$\widehat{\text{Var}}(e^{\text{mid}}) = \frac{1}{|\mathcal{E}_{\text{mid}}| - 1} \sum_{e \in \mathcal{E}_{\text{mid}}} (e - \bar{e})^2. \quad (6.46)$$

Finally, due to the assumption of uncorrelated measurement errors between questionnaires and single-item trust probes, we can disregard covariance terms and define further the variances of these residuals as

$$\text{Var}(e^{\text{anchor}}) = \text{Var}(\varepsilon^{(40)} - \varepsilon^{(1)}) \approx \sigma_{40}^2 + \sigma_1^2, \quad (6.47)$$

$$\text{Var}(e^{\text{mid}}) = \text{Var}(\varepsilon^{(14)} - \varepsilon^{(1)}) \approx \sigma_{14}^2 + \sigma_1^2. \quad (6.48)$$

These quantities directly measure how uncertain the single-item trust probe is when used as a surrogate for the 40-item and 14-item calibrated references, respectively.

We are left with an undefined set of 3 unknowns (σ_1^2 , σ_{14}^2 , and σ_{40}^2) and 2 equations. We can return to our conservative assumption for the relative weighting of 14-item and 40-item questionnaires to fully define this problem.

$$\begin{cases} \text{Var}(e^{\text{anchor}}) = \sigma_{40}^2 + \sigma_1^2, \\ \text{Var}(e^{\text{mid}}) = \sigma_{14}^2 + \sigma_1^2, \\ \sigma_{14}^2 = \frac{40}{14} \sigma_{40}^2. \end{cases} \quad (6.49)$$

Using our weight definition, the solution of this set of equations yields a triple of relative weights (w_{40} , w_{14} and w_1), where

$$w_{40} = \frac{26}{14(\text{Var}(e^{\text{mid}}) - \text{Var}(e^{\text{anchor}}))}, \quad (6.50)$$

$$w_{14} = \frac{26}{40(\text{Var}(e^{\text{mid}}) - \text{Var}(e^{\text{anchor}}))}, \quad (6.51)$$

$$w_1 = \frac{26}{40\text{Var}(e^{\text{anchor}}) - 14\text{Var}(e^{\text{mid}})}. \quad (6.52)$$

Normalization of Reliability Weights and Choice of Reference Weight

In the weighted least-squares (WLS) objective used for model identification, only the relative magnitudes of the weights matter. Multiplying all weights by a common positive constant scales the objective uniformly but does not change its minimizer. Consequently, the absolute scale of the weights can be fixed by convention.

In this work, the 40-item questionnaire is selected as the reference instrument. Therefore, it is convenient to assign this instrument a unit weight,

$$w_{40} \equiv 1. \quad (6.53)$$

All remaining instrument weights are then normalized and interpreted as relative confidence levels with respect to the 40-item reference. The resulting normalized weights, as a function of the empirically estimated instrument variances, are:

$$w_{40} = 1, \quad (6.54)$$

$$w_{14} = \frac{14}{40}, \quad (6.55)$$

$$w_1 = \frac{14\text{Var}(e^{\text{mid}}) - 14\text{Var}(e^{\text{anchor}})}{40\text{Var}(e^{\text{anchor}}) - 14\text{Var}(e^{\text{mid}})}. \quad (6.56)$$

The resulting weighting scheme reflects three complementary principles:

- longer questionnaires yield more reliable trust estimates,
- shorter instruments contribute useful information but with higher uncertainty,
- high-frequency probes increase temporal resolution without overwhelming the identification process.

By explicitly modeling and estimating measurement uncertainty, the calibration framework enables principled integration of heterogeneous trust observations into a unified identification pipeline.

6.6.8. Measurement Calibration and Weighting Summary

The calibration and weighting procedure can be understood as a two-step attempt to (i) express all trust measurements on the same numerical scale and (ii) prevent noisier instruments from disproportionately influencing model identification.

First, each instrument is mapped onto the 40-item reference scale so that all observations can be compared directly to the simulated trust trajectory. Second, relative reliability weights are derived from the typical disagreement observed between instruments when they are measured at the same conceptual moment.

At anchor times, the difference between the calibrated probe and the 40-item questionnaire reflects how noisy the single-item probe is relative to a high-fidelity reference; at midpoints, the difference between the calibrated probe and the 14-item questionnaire provides an additional measure of probe variability when the full questionnaire is not available. Because the 14-item score is structurally related to the 40-item score at anchors (it is computed from a subset of the same responses), its relative weight is defined conservatively based on item count rather than inferred from anchor differences.

Together, these steps yield a transparent set of relative weights that emphasize longer questionnaires when available, while still allowing shorter questionnaires and probes to contribute information without dominating the estimation process.

6.7. Participant Recruitment, Ethics, and Data Handling

This section documents the participant recruitment strategy, ethical approval process, and data handling procedures underlying the experimental study. These aspects are reported to ensure transparency, reproducibility, and compliance with institutional research standards, and to clarify the conditions under which the empirical data used for trust model identification were collected. Particular emphasis is placed on informed consent, participant autonomy, and data anonymity, as these considerations directly constrain experimental design choices and post-hoc data processing.

6.7.1. Participant Recruitment and Eligibility

Participants were recruited through personal and professional networks. They were provided with only general information about the experiment prior to participation and were not informed about the specific research hypotheses or modeling objectives, in order to minimize expectancy effects and response bias. Participation was voluntary, and no financial or material compensation was offered.

Eligibility was restricted to individuals with sufficient proficiency in English language to ensure correct comprehension of task instructions and accurate completion of the trust questionnaires administered during the experiment. The total duration of the experiment was communicated in advance and was estimated at approximately 40 minutes.

6.7.2. Ethical Approval and Consent Procedure

The experimental protocol was reviewed and approved by the TU Delft Human Research Ethics Committee (HREC). The study involved no deception and posed no physical risk to participants.

Informed consent was obtained through a formal Opening Statement, following TU Delft ethical guidelines. This Opening Statement constituted the first screen displayed when participants accessed the experiment URL. It provided a concise explanation of the task context, the nature of participation, data handling practices, and participants' rights. The statement concluded with an explicit "*I consent*" button. Only after providing consent did participants enter fullscreen mode, and only then did experimental data logging commence.

For transparency and reproducibility, the full text of the Opening Statement is provided in Appendix B.

6.7.3. Data Handling, Anonymity, and Participant Rights

All participants were assigned fully anonymous random identifiers at the start of each experiment. No personally identifying information was collected at any stage of the experiment. Experimental responses were stored in anonymized form using a cloud-based logging infrastructure linked directly to the web-based experiment.

Participants were explicitly informed that they could withdraw from the experiment at any point during task execution via a persistent on-screen control. Upon withdrawal, they were presented with three options: (i) withdraw and delete all data collected up to that point, (ii) withdraw while allowing previously collected data to be retained, or (iii) resume the experiment. These options ensured that participants retained agency over their participation and data during the experiment.

Due to the fully anonymous nature of the collected data, participants were also informed that it would not be possible to identify or delete their individual data after completion of the experiment. This limitation was clearly communicated as a direct consequence of the anonymity safeguards in place.

Together, these procedures ensure compliance with institutional ethical standards while preserving participant autonomy, data privacy, and the integrity of the collected dataset.

6.8. Synthesis and Implications for Research Question 2

This section synthesizes the contributions of chapter 6 and explicitly addresses the second research question posed in this thesis, which concerns the identification of the parameters of the proposed trust dynamics model from empirical human-robot interaction data.

Research Question 2: *How can the parameters of the proposed trust dynamics model be identified from empirical human-robot interaction data?*

- RQ2.a** How can a controlled human-subject experiment be designed specifically to elicit trust dynamics that are informative for parameter identification, while remaining consistent with established findings in the human-robot trust literature?
- RQ2.b** How can noisy, heterogeneous trust measurements (e.g., self-reports and behavioral indicators) be incorporated into a principled estimation framework for identifying latent trust dynamics?

Identification-oriented experimental design (RQ2.a). This chapter demonstrates how a controlled human-subject experiment can be purpose-designed to elicit trust dynamics that are informative for parameter identification. Rather than relying on randomized task outcomes, the experiment employs predefined trial sets encoding characteristic temporal patterns of drone reliability and environmental risk, such as early failure, gradual stabilization, volatility, and recovery. These patterns are grounded in well-established empirical findings on trust formation and degradation, ensuring theoretical consistency with the trust literature while deliberately exciting different trust dynamics modes. By prescribing outcome sequences and tailoring trust probe placement to expected regime changes, the design maximizes information content per participant, enabling identification-relevant trust trajectories to be collected within realistic constraints on experiment duration and sample size.

Integration of heterogeneous trust measurements (RQ2.b). The chapter further establishes how heterogeneous trust measurements can be incorporated into a principled identification framework. Trust is measured using a combination of validated multi-item questionnaires, repeated single-item trust probes, and behavioral indicators such as reliance decisions and response timing. These measures are temporally aligned with interaction events and environmental conditions, yielding time-resolved observations suitable for dynamic model estimation. Explicit data validation criteria, preprocessing steps, and inclusion rules are defined to manage noise, missing responses, and inter-individual variability. Together, these procedures transform heterogeneous, imperfect observations into an identification-ready dataset that supports estimation of latent trust dynamics without conflating measurement noise with model structure.

Taken together, the experimental design, measurement strategy, and data handling procedures presented in this chapter provide a complete and principled answer to Research Question 2. The resulting dataset is explicitly constructed to support parameter identification of the proposed trust dynamics model, forming a necessary bridge between the formal model definition in chapter 5 and the identification and validation procedures developed in chapter 7.

Identification of the Proposed Trust Dynamics Model

This chapter constitutes the empirical core of the thesis and addresses Research Question 3, which concerns the empirical identifiability and validation of the proposed trust dynamics model. Building on the formal model definition presented in chapter 5 and the experimental design and measurement framework described in chapter 6, the chapter evaluates whether the model can be meaningfully identified from human-robot interaction data and to what extent its parameters and predictions generalize to unseen participants.

To this end, the chapter is structured as a sequential identification pipeline. It begins by analyzing the collected dataset to verify structural consistency, empirical plausibility, and suitability for parameter estimation, including validation of participant trajectories, measurement completeness, and balance across experimental conditions. The chapter then presents the identification methodology, followed by parameter estimation results, out-of-sample validation, robustness analyses, and contextual comparison with representative baseline trust models from the literature. The chapter concludes with a focused discussion that synthesizes these findings and provides an explicit answer to the third research question.

7.1. Overview of the Identification Pipeline and Research Question

The primary purpose of this chapter is to assess to what extent the proposed trust dynamics model can be empirically identified and validated using the data collected in the online human-robot interaction experiment. In terms of the thesis structure, this chapter directly addresses Research Question 3 and constitutes the transition from model and experimental design to empirical evaluation.

The analysis presented here assumes the trust model structure defined in chapter 5 and the experimental protocol and measurement instruments described in chapter 6. The focus of this chapter is therefore not on theoretical model development or experimental design decisions, but on empirical inference: determining whether the available data supports reliable estimation of the model parameters, whether the inferred trust trajectories reproduce observed measurements, and how robust these conclusions are under different evaluation conditions.

This chapter does not address behavioral coupling or closed-loop interaction effects, which are treated separately in chapter 8, nor does it aim to establish general psychological claims about human trust. Instead, its scope is deliberately restricted to evaluating the empirical viability and limitations of the proposed trust dynamics model within the defined modeling and identification framework.

7.1.1. Trust Model Identification Problem Formulation

From a system identification perspective, the trust modeling problem addressed in this chapter can be framed as the estimation of a latent, time-varying internal state from indirect and noisy observations. Trust is modeled as an unobserved dynamic state whose evolution is governed by the trust dynamics model defined in chapter 5.

The inputs to this system consist of exogenous interaction events and contextual variables generated during the experimental task, including robot performance outcomes, task-specific events, and pre-interaction information such as reputation and baseline expectations. These inputs drive the temporal evolution of the latent trust state according to a shared set of model parameters.

Observations of trust are obtained through a combination of intermediate trust probes and questionnaire-based measurements administered at discrete time points. These measurements are treated as noisy, partial observations of the underlying trust state rather than as direct ground-truth labels. As a result, the identification problem involves jointly accounting for trust dynamics and measurement uncertainty.

The unknown quantities in this formulation include the global model parameters governing trust evolution, as well as participant-specific initial conditions or offsets where applicable. The identification objective is to estimate these parameters from a subset of participant trajectories and to evaluate the resulting model on held-out participants, thereby assessing both parameter identifiability and predictive generalization.

7.1.2. Identification Pipeline Overview

To address this identification problem, the chapter adopts a structured, end-to-end identification pipeline that explicitly separates data readiness, parameter estimation, and empirical evaluation. This pipeline is designed to ensure that model identification and validation are grounded in empirically plausible data and that the resulting conclusions are interpretable and reproducible.

At a high level, the pipeline consists of the following stages: (i) dataset overview and structural validation, (ii) exploratory characterization of trust measurements and experimental balance, (iii) participant-level partitioning into training and validation subsets, (iv) parameter identification on the training set, (v) out-of-sample validation on held-out participants, (vi) robustness and sensitivity analysis, and (vii) contextual comparison with representative baseline trust models from the literature.

A schematic overview of this pipeline is provided in Figure 7.1, which illustrates the flow from raw experimental data to validated identification results. This structure motivates the organization of the chapter and explains why dataset validation and exploratory analyses precede formal model fitting.

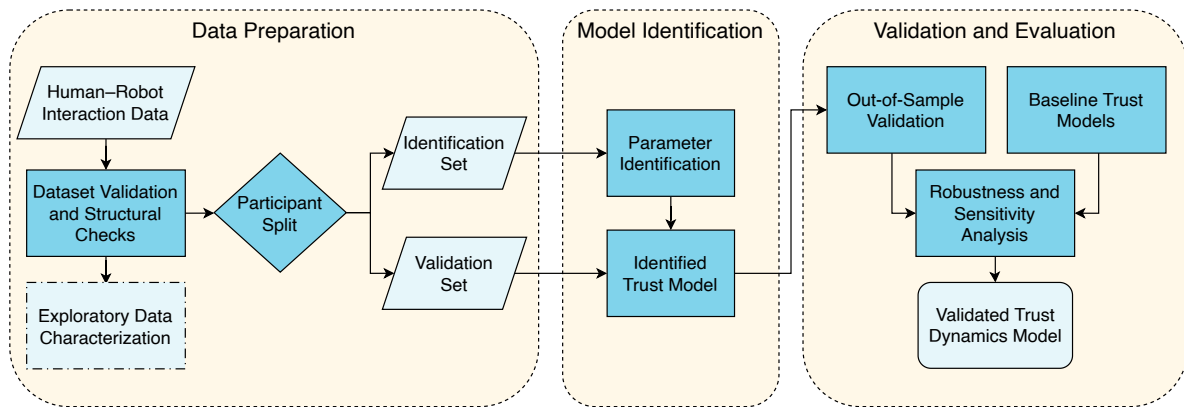


Figure 7.1: End-to-end identification and evaluation pipeline for the proposed trust dynamics model. Experimental data are first validated and characterized before participant-level partitioning into identification and validation subsets. Model parameters are estimated using the identification set to produce an identified trust model, which is then evaluated out-of-sample. Robustness, sensitivity, and comparative analyses against baseline trust models support the empirical validation of the final trust dynamics formulation.

7.1.3. Dataset Validation as a Precondition for Identification

Before attempting to identify the parameters of the proposed trust dynamics model, it is essential to establish that the collected dataset is suitable for empirical inference. For this reason, the first part of this chapter is devoted to dataset validation and exploratory characterization rather than to model fitting.

These analyses serve several purposes. First, they verify the structural consistency and completeness of participant trajectories, ensuring that required task events and measurement instruments are present and correctly aligned in time. Second, they assess whether observed trust measurements exhibit empirically plausible patterns, such as variability across participants and evolution over the course of the interaction. Third, they examine balance across experimental conditions and participant attributes to identify potential sources of confounding.

Importantly, the analyses presented in Sections 7.2–7.6 do not constitute tests of the proposed trust dynamics model. Instead, they establish the empirical substrate on which identification is performed and provide the necessary context for interpreting the identification and validation results reported later in the chapter.

7.1.4. Evaluation Criteria and Success Conditions

The empirical evaluation of the trust dynamics model in this chapter is guided by a set of explicit criteria aligned with Research Question 3. These criteria are intended to assess not only goodness of fit, but also identifiability, generalization, and robustness. It is important to restate the challenges inherent in evaluating trust dynamics models. Unlike physical system states, trust is not directly observable and must be inferred from noisy, heterogeneous measurements such as self-reports and behavioral indicators. As a result, validation cannot rely on direct state comparison, but instead must be framed in terms of trajectory-level agreement, parameter identifiability, robustness, and consistency with established empirical trust phenomena. These considerations define the evaluation criteria used throughout this chapter.

First, the model is evaluated on its ability to reproduce observed trust trajectories over time, as reflected in intermediate trust probes and questionnaire measurements. Second, generalization is assessed through out-of-sample evaluation on participants not used during parameter identification, with the goal of determining whether inferred parameters capture shared trust dynamics rather than individual-specific noise. Third, parameter identifiability is examined by analyzing estimation uncertainty, parameter stability across runs, and the presence of weakly identifiable or degenerate parameters. Fourth, robustness is assessed via sensitivity analyses with respect to modeling assumptions, measurement subsets, and preprocessing choices. Finally, the behavior of the proposed model is contextualized through comparison with representative baseline trust models from the literature. These comparisons are used to situate the empirical performance of the model and to aid interpretation.

7.2. Dataset Overview and Validation

This section summarizes the outcomes of the data collection process for the online SaR trust experiment introduced in chapter 6 and defines the resulting datasets used in the remainder of this thesis. We first provide an overview of the raw event logs collected from all recorded sessions and the corresponding measurement instruments. We then report the results of the validation and participant screening procedure that yields the cleaned analytic participant set used for model identification and validation in subsequent sections. We present the dataset at a conceptual level (i.e., as instruments, events, and participant-level records), while implementation details of the data preprocessing and analysis code are documented separately in the appendices.

7.2.1. Raw Dataset Overview

The experiment produced a raw event log in CSV format in which each row corresponds to a timestamped experimental event associated with a participant session. After normalization of the raw log, the dataset contains 10 distinct event types, corresponding to demographic information, repeated task trials, trust measurements, and questionnaire instruments. A total of 50 participant sessions were recorded (as indicated by the number of demographics events), yielding 2455 logged door trial events across all participants.

Table 7.1 summarizes the raw counts per event type. The door trial events represent the repeated decision task trials, while trust probes specifically correspond to repeated single-item trust probes administered during the task. Longer-form trust questionnaires were administered before (pre) and after (post) the interaction, and condensed questionnaires were administered at intermediate time points (mid1 and mid2). In addition, the dataset contains reputation item and reputation probe events related to the pre-interaction reputation/expectation elicitation, and an emergency trial event capturing the participant's choice in the emergency scenario.

Table 7.1: Raw event counts by event type in the collected log.

Event type	Count
Demographics	50
Door trial	2455
Emergency trial	45
Mid1 14-item questionnaire	43
Mid2 14-item questionnaire	40
Pre 40-item questionnaire	45
Post 40-item questionnaire	38
Reputation item	80
Reputation probe	35
Trust probe	530

The raw counts in Table 7.1 already indicate that not all instruments were completed by all recorded participants. For instance, while all sessions include demographic information ($N_{\text{raw}} = 50$), only 45 sessions contain a recorded emergency trial response, and only 38 sessions contain the post-interaction 40-item questionnaire. These inconsistencies motivate the validation and inclusion/exclusion procedure described in a later subsection, whose goal is to produce a cleaned participant set in which all required instruments and trial sequences are present and internally consistent.

7.2.2. Measurement Calibration Outputs

As described in Chapter 6, the heterogeneous trust measurements collected during the experiment were calibrated to enable principled aggregation, comparison, and use in trust dynamics identification. This calibration yields two sets of outputs that are used throughout the remainder of this chapter: (i) relative weights assigned to different trust measurement instruments, and (ii) global linear mappings that express single-item and 14-item questionnaire scores on the same numerical scale as the 40-item trust questionnaire.

The final calibrated weights assigned to each measurement instrument are:

$$w_{40} = 1, \quad w_{14} = 0.35, \quad w_1 = 0.00764. \quad (7.1)$$

The 40-item questionnaire is defined as the reference instrument and is therefore assigned unit weight. Intermediate 14-item questionnaires receive a lower but still substantial weight, following the conservative proportional rule $w_{14} = 14/40 w_{40}$, which reflects their reduced length and information content relative to the full questionnaire. In contrast, repeated single-item trust probes are assigned a much smaller weight, consistent with their substantially higher noise, contextual sensitivity, and lower reliability when considered individually.

These weights encode an explicit measurement hierarchy: questionnaire-based trust assessments anchor the latent trust state, while single-item probes contribute information primarily in aggregate rather than as precise point estimates. All weights are held fixed for model identification, validation, and baseline comparisons, ensuring that subsequent results reflect trust dynamics modeling rather than implicit re-weighting of measurements.

To ensure numerical comparability across instruments, single-item probe scores and 14-item questionnaire scores are linearly mapped to the 40-item questionnaire scale using global ordinary least-squares fits computed during calibration. The resulting mappings are:

$$\hat{Q}_{\text{probe}}^{(40)} = a_1 Q^{(1)} + b_1, \quad \hat{Q}_{\text{mid1,mid2}}^{(40)} = a_{14} Q^{(14)} + b_{14}, \quad (7.2)$$

with calibrated coefficients:

$$a_1 = 0.4874, \quad b_1 = 31.9276, \quad a_{14} = 0.8250, \quad b_{14} = 5.9654. \quad (7.3)$$

These coefficients indicate that both shorter instruments systematically compress the dynamic range of trust scores relative to the 40-item questionnaire and require affine correction to align their means and variances. The mappings are applied uniformly across all participants and time points and are not re-estimated during trust dynamics identification.

Together, the calibrated weights and score mappings define an explicit measurement model that underpins all identification and validation results reported in this chapter. While alternative calibration choices are possible, the values reported here represent a principled compromise between instrument reliability, interpretability, and data efficiency. Importantly, these calibration outputs are treated as fixed inputs to the trust dynamics identification pipeline, ensuring a clear separation between measurement modeling and dynamic model estimation. Their explicit reporting here clarifies the assumptions under which all subsequent quantitative and qualitative results should be interpreted.

7.2.3. Dataset Validation and Inclusion Criteria

Applying the validation and inclusion criteria described in Section 6.5 resulted in the exclusion of 15 participant sessions from the raw dataset, yielding a final cleaned analytic dataset of 35 participants. These exclusions reflect protocol non-compliance or incomplete instrumentation in the raw logs, rather than any post-hoc selection based on behavioral outcomes or model fit. In the cleaned set, all retained participants contain complete and consistent door-trial task trajectories, the full trust probe schedule, and all questionnaire instruments required by the measurement calibration and trust dynamics identification pipeline.

The excluded sessions exhibited one or more violations of the hard inclusion criteria. In the raw validation report, the most common failure modes included: (i) absence of door-trial task data (i.e., no recorded door trials), (ii) incomplete trust probe schedules (e.g., fewer than the expected 13 intermediate probes), (iii) missing questionnaire instruments (most frequently the post-interaction 40-item questionnaire), and (iv) structural inconsistencies in the reconstructed block-level task sequence and probe alignment.¹

After applying all hard inclusion criteria, the validated analytic dataset ($N = 35$) was examined for soft validation flags related to demographics completeness and emergency-trial responses. No participants in the cleaned sample exhibited missing demographic fields or invalid emergency-trial responses. As a result, all validated participants contribute fully to both descriptive characterization and auxiliary behavioral analyses.

The remaining 35 participants constitute the validated analytic dataset used throughout the subsequent chapters. By construction, every participant in this set contains:

- a complete door-trial task trajectory with valid `suggestion` and `choice` fields for all trials,
- a complete intermediate trust probe schedule with 13 valid probe values in $[0, 100]$,
- all questionnaire instruments required by the measurement calibration and trust modeling pipeline with valid numeric fields, and
- verified alignment between `after_door` trust probes and existing door events in the participant timeline.

As a consequence, the cleaned dataset exhibits a uniform experimental structure across participants: differences between participants arise from assigned experimental condition sequences, individual differences captured by baseline measures (e.g., pre-interaction questionnaire and review expectations), and behavioral variability during the task, rather than from missingness or structural inconsistencies.

Table 7.2 reports an aggregated breakdown of exclusion reasons across the discarded sessions. Each excluded participant may contribute to multiple categories (e.g., a session may simultaneously lack door trials and have missing questionnaires). This breakdown provides a concise accounting of the dominant sources of data loss in the raw collection and motivates the use of the defined strict inclusion criteria to ensure an analyzable dataset.

Table 7.2: Aggregated exclusion breakdown across the discarded participant sessions. Counts reflect the number of excluded sessions exhibiting each issue; categories are not mutually exclusive.

Issue category	Count
Missing door trials	6
Incorrect number of trust probes	15
Missing pre 40-item questionnaire	5
Missing post 40-item questionnaire	12
Probe-to-door alignment failure	2

7.2.4. Validated Sample Characterization

This subsection characterizes the final validated analytic sample of participants retained after preprocessing and validation. The objective is twofold: (i) to provide transparency regarding the composition of the participant pool used for model identification and validation, and (ii) to assess whether the cleaned dataset exhibits major imbalances across experimental conditions or participant attributes that could bias subsequent analyses. All results reported in this subsection are computed exclusively on the cleaned dataset described in Section 7.2.3.

A total of 35 participants completed the study, with the majority using desktop devices (94%). Participants were distributed across eight door-trial sequence sets with comparable baseline trust levels. No participants exhibited extremely low initial trust (below 40%), and baseline trust distributions were broadly similar across sets. Figure 7.2 and Table 7.3 summarize the age-range and self-reported gender distributions in the validated sample, respectively. Age was collected in categorical ranges rather than as a continuous variable, and gender responses include both fixed-choice and self-describe options, as permitted by the experimental interface.

¹ The emergency trial response was treated as a soft validation criterion in the default preprocessing configuration. Nevertheless, several excluded sessions also exhibited missing or invalid emergency choices, typically coinciding with broader session incompleteness.

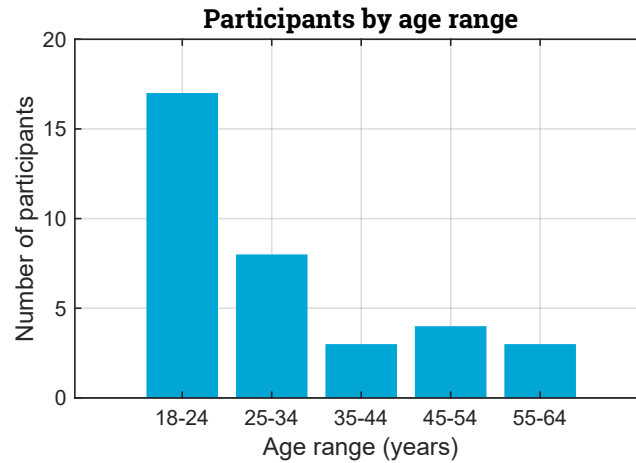


Figure 7.2: Age-range distribution of the validated analytic sample.

In the emergency decision task, 80% of participants opted for self-reliance rather than robot assistance, indicating a conservative baseline reliance tendency. This split is reported in Table 7.4.

Table 7.3: Gender distribution of the validated analytic sample.

Gender	Count
Man	22
Woman	10
Gender-fluid	1
Non-binary	1
Prefer not to say	1

Table 7.4: Participant emergency decision choice.

Emergency choice	Count
Self-reliance	28
Robot-assisted	7

No demographic filtering was applied during preprocessing; therefore, these distributions reflect the natural composition of the recruited sample after protocol compliance filtering.

Participants were randomly assigned to one of the eight predefined door-trial sequence sets. Figure 7.3 graphically shows the distribution of validated participants per set.

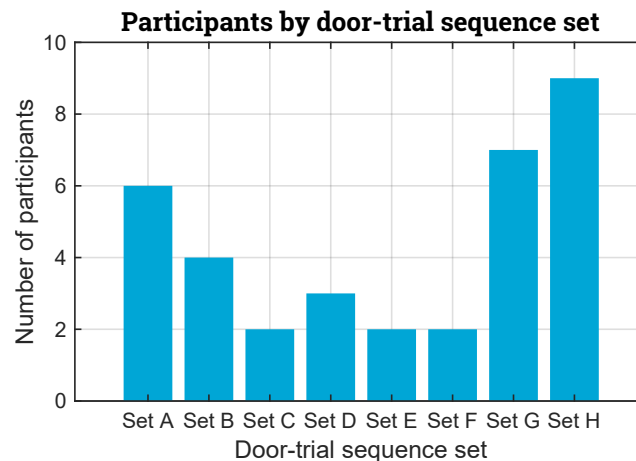


Figure 7.3: Distribution of validated participants across door-trial sequence sets.

While perfect balance across sets was not enforced at recruitment time, the resulting distribution shows that all experimental sets are represented in the cleaned dataset, with at least 2 participants per set condition. Set-level balance is further examined below through cross-tabulations with participant attributes and baseline trust measures.

Before interacting with the task, participants provided trust-related measures via (i) a reputation/review expectation item and (ii) a 40-item pre-interaction trust questionnaire. Figure 7.4 shows the distribution of expected review scores, while Figure 7.5 contrasts expected versus reported review outcomes.

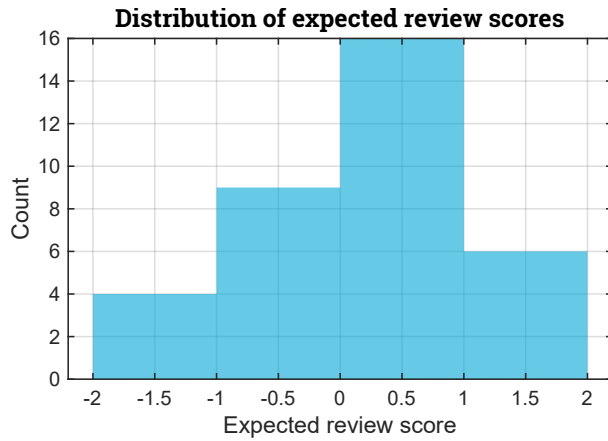


Figure 7.4: Distribution of expected review scores prior to task interaction.

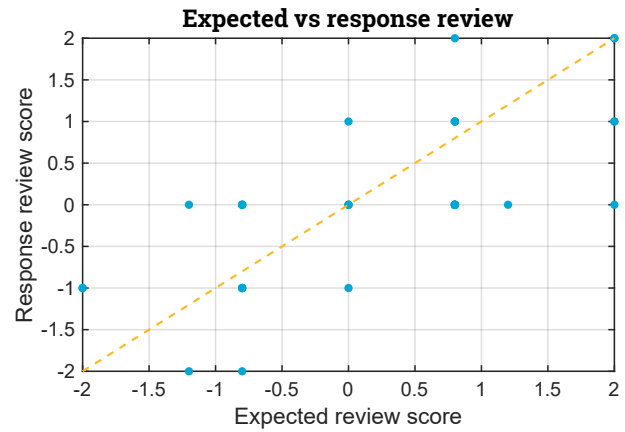


Figure 7.5: Expected versus response review scores across validated participants.

Figure 7.6 summarizes the distribution of pre-interaction questionnaire scores.

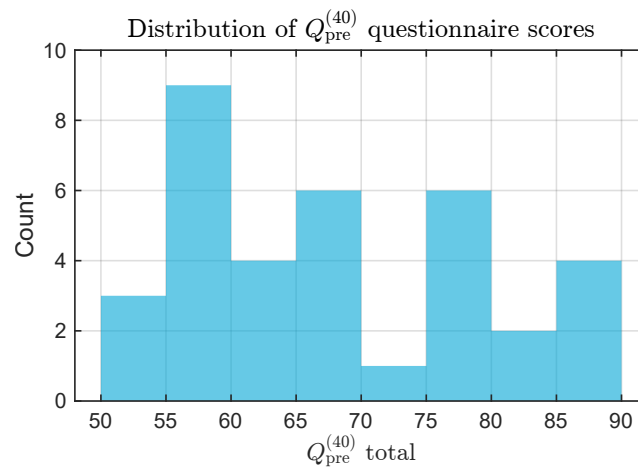


Figure 7.6: Distribution of pre-interaction trust questionnaire scores.

Of particular notice is the fact that pre-interaction 40-item questionnaire scores are all above 50%, indicating that, without much available information on the drone, participants generally tend to have a positive disposition to trust it. Moreover, these baseline measures reveal substantial inter-individual variability prior to task exposure, which motivates the inclusion of initial trust and reputation-related terms in the proposed trust dynamics model.

7.2.5. Exploratory Analysis of Reputation Effects

This subsection reports the empirical distribution of review-based reputation conditions in the validated analytic sample and the observed participant responses to the reputation prime. The metrics defined in Section 6.4.2 are applied here to evaluate whether participant responses exhibit evidence of negativity bias.

Across the validated analytic sample ($N = 35$), the system-assigned expected reputation took values in

$$v \in \{-2, -1.2, -0.8, 0, +0.8, +1.2, +2\}, \quad (7.4)$$

with the following counts per condition:

$$\begin{aligned} v = -2.0 : & \quad 2 \\ v = -1.2 : & \quad 2 \\ v = -0.8 : & \quad 9 \\ v = 0 : & \quad 5 \\ v = +0.8 : & \quad 11 \\ v = +1.2 : & \quad 1 \\ v = +2.0 : & \quad 5 \end{aligned} \quad (7.5)$$

The mean assigned expected reputation across participants was

$$\bar{v} = +0.183, \quad (7.6)$$

indicating a slightly positive distribution of review tones.

Participant responses were encoded on the five-level scale $s \in \{-2, -1, 0, +1, +2\}$. The mean response across validated participants was

$$\bar{s} = +0.029. \quad (7.7)$$

Figure 7.5 contrasts expected review values and participant responses at the individual level.

Applying the diagnostics defined in Section 6.4.2, the mean absolute response magnitudes were:

$$\mu_+ = 0.647, \quad \mu_- = 0.769, \quad (7.8)$$

yielding a raw negativity-bias ratio

$$\eta_{\text{raw}} = 1.189. \quad (7.9)$$

The normalized influence analysis yielded:

$$\bar{I}_+ = 0.544, \quad \bar{I}_- = 0.782, \quad \eta_{\text{norm}} = 1.437. \quad (7.10)$$

For the neutral review condition ($v = 0$), the mean response and mean absolute response were:

$$\bar{s}_0 = 0, \quad \overline{|s|}_0 = 0.4. \quad (7.11)$$

Overall, these results indicate that possible negativity bias effects on expected reputation values are reduced. This is supported by the relatively similar influence analysis for both positive and negative expected review conditions, as well as by the additional mean response of 0 for neutral review conditions (that contained equal number of positively and negatively skewed individual reviews). Even though some level of imbalance was found, it is not considered strong enough to support claims of a bias shift on participants' perceptions of the shown reviews. Consequently, the reputation initialization used in the trust dynamics model remains deterministically tied to the system-assigned expected reputation value v_i , without incorporating participant response values. This choice preserves interpretability and avoids introducing additional variability into model identification.

7.2.6. Questionnaire Evolution During the Task

Intermediate and post-interaction questionnaires were collected to capture trust evolution over the course of the interaction. Figures 7.7, 7.8 and 7.9 illustrate the distributions of these questionnaire scores among all validated participants.

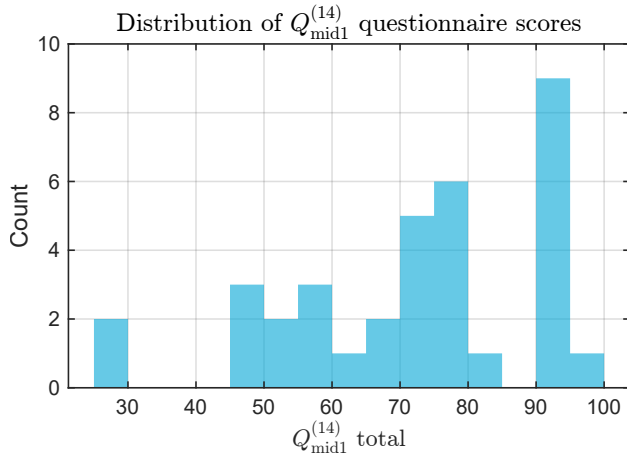


Figure 7.7: First intermediate trust questionnaire scores distribution.

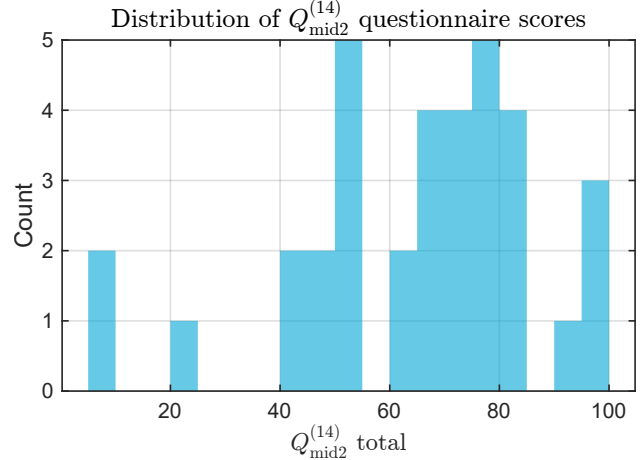


Figure 7.8: Second intermediate trust questionnaire scores distribution.

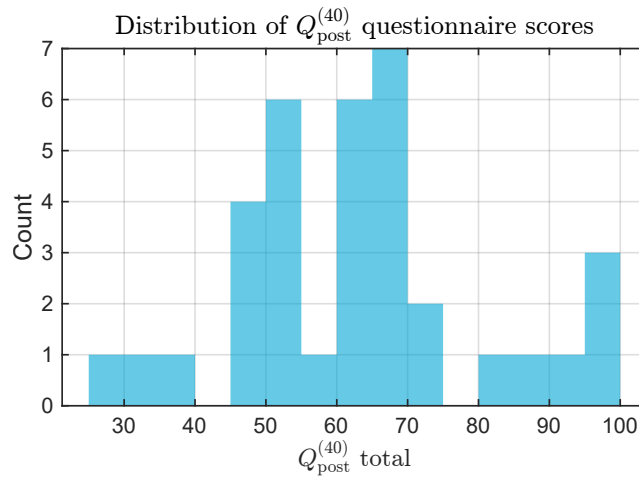


Figure 7.9: Post-interaction trust questionnaire scores distribution.

These metrics show that mid-block and post door trials questionnaires have a much wider range of recorded values than the previously reported pre-interaction questionnaire scores (Figure 7.6). This is strong evidence pointing to the fact that the experimental setup was successful in inducing relevant shifts in participants' trust levels during the interaction task. Additionally, we see that all scores reported in Figures 7.7, 7.8 and 7.9 peak at around 70%, once again indicating a general tendency from participants to trust the robot.

7.2.7. Balance Across Experimental Sets

To assess whether participant attributes and baseline trust measures are unevenly distributed across experimental condition sets, a series of balance matrices were computed. These cross-tabulate experimental set assignment against age range, gender, binned pre-interaction trust scores, expected review scores, and emergency-choice outcomes.

For brevity purposes, only 2 of these checks are reported on in this section, namely experimental set assignment against binned pre-interaction trust scores, in Figure 7.10, and expected review scores, in Figure 7.11.

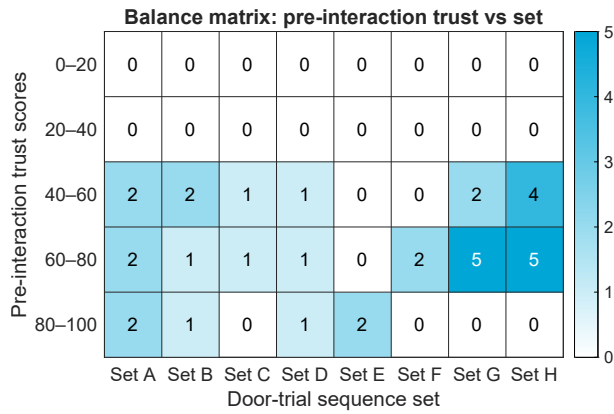


Figure 7.10: Balance of participant binned pre-interaction trust scores across experimental sets.

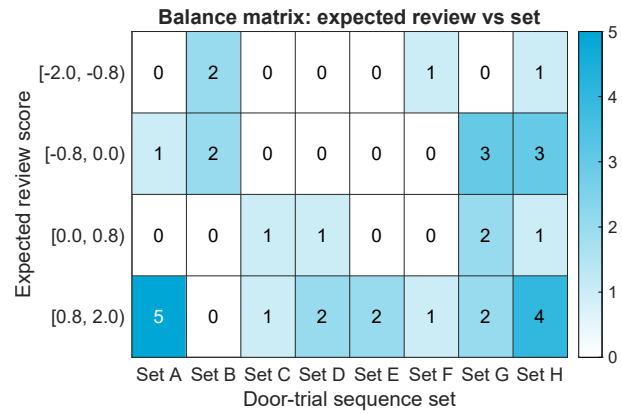


Figure 7.11: Balance of participant expected review scores across experimental sets.

Overall, the validated analytic sample exhibits representation across all experimental condition sets and a broad range of demographic and baseline trust profiles. While minor imbalances are inevitable given the sample size, no single set is dominated by a narrow subset of participant characteristics, supporting the use of the cleaned dataset for subsequent trust dynamics identification and validation.

7.2.8. Training and Validation Partitioning

To enable model identification and out-of-sample evaluation, the validated analytic dataset was further partitioned into independent training and validation subsets. This split was performed after all preprocessing, validation, and time-enrichment steps, ensuring that both subsets consist exclusively of complete and internally consistent participant trajectories.

From the 35 validated participants, an approximately 70/30 split was used for training and validation, respectively. Therefore, 25 participants were assigned to the training set and 10 were reserved as a held-out validation set. The split was defined explicitly at the participant level by index and applied deterministically, without random shuffling. This approach ensures full reproducibility of all reported results and allows the same participants to be reused consistently across different model variants and identification experiments.

The training set is used exclusively for parameter identification and calibration of the proposed trust dynamics model presented in chapter 5, while the validation set is used only for out-of-sample evaluation, diagnostic analysis, and qualitative inspection of model behavior. No participant appears in both sets.

Given the moderate sample size and the relatively high dimensionality of the proposed trust dynamics model, a participant-level split was chosen over trial-level or time-slice-based splits. This prevents information leakage across sets and ensures that validation results reflect generalization to truly unseen individuals, rather than interpolation within a participant's own trajectory. The manual selection of participants was done to ensure that both training and validation sets had each condition set and as many review conditions as possible reflected in their composition. This was done to ensure balance between both sets. Random selection of the participant split would risk biasing training results by over-representing some conditions in detriment of others that would not even be present, due to the reduced sample size. Moreover, this allows for the explicit testing of the trained model's ability to be generalized to held-out participants of every condition set.

After defining the training and validation subsets, balance checks analogous to those reported for the full validated sample were performed separately on each subset. Specifically, the following distributions were examined for both the training and validation sets:

- assignment to door-trial sequence sets (SetA–SetH),
- demographic attributes (age range and gender),
- baseline trust measures (pre-interaction 40-item questionnaire scores),
- expected review scores from the reputation prime, and
- emergency choice outcomes.

Of special relevance is the fact that the resulting train and validation sets each have all door-trial sequence sets represented. Moreover, the training has all 7 expected review scores present in its composition, while the validation set has 6 out of 7 represented. Priority was given to the training set for review scores completeness, in order to have as much variations as possible while fitting the model. Beyond these metrics, no systematic collapse of demographic or baseline trust categories into a single subset was observed. While perfect stratification was not possible, particularly for the smaller validation set, the resulting split preserves sufficient diversity in experimental conditions and participant attributes to support meaningful out-of-sample evaluation of the identified trust dynamics model.

All parameter identification, optimization, and model selection results reported in this thesis are computed exclusively on the training set, unless explicitly stated otherwise. Therefore, the validation metrics, trajectory comparisons, and diagnostic plots reported for the validation set reflect genuine generalization performance. The explicit definition and documentation of the split ensure that future work can reproduce, extend, or revise the identification pipeline using the same participant partitions.

7.3. Parameter Identification Problem Formulation for Trust Dynamics

This section formulates the parameter identification problem for the proposed trust dynamics model. Building on the model structure introduced in chapter 5, the objective is to estimate a set of global model parameters that best explain observed trust measurements across participants.

7.3.1. Model State, Inputs, Outputs, and Parameters

The trust dynamics model describes the evolution of a latent trust state

$$\tau(t) \in [0, 1], \quad (7.12)$$

which represents a participant's instantaneous level of trust in the robotic system. The trust state evolves continuously in time and is influenced by discrete task-related events and latent dynamics between events.

The model is driven by participant-specific input data, including discrete task events (e.g., door outcomes and associated risk levels), and initial dispositional and reputation trust estimates derived from pre door tasks questionnaires and probes.

The output of the model is a simulated trust trajectory that follows the rules described in chapter 5. This output is compared to observed trust measurements collected during the experiment, consisting of post-experiment questionnaires, mid-block questionnaires, and single-item trust probes. These measurements provide noisy observations of the latent trust state at discrete time instants.

The behavior of the trust dynamics model is governed by the previously defined global parameter vector

$$\theta = \left[\lambda^{\text{rep}} \quad \alpha^{\text{sit}} \quad \lambda^{\text{sit}} \quad \phi^{\text{fail}} \quad \phi^{\text{succ}} \quad a^{\text{succ}} \quad \lambda^{\text{lat}} \quad \kappa^{\text{lat}} \right]^{\top}, \quad (7.13)$$

whose components control reputation decay, situational trust modulation, event-driven experience updates, and latent trust evolution. These parameters are assumed to be shared across all participants and constant over time.

7.3.2. Trust Simulation and Measurement Alignment

Given a candidate parameter vector θ and a participant's task data, the trust dynamics model is simulated forward in time to produce a continuous trust trajectory $\tau(t; \theta)$. The simulation is performed on a uniform internal time grid that captures the temporal ordering of task events and latent trust evolution.

Trust measurements occur at irregular and participant-specific times that do not generally coincide with the internal simulation grid. To enable comparison between model predictions and observed data, predicted trust values are obtained by sampling the simulated trust trajectory at the measurement times:

$$\hat{\tau}_{i,m}(\theta) = \tau(t_{i,m}; \theta), \quad (7.14)$$

where $t_{i,m}$ denotes the time of the m -th trust measurement for participant i .

With the exception of the pre-task 40-item questionnaire used for dispositional trust assessment, trust measurements are treated as passive observations: they do not alter the trust state and serve solely as read-outs of the latent trust trajectory.

7.3.3. Weighted Least-squares Identification Objective

Let $\hat{q}_{i,m} \in [0, 1]$ denote the observed trust value for participant i at measurement index m , and let $\hat{\tau}_{i,m}(\theta)$ be the corresponding model prediction. The residual associated with this measurement is defined as

$$e_{i,m}(\theta) = \hat{q}_{i,m} - \hat{\tau}_{i,m}(\theta). \quad (7.15)$$

Because trust measurements are obtained using instruments with different reliability and temporal resolution, each residual is assigned a positive weight $w_{i,m}$ determined by the measurement type. These weights reflect the relative confidence placed in each measurement modality.

The weighted least-squares cost for a single participant is defined as

$$J_i(\theta) = \sum_{m=1}^{N_m} w_{i,m} e_{i,m}(\theta)^2, \quad (7.16)$$

where N_m is the number of trust measurements available for participant i .

The global identification objective aggregates the participant-level costs across all participants:

$$J(\theta) = \sum_{i=1}^{N_p} J_i(\theta), \quad (7.17)$$

where N_p denotes the total number of participants used for identification. The parameter identification problem is thus formulated as the minimization problem

$$\theta^* = \arg \min_{\theta} J(\theta). \quad (7.18)$$

This formulation corresponds to a global weighted least-squares estimation problem in which a single parameter vector θ is fitted to explain trust measurements across all participants simultaneously.

7.3.4. Constraints and Modeling Assumptions

The identification problem is solved under a set of structural constraints and modeling assumptions. Model parameters are constrained to ranges that reflect their physical or psychological interpretation, such as non-negativity of decay rates and bounded magnitude of event-driven trust updates.

Dispositional trust is treated as an exogenous quantity derived directly from questionnaire data and is not estimated as part of θ . Several secondary shaping parameters governing latent trust behavior are fixed a priori based on model design considerations and are not included in the identification problem.

All parameters are assumed to be time-invariant and shared across participants. Inter-individual variability in trust trajectories arises solely from differences in participant data (event sequences, risk exposure, and measurement timing), not from participant-specific parameter values.

7.4. Summary of Experimental Data Used for Identification

This section summarizes the experimental data used to identify the parameters of the proposed trust dynamics model. The focus is on defining the nature, structure, and partitioning of the data entering the identification problem, without reference to implementation-specific details.

7.4.1. Data Sources and Unit of Analysis

The unit of analysis for model identification is the individual participant. For each participant, the available data consist of two primary components: a temporally ordered sequence of task-related events and a set of trust measurements collected at discrete time points during the experiment.

Task-related data include the sequence of interaction events experienced by the participant, such as decision points and associated outcomes, together with task-specific attributes (e.g., perceived risk levels). These events define the external inputs driving the trust dynamics model. Trust measurements provide observations of the participant's latent trust state at specific times and are used as the empirical basis for parameter identification.

All identification and validation procedures operate at the participant level. Model parameters are estimated globally across participants, while participant-specific variability arises from differences in event sequences and measurement timing.

7.4.2. Trust Measurement Modalities and Temporal Structure

Trust measurements are obtained using multiple instruments with different temporal resolutions and response formats. These include comprehensive post-experiment questionnaires, shorter mid-block questionnaires administered during the task, and single-item trust probes collected at specific moments.

Measurement times are irregular and participant-dependent, reflecting the experimental protocol and participant progression through the task. To enable joint use of all available trust measurements, values from different instruments are mapped onto a common normalized trust scale prior to model identification. This mapping ensures consistency across measurement modalities while preserving relative trust levels.

Trust measurements are treated as passive observations of the latent trust state. They do not influence the evolution of trust in the model and are used solely to evaluate the discrepancy between model predictions and observed data at their respective time points.

7.4.3. Training and Validation Participant Sets

To assess the generalization capability of the identified trust dynamics model, the available participant dataset is partitioned into separate training and validation sets. The split is performed at the participant level, such that all data associated with a given participant belong exclusively to either the training set or the validation set.

The training set is used to estimate model parameters and to derive any global data mappings required for identification. The validation set is held out entirely from the identification process and is used only for performance evaluation and robustness analysis. This separation prevents information leakage and enables an assessment of how well the identified model generalizes to previously unseen participants.

7.4.4. Data Inclusion and Consistency Assumptions

All participants entering the model identification stage satisfy a set of consistency requirements established during earlier preprocessing and experimental validation steps. As a result, the dataset used for identification is structurally homogeneous across participants.

Specifically, each participant included in the identification and validation sets provides the same types, ordering, and number of trust measurements, as well as a complete sequence of task-related events required to simulate the trust dynamics model. No additional data exclusion or imputation is performed at the identification stage.

The primary source of variability across participants lies in the timing of trust measurements and task events. Measurement times depend on individual differences in task progression speed and interaction behavior, leading to participant-specific temporal alignment between simulated trust trajectories and observed trust values. This variability is explicitly accounted for in the identification formulation by evaluating model predictions at participant-specific measurement times, as described in Section 7.3.

7.5. Model Validation and Evaluation Metrics

This section defines the evaluation criteria used to assess the quality of the identified trust dynamics model. The metrics quantify goodness of fit at measurement times, systematic prediction bias, and generalization performance across participant splits. All criteria introduced here are used consistently in Section 7.9.

7.5.1. Weighted Goodness-of-fit Metrics

Overall fit quality is quantified using weighted error norms computed over the residual set of a given split (training or validation). Let $\mathcal{J}$ denote the index set of all usable measurements in that split, and define e_n and w_n as the residual and weight for the n -th element in $\mathcal{J}$. The weighted root-mean-square error (RMSE) is

$$\text{wRMSE} \triangleq \sqrt{\frac{\sum_{n \in \mathcal{J}} w_n e_n^2}{\sum_{n \in \mathcal{J}} w_n}}, \quad (7.19)$$

and the weighted mean absolute error (wMAE) is

$$\text{wMAE} \triangleq \frac{\sum_{n \in \mathcal{J}} w_n |e_n|}{\sum_{n \in \mathcal{J}} w_n}. \quad (7.20)$$

These metrics are used as primary goodness-of-fit indicators, with emphasis on validation-set values to assess generalization.

7.5.2. Bias Metrics

Systematic over or under-prediction is assessed using both unweighted and weighted bias measures. The (unweighted) bias is defined as

$$\text{bias} \triangleq \frac{1}{|\mathcal{J}|} \sum_{n \in \mathcal{J}} e_n, \quad (7.21)$$

and the weighted bias (wBias) is defined as

$$\text{wBias} \triangleq \frac{\sum_{n \in \mathcal{J}} w_n e_n}{\sum_{n \in \mathcal{J}} w_n}. \quad (7.22)$$

Bias metrics complement wRMSE/wMAE by distinguishing symmetric dispersion around zero from consistent directional error.

7.5.3. Participant-level Performance Metrics

To characterize inter-participant variability, the weighted metrics above are also computed per participant. For participant i with measurement index set $\mathcal{J}_i$, participant-level metrics wRMSE_i , wMAE_i , bias_i , and wBias_i are computed by restricting $\mathcal{J}$ to $\mathcal{J}_i$. This provides a distribution of performance values across participants, enabling identification of outliers and assessing whether overall performance is driven by a small subset of individuals.

7.5.4. Instrument-level Residual Diagnostics

Residuals are additionally summarized by measurement instrument type to verify that performance is not dominated by a single instrument stream. For each type κ , residual subsets are formed and the following descriptive statistics are reported:

- sample count N_κ ,
- mean residual and standard deviation,
- weighted RMSE computed within the type subset,
- mean absolute residual (as a magnitude summary).

These diagnostics support interpretation of model fit across questionnaire-derived measurements and probe measurements, and help identify systematic discrepancies associated with specific instrument classes.

7.5.5. Train-validation Generalization Assessment

Generalization is assessed by reporting the full set of overall metrics (RMSE, wMAE, bias, wBias, and measurement count) separately on the training and validation participant splits. The validation-set metrics are treated as the primary indicators of generalization performance, while train-validation gaps are used to detect overfitting.

7.5.6. Metric-based Ranking of Candidate Parameter Estimates

When multiple candidate parameter estimates are available (e.g., from different optimization runs), candidates are ranked using a validation-first criterion. The primary ranking objective is minimization of validation wRMSE, with secondary criteria applied as tie-breakers:

1. minimize validation wMAE,
2. minimize $|\text{wBias}|$ on the validation set,
3. minimize training wRMSE.

This ranking emphasizes predictive accuracy on held-out participants while discouraging solutions with large systematic bias.

7.6. Robustness and Sensitivity Analysis Framework

This section defines the robustness and sensitivity analyses used to assess the stability of the identified trust dynamics model. The goal is to determine whether (i) the model's predictive accuracy and (ii) the inferred parameter values remain consistent under controlled variations. These analyses are diagnostic, as they provide empirical evidence about practical stability and local identifiability of the fitted model, without introducing additional modelling assumptions or changing the model structure.

Throughout this section, robustness is assessed with respect to a chosen validation objective computed on a held-out subset of participants. The primary objective used in the analyses is the weighted root-mean-square error between observed trust measurements and model predictions. Sensitivity analyses examine how this objective changes when the parameter vector is perturbed in controlled ways.

7.6.1. Purpose and Definitions

Robustness in this chapter is evaluated along two complementary dimensions:

- **Predictive robustness:** the stability of predictive performance on held-out data. A model is considered predictively robust if its validation error remains similar under reasonable methodological variations.
- **Parameter robustness:** the stability of the identified parameter values. A model is considered parameter-robust if different fitting procedures lead to similar parameter vectors, suggesting that the data constrain the parameters consistently.

7.6.2. Robustness Across Identification Methods

Robustness across identification methods is assessed by fitting the same model using multiple optimisation algorithms and comparing the resulting solutions. This comparison serves two roles:

- it probes whether the selected solution is a numerical artefact of a specific solver, and
- it provides evidence about the shape of the objective landscape (e.g., whether multiple distinct parameter vectors yield similar validation performance).

Let $\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(K)}\}$ denote the set of candidate parameter vectors produced by the different optimisation algorithms. For each candidate $\theta^{(k)}$, the following quantities are computed:

1. **Overall performance metrics** on training and validation sets: $\text{wRMSE}(\theta^{(k)})$, $\text{wMAE}(\theta^{(k)})$, $\text{bias}(\theta^{(k)})$, and $\text{wBias}(\theta^{(k)})$.
2. **Measurement-type stratified metrics** on training and validation sets, computed by restricting evaluation to each measurement type (e.g., post-task questionnaire, mid-task questionnaires, probes).
3. **Pairwise parameter distances** between candidates:

$$d_{ab} = \|\theta^{(a)} - \theta^{(b)}\|_2, \quad (7.23)$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Small distances indicate that different optimisers converge to similar parameter values; large distances indicate solver-dependent parameter regimes.

The cross-optimizer analysis is interpreted as a robustness measure: if multiple optimisers produce candidates with similar validation performance and similar parameter values, this supports the stability of the identification. Conversely, if validation performance is similar but parameter distances are large, this suggests potential parameter non-identifiability or compensatory parameter trade-offs (different parameters producing similar predictions).

7.6.3. Local One-at-a-time Sensitivity Around the Selected Parameters

Local sensitivity analysis evaluates how the validation objective changes when parameters are perturbed around a reference solution. Let θ^* denote the selected parameter vector used as the baseline for sensitivity analysis (selected according to the validation-based protocol defined earlier in the chapter). Let $f(\theta)$ denote the validation objective, primarily $f(\theta) = \text{wRMSE}(\theta)$.

For each parameter index $i \in \{1, \dots, p\}$ and for each relative perturbation magnitude ϵ in a predefined set (e.g., $\epsilon \in \{\epsilon_1, \epsilon_2, \dots\}$), two perturbed vectors are constructed:

$$\theta_{i,+}^*(\epsilon) = \theta^* \text{ with } \theta_i^* \leftarrow \theta_i^*(1 + \epsilon), \quad \theta_{i,-}^*(\epsilon) = \theta^* \text{ with } \theta_i^* \leftarrow \theta_i^*(1 - \epsilon), \quad (7.24)$$

with appropriate handling for near-zero parameters (where additive perturbations are used to keep perturbations meaningful). Perturbed vectors are constrained to remain within admissible parameter bounds consistent with the model definition.

For each perturbation, the change in objective relative to the baseline is computed:

$$\Delta f_{i,\pm}(\epsilon) = f(\theta_{i,\pm}^*(\epsilon)) - f(\theta^*). \quad (7.25)$$

A large $|\Delta f_{i,\pm}(\epsilon)|$ indicates that small changes in parameter i substantially affect validation performance, i.e., parameter i is locally influential near θ^* .

In addition to overall objective changes, the same Δ analysis is computed for measurement-type stratified wRMSE. This detects whether a parameter primarily affects prediction quality for a particular measurement modality.

To summarize sensitivity per parameter across perturbation magnitudes and directions, parameters are ranked using an aggregate statistic such as the median absolute objective change:

$$S_i^{\text{OAT}} = \text{median} \left(\{|\Delta f_{i,+}(\epsilon)|, |\Delta f_{i,-}(\epsilon)|\}_\epsilon \right). \quad (7.26)$$

This provides an interpretable ordering of parameters by local influence on validation performance.

7.6.4. Local Screening Within a Bounded Neighborhood

One-at-a-time sensitivity is simple to interpret but may miss effects that depend on interactions among parameters. To obtain a more global local picture (within a neighborhood), a Morris-like screening procedure is used. This method estimates elementary effects, which measure how the objective changes when stepping along one parameter direction while the other parameters vary within a bounded region.

A bounded neighborhood is defined around θ^* by allowing each component to vary within a specified relative range:

$$\theta_i \in [\theta_i^*(1 - \rho), \theta_i^*(1 + \rho)], \quad (7.27)$$

with a near-zero handling rule for parameters close to zero, and with intersection against admissible global bounds to ensure all sampled vectors are valid.

The method samples multiple random trajectories through this neighborhood. Along each trajectory, one parameter is perturbed at a time by a small step (typically a fixed relative step size). If θ is the current point and θ' is the point obtained by stepping only along parameter i , the elementary effect for that step is:

$$\text{EE}_i = \frac{f(\theta') - f(\theta)}{\delta}, \quad (7.28)$$

where δ is the step size (expressed in the same normalized units used to define the perturbation). Repeating this process yields a distribution of elementary effects for each parameter.

For each parameter i , two standard summaries are computed:

$$\mu_i^* = \mathbb{E} [|\text{EE}_i|], \quad (7.29)$$

$$\sigma_i = \text{Std} [\text{EE}_i]. \quad (7.30)$$

Here, μ_i^* captures overall influence (parameters with large μ_i^* tend to change the objective strongly), while σ_i captures variability of influence across the neighborhood. High σ_i may indicate nonlinear effects and/or interactions with other parameters.

The same screening summaries can additionally be computed for measurement-type stratified objectives (e.g., wRMSE restricted to a given measurement type), supporting instrument-specific sensitivity interpretation.

7.6.5. Pairwise Response Mapping for Influential Parameters

Screening identifies which parameters have the largest local influence, but it does not directly visualize interaction structure. To probe interactions, a pairwise response map is computed for the two most influential parameters identified by the screening summaries.

Let i and j denote the two selected parameter indices. A grid of multiplicative factors $\{g_1, \dots, g_m\}$ is defined, and parameter pairs are constructed as:

$$\theta_i = \theta_i^* g_a, \quad \theta_j = \theta_j^* g_b, \quad \text{for } a, b \in \{1, \dots, m\}, \quad (7.31)$$

with appropriate near-zero handling and enforcement of admissible parameter bounds. All other parameters remain fixed at their baseline values in θ^* .

For each grid point, the objective value $f(\theta)$ is computed on the validation set. The resulting matrix of values defines a local response surface, which can be examined in terms of absolute objective levels $f(\theta)$ across the grid, and objective changes relative to baseline $f(\theta) - f(\theta^*)$.

The same grid evaluation can be repeated for measurement-type stratified objectives, providing a view of whether apparent interactions are driven by a specific instrument category.

7.6.6. Measurement-type Stratified Robustness Diagnostics

All robustness and sensitivity computations described above are performed both overall, pooling residuals across all measurement types, and by measurement type, restricting the evaluation to one measurement category at a time.

This stratification is important because a model can appear robust overall while being fragile for a particular instrument type. By reporting per-type wRMSE (and related metrics), the analysis checks whether robustness conclusions are consistent across instruments.

7.6.7. Scope and Limitations

The robustness and sensitivity analyses defined in this section provide empirical evidence about stability across numerical identification methods and local stability around the selected parameter vector. They do not provide a formal proof of structural identifiability, nor do they produce full uncertainty bounds on parameters. Instead, they are intended to (i) detect solver-dependent solutions, (ii) identify locally influential parameters, and (iii) reveal potential parameter trade-offs and interactions that could affect the interpretation of identification results.

7.7. Baseline Models and Comparative Evaluation Protocol

This section defines the baseline predictors used to contextualize the performance of the proposed trust dynamics model, and it specifies the evaluation protocol used to compare all methods under identical information assumptions. Baseline models serve two complementary roles: (i) they provide interpretable reference points spanning increasingly structured assumptions about trust, and (ii) they quantify how much predictive value is gained by the proposed dynamics model beyond simple heuristics and established literature baselines.

All baselines are evaluated in the same experimental context as the proposed model, using the same discretized time grid, the same event alignment, and the same measurement sampling scheme. Baseline parameters that require calibration are fitted on the training split only, and generalization is assessed on the validation split.

7.7.1. Objective of the Baseline Comparison

The baseline comparison is designed to answer the following questions:

1. **Added value over static predictors:** does a dynamical model outperform constant-trust assumptions such as dispositional trust or a global mean?
2. **Added value over lightweight event-driven rules:** does explicitly modeling trust dynamics provide improvements beyond simple rules that update trust only when salient task outcomes occur (e.g., success/failure events)?
3. **Added value over a principled Bayesian inference baseline:** how does the proposed model compare against a probabilistic observer that infers latent trust from outcomes and sporadic trust reports?

To ensure that differences in performance reflect differences in modeling assumptions (rather than differences in available inputs or scoring procedures), all methods are compared using the same residual definition, the same measurement weights, and the same measurement-aligned evaluation protocol.

7.7.2. Baseline Model Families

Three baseline families are considered: constant predictors, event-driven “bump” models, and a Bayesian filtering baselines. The two literature baselines are employed for relevant comparisons. Here, they are described in the specific application context of this study, with emphasis on implementation assumptions, information restrictions (to avoid circularity), and evaluation-alignment details.

Constant predictors

Constant predictors assume trust does not evolve during the experiment; instead, they output a single value at all measurement times. Despite their simplicity, these predictors establish a meaningful reference: a dynamical model should improve upon the best constant explanation of the data.

Let $\hat{\tau}(t_m)$ denote the predicted trust at the m -th measurement time t_m for participant i .

Dispositional constant. This baseline predicts a participant-specific constant trust level equal to the participant’s dispositional (trait) trust value. Formally,

$$\hat{\tau}(t_m) = \tau_i^{\text{disp}}, \quad (7.32)$$

where τ_i^{disp} is participant i ’s dispositional trust.

Global training mean constant. This baseline predicts a single constant value for all participants and all times, computed as a weighted mean of aligned measurements on the training set:

$$\hat{\tau}(t_m) = \mu_{\text{TRAIN}}, \quad (7.33)$$

where μ_{TRAIN} is computed using the same measurement weighting scheme employed in the main objective. This ensures the constant is consistent with the evaluation metric (i.e., it is optimal under the weighted criterion among global constants).

Oracle participant mean constant. This baseline predicts a constant that is equal to each participant’s own weighted mean trust within the evaluated split. For participant i ,

$$\hat{\tau}(t_m) = \mu_i \quad \text{with} \quad \mu_i = \frac{\sum_{m \in \mathcal{M}_i} w_m \tau(t_m)}{\sum_{m \in \mathcal{M}_i} w_m}, \quad (7.34)$$

where $\mathcal{M}_i$ is the set of measurement indices available for participant i in the evaluated split (training or validation), and w_m are the measurement weights. This is an “oracle” because it uses the participant’s own measurements to define the constant; it is therefore not deployable, but it provides a strong upper bound among constant predictors and helps interpret the achievable error levels when dynamics are ignored.

Event-driven bump models

Event-driven bump models update trust only when discrete task outcomes occur. In this particular context, that relates to door events with success or failure outcomes. Their theoretical motivation and canonical forms are beyond the scope of this thesis report, but can be consulted by following the corresponding references. In this section, we specify how these models are instantiated for the present task and how they are aligned to the measurement grid used throughout the analysis pipeline.

Let k index the discretized simulation grid and let $u_k \in \{-1, 0, +1\}$ denote the door outcome input at index k , where $u_k = +1$ indicates success, $u_k = -1$ indicates failure, and $u_k = 0$ indicates no door event (or unknown outcome) at that step. The trust state is denoted $T_k \in [0, 1]$.

For the specific application case, we work based on three assumptions, namely that each grid step uses at most one door event, that after each update trust is constrained to $[0, 1]$, and that initial trust is set to the dispositional trust level.

Asymmetric bump with saturation. Trust increases by δ_+ after a success and decreases by δ_- after a failure:

$$T_{k+1} = \text{clip}_{[0,1]}(T_k + \delta_+ \mathbb{I}[u_k = +1] - \delta_- \mathbb{I}[u_k = -1]), \quad (7.35)$$

with constraints $\delta_+ \geq 0$ and $\delta_- \geq 0$. The parameters (δ_+, δ_-) are exclusively calibrated on the training set using the same weighted residual objective employed elsewhere. This model is taken from the literature, specifically from its application in the work by Wilmes and Jamshidinejad (2025).

Symmetric bump with saturation. Trust changes by a single step magnitude Δ with sign determined by the outcome:

$$T_{k+1} = \text{clip}_{[0,1]}(T_k + \Delta u_k), \quad (7.36)$$

with $\Delta \geq 0$ calibrated on the training set of participants. This baseline captures a minimal “reward/punishment” update rule with equal magnitude and opposite sign for success and failure. It is an intentional simplification of the asymmetric model, to test the necessity of different trust steps for successes and failures.

Bayesian trust inference (OPTIMo-lite)

OPTIMo-lite is included as a baseline trust inference model, not as a behavioral model. Its role in the pipeline is to answer the question:

Given robot performance outcomes and sporadic trust reports, how would a principled Bayesian observer infer the human’s latent trust over time?

This places OPTIMo-lite in the same conceptual category as the other deterministic trust dynamics models. It produces an inferred trust trajectory from observed outcomes and trust measurements. Theoretical background on OPTIMo as originally proposed (Xu and Dudek, 2015) is beyond the scope of this thesis report; here the specific “lite” configuration is explained, as well as its adaptation to the present experimental setting.

A key feature of OPTIMo is that evidence channels may be present or absent depending on the application, and inference remains well-defined. In its original formulation (Xu and Dudek, 2015), evidence can include robot performance, trust feedback, trust-change reports, human interventions, and extraneous causes; OPTIMo-lite uses only the evidence sources available and appropriate for fair comparison in this study.

OPTIMo-lite does not mean “no evidence”; rather, it means excluding behavioral/intervention evidence (and any other unavailable channels) while retaining outcome and trust-probe evidence. This choice is justified by two methodological requirements:

1. **Avoiding circularity:** In this thesis, trust estimation and behavioral modeling are kept explicitly separate. If a trust inference baseline were allowed to use behavioral evidence (follow/override), then inferred trust would already be conditioned on the behaviors that the behavioral model later tries to explain, resulting in circular reasoning and implicit information leakage.
2. **Fair comparison across baselines:** The other baselines in this section (constants and bump models) do not use behavioral evidence (follow/not follow), only interaction outcomes (success/failure). Allowing OPTIMo to use additional inputs would change the information assumptions and make the comparison reflect access to extra signals rather than differences in trust inference structure.

Accordingly, OPTIMo-lite uses only performance outcomes from the task (door success/failure events), and the trust measurements available during the experiment. It is important to note that this constitutes a fundamental difference from the proposed model, which does not have access to any of the interaction trust measurements. All other channels (interventions, trust-change reports, extraneous causes) are excluded. There is direct precedent for using a reduced OPTIMo configuration in comparative trust-dynamics studies.

Importantly, this model simplification has precedent in the established literature. Namely, Guo and Yang (2021) use OPTIMo as a baseline in a way that is adapted for comparability with other models (i.e., the model “uses the robot’s performance history, but not other behavioral variables [like human agent’s intervention behaviors]”). The key methodological principle is the same here: to match information assumptions across trust models, so that the comparison is about trust dynamics, and not about who got more inputs, and therefore has an unfair advantage.

OPTIMo typically uses a performance summary variable that increases with worse performance (Xu and Dudek, 2015). To match the door-based task while preserving this interpretation, we define a binary performance indicator:

$$p_k = \begin{cases} 1, & \text{if the robot fails at door step } k, \\ 0, & \text{if the robot succeeds at door step } k, \\ 0, & \text{if no door outcome is defined at step } k \text{ (conservative default).} \end{cases} \quad (7.37)$$

This corresponds to using a window size of one event, consistent with the discrete nature of the task outcomes.

OPTIMo-lite is implemented as a grid-based Bayesian filter (histogram filter) over bounded trust $\tau \in [0, 1]$, which provides a practical and defensible approximation of continuous Bayesian inference. Its implementation follows the steps described below:

1. **Representation.** Discretize $[0, 1]$ into a fixed grid $g_1, \dots, g_{n_g}$ and represent belief as

$$b_k(j) \approx P(\tau_k = g_j \mid \text{observations up to } k). \quad (7.38)$$

The belief is a probability vector satisfying $b_k(j) \geq 0$ and $\sum_j b_k(j) = 1$.

2. **Prediction (transition) step.** Belief is propagated using a Markov transition model derived from OPTIMo's transition mean structure (see Theoretical Background). A convenient specialization for this study is:

$$\tau_k \mid \tau_{k-1} = t \sim \mathcal{N}(\mu_k(t), \sigma_t^2), \quad \mu_k(t) = t + \omega_{tb} + \omega_{tp}p_k \quad (+ \omega_{td}(p_k - p_{k-1}) \text{ if enabled}), \quad (7.39)$$

where ω_{tb} is a baseline drift term, ω_{tp} captures the effect of current failure, and ω_{td} (optional) captures a trend effect. The prediction step computes the prior belief $\tilde{b}_k$ by integrating the transition against the previous belief; on a grid this becomes a matrix multiplication.

3. **Boundary handling for bounded trust.** Because trust is bounded to $[0, 1]$, probability mass implied by the Gaussian transition outside the interval must be handled. OPTIMo-lite applies a consistent boundary rule (e.g., truncation and renormalization) at every step. The exact rule is held fixed for both training (parameter fitting) and evaluation.
4. **Measurement (probe) update step.** When a trust probe is available near step k , the belief is updated via Bayes' rule:

$$b_k(t) \propto P(f_k \mid \tau_k = t) \tilde{b}_k(t), \quad (7.40)$$

using a Gaussian observation model consistent with the trust scale:

$$f_k \mid \tau_k = t \sim \mathcal{N}(t, \sigma_f^2), \quad (7.41)$$

where f_k is the observed trust report mapped to $[0, 1]$, and σ_f controls the assumed noise of trust reports (i.e., how tightly probes constrain the latent trust).

To compare OPTIMo-lite with other methods using residual metrics, the belief distribution must be converted into a single predicted trust value at each measurement time. OPTIMo-lite uses the posterior mean:

$$\hat{\tau}_k = \mathbb{E}[\tau_k \mid \cdot] \approx \sum_{j=1}^{n_g} g_j b_k(j), \quad (7.42)$$

which is a standard estimator under squared-error loss and is robust to multi-modality relative to maximum a posteriori (MAP) estimates. Predicted values at measurement times are obtained by sampling $\hat{\tau}$ according to the same measurement alignment used for all methods.

OPTIMo-lite is used as a global trust inference baseline, matching the training and validation split used for the other baselines and the proposed model. Participant-specific fitting would constitute a personalization upper bound rather than a baseline and is therefore not considered here.

Additionally, to isolate the impact of incorporating trust probes and questionnaires (which the proposed model does not use as input) into the filter, an ablation variant of OPTIMo-lite is also evaluated in which the measurement update step is disabled (i.e., the belief evolves using outcomes only, similarly to the proposed model). This yields a purely outcome-driven Bayesian predictor that remains consistent with the no-behavior-evidence restriction.

7.7.3. Parameter Fitting for Baselines

Some baselines contain free parameters and must be calibrated. To ensure comparability and avoid information leakage, all baseline parameters are fitted on the same training participants. The fitting objective is consistent with the global pipeline objective: it minimizes the weighted discrepancy between observed and predicted trust at aligned measurement times (RMSE). Table 7.5 collects the baseline parameters that are fitted on the training set of participants.

Table 7.5: Baseline parameters calibrated on the training set of participants. For asymmetric and symmetric bump parameters, non-negativity constraints are enforced.

Model	Fitted Parameters
Symmetric Bump	Δ
Asymmetric bump ¹	δ_+, δ_-
OPTIMo-lite ²	$\omega_{tb}, \omega_{tp}, \sigma_t$

¹ Wilmes and Jamshidnejad (2025).

² Xu and Dudek (2015); Guo and Yang (2021).

All fitted parameters are subsequently held fixed when evaluating baselines on the validation set.

7.7.4. Leakage Prevention for Filter-based Baselines

Filter-based methods update their internal state using incoming measurements. If the prediction at a measurement time is recorded after assimilating that measurement, the method would effectively be “graded on information it has already seen”, producing overly optimistic error estimates. This issue is commonly referred to as information leakage.

To prevent leakage, predictions at measurement times are computed using the filter state before applying the measurement update for that time. This yields a fair “predict-then-update” evaluation consistent with how an online estimator would operate in deployment.

7.7.5. Metrics and Baseline Comparisons

All baseline methods are compared using the same residual-based metrics computed on aligned measurement samples. In addition to overall performance, results are stratified by measurement kind and by participant. These include weighted RMSE, MAE and bias metrics, following the same definitions introduced in section 7.5.

Additionally, for each baseline method, the improvement of the proposed trust dynamics model is computed as:

$$\Delta \text{wRMSE} = \text{wRMSE}_{\text{baseline}} - \text{wRMSE}_{\text{model}}, \quad (7.43)$$

$$\% \Delta \text{wRMSE} = 100 \cdot \frac{\text{wRMSE}_{\text{baseline}} - \text{wRMSE}_{\text{model}}}{\text{wRMSE}_{\text{baseline}}}, \quad (7.44)$$

and analogously for other metrics. Once again, improvements are computed both overall and per measurement kind to preserve interpretability.

7.8. Summary of Identification Methodology

This section summarizes the end-to-end methodology used to identify the parameters of the proposed trust dynamics model and to evaluate the resulting model prior to reporting numerical outcomes. The purpose is to provide a single, coherent overview that connects the identification formulation (section 7.3), the data used (section 7.4), the evaluation metrics (section 7.5), the robustness and sensitivity framework (section 7.6), and the baseline comparison protocol (section 7.7). Throughout, the emphasis is on methodological consistency and prevention of information leakage between training, validation, and downstream behavioral modeling.

7.8.1. Overview and Scope

The identification task in this chapter concerns the dynamics of trust as a latent cognitive state. A single global parameter vector $\theta \in \mathbb{R}^8$ is estimated to explain trust measurements across all participants, under the assumption that the underlying trust dynamics mechanism is shared across individuals. Participant-specific variability in predicted trust trajectories arises from differences in each participant's event sequence, timing, and initial dispositional trust input, rather than from participant-specific parameter values.

Crucially, the identification stage is restricted to trust inference and does not model behavioral decisions or interventions. Trust is treated as an upstream latent-state generator: its estimation and validation are performed independently of the behavioral modeling and analysis presented in the subsequent chapter. This separation is essential both for interpretability (trust dynamics are validated on their own terms) and for avoiding circularity (later behavior models should not be asked to explain actions using a trust estimate that already conditions on those same actions).

7.8.2. Data and Evaluation Setup

The unit of analysis is the individual participant. For each participant, the dataset contains (i) a temporally ordered sequence of task-related events that drives the trust dynamics model, and (ii) discrete trust measurements collected during the experiment using multiple instruments with different temporal structure and reliability. All trust measurements are mapped to a common normalized scale in $[0, 1]$ to enable joint use in a single identification objective.

To evaluate generalization, the full dataset is partitioned into disjoint training and validation participant sets. The split is performed at the participant level, ensuring that all data for a participant appear in exactly one split. The training split is used for parameter estimation and any required calibration steps, while the validation split is held out for performance evaluation, robustness diagnostics, and comparative reporting. This protocol supports an assessment of generalization to previously unseen participants and prevents leakage of validation information into the identification process.

7.8.3. Identification Workflow: Simulation, Alignment, and Objective

For a given parameter vector θ , the proposed trust dynamics model defines a latent trust trajectory $\tau(t; \theta) \in [0, 1]$ that evolves in continuous time under the influence of participant-specific task events. We reuse the notation introduced in subsection 7.3.2 to define model prediction samples at each participant's measurement times:

$$\hat{\tau}_{i,m}(\theta) = \tau(t_{i,m}; \theta), \quad (7.45)$$

where $t_{i,m}$ denotes the time stamp of measurement m for participant i . Measurements are treated as passive observations: they do not update or alter the model state during simulation, and serve only as read-outs of the latent trust state at specific time instants.

Model fit is evaluated using measurement-aligned residuals

$$e_{i,m}(\theta) = \hat{q}_{i,m} - \hat{\tau}_{i,m}(\theta), \quad (7.46)$$

where $\hat{q}_{i,m}$ is the observed trust value. To reflect the differing reliability and granularity of measurement modalities, residuals are weighted by instrument-specific positive weights. The global identification objective is a weighted least-squares criterion aggregated across participants:

$$J(\theta) = \sum_{i=1}^{N_p} \sum_{m=1}^{N_m} w_{i,m} e_{i,m}(\theta)^2, \quad \theta^* = \arg \min_{\theta} J(\theta), \quad (7.47)$$

with appropriate parameter constraints and modeling assumptions applied as described in Section 7.3.

7.8.4. Performance Assessment and Reporting Principles

The quality of the identified model is assessed using residual-based metrics computed at measurement times. The primary goodness-of-fit indicators are the weighted root-mean-square error (RMSE) and weighted mean absolute error ($wMAE$), complemented by bias measures (bias and weighted bias, $wBias$) to detect systematic over- or under-prediction.

Metrics are reported at multiple aggregation levels to support interpretability:

- **Overall (split-level) metrics:** residuals pooled across all participants within training or validation sets to quantify global fit and generalization.
- **Participant-level metrics:** metrics computed per participant to characterize inter-individual variability and identify potential outliers.
- **Instrument-level diagnostics:** metrics computed separately for each measurement modality to verify that performance is not dominated by a single instrument stream and to identify modality-specific discrepancies.

When multiple candidate parameter estimates are available (e.g., from different optimization attempts), a validation-first ranking protocol is applied (subsection 7.5.6). Candidates are primarily ranked by validation $wRMSE$, with secondary criteria such as validation $wMAE$ and absolute bias used as tie-breakers. This selection protocol prioritizes generalization performance on held-out participants and discourages solutions that achieve low error by introducing large systematic bias.

7.8.5. Robustness and Sensitivity Checks

Beyond goodness-of-fit, the identification process includes diagnostic analyses that probe the stability of both performance and parameters. These analyses are intended to provide empirical evidence of practical robustness and local identifiability in the data regime of the study.

First, robustness is assessed across identification methods by fitting the same model using multiple optimization algorithms and comparing the resulting candidate solutions. This comparison can be interpreted as a robustness metric: if distinct optimizers converge to similar parameter vectors and yield similar validation performance, the identification is less likely to be a numerical artifact of a particular solver. Conversely, if validation performance is similar but parameter vectors differ substantially, this suggests the presence of compensatory parameter trade-offs or weak practical identifiability.

Second, local sensitivity around the selected solution θ^* is evaluated using controlled perturbations of parameters and observing the induced change in the validation objective (primarily validation $wRMSE$). One-at-a-time perturbations provide an interpretable ranking of which parameters most strongly affect predictive performance locally. A Morris-like screening procedure extends this idea by exploring elementary effects along random trajectories within a bounded neighborhood, which can reveal nonlinearities and interaction-like behavior. Finally, pairwise response mapping for the most influential parameters provides a localized visualization of how validation performance varies over a two-parameter grid.

All robustness and sensitivity analyses are computed both overall and stratified by measurement type. This stratification is important because apparent stability in pooled metrics can mask fragility for a particular measurement modality.

7.8.6. Comparative Baseline Protocol and Leakage Prevention

To contextualize the performance of the proposed trust dynamics model, a family of baselines is evaluated under identical information assumptions and scoring rules. Baselines span constant predictors, event-driven bump models, and a Bayesian trust inference baseline (OPTIMo-lite). The literature baselines (asymmetric bump and OPTIMo) are applied in the specific experimental setting with explicit assumptions required for alignment and fair comparison.

Two principles govern the baseline protocol. First, training-only calibration: baseline parameters that require fitting are calibrated exclusively on the training set using the same weighted residual objective used throughout the pipeline. The fitted baseline parameters are then held fixed when evaluating both training and validation sets. Second, measurement-aligned evaluation: every method is evaluated on the same discretized time grid, the same event alignment, and the same set of measurement instances, so that differences in metrics reflect modeling assumptions rather than differences in sampling or scoring.

For filter-based baselines such as OPTIMo-lite, leakage prevention is critical. Because Bayesian filters update their internal belief using measurements, scoring a prediction after assimilating the measurement

would overestimate performance by implicitly using the ground-truth observation being predicted. Therefore, predictions at measurement times follow a predict-then-update policy: the point prediction used for residual computation is produced from the filter state before applying the measurement update at that time index. This one-step-ahead evaluation ensures that the baseline is compared fairly against methods that do not assimilate measurements online, and it matches how an online estimator would operate in deployment.

In summary, the identification methodology combines (i) global weighted least-squares fitting with participant-level simulation and measurement-time alignment, (ii) validation-first selection and multi-level residual diagnostics, (iii) robustness and sensitivity analyses to probe stability and identifiability, and (iv) a carefully aligned baseline comparison protocol with explicit leakage prevention for filter-based methods. The following sections report the resulting parameter estimates and quantitative performance outcomes.

7.9. Identification Results

This section presents the results of the trust model identification and validation, organized to support the results-oriented subquestions of Research Question 3. Specifically, quantitative fit metrics and qualitative trust trajectory predictions address the model's out-of-sample predictive capability (RQ3.b), sensitivity and interaction analyses provide empirical evidence regarding the practical identifiability and learnability of model parameters (RQ3.c), and robustness analyses with respect to estimation assumptions and parameter perturbations address robustness considerations (RQ3.d).

All results presented here correspond directly to the identification formulation, evaluation metrics, robustness analyses, and baseline comparison protocols introduced earlier in this chapter. Numerical values, tables, and figures are reported with reference to the training and validation participant splits as defined in subsection 7.2.8, with emphasis on validation performance when assessing generalization.

7.9.1. Selected Parameter Estimates

This subsection reports the parameter vector selected as the final estimate of the proposed trust dynamics model and summarizes its numerical properties. The selected parameters are obtained following the validation-first ranking protocol defined in Section 7.5, using the weighted root-mean-square error (wRMSE) on the held-out validation set as the primary criterion. Four candidate parameter vectors were derived from four distinct optimization runs using a genetic algorithm (twice), gradient-based and derivative-free solvers, implemented in MATLAB using the `ga`, `fmincon` and `patternsearch` solvers, respectively.

To balance global exploration of the parameter space with computational efficiency, a hybrid optimization strategy was adopted. First, the genetic algorithm was executed twice using an extensive configuration to mitigate the effects of stochastic variability and to explore multiple regions of the nonconvex objective landscape. The best-performing solution across these global runs (showing lowest objective function cost) was then used to initialize two complementary local refinement procedures: a gradient-based solver and a derivative-free pattern search algorithm, both configured with similarly extensive computational budgets. Across all optimization runs, the total number of objective function evaluations was on the order of 10^4 , reflecting a substantial exploration of the parameter space given the cost of each evaluation. This sequential global-local approach enabled broad exploration of candidate solutions while leveraging efficient local convergence, and allows consistency across fundamentally different optimization paradigms to be assessed.

Among the candidate parameter vectors produced by the different optimization procedures, the solution obtained using the gradient-based solver was selected. This solution achieved the lowest validation-set wRMSE while maintaining competitive performance on all secondary metrics (wMAE, bias, and weighted bias), and did not exhibit signs of numerical instability or overfitting.

The selected parameter vector is denoted by

$$\theta^* = \left[\lambda^{\text{rep}} \quad \alpha^{\text{sit}} \quad \lambda^{\text{sit}} \quad \phi^{\text{fail}} \quad \phi^{\text{succ}} \quad a^{\text{succ}} \quad \lambda^{\text{lat}} \quad \kappa^{\text{lat}} \right]^\top. \quad (7.48)$$

The numerical values of the selected parameters are reported in Table 7.6.

Lower and upper bounds were selected to be as permissive as possible while remaining consistent with the intended role and physical interpretation of each parameter in the trust dynamics model. Parameters representing normalized weights (α^{sit}) or absolute trust variations (ϕ^{fail} and ϕ^{succ}) were bounded within the interval $[0, 1]$ to enforce physically meaningful values.

For the latent trust dynamics parameters (λ^{lat} and κ^{lat}), relatively small upper bounds were imposed to reflect the intended slow evolution of latent trust compared to other components of the model. Similarly,

Table 7.6: Numerical values of the selected parameters following the WLS optimization solution.

Parameter	Estimated value	Lower bound	Upper bound	Proximity to bounds ¹
λ^{rep}	0.0522	1e-4	0.10	0.1926
α^{sit}	0.0185	0	1.00	0.0051
λ^{sit}	1.0097	1e-1	2.00	0.2993
ϕ^{fail}	0.0269	0	1.00	0.0116
ϕ^{succ}	0.0109	0	1.00	0.0034
a^{succ}	0.3130	1e-2	0.65	0.0214
λ^{lat}	0.0073	1e-6	0.01	0.1444
κ^{lat}	0.0016	1e-6	0.01	0.0461

¹ Computed as the normalized distance to the closest bound, relative to the full admissible parameter range (lower and upper bounds).

the bounds on the situational decay parameter λ^{sit} were chosen conservatively. Since the situational trust component receives a normalized risk input in the range $[0, 1]$, values of λ^{sit} approaching the upper bound already correspond to very strong decay under maximal risk conditions.

From a purely mathematical validity standpoint, the success curve scaling parameter a^{succ} could be bounded within $[0, 1]$. However, this parameter governs the initial slope of the personal experience success curve, and therefore directly influences how rapidly trust can be gained from successful interactions. To encode the modeling assumption that trust is accumulated gradually rather than rapidly, the upper bound on a^{succ} was restricted to 0.65.

The proximity to bounds indicates how close each estimated parameter lies to the enforced optimization limits, serving as a first-order diagnostic of potential boundary effects in the identification process.

All parameters lie strictly within their admissible bounds and satisfy the imposed structural constraints, including non-negativity and relative ordering constraints (e.g., $\phi^{\text{succ}} \leq \phi^{\text{fail}}$). In general, parameters are not driven to boundary values, suggesting that the obtained solution is not trivially determined by constraint saturation. In particular, the situational trust weighting parameter α^{sit} is estimated to be close to the lower end of its admissible range, reflecting a reduced contribution of the situational trust component within the identified parameterization.

These observations motivate a more detailed analysis of parameter influence and interaction effects, presented in the following sensitivity and identifiability analyses. Unless stated otherwise, all results reported in the remainder of this chapter are based on simulations performed using θ^* with a simulation time step of $\text{dt} = 1 \text{ s}$.

7.9.2. Overall Model Fit and Out-of-Sample Generalization

This subsection reports the overall predictive performance of the identified trust dynamics model on the training and validation participant sets. Performance is evaluated using the residual-based metrics defined in section 7.5, with emphasis on validation-set results as indicators of generalization. Table 7.7 summarizes the overall performance of the identified model on the training and validation sets, with N_{meas} defining the number of measurements in each training set. Additionally, Figure 7.12 shows the distribution of participant-level wRMSE values on the training set.

Table 7.7: Overall identification performance of the selected trust dynamics model on training and validation participant sets.

Metric	Training	Validation
wRMSE	0.1335	0.1357
wMAE	0.1042	0.1070
Bias	0.0101	-0.0124
wBias	-0.0072	-0.0091
N_{meas}	300	120

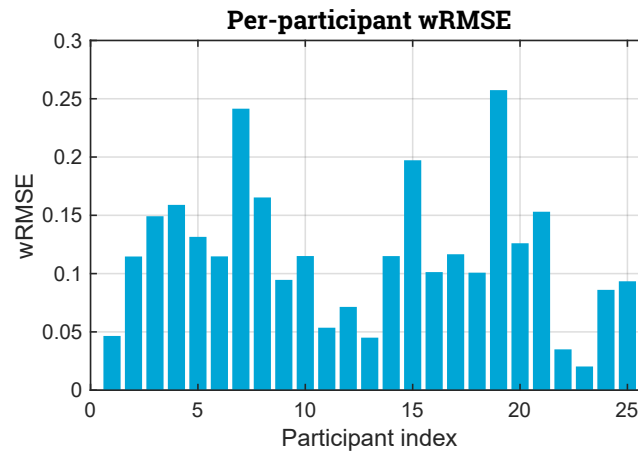


Figure 7.12: Distribution of participant-level weighted RMSE on the training set.

The weighted RMSE and MAE values indicate a comparable level of predictive accuracy on the training and validation sets. The relatively small gap between training and validation errors suggests that the identified model generalizes well to unseen participants and does not exhibit strong overfitting to the training data.

7.9.3. Participant-level Performance Variability

To assess inter-individual variability in model performance, participant-level evaluation metrics are computed on the validation set. Figure 7.13 summarizes the distribution of participant-level wRMSE values on the validation split.

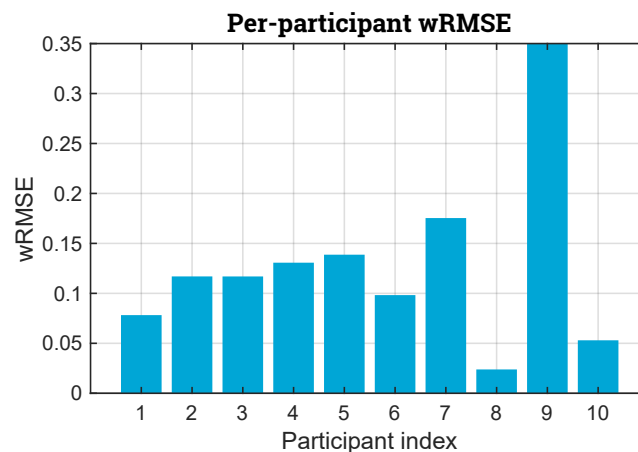


Figure 7.13: Distribution of participant-level weighted RMSE on the validation set.

The majority of participants exhibit moderate prediction errors clustered within a relatively narrow range, indicating that the identified trust dynamics model achieves comparable predictive accuracy across most individuals. At the same time, a single participant displays substantially higher wRMSE value (0.35) than the rest of the validation set. This case contribute disproportionately to the overall validation error but does not dominate the aggregate performance metrics.

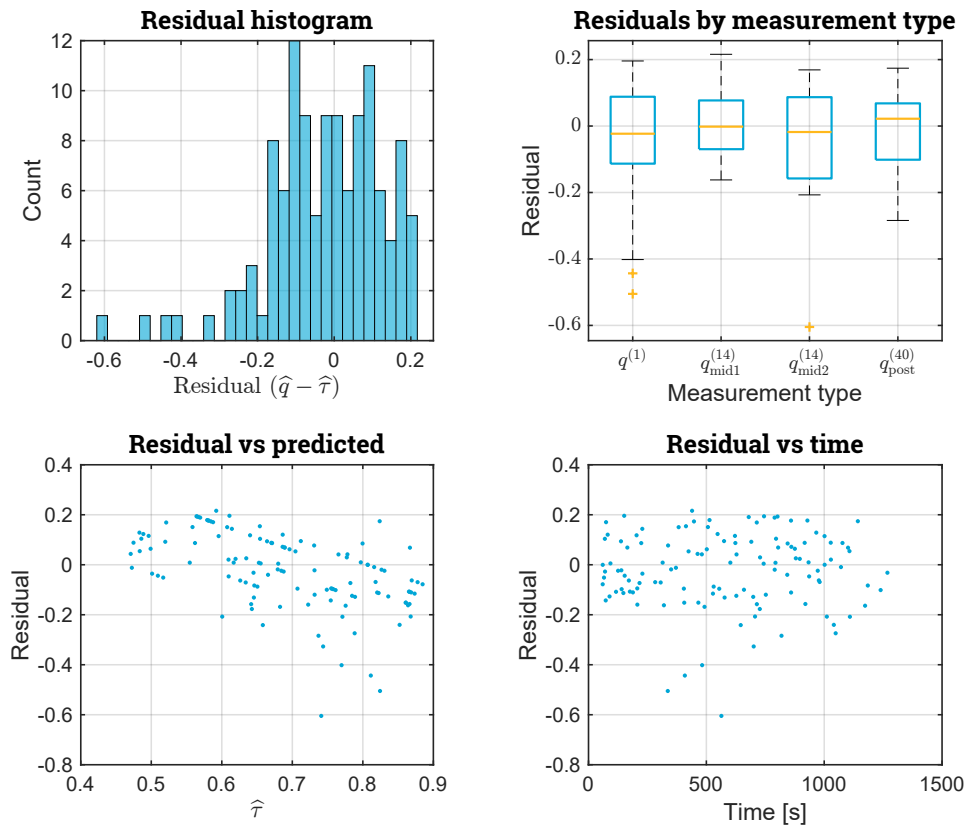
7.9.4. Performance Stratified by Measurement Type

To examine whether model performance varies across different trust measurement modalities, residual-based evaluation metrics are stratified by measurement instrument type on the validation set. This breakdown assesses whether predictive accuracy is consistent across questionnaires and probes, or whether specific measurement types exhibit systematically higher or lower errors. Table 7.8 summarizes instrument-level performance on the validation set.

Table 7.8: Overall identification performance stratified by measurement-type on validation participant set.

Metric	$q^{(1)}$	$q_{\text{mid1}}^{(14)}$	$q_{\text{mid2}}^{(14)}$	$q_{\text{post}}^{(40)}$
wRMSE	0.1481	0.1094	0.2223	0.1368
wMAE	0.1178	0.0860	0.1528	0.1080
Bias	-0.0249	0.0135	-0.0630	-0.0170
N_{meas}	90	10	10	10

Additionally, Figure 7.14 visually shows a residual distribution (histogram), as well as plots of residuals as function of both simulation time and predicted trust. Residual boxplots by measurement type are also reported.

**Figure 7.14:** Residual distributions and error metrics stratified by measurement type on the validation set.

Across all measurement types, residual distributions are centered near zero, indicating limited systematic bias. Error magnitudes remain comparable across instruments, suggesting that the identified trust dynamics model generalizes consistently across questionnaire-based measurements and probe-based measurements. While central tendencies are similar across instrument types, the spread of residuals differs slightly between instruments. Some measurement types exhibit heavier tails or isolated outliers, which contribute to increased dispersion without substantially shifting mean performance metrics. These effects are reflected in wRMSE values rather than bias measures.

7.9.5. Robustness Across Identification Methods

This subsection reports robustness results obtained by identifying the proposed trust dynamics model using multiple optimization methods. The objective is to assess whether the selected solution is sensitive to the choice of optimizer and to characterize the variability of both validation performance and estimated parameters across candidate solutions.

Table 7.9 summarizes the performance of all candidate parameter estimates obtained during the multi-stage identification process. For each optimizer, weighted error metrics are reported on both the training and validation splits, together with an indicator of the solution selected for subsequent analysis.

Table 7.9: Validation performance of candidate parameter estimates from different identification methods. Subscript t indicates the training participant set, v indicates the validation participant set.

Optimizer	wRMSE _t	wRMSE _v	wMAE _t	wMAE _v	wBias _t	wBias _v
Genetic Algorithm (ga 1)	0.130	0.148	0.104	0.109	-0.004	-0.020
Genetic Algorithm (ga 2)	0.130	0.154	0.103	0.113	-0.005	-0.020
Gradient-based (fmincon)	0.134	0.136	0.104	0.107	-0.007	-0.009
Derivative-free (patternsearch)	0.132	0.137	0.105	0.109	-0.002	-0.004

Across identification methods, validation performance is of comparable magnitude, with validation wRMSE values lying within a narrow range. The selected solution, obtained using a gradient-based constrained optimizer, achieves the lowest wRMSE among all validation candidates while maintaining similar training performance, supporting its use as the reference parameter vector in subsequent analyses.

To complement performance-based robustness assessment, similarity between candidate parameter vectors is evaluated using pairwise Euclidean distances in parameter space. Table 7.10 reports the pairwise distances between all candidate solutions.

Table 7.10: Pairwise Euclidean distances between candidate parameter vectors. Values are reported for diagnostic comparison of solution similarity.

	ga 1	ga 2	fmincon	patternsearch
ga 1	—	1.3293	1.0330	1.9250
ga 2	1.3293	—	0.4486	0.6494
fmincon	1.0330	0.4486	—	0.9109
patternsearch	1.9250	0.6494	0.9109	—

Overall, the reported distances indicate a high degree of consistency between solutions obtained using different optimization strategies. In particular, the relatively small distances between the solutions produced by the gradient-based and derivative-free local solvers suggest convergence to a similar region of the parameter space, despite their fundamentally different optimization mechanisms. This reduces the likelihood that the selected solution is an artifact of a specific solver or numerical procedure.

The largest distance is observed between the parameter vector obtained from the first genetic algorithm run and the pattern search solution, which is consistent with the exploratory nature of early-stage global search compared to subsequent local refinement. Taken together, these results support the conclusion that the candidate solutions correspond to closely related optima rather than diverse local minima.

Overall, the results indicate that the identification procedure yields solutions with broadly consistent validation performance across optimization methods, while still exhibiting non-negligible variation in the estimated parameter vectors. This motivates the robustness and sensitivity analyses reported in subsequent subsections.

To assess the robustness of the optimization results and reduce the risk of missing a rare low-cost solution, additional genetic algorithm runs were executed beyond the two runs used in the main identification. Due to computational time constraints, these additional runs were not incorporated into the primary optimization procedure and were instead executed in the background as a verification step. Approximately 20 independent GA optimizations were initialized with different random seeds under the same configuration and data splits. The resulting objective values consistently remained within the same order of magnitude as those reported for the selected solution, and no substantially lower-cost outliers were observed. This provides empirical confidence that the reported parameter estimates are not the result of a favorable initialization and are representative of the attainable performance of the adopted optimization setup.

7.9.6. Sensitivity and Parameter Interaction Analysis for Practical Identifiability

To assess which parameters are practically identifiable from the available data and which exert a dominant influence on validation performance, a sensitivity and interaction analysis was conducted around the identified parameter vector. Local one-at-a-time (OAT) perturbation analysis reveals a clear hierarchy in parameter influence on validation performance. Table 7.11 reports the ranking of parameters based on the median absolute change in validation wRMSE induced by small relative perturbations around the selected parameter vector.

Table 7.11: One-at-a-time sensitivity rankings based on median absolute validation wRMSE change.

Parameter	Median $ \Delta \text{wRMSE} $
ϕ^{fail}	2.7939e-03
ϕ^{succ}	1.8853e-03
κ^{lat}	7.5150e-04
λ^{lat}	2.1564e-04
a^{succ}	1.0941e-04
α^{sit}	8.8874e-05
λ^{sit}	9.7110e-06
λ^{rep}	8.7771e-06

The strongest sensitivities are observed for the event-driven trust update parameters ϕ^{fail} and ϕ^{succ} , which dominate the ranking by a substantial margin. These parameters consistently produce the largest changes in validation wRMSE under perturbation, indicating that model performance is highly sensitive to how strongly trust responds to success and failure events. Secondary influence is observed for the latent dynamics parameters κ^{lat} and λ^{lat} , while situational weighting and reputation-related parameters exhibit comparatively weaker local effects. The least influential parameters under OAT perturbations are λ^{sit} and λ^{rep} , whose perturbations induce only marginal changes in validation performance within the explored range.

The Morris-like screening analysis provides a complementary view of parameter influence by accounting for variability across a bounded local neighborhood. Table 7.12 reports the screening summaries μ^* (overall influence) and σ (variability of influence) for each model parameter.

Table 7.12: Morris-like screening summaries (μ^* , σ) for all model parameters.

Parameter	μ^*	σ
λ^{rep}	0.00017	0.00018
α^{sit}	0.00132	0.00139
λ^{sit}	0.00117	0.00144
ϕ^{fail}	0.04920	0.05231
ϕ^{succ}	0.03056	0.02897
a^{succ}	0.00214	0.00279
λ^{lat}	0.00428	0.00546
κ^{lat}	0.01291	0.01465

Consistent with the OAT analysis, ϕ^{fail} and ϕ^{succ} exhibit the largest μ^* values, confirming their dominant role in determining validation performance. These parameters also show elevated σ values, indicating that their influence varies across the local neighborhood and may involve nonlinear effects or interactions with other parameters.

Among the remaining parameters, κ^{lat} and λ^{lat} display moderate μ^* and σ values, suggesting meaningful but less dominant influence. Parameters associated with reputation and situational modulation remain comparatively weak, with low μ^* values and limited variability. Stratified screening by measurement type (not shown here for brevity) confirms that the dominance of event-driven parameters persists across all trust measurement modalities, although the relative influence of latent dynamics parameters varies with measurement

type.

To further examine potential interactions between influential parameters, a localized pairwise response analysis was conducted for the two most influential parameters identified by the screening analyses, namely ϕ^{fail} and ϕ^{succ} . Figure 7.15 shows the validation wRMSE evaluated over a grid of multiplicative perturbations applied jointly to these parameters.

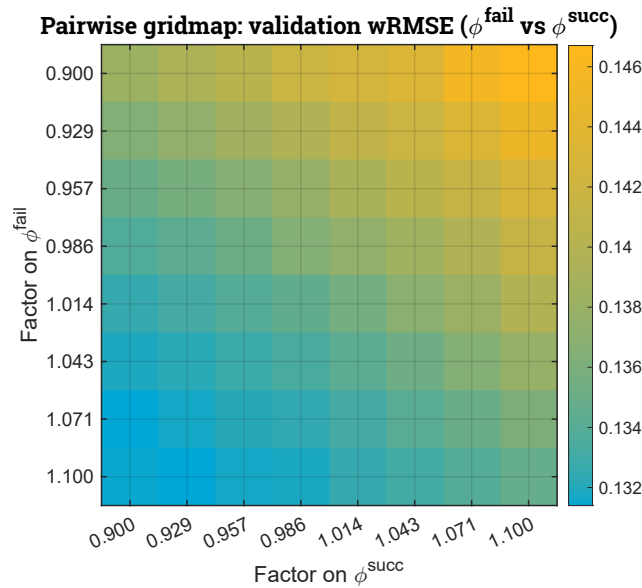


Figure 7.15: Pairwise validation wRMSE response surface for ϕ^{fail} and ϕ^{succ} .

The resulting response surface exhibits a smooth, monotonic structure without abrupt discontinuities, indicating that validation performance varies gradually with coordinated changes in ϕ^{fail} and ϕ^{succ} . The surface reveals a shallow valley region rather than a sharply localized optimum, suggesting partial trade-offs between these parameters: multiple combinations yield comparable validation performance within the explored neighborhood. This interaction structure is consistent with the screening results, which indicated both strong influence and non-negligible variability for these parameters.

7.9.7. Identification Performance Relative to Baseline Trust Predictors

This subsection compares the overall validation performance of the proposed trust dynamics model against the baseline families defined in Section 7.7. All methods are evaluated on the same measurement-aligned samples and scored using the weighted residual metrics defined in Section 7.5. Unless stated otherwise, reported values correspond to the validation participant split ($N_{\text{meas}} = 120$ aligned measurement samples).

Table 7.13: Overall baseline comparison on the validation set. All comparisons are made using the same number of measurements in each model ($N_{\text{meas}} = 120$).

Trust Dynamics Model	wRMSE	wMAE	Bias	wBias
Proposed model	0.1357	0.1070	-0.0124	-0.0091
Dispositional constant	0.1668	0.1187	-0.0387	-0.0445
Global TRAIN-mean constant	0.1778	0.1323	0.0531	0.0472
Oracle participant-mean constant	0.1126	0.0775	0.0058	≈ 0
Symmetric bump model	0.2956	0.2607	-0.1849	-0.2025
Asymmetric bump model ¹	0.1654	0.1250	-0.0349	-0.0350
OPTIMo-lite ²	0.1245	0.0979	0.0176	0.0208
OPTIMo-lite (outcome-only ablation)	0.1848	0.1544	0.0633	0.0962

¹ Wilmes and Jamshidnejad (2025).

² Xu and Dudek (2015); Guo and Yang (2021).

Table 7.13 summarizes overall metrics on the validation set. The proposed model achieves an overall validation error of $wRMSE = 0.1357$ and $wMAE = 0.1070$, with small negative bias ($wBias = -0.0091$). Relative to constant baselines, the model improves upon the dispositional-trust constant ($wRMSE = 0.1668$) and the global training-mean constant ($wRMSE = 0.1778$). Among event-driven baselines, the model improves upon the asymmetric bump model ($wRMSE = 0.1654$) and substantially outperforms the symmetric bump model ($wRMSE = 0.2956$). OPTIMo-lite attains a lower overall validation $wRMSE$ (0.1245) than the proposed model under the present evaluation setting. The OPTIMo-lite baseline differs from the proposed model in that it explicitly uses previously observed trust measurements for each participant as an input to the prediction process. When this input is removed, the ablated OPTIMo-lite variant consistently underperforms the proposed trust dynamics model across evaluation metrics.

Table 7.14: Validation $wRMSE$ by measurement kind.

Trust Dynamics Model	$q^{(1)}$	$q_{mid1}^{(14)}$	$q_{mid2}^{(14)}$	$q_{post}^{(40)}$
Proposed model	0.1378	0.1098	0.1909	0.1195
Dispositional constant	0.1486	0.1187	0.2245	0.1585
Asymmetric bump model ¹	0.1563	0.1106	0.2386	0.1500
OPTIMo-lite ²	0.1117	0.1136	0.1263	0.1283

¹ Wilmes and Jamshidnejad (2025).

² Xu and Dudek (2015); Guo and Yang (2021).

To verify that aggregate performance is not driven by a single measurement stream, Table 7.14 reports validation $wRMSE$ per measurement kind. The proposed model yields: (i) single probe-level² $wRMSE = 0.1378$ ($N_{meas} = 90$), (ii) mid-block questionnaire $wRMSE = 0.1098$ for the first 14-item questionnaire³ ($N_{meas} = 10$) and $wRMSE = 0.1909$ for the second 14-item questionnaire⁴ ($N_{meas} = 10$), and (iii) post-task questionnaire⁵ $wRMSE = 0.1195$ ($N_{meas} = 10$). OPTIMo-lite exhibits lower $wRMSE$ than the proposed model on single probes and $q_{mid2}^{(14)}$, while the proposed model attains lower $wRMSE$ than OPTIMo-lite on $q_{mid1}^{(14)}$ and $q_{post}^{(40)}$.

Table 7.15: Overall improvement of the proposed model relative to baselines on the validation set (positive values indicate lower error for the proposed model).

Baseline Model	Baseline $wRMSE$	Model $wRMSE$	$\Delta wRMSE$	% $\Delta wRMSE$
Dispositional constant	0.1668	0.1357	0.0311	18.6
Global TRAIN-mean constant	0.1778	0.1357	0.0421	23.7
Oracle participant-mean constant	0.1126	0.1357	-0.0231	-20.5
Symmetric bump model	0.2956	0.1357	0.1599	54.1
Asymmetric bump model ¹	0.1654	0.1357	0.0297	17.9
OPTIMo-lite ²	0.1245	0.1357	-0.0112	-9.0
OPTIMo-lite (outcome-only)	0.1848	0.1357	0.0491	26.6

¹ Wilmes and Jamshidnejad (2025).

² Xu and Dudek (2015); Guo and Yang (2021).

Table 7.15 reports absolute and percentage improvements of the proposed model relative to each baseline. Overall, the proposed model improves $wRMSE$ 18.6% relative to the dispositional constant and by 17.9% relative to the asymmetric bump model, while the outcome-only OPTIMo-lite ablation is worse than the proposed model by 26.6%. The proposed model only reports a lower $wRMSE$ score when compared to the OPTIMo-lite model at -9.0%. For reference, the oracle participant-mean constant yields a lower $wRMSE$ than the proposed model; this baseline is not deployable but provides an upper bound among constant predictors.

² Referenced as $q^{(1)}$. ³ Referenced as $q_{mid1}^{(14)}$. ⁴ Referenced as $q_{mid2}^{(14)}$. ⁵ Referenced as $q_{post}^{(40)}$.

7.9.8. Qualitative Trust Trajectory Predictions on Held-Out Participants

To complement the aggregate and participant-level error metrics reported above, this subsection presents representative examples of simulated trust trajectories on the validation split. For each selected participant, model predictions are compared against baseline trust predictors to illustrate qualitative differences in temporal behavior and event sensitivity.

The participants shown below are selected to reflect distinct regions of the participant-level validation error distribution: (i) a best-performing case (lowest validation $wRMSE$), (ii) a typical case (near-median validation $wRMSE$), (iii) a worst-performing case (highest validation $wRMSE$), and (iv) an additional representative case with mixed performance across measurement types. Due to the small set of validation participants, these trajectories also represent different door trial set conditions.

Best-performing Participant

Figure 7.16 shows simulated trust trajectories for the best-performing validation set participant across all baseline methods and the proposed model. For this participant, the proposed model closely tracks observed trust updates at the first mid-block point and at the final trust measurement, while underestimating between-doors measurements and the second mid-block measurement.

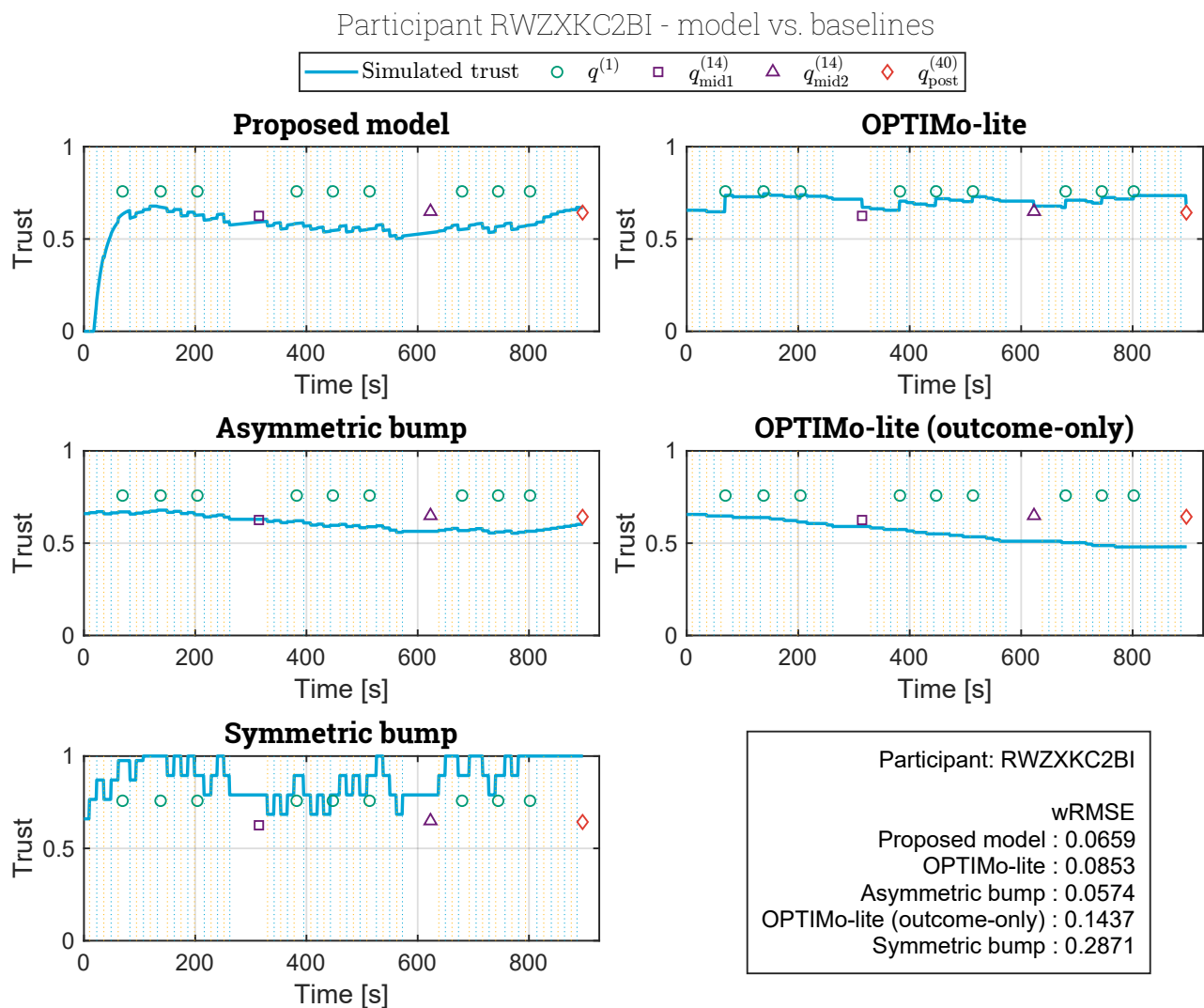


Figure 7.16: Best-performing validation set participant (noticeably low $wRMSE$): simulated trust trajectories for all baselines and the proposed model.

Typical Participant

Figure 7.17 illustrates trust trajectories for a participant whose validation wRMSE is close to the median of the validation distribution. The proposed model reproduces the overall trend of trust evolution and event-related deviations while remaining comparable to, or slightly better than, baseline predictors. In particular, it estimates trust measurements from the first and third door-trial blocks particularly well, while overestimating the final trust measurement.

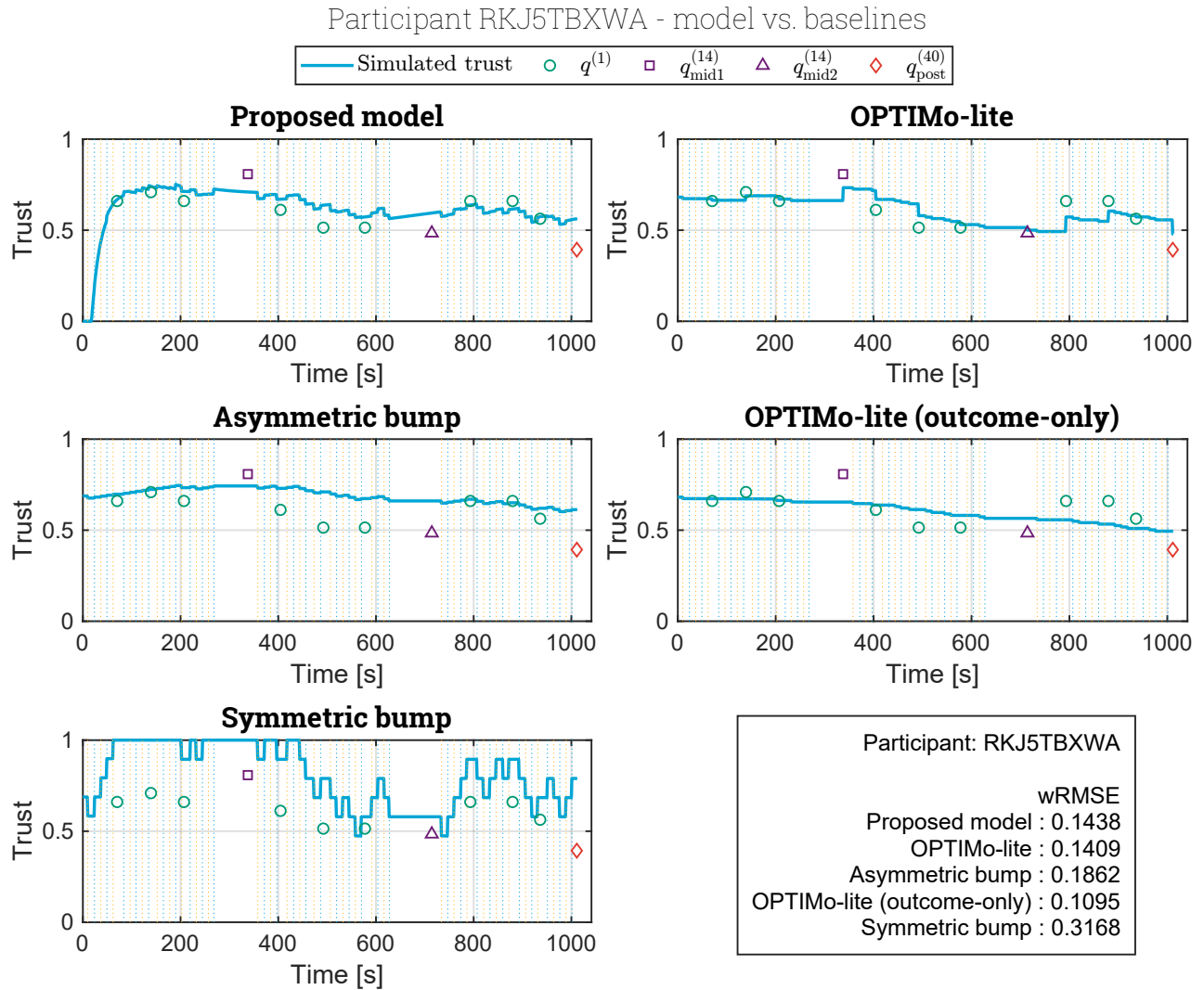


Figure 7.17: Typical validation set participant (near-median wRMSE): simulated trust trajectories for all baselines and the proposed model.

Worst-performing Participant

Figure 7.18 shows trust trajectories for the worst-performing validation participant in terms of $wRMSE$. In this case, all methods (including the proposed model) exhibit noticeable deviations from observed trust measurements, possibly related to the extreme variation in the trust measurements recorded for this participant.

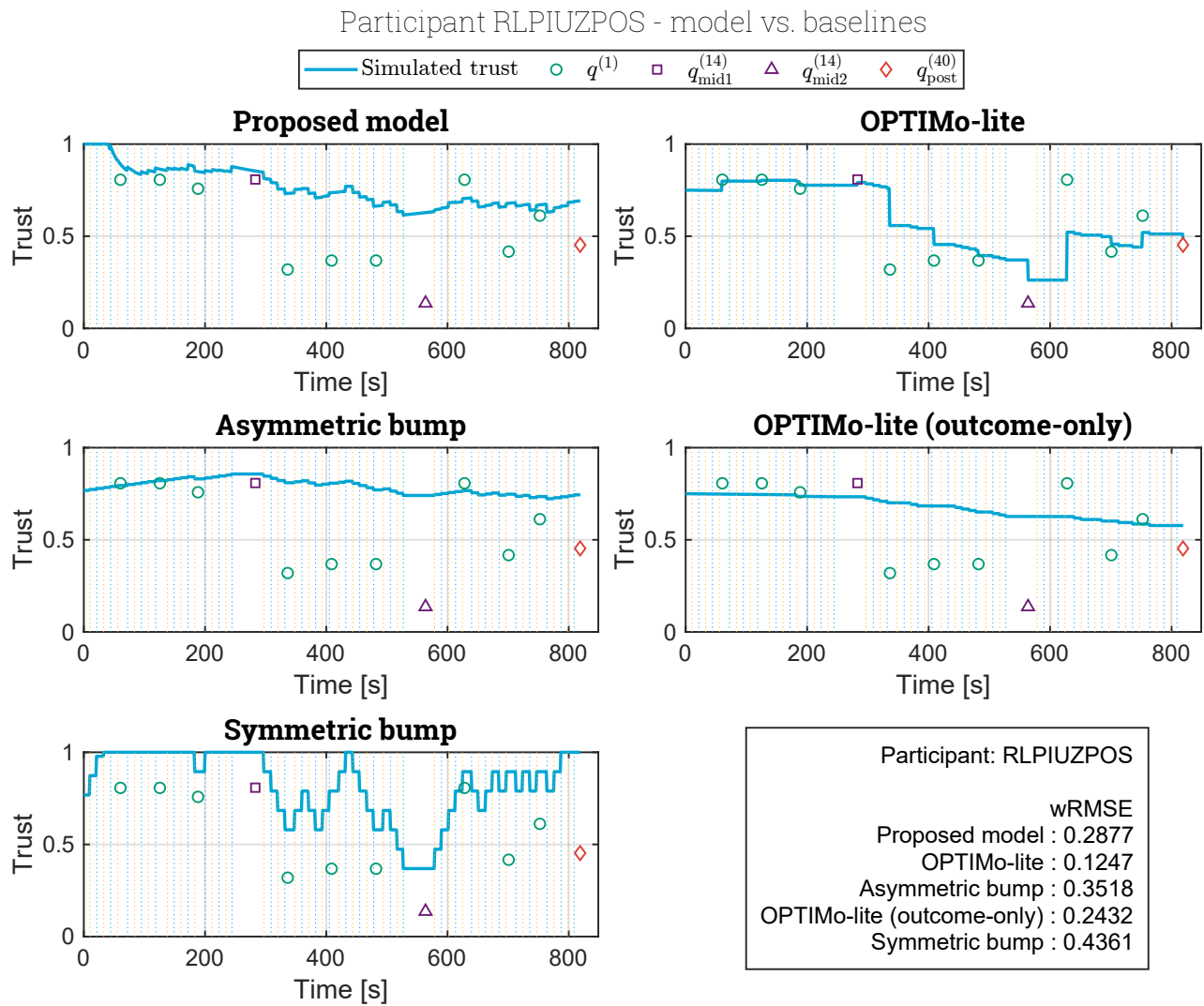


Figure 7.18: Worst-performing validation set participant (highest $wRMSE$): simulated trust trajectories for all baselines and the proposed model.

Additional Representative Participant

Finally, Figure 7.19 presents an additional representative participant exhibiting mixed performance across measurement types. While some measurement points are well captured by the proposed model, others show residual discrepancies comparable to those of baseline predictors. Moreover, the model does capture the general upward tendency of the trust measurements, particularly of mid-block and final measurements. This is particularly visible when we overlay the predicted trust trajectories for this participant using the proposed and OPTIMo-lite models, as shown in Figure 7.20.

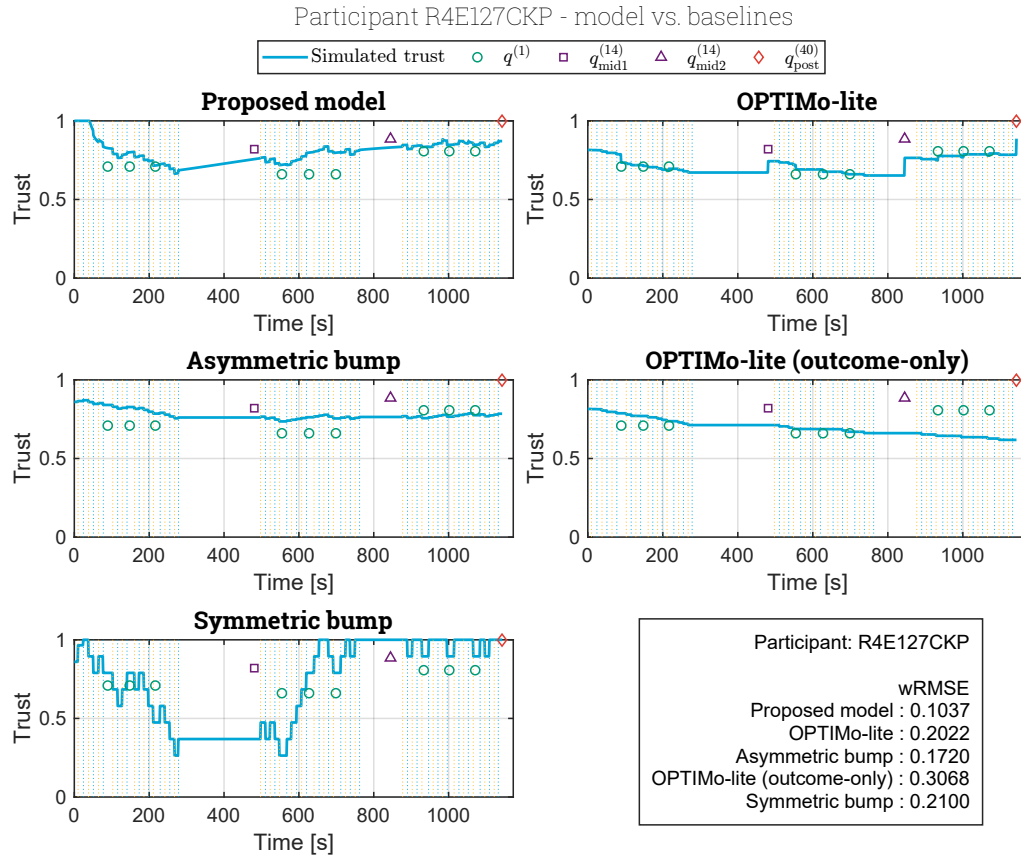


Figure 7.19: Additional representative validation set participant: simulated trust trajectories for all baselines and the proposed model.

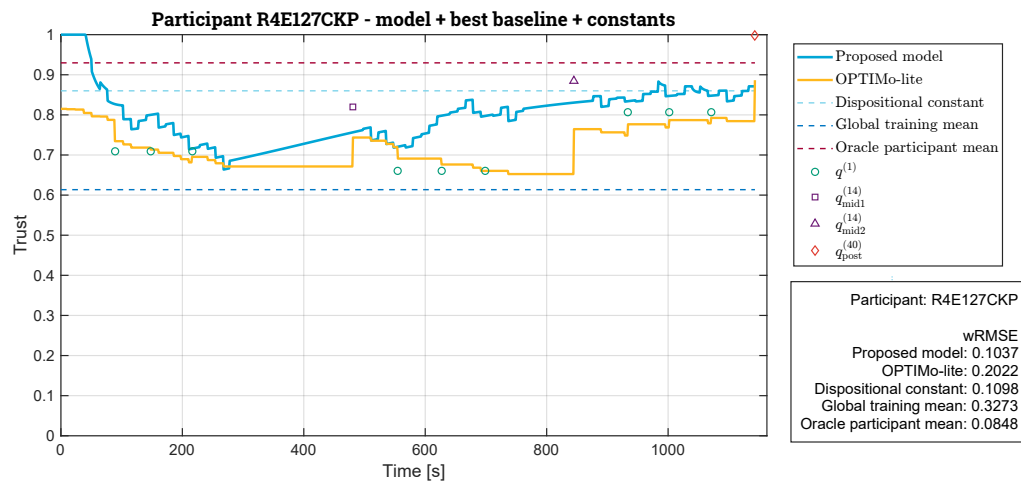


Figure 7.20: Additional representative validation set participant: proposed model versus best baseline.

7.9.9. Summary of Identification Results

This section presented the results of identifying the proposed trust dynamics model using experimentally collected data. Model performance was evaluated on a held-out validation set using multiple complementary metrics, including weighted error measures, bias indicators, and participant-level statistics.

At the aggregate level, the selected parameter estimate achieved consistently low validation errors, with validation performance comparable to or better than alternative candidate solutions obtained using different identification methods. Error distributions exhibited limited bias and no pronounced systematic trends across prediction magnitude or time, indicating stable overall model behavior.

Participant-level analyses revealed moderate to low inter-individual variability in validation performance. The majority of participants showed comparable error levels, suggesting that model performance was not dominated by a small subset of individuals. Qualitative trajectory comparisons further illustrated that the model is capable of capturing different trust evolution patterns across a range of participant behaviors, while also highlighting cases in which participant-specific dynamics remain difficult to reproduce with a shared-parameter model.

Robustness analyses demonstrated that multiple identification methods yielded parameter estimates with similar validation performance, indicating limited solver dependence and a well-behaved optimization landscape. Sensitivity and screening analyses identified a subset of parameters exerting dominant influence on model performance, while others exhibited comparatively weaker effects, supporting the practical identifiability of the most influential trust dynamics parameters.

Finally, comparative evaluations against baseline trust models showed that the proposed model provides competitive or improved validation performance across most metrics and measurement types, especially those operating without access to past trust measurements, while maintaining a unified dynamical structure. These results collectively indicate that the proposed trust dynamics model can be reliably identified from the available data and exhibits adequate generalization properties within the experimental context considered.

Collectively, these results provide quantitative and qualitative evidence supporting the generalization performance, robustness properties, and partial parameter identifiability of the proposed trust dynamics model under the considered experimental conditions.

7.10. Discussion of Trust Model Identification Results

This section discusses the identification and validation results presented in the previous section, with the aim of interpreting their implications for the empirical identifiability, generalization, and robustness of the proposed trust dynamics model. Rather than reiterating numerical outcomes, the discussion focuses on what the results establish about the structure of the identified model, which trust components are reliably supported by the available data, and under what conditions the model's predictions can be considered robust. The discussion is organized around four complementary perspectives: preliminary evidence of identifiability from the estimated parameter vector, practical identifiability and parameter interactions revealed through sensitivity analyses, out-of-sample generalization and error structure across participants and measurement types, and comparative performance assessed through baseline models and qualitative trust trajectory analysis.

7.10.1. Parameter Estimates and Identifiability Preliminaries

The estimated parameter vector reported in Table 7.6 provides a first indication that the proposed trust dynamics model can be identified under the imposed structural and bound constraints without degenerating to trivial or constraint-saturated solutions. All parameters remain strictly within their admissible ranges, suggesting that the weighted least-squares optimization does not rely on boundary effects to achieve predictive performance. This is a necessary, though not sufficient, condition for practical identifiability, as it indicates that the available data are informative enough to locate a feasible interior solution in the parameter space.

At the same time, the relative proximity of individual parameters to their bounds indicates heterogeneity in how strongly different components of the model are constrained by the data. In particular, parameters associated with the situational trust module exhibit consistently low proximity values, most notably the situational trust weighting parameter α^{sit} . This suggests that, within the present experimental and measurement regime, the contribution of situational trust to the integrated trust dynamics is weakly expressed in the identified parameterization. Importantly, this observation should not be interpreted as evidence that situational trust is intrinsically unimportant, but rather that its influence is either small relative to other modules or, more likely, that this module is insufficiently excited by the experimental conditions to be robustly learned. This situational trust result highlights the difficulties associated with modeling risk in ethically safe and controlled experimental

setups, where participants are not truly placed in danger.

More generally, weak parameter sensitivity in the present study should be interpreted as a statement about practical identifiability under the explored experimental regime, not as a statement about the conceptual relevance of the corresponding trust mechanism. Sensitivity analyses quantify how perturbations affect the validation cost within the tested context and for the selected measurement assumptions; they do not establish that a module cannot be influential under different operational conditions. The proposed formulation explicitly allows situational trust to modulate the integrated trust state as a function of contextual risk. However, because the experiment was necessarily constrained to remain ethically safe, the induced risk variation may not have been sufficiently strong or separable from event-driven effects to robustly constrain situational parameters. Consequently, situational trust remains a meaningful component of the model architecture even though it is only weakly informed by the present dataset.

Additionally, the bound proximity metrics should be interpreted as a coarse diagnostic rather than a definitive identifiability assessment. Parameters that lie well within their admissible ranges may still be subject to trade-offs or compensatory effects with other parameters, while parameters near bounds may reflect either genuine model behavior or limited excitation in the data. As such, these estimates primarily establish that the identification problem is well-posed and numerically stable, while motivating a deeper examination of parameter sensitivity and interaction effects to assess which parameters are practically identifiable and which remain weakly constrained.

From the perspective of Research Question 3.c, this result demonstrates that the proposed identification framework yields a coherent and structurally consistent parameterization, but it does not, by itself, resolve the identifiability of individual trust components. Instead, it sets the stage for subsequent sensitivity and interaction analyses, which are required to determine how variations in individual parameters influence model predictions and to distinguish genuinely identifiable parameters from those that are only weakly informed by the data.

7.10.2. Practical Identifiability and Parameter Interactions

The combined one-at-a-time perturbation analysis, Morris-like screening, and localized pairwise response surface provide complementary evidence regarding which parameters of the trust dynamics model are practically identifiable from the available data and which are only weakly constrained. As a whole, these results indicate a clear hierarchy in parameter influence on validation performance, with implications for both parameter learnability and model interpretation.

Across both local perturbation and neighborhood-based screening analyses, the event-driven trust update parameters ϕ^{fail} and ϕ^{succ} consistently emerge as the most influential contributors to validation error. Their dominance suggests that the data strongly constrain how trust responds to success and failure events, making these parameters practically identifiable under the present experimental conditions. This is a meaningful result, as these parameters directly govern the discrete trust update mechanism that links observable task outcomes to latent trust dynamics. This also aligns with the literature findings that direct experience and interaction outcomes are one of the most influential factors governing human-robot trust dynamics.

In contrast, parameters associated with situational modulation and reputation effects exhibit weak sensitivity across all analyses. The low influence of these parameters indicates that variations within their admissible ranges have little impact on predictive performance, suggesting that they are either weakly identifiable or play a secondary role relative to event-driven and latent dynamics in the current dataset. Importantly, this should not be interpreted as evidence that situational or reputational trust components are irrelevant in general, but rather that their effects are not strongly expressed or separable given the available measurements and experimental excitation. In particular, the situational trust component results are aligned with the preliminary analysis of the previous subsection. As for reputation components, this might be a limitation of the experimental setup. Since we define reputation as a fast decaying module of trust as personal experience accumulates, it is possible that it still plays a role in the trust dynamics, which is simply not captured by the time we reach the first trust measurement.

The Morris-like screening results further refine this interpretation by revealing not only parameter influence but also variability of influence across the local neighborhood. The elevated σ values observed for ϕ^{fail} and ϕ^{succ} indicate that their impact on model performance is not purely linear and may depend on interactions with other parameters. This motivates the pairwise response analysis, which explicitly examines coordinated variations in these parameters.

The resulting response surface exhibits a shallow valley rather than a sharply localized optimum, indicating that multiple combinations of ϕ^{fail} and ϕ^{succ} yield comparable validation performance. This structure implies partial parameter trade-offs: while the overall strength of event-driven trust updates is well constrained, the

precise allocation of influence between success and failure pathways is less uniquely determined. From an identifiability perspective, this suggests that these parameters are practically identifiable at the level of their combined effect on trust dynamics, but only partially separable from one another.

Overall, these findings indicate that the identification framework successfully recovers the dominant mechanisms governing trust evolution—particularly event-driven updates and latent dynamics—while also revealing limitations in the unique identification of secondary pathways. This distinction is central to Research Question 3.c: the model is identifiable in the sense that its predictive behavior is well constrained by the data, but not all parameters are equally learnable in isolation. The observed parameter hierarchy and interaction structure therefore reflect both the strengths and inherent limitations of identification under the considered experimental and measurement regime.

The comparison of candidate parameter estimates obtained using different identification methods provides insight into the robustness of the proposed identification framework with respect to estimation assumptions, as required by Research Question 3.d. Across all considered optimizers, validation performance remains within a narrow range, indicating that the predictive capability of the model is not contingent on a specific solver or optimization trajectory. This consistency suggests that the identification problem is numerically well behaved and that the selected solution is not an artifact of a particular optimization method.

At the same time, the observed variation in parameter vectors across optimization methods highlights an important distinction between robustness of predictive performance and uniqueness of parameter estimates. While all candidate solutions achieve comparable validation accuracy, their locations in parameter space differ non-negligibly. This indicates that multiple parameterizations can yield similar predictive behavior, reflecting the presence of compensatory effects or trade-offs among model parameters. Such behavior is common in dynamical models with interacting components and does not undermine predictive validity, but it does limit the extent to which individual parameters can be uniquely interpreted.

From an identifiability perspective, these results reinforce the conclusions drawn from the sensitivity and interaction analyses. Parameters that exert dominant influence on validation performance, such as the event-driven update gains, are likely constrained across candidate solutions, whereas weaker parameters may vary more freely without substantially affecting model predictions. The existence of multiple near-equivalent solutions therefore points to practical non-uniqueness in certain regions of the parameter space rather than instability of the identification procedure.

Importantly, this optimizer-level robustness supports the use of the selected parameter vector as a representative solution for subsequent analyses, including sensitivity screening and qualitative trajectory evaluation. The choice of a gradient-based constrained optimizer as the reference solution is justified by its favorable validation performance, but the broader result is that the model's predictive behavior is stable across estimation assumptions. In this sense, the robustness analysis strengthens confidence in the identified trust dynamics at the behavioral level, while simultaneously motivating a cautious interpretation of individual parameter values.

Overall, these findings address Research Question 3.d by demonstrating that the proposed identification approach yields consistent predictive performance across optimization methods, while also clarifying the limits of parameter uniqueness inherent to the model and dataset. This distinction between robust prediction and partial parameter non-identifiability is central to understanding both the strengths and limitations of the identified trust model.

7.10.3. Generalization Performance and Error Structure

The overall identification metrics on the training and validation sets provide the primary quantitative evidence that the proposed trust dynamics model generalizes beyond the participants used for parameter estimation. The close agreement between training and validation error levels indicates that the identification procedure does not result in strong overfitting to the training data. In particular, the absence of a pronounced error increase on the validation set suggests that the identified parameterization captures structural regularities in trust evolution that are shared across participants, rather than idiosyncratic patterns specific to the training cohort. At the same time, comparable aggregate errors do not imply that predictive performance is uniform across individuals or measurement modalities, motivating a more detailed examination of error structure.

Participant-level validation metrics reveal that most individuals cluster within a relatively narrow range of weighted prediction error, indicating broadly consistent model performance across the majority of the validation set. This concentration supports the interpretation that the identified dynamics generalize at the population level, rather than relying on a small subset of well-modeled individuals to achieve favorable aggregate metrics. However, the presence of a single participant with substantially higher validation error highlights an important limitation: while the model generalizes well on average, it does not perform uniformly across all

behavioral regimes. This outlier contributes disproportionately to the overall validation error and illustrates that certain trust trajectories (particularly those exhibiting extreme variability or apparent inconsistency) are not well captured by the assumed latent dynamics. Importantly, the existence of such cases does not invalidate the generalization claim, but does delimit its scope by emphasizing that predictive validity holds for most, but not all, participants. Moreover, this can also be the result of the limited training dataset. Having more experimental data would inherently increase the trust trajectory variations used in the model training phase, possibly allowing for it to capture more diverse trust trajectories in unseen individuals.

Stratification of performance by measurement type provides additional insight into how generalization manifests across different trust assessment instruments. Residual distributions across probes, mid-block questionnaires, and post-task questionnaires are all centered near zero, indicating limited systematic bias in the predictions. This suggests that the identified model does not consistently overestimate or underestimate trust for any particular measurement modality. Nevertheless, differences in residual dispersion across instruments are evident, reflecting heterogeneity in both measurement noise and temporal resolution. Mid-block and post-task questionnaires, which are fewer in number but semantically richer, exhibit different error characteristics than the more frequent probe measurements, which are inherently noisier and temporally localized. These results are expected, given the multi-resolution framework used for the trust measurements.

These differences underscore an important point for interpreting both quantitative metrics and qualitative plots: predictive accuracy should not be understood as uniform closeness to all trust measurements. Instead, the identification and evaluation framework explicitly accounts for the relative reliability of each measurement type through weighting. As a result, deviations at low-reliability probe measurements may contribute less to aggregate error than smaller deviations at questionnaire measurements. The residual diagnostics corroborate this interpretation by showing that increased dispersion in certain measurement types (such as single trust probes) manifests primarily as heavier tails rather than as systematic bias. This structure indicates that errors are driven by variability rather than by consistent model miscalibration.

Residual analyses further support the absence of gross model misspecification. The lack of strong trends in residuals as a function of predicted trust or time suggests that the model does not suffer from obvious state-dependent or temporal bias across the validation set. While isolated large residuals are present (consistent with the participant-level outlier discussed above), there is no evidence of widespread heteroscedasticity or drift that would indicate a failure of the underlying trust dynamics formulation. Instead, the residual structure is consistent with a model that captures the dominant temporal patterns of trust evolution while remaining subject to noise and unmodeled individual-specific effects.

Taken together, these results support Research Question 3.b by demonstrating that the proposed trust dynamics model achieves stable out-of-sample performance at the aggregate level, with generalization that extends across participants and measurement types. At the same time, they motivate caution against overly strong claims: generalization is robust in the average sense, but performance varies across individuals, and certain behavioral regimes remain challenging. This quantitative perspective provides essential context for the subsequent baseline comparisons and qualitative trajectory analyses, which further examine how predictive accuracy and model behavior manifest under different information assumptions and across representative individual cases.

7.10.4. Comparative Performance and Qualitative Trust Trajectories

The baseline comparisons provide an essential contextualization of the predictive performance of the proposed trust dynamics model and clarify the conditions under which its performance should be interpreted. Relative to constant predictors and simple event-driven baselines, the proposed model consistently achieves lower validation error, indicating that the identified trust dynamics capture structure beyond static dispositions or instantaneous outcome effects. Additionally, when comparing symmetric and asymmetric bump baselines, results show a clear advantage towards asymmetric trust gain and loss regimes. This further supports the proposed modeling choice of having different trust gain and loss equations for the personal experience module. These improvements demonstrate that the model successfully integrates temporal dynamics and event history in a way that yields meaningful predictive gains, directly supporting Research Question 3.b.

The comparison against OPTIMo-lite requires more careful interpretation, as this baseline operates under different information assumptions. OPTIMo-lite explicitly incorporates previously observed trust measurements for each participant as an input to the prediction process, effectively conditioning future trust estimates on past self-reports. Under this setting, OPTIMo-lite attains lower aggregate validation error than the proposed model. However, when this additional information channel is removed, the outcome-only ablation of OPTIMo-lite performs substantially worse than the proposed trust dynamics model. This contrast highlights

that OPTIMo-lite's advantage arises from access to historical trust measurements rather than from superior modeling of trust dynamics based solely on task outcomes and latent processes.

From a modeling perspective, this distinction is central. The proposed model operates without access to any prior trust measurements at inference time, relying exclusively on observed task events and internal state evolution. That its performance approaches that of a model with explicit trust feedback, while substantially outperforming comparable outcome-only baselines, indicates that the identified dynamics encode a nontrivial approximation of trust evolution in the absence of direct trust observations. This result supports the claim that the model generalizes meaningfully under realistic deployment constraints where continuous trust measurements are unavailable.

The oracle participant-mean constant further clarifies the performance envelope. While it yields lower error than the proposed model, it relies on participant-specific information that is unavailable in deployment and therefore serves only as a reference upper bound rather than a competing method. Its inclusion emphasizes that the proposed model operates in a more constrained information regime than several baselines, reinforcing the importance of interpreting performance in light of information accessibility rather than absolute error magnitude alone.

Stratified evaluation by measurement type further indicates that aggregate performance is not driven by a single trust measurement stream. While OPTIMo-lite achieves lower error on probe-level measurements and one mid-block questionnaire, the proposed model performs competitively or better on other measurement types, including post-task trust assessments. Given the limited sample size per condition, these differences should be interpreted cautiously; nevertheless, they suggest that the relative strengths of the models depend on the temporal and semantic characteristics of the trust measurement rather than reflecting uniform dominance by any single approach.

Overall, the baseline comparisons support Research Question 3.b by demonstrating that the proposed model achieves robust predictive performance relative to deployable baselines under comparable information assumptions. At the same time, they delimit the scope of the contribution by showing that models with direct access to historical trust measurements can outperform purely outcome-driven approaches.

Beyond scalar error metrics, the qualitative shape of predicted trust trajectories is an important indicator of whether a model behaves plausibly as a latent cognitive state. In particular, trust is typically assumed to exhibit inertia: in the absence of salient interaction events, it should evolve gradually rather than oscillate rapidly between successive time steps. Several baseline predictors can achieve competitive aggregate error while producing comparatively "bumpy" trajectories, i.e., high-frequency fluctuations that are mathematically admissible but cognitively less interpretable and potentially indicative of overreacting to noisy measurements. By contrast, the selected trust dynamics model tends to generate smooth latent evolution punctuated by discrete event-driven updates, aligning with the modeling premise of a slowly varying internal belief modulated by task outcomes. This qualitative smoothness is therefore a strength of the proposed formulation, as it improves interpretability and yields a more stable substrate for downstream behavioral coupling and closed-loop simulation.

Figures 7.16–7.20 provide a qualitative view of out-of-sample trust prediction for four held-out participants, spanning best-case performance, near-median performance, worst-case performance, and an additional representative mixed case. These plots serve a distinct role relative to aggregate metrics: they allow visual inspection of how each method generates temporal trust dynamics, whether changes are smooth or abrupt, whether the timing of deviations aligns with observed trust reports, and which portions of the trial produce systematic residual errors. In this sense, the trajectory figures are an important complement to $wRMSE/wMAE$ summaries, because they make the model behavior interpretable at the level of time evolution rather than as a single scalar score.

However, the interpretability of these figures comes with a major caveat: visual closeness of a simulated trajectory to all observed trust measurements can be misleading if the differing uncertainty levels (and corresponding weights) of the measurement modalities are not kept in mind. The model is identified and evaluated under a weighted error metric in which end-of-task questionnaire measurements are treated as the most reliable, mid-block questionnaire measurements are weighted substantially lower than end-of-task (but still meaningfully), and individual probe reports carry much smaller weight than either questionnaire type. Consequently, a trajectory may appear to "miss" many probe points while still achieving low weighted error, because these probe measurements are modeled as comparatively noisy indicators. Conversely, a trajectory that visually tracks probe-level fluctuations but misses questionnaire points may appear qualitatively appealing while being penalized more strongly under the adopted evaluation assumptions. These figures should therefore be interpreted as illustrating the temporal structure of predictions under a measurement hierarchy, rather than as

an unweighted measure of closeness to every plotted marker.

This measurement hierarchy is particularly relevant for understanding why some apparent discrepancies do not necessarily imply model failure. For example, in the best-performing participant (Figure 7.16), the proposed model aligns closely with the mid-block and final questionnaire measurements, while remaining systematically below many probe measurements and underestimating the second mid-block questionnaire. Under the adopted weighting, accurate alignment at the most reliable measurement points can dominate the overall score even if probe-level deviations persist. This case illustrates the intended behavior of the identification objective: the model prioritizes capturing the major trust state reflected in questionnaires, while treating probe reports as low-confidence, high-variance observations that are informative in aggregate but not deterministically followed at every instance. Importantly, this does not imply that probe information is ignored; rather, it clarifies that the model is not incentivized to match probe-level “noise” at the expense of reliable questionnaire anchors.

The typical participant case (Figure 7.17) reinforces this interpretation under near-median performance. Here, the proposed model reproduces the overall trend of trust evolution and exhibits event-related deviations that are qualitatively plausible and comparable to baseline predictors. Visually, it tracks measurements from earlier blocks more closely while deviating at the final questionnaire point, consistent with the notion that residual bias can manifest at high-weight measurements and therefore contribute significantly to $wRMSE$ even if intermediate behavior appears reasonable. This case is valuable because it demonstrates that the identified dynamics do not only perform well in cherry-picked low-error scenarios: the model continues to produce structured temporal evolution that remains competitive relative to baselines in a participant closer to the central tendency of validation performance.

The worst-performing participant (Figure 7.18) provides an important counterpoint and establishes the boundaries of what the model can capture. In this case, all methods (including the proposed model and baselines) exhibit substantial deviations from observed trust measurements. The defining characteristic of this participant is the extreme variability and apparent inconsistency of the recorded trust reports over time. Such behavior may reflect idiosyncratic response patterns, intermittent disengagement, inconsistent internal scaling of trust reports, or other unmodeled cognitive factors. The fact that multiple modeling assumptions fail simultaneously is informative: it suggests that this scenario is not a localized deficiency of a particular estimator or model family, but rather a regime in which the available observations do not support a coherent low-dimensional trust trajectory under the assumed model structure and measurement interpretation. From the perspective of Research Question 3.b, this case appropriately limits claims of predictive validity to participants for whom trust reports behave consistently enough to be represented as a smooth latent process with event-driven updates. From the perspective of Research Question 3.d, it also underscores that robustness is not equivalent to universal fit: stable performance across methods does not prevent systematic failure in outlier behavioral regimes.

The additional representative participant (Figures 7.19 and 7.20) illustrates mixed performance across measurement modalities within a single trial. In this case, the proposed model captures an overall upward tendency that is most clearly reflected in mid-block and final questionnaire measurements, while still exhibiting residual discrepancies at other measurement points that are comparable to those of baseline predictors. The overlay comparison against the best baseline (Figure 7.20) is particularly informative because it emphasizes that competitive scalar performance can correspond to qualitatively different trajectory shapes. In other words, two methods may achieve similar aggregate error while embodying different assumptions about how trust evolves between questionnaire anchors. This highlights a key benefit of the proposed model: it generates a structured, mechanistically interpretable trajectory rather than a purely interpolative or memory-driven estimate, while still remaining competitive with strong baselines under the adopted information constraints.

Collectively, the four trajectory cases highlight several consistent qualitative patterns. First, the proposed model tends to produce smooth latent evolution punctuated by event-driven adjustments, which aligns with the modeling premise of a slowly varying trust state modulated by discrete outcomes. Second, the largest residual errors often occur in regions where probe measurements exhibit high variability, suggesting that a significant portion of the apparent mismatch is attributable to observation noise rather than to gross model misbehavior. Third, the final questionnaire measurement plays a disproportionately important role in perceived performance due to its high reliability weight; therefore, over- or underestimation at the end-of-task point can dominate error even when intermediate behavior appears qualitatively reasonable. These patterns clarify that qualitative inspection should be anchored in the same uncertainty-aware perspective used in quantitative evaluation.

Finally, it is worth noting that an exploratory robustness check was performed by repeating the simulations and identification under an alternative probe weighting scheme that treated probe measurements as substantially more informative than in the calibrated weighting (following the same proportional rule used for mid-block questionnaires). Under this exploratory setting ($w_{40} = 1$, $w_{14} = 14/40$ and $w_1 = 1/40$), validation metrics and identified parameter values remained broadly similar (results omitted for brevity). While this provides some indication that the identification results are not highly sensitive to modest changes in probe weighting, the interpretation of such robustness must be made cautiously. Similar aggregate metrics under different weights primarily indicate that multiple weighting assumptions can lead to comparably good fits under their respective objectives; this does not necessarily imply that the resulting trajectories are objectively “closer” to the true underlying trust process, particularly when the probe reports remain noisy and when increasing their weight does not change the underlying measurement uncertainty. As such, these qualitative plots should be viewed as a transparent window into model behavior under explicit measurement assumptions, rather than as definitive evidence that the simulated trajectory approximates all trust measurements equally well.

In summary, the trajectory figures substantiate the quantitative validation results by showing that the proposed model can generate coherent, interpretable trust dynamics that align well with high-reliability measurement points for both best-case and typical participants, while also revealing realistic failure modes in outlier cases. They also make explicit the practical consequence of the adopted measurement hierarchy: qualitative closeness must be interpreted through the lens of measurement uncertainty and the corresponding weighting used in identification and evaluation.

7.10.5. Limitations and Scope of the Identification Results

The results presented in this chapter should be interpreted in light of several important limitations that define the scope of the identification and validation claims. These limitations are inherent to the experimental setting, measurement assumptions, and modeling objectives, and do not detract from the internal consistency of the results; instead, they clarify the conditions under which the conclusions of Research Question 3 hold.

A primary limitation of the present study is the relatively small number of participants available for identification and validation. While this constraint reflects practical considerations common to controlled human-subject experiments, it necessarily limits the diversity of observed trust trajectories and behavioral regimes. As a result, the identified parameters and validation results should be understood as characterizing population-level trust dynamics within the explored experimental conditions, rather than as providing exhaustive coverage of all possible trust behaviors. At the same time, this limitation also highlights a key strength of the proposed approach: despite the limited sample size, the identification framework yields stable parameter estimates, meaningful out-of-sample generalization, and coherent trust trajectory predictions for most participants. This indicates that the proposed model structure and estimation pipeline are sufficiently informative and data-efficient to support empirical identification under realistic experimental constraints.

A second limitation concerns parameter identifiability. The analyses demonstrate practical identifiability of dominant behavioral mechanisms, particularly event-driven trust updates and latent dynamics, but do not establish theoretical or structural identifiability of all individual parameters. Several parameters are shown to be weakly constrained or identifiable primarily through their combined effects. This reflects both the interacting nature of the model and the level of excitation provided by the experimental design. Consequently, parameter values should be interpreted as components of a validated predictive model rather than as unique or directly interpretable psychological quantities.

A further limitation concerns the absence of direct ground-truth observations of human trust, which is not specific to this study but inherent to empirical human-robot interaction trust research more broadly. Trust is a latent cognitive construct that can only be accessed indirectly through self-reports, behavioral proxies, or performance-related indicators, all of which are subject to noise, bias, and context dependence. Consequently, identification and validation of trust models necessarily rely on explicit assumptions about measurement reliability and interpretation, rather than comparison against an observable ground truth. The present work addresses this limitation by adopting a transparent measurement model with explicitly reported score mappings and a clearly defined weighting scheme (with the 40-item questionnaire as reference, a conservative proportional weight for the 14-item instrument, and a calibrated weight for the single-item probes) and by interpreting both quantitative metrics and qualitative trajectories within that framework. As such, the reported results should be understood as validating the internal consistency and predictive utility of the proposed trust dynamics model under explicit measurement assumptions, rather than as recovering a noise-free or directly observable ground truth for human trust.

In light of these considerations, the limitations discussed above delineate the scope of the present contribution while reinforcing its methodological value. Beyond the specific numerical results, this chapter validates a complete identification and validation pipeline for dynamic trust modeling that can be reused, extended, and refined in future studies with larger participant samples, richer excitation, or alternative measurement strategies. In this sense, the primary contribution lies not only in the identified parameter values, but in the demonstrated viability of the proposed modeling and identification approach.

7.11. Synthesis and Implications for Research Question 3

This section synthesizes the contributions of chapter 7 and explicitly addresses the third research question posed in this thesis, which concerns the empirical identification and validation of the proposed trust dynamics model using human-robot interaction data.

Research Question 3: *To what extent can the proposed trust dynamics model be empirically identified and validated?*

- RQ3.a** How can an identification strategy be formulated that enables the estimation of shared trust model parameters while accounting for inter-individual variability and supporting generalization beyond the training data?
- RQ3.b** Can the model reproduce observed trust trajectories across participants when evaluated on held-out validation data?
- RQ3.c** Which model parameters are consistently identifiable from the available data, and which exhibit weak identifiability or high uncertainty?
- RQ3.d** How robust are the identified parameters and model predictions to variations in experimental conditions, measurement subsets, and estimation assumptions?

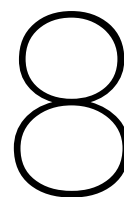
Identification strategy with shared parameters and generalization support (RQ3.a). This chapter establishes an identification strategy that estimates a single set of shared trust model parameters while explicitly evaluating generalization beyond the training data. The identification problem is formulated as a constrained parameter estimation task applied to a pooled dataset, with shared dynamics assumed across participants and inter-individual variability expressed through differences in observed trust trajectories rather than individualized parameters. Generalization is explicitly enforced and assessed by separating training and validation participant sets, ensuring that parameter estimation is driven by population-level structure rather than participant-specific overfitting. The resulting framework enables empirical identification of shared trust dynamics while remaining compatible with realistic sample sizes and experimental constraints.

Out-of-sample reproduction of trust trajectories (RQ3.b). The validation results demonstrate that the identified trust dynamics model can reproduce observed trust trajectories across participants when evaluated on held-out data. Quantitative metrics show comparable performance on training and validation sets, indicating that predictive accuracy is preserved out of sample. Participant-level analyses further reveal that most validation participants exhibit moderate and comparable prediction errors, while qualitative trajectory visualizations illustrate that the model generates coherent trust evolution patterns for best-performing, typical, and representative participants. Although certain participants exhibit substantial deviations that challenge all evaluated models, these cases bound the generalization claims without undermining the overall ability of the model to reproduce trust trajectories at the population level.

Parameter identifiability and uncertainty structure (RQ3.c). Sensitivity analyses, screening methods, and interaction studies provide insight into which model parameters are consistently identifiable from the available data. Event-driven trust update parameters and latent trust dynamics parameters emerge as the most influential and well-constrained components, indicating that the data strongly inform how trust responds to task outcomes and evolves over time. In contrast, parameters associated with situational modulation and reputation effects exhibit weak sensitivity and proximity to boundary values, suggesting limited identifiability or high uncertainty under the present experimental conditions. Importantly, this weak identifiability should be interpreted as context-limited excitation rather than as evidence that these modules are structurally irrelevant to the proposed model. Interaction analysis further reveals partial trade-offs among influential parameters, indicating that identifiability is achieved primarily in terms of combined behavioral effects rather than unique recovery of all individual parameter values.

Robustness of identified parameters and predictions (RQ3.d). The robustness analyses demonstrate that both the identified parameters and resulting model predictions are stable under a range of estimation assumptions. Comparable validation performance is obtained across multiple optimization methods, indicating that the identification results are not artifacts of a particular solver or initialization. While the estimated parameter vectors exhibit non-negligible variation across identification runs, predictive performance remains consistent, reflecting practical non-uniqueness rather than instability. Additional robustness considerations arise from baseline comparisons and qualitative trajectory analyses, which show that the proposed model remains competitive with deployable alternatives operating under similar information constraints. Exploratory analyses further suggest limited sensitivity to moderate variations in measurement weighting, although such robustness is interpreted cautiously given the inherent noise and uncertainty in trust measurements.

Taken together, the identification, validation, and robustness results presented in this chapter provide a principled and qualified answer to Research Question 3. The findings demonstrate that the proposed trust dynamics model can be empirically identified using shared parameters, reproduces trust trajectories across unseen participants, and yields interpretable insight into which trust mechanisms are strongly supported by the data and which remain weakly constrained. At the same time, the results clarify the limits imposed by inter-individual variability and measurement uncertainty, establishing a realistic foundation for the subsequent use of the identified model in simulation-based evaluation and closed-loop human-robot interaction studies.



Trust-Aware Behavioral Model

This chapter extends the trust modeling and identification pipeline developed in chapters 5 through 7 by introducing a minimal behavioral modeling layer that connects inferred trust states to observable reliance decisions in the experimental task. While the Trust Dynamics Model is evaluated primarily through its ability to reproduce reported trust measurements over time, trust-aware interaction analysis also requires a principled way to relate latent trust to discrete actions. The purpose of this chapter is therefore to explore, under deliberately simplified assumptions, whether the trust trajectories inferred by the identified model exhibit consistent explanatory signal for reliance behavior under held-out conditions, without aiming to provide a comprehensive or psychologically complete account of human decision-making.

The chapter first clarifies the behavioral layer as an auxiliary, exploratory probe that links inferred trust to reliance decisions without treating behavioral performance as validation of the upstream trust dynamics. It then introduces a compact family of nested trust-to-action mappings within a shared probabilistic decision framework, defines a decision-level dataset with strict causal alignment by pairing each decision with pre-decision inferred trust, and evaluates predictive performance under a participant-level train/validation split using likelihood-based estimation, proper scoring rules, and trust-free baselines. The analysis then shifts to descriptive individual differences through per-participant mechanism selection with identifiability and robustness diagnostics, before moving to closed-loop rollouts that impose an explicit behavior–outcome coupling to stress-test the interaction patterns induced by loop closure using override-centric signatures and distributional comparisons. Finally, trajectory-level sanity checks and guardrails quantify how coupling perturbs trust relative to replay and document when personalized behavioral parameterizations are unreliable, after which results are consolidated, interpreted diagnostically, and synthesized in relation to Research Question 4.

8.1. Purpose and Scope of Trust-Aware Behavioral Modeling

The Trust Dynamics Model developed and tested in the previous chapters estimates a continuous latent trust state that evolves over time as a function of interaction history and contextual factors. While trust is widely understood as a central determinant of reliance in human-robot interaction, it is not itself an observable behavior. Instead, trust is commonly interpreted as an internal attitude that influences decisions, without uniquely determining them. As a result, validating a trust dynamics model requires a clear conceptual separation between the estimation of latent trust and the modeling of observable reliance behavior.

The purpose of this chapter is to introduce a minimal, trust-aware behavioral modeling layer that maps the inferred trust state to observable human actions in the experimental task. This behavioral layer serves as an auxiliary component of the overall modeling pipeline. It is not intended to provide a comprehensive or psychologically complete model of human decision-making. Rather, it functions as a controlled probe to assess whether the inferred trust state carries sufficient information to explain key patterns of reliance behavior observed in the data.

A critical distinction is therefore maintained throughout this chapter. The Trust Dynamics Model is responsible for predicting the evolution of an internal cognitive state, and its validation is grounded in its ability to reproduce reported trust measurements, temporal structure, and cross-participant regularities. The Behavioral Model, by contrast, specifies a probabilistic relationship between this latent trust state and discrete reliance decisions. Performance of the behavioral model should not be interpreted as a direct measure of the

accuracy of the trust dynamics model. Limited behavioral predictability reflects the inherently multifactorial nature of human decision-making, in which trust is a necessary but not sufficient determinant of action.

Within this framing, behavioral modeling is used in three complementary ways. First, it enables a principled evaluation of whether trust estimates derived from the Trust Dynamics Model exhibit predictive signal for observed reliance decisions under held-out conditions. Second, it allows the exploration of individual differences in how trust translates into behavior, highlighting heterogeneity in reliance strategies across participants. Third, when embedded in closed-loop simulations, it provides a means to assess whether the coupled trust–behavior system generates interaction patterns that are qualitatively consistent with those observed in human-robot interaction.

Throughout this chapter, the behavioral models are deliberately kept simple and interpretable, and multiple candidate mappings are considered rather than a single definitive model. This design choice reflects both the limited observability of human decision variables in the present dataset and the broader objective of isolating the explanatory role of trust within a larger, partially unobserved decision process. Claims of behavioral realism are therefore conditional on the stated modeling assumptions, and the results of the behavioral analyses are interpreted as diagnostic rather than confirmatory evidence of trust model validity.

Accordingly, failure of a given trust-to-action mapping to predict or reproduce behavior should be interpreted not as failure of the trust dynamics model itself, but as evidence that trust alone is insufficient to explain action under the tested assumptions.

8.2. Candidate Trust-to-Action Mappings

To examine how inferred trust states relate to observable reliance decisions, this chapter considers a small family of trust-to-action mappings with increasing expressive power. Rather than committing to a single behavioral theory, multiple candidate mappings are introduced to probe different hypotheses about the role of trust in decision-making. Each mapping defines a probabilistic relationship between the latent trust state and the likelihood of following the robot’s recommendation, while deliberately abstracting away many other unobserved cognitive factors.

The guiding design principle is parsimony. The behavioral mappings are intentionally simple, interpretable, and nested, allowing differences in predictive or generative performance to be attributed to specific modeling assumptions. Collectively, these mappings span a spectrum ranging from a minimal diagnostic baseline to a more flexible formulation that accommodates individual variability and decision noise.

All mappings considered in this chapter share a common probabilistic decision framework, which is introduced first. The subsequent subsections then describe each trust-to-action hypothesis in turn.

8.2.1. General Probabilistic Decision Framework

At each decision point, the human agent produces a binary action indicating whether the robot’s recommendation is followed or overridden. Let $y \in \{0, 1\}$ denote the observed action, where $y = 1$ corresponds to following the recommendation and $y = 0$ to overriding it. The behavioral model specifies the probability of following as a function of the inferred trust state and a small number of interpretable parameters:

$$p = \Pr(y = 1 \mid \cdot). \quad (8.1)$$

To accommodate variability in human decision-making and to enable statistical inference, the action is modeled probabilistically rather than deterministically. The observed decision is treated as a Bernoulli random variable with parameter p , reflecting the fact that similar internal states may occasionally lead to different actions due to noise, uncertainty, or unmodeled influences.

Across all candidate mappings, the probability of following is expressed using a smooth sigmoidal function. This choice replaces a hard decision threshold with a continuous transition, capturing graded sensitivity near points of indifference and yielding a well-behaved likelihood function for subsequent analysis. Differences between the candidate mappings arise from how trust, self-confidence, and additional variability are incorporated into the argument of this sigmoid.

8.2.2. Trust-as-probability Mapping

The first mapping treats the inferred trust state itself as a direct proxy for the probability of following the robot’s recommendation. Under this hypothesis, trust is assumed to be numerically calibrated such that higher trust corresponds directly to a higher likelihood of reliance, without the need for additional thresholds or transformation parameters.

This mapping is intentionally strong and minimal. It introduces no free parameters and asserts that the trust dynamics model already produces a quantity that can be interpreted probabilistically. As such, it serves primarily as a diagnostic baseline: if even this direct interpretation of trust fails to explain observed actions, more structured mappings are unlikely to succeed without introducing additional assumptions.

While psychologically simplistic, this mapping provides a useful reference point for evaluating whether inferred trust values align at all with observed reliance frequencies.

8.2.3. Thresholded Reliance Mapping

The second mapping introduces an explicit decision threshold by modeling reliance as a function of trust relative to the human operator's self-confidence. In this formulation, trust does not directly encode the probability of following; instead, the likelihood of reliance depends on whether trust exceeds an internal reference level that reflects how much confidence the individual requires before accepting robotic guidance.

This hypothesis is motivated by the idea that trust must surpass a subjective threshold before influencing behavior, and that this threshold may vary across individuals. A single sensitivity parameter controls how sharply the probability of following transitions as trust crosses this reference level, allowing behavior to range from nearly deterministic switching to highly stochastic decision-making.

Compared to the trust-as-probability mapping, this formulation relaxes the assumption that trust is numerically calibrated while retaining a high degree of interpretability. It captures the notion that trust guides behavior primarily through relative comparisons rather than absolute values.

8.2.4. Thresholded Reliance with MODulation and Lapse

The third mapping further extends the thresholded formulation by allowing additional modulation and decision noise. In this variant, the influence of trust on reliance can be scaled or shifted by individual characteristics, and a lapse component is introduced to account for occasional trust-independent decisions.

The lapse term represents unmodeled factors such as momentary distraction, misunderstanding, or strategic deviations that cause the observed action to deviate from what trust alone would predict. Importantly, this component does not negate the role of trust; instead, it acknowledges that even high or low trust states do not deterministically fix behavior.

This formulation is the most flexible of the three and therefore carries the highest risk of overfitting. For this reason, it is treated as an exploratory extension rather than a definitive behavioral theory. Its inclusion allows assessment of whether modest increases in behavioral expressiveness meaningfully improve explanatory power or generative realism, relative to simpler trust-to-action mappings.

8.3. Behavioral Dataset Construction and Causal Alignment

Behavioral modeling in the present context requires pairing discrete human reliance decisions with the internal trust state that precedes them. Because trust evolves continuously over time while actions occur at specific decision events, careful temporal alignment is necessary to ensure that trust is used as a legitimate predictor of behavior rather than as a post-hoc correlate. This section defines the construction of a decision-level behavioral dataset that satisfies this causal requirement and serves as a common foundation for all subsequent behavioral analyses.

Each interaction in the experimental task contains a finite set of decision events at which the human participant chooses whether to follow or override the robot's recommendation. These decision events constitute the atomic units of behavioral analysis. For each event, the observed action is represented as a binary variable indicating whether the recommendation was followed. No assumptions are made at this stage about the cognitive processes underlying this decision, nor about the form of the trust-to-action mapping to be applied later.

To preserve causal consistency, the trust value associated with a given decision event is defined as the inferred trust state immediately prior to that event. Specifically, trust is evaluated at the last time point preceding the presentation of the robot's recommendation, before any outcome information related to the decision becomes available. This choice ensures that the trust variable reflects the participant's internal state at the moment of decision, uncontaminated by the consequences of that decision. Using post-decision or post-outcome trust values would artificially inflate apparent predictive performance and is therefore avoided.

In addition to the inferred trust state, each decision event is associated with a small set of contextual covariates that are directly observable in the experimental task. These include the human's self-reported confidence, task-specific risk indicators, and the robot's recommendation at that event. These variables are

retained without transformation to allow later behavioral models to condition on them explicitly or to marginalize over them as needed. Importantly, the dataset construction itself does not encode any behavioral assumptions about how these covariates interact with trust.

The resulting behavioral dataset is thus defined at the level of individual decision events, with each entry consisting of a causally aligned trust estimate, an observed reliance action, and a consistent set of contextual descriptors. This dataset is shared across all behavioral analyses in this chapter, including predictive evaluation, individual-level modeling, and closed-loop simulation. Differences between these analyses arise from modeling choices rather than from differences in data construction, ensuring that comparisons across behavioral hypotheses remain meaningful and internally consistent.

8.4. Behavioral Model Estimation and Probabilistic Evaluation

This section evaluates whether the inferred trust trajectories provide predictive signal for reliance decisions under held-out conditions, conditional on the candidate trust-to-action mappings introduced in section 8.2. The objective is not to validate a comprehensive model of human decision-making, but to assess the extent to which a trust-derived latent state is sufficient to explain observed follow/override behavior when embedded in simple probabilistic decision rules. As emphasized in section 8.1, behavioral predictability should not be interpreted as a direct measure of trust estimation accuracy.

8.4.1. Training and Evaluation Protocol

Behavioral mappings are evaluated under a strict separation between parameter estimation and performance assessment. Parameters (where applicable) are estimated using a training participant set, and predictive performance is then evaluated on a disjoint validation participant set. This separation ensures that reported performance reflects generalization across individuals rather than within-participant curve fitting.

To contextualize performance, two trust-free baseline predictors are included. First, a global base-rate predictor uses the training-set follow frequency as a constant follow probability for all validation trials. Second, an uninformed predictor uses a constant probability of 0.5. These baselines provide references for interpreting whether trust-conditioned mappings yield meaningful improvements over simple heuristics.

8.4.2. Likelihood-based Parameter Estimation

For mappings that include free parameters, estimation is performed by maximum likelihood under a Bernoulli decision model. Let $y \in \{0, 1\}$ denote the observed action at a decision event, with $y = 1$ indicating follow and $y = 0$ indicating override. For a predicted follow probability $p \in (0, 1)$, the negative log-likelihood (NLL) over a set of N decision events is

$$\text{NLL} = - \sum_{d=1}^N [y_d \log p_d + (1 - y_d) \log(1 - p_d)]. \quad (8.2)$$

This objective yields probabilistically meaningful parameter estimates and supports comparison across models using proper scoring rules.

The trust-as-probability mapping contains no free parameters and therefore requires no estimation. The thresholded and lapse-augmented mappings are fit on the training set only, producing fixed parameter values that are applied unchanged on the validation set.

8.4.3. Probabilistic Evaluation Metrics

Evaluation is primarily probabilistic rather than purely classificatory. Since the behavioral mappings produce probabilities, performance is assessed using proper scoring rules that reward both correctness and calibration.

Two primary probabilistic metrics are reported on the validation set:

- **Negative log-likelihood (NLL):** reported as both total NLL and mean NLL (per decision event). Lower values indicate better probabilistic prediction.
- **Brier score:** the mean squared error between predicted follow probabilities and binary outcomes. Lower values indicate better probabilistic accuracy and calibration.

To assess whether predicted probabilities correspond to empirical frequencies, calibration is evaluated via reliability binning on the validation set. Predicted follow probabilities are grouped into bins on $[0, 1]$, and for each bin the mean predicted probability is compared against the empirical follow rate. Uncertainty in the empirical rate is summarized using confidence intervals for binomial proportions. A scalar summary of

miscalibration is reported using the expected calibration error (ECE), which aggregates absolute bin-wise deviations weighted by bin occupancy.

For completeness and interpretability, a simple thresholded accuracy is reported by converting probabilities to hard decisions at $p \geq 0.5$. Because accuracy discards probability information and can be misleading under class imbalance, it is treated as secondary to NLL and Brier score.

In addition, a receiver operating characteristic area under the curve (ROC-AUC) is reported as a threshold-free, rank-based measure of discrimination. Here, ROC-AUC quantifies how well the predicted follow probabilities rank follow events above override events, independent of any particular decision threshold. Since ROC-AUC evaluates ordering rather than calibration, it is interpreted as a complementary diagnostic rather than a primary metric.

In the experimental task, overrides are typically less frequent than follows and are often the behavior of greatest practical interest. Therefore, additional diagnostics treat override as the positive class. Specifically, if p_d denotes the predicted probability of following, then the predicted probability of overriding is $1 - p_d$, and metrics are computed with override as the positive label. The reported override-focused diagnostics include precision, recall, and F1-score for overrides, as well as balanced accuracy (mean of follow and override recall). When informative (i.e., both classes present), a precision–recall area-under-curve (PR-AUC) for overrides may also be reported. These metrics are used diagnostically to characterize model behavior under imbalance; they do not replace proper scoring rules.

To better understand where probabilistic errors arise, NLL and Brier score are additionally decomposed by the true class. This yields separate mean scores for follow events and override events, highlighting whether a mapping is disproportionately miscalibrated or overconfident for the rarer override decisions.

8.4.4. Baseline Comparisons and Interpretive Framing

To make improvements interpretable, results are reported both in absolute terms and relative to baselines. In particular, differences in key metrics (e.g., mean NLL, Brier score, accuracy) are computed with respect to the global base-rate predictor and the uninformed 0.5 predictor. Relative reporting clarifies whether trust-conditioned mappings yield practical improvements beyond what could be achieved without trust.

Finally, evaluation is reported at two complementary aggregation levels, namely pooled evaluation, aggregating all validation decisions into a single performance summary per mapping, and participant-level evaluation, summarizing metrics per validation participant to expose heterogeneity and to avoid a small number of participants dominating pooled results.

This dual view supports cautious interpretation: strong pooled performance without consistent participant-level improvements may indicate that predictive signal is concentrated in a subset of individuals, while uniformly weak performance suggests that trust alone is insufficient to explain reliance decisions under the tested mappings.

8.5. Individual Differences in Reliance Mechanisms

Pooled analyses assess whether a small set of candidate reliance mechanisms can explain behavior on average and generalize across participants. In contrast, the goal of this section is to characterize mechanistic heterogeneity: the same candidate mechanisms are fitted separately for each participant and an analysis is made on which mechanism provides the most parsimonious explanation of that participant's decisions.

Importantly, this per-participant analysis is descriptive rather than predictive. It is conducted on the validation participants only and is intended to reveal individual differences in how reliance relates to trust and self-confidence, not to claim cross-participant generalization.

8.5.1. Per-participant Model Selection

For each participant, three increasingly flexible mappings from the inferred trust-at-decision τ_t and self-confidence β to the probability of following the robot recommendation are compared. For compactness, they are referred by mechanism:

1. **Trust-as-probability:** $p(\text{follow}_t) = \tau_t$ (no fitted parameters).
2. **Threshold on trust–confidence margin:** $p(\text{follow}_t) = \sigma(k(\tau_t - \beta))$ with $k \geq 0$.
3. **Confidence offset + lapse:** $p(\text{follow}_t) = (1 - \varepsilon) \sigma(k\tau_t + \alpha(\beta - 0.5)) + \varepsilon \cdot 0.5$, with $k \geq 0$, $\varepsilon \in [0, 0.5]$.

These mechanisms span a spectrum from a parameter-free interpretation (trust directly as compliance probability), through a sharpness-of-threshold effect in the trust–confidence margin, to a more flexible account in which confidence modulates reliance with an additional lapse component capturing occasional uninformed choices.

For each participant, parameters (when present) are estimated by maximum likelihood on that participant’s full sequence of door-level decisions (60 trials). The primary goodness-of-fit measure is the Bernoulli negative log-likelihood (NLL) of follow/override outcomes under the predicted follow probability. As a complementary proper scoring rule, the Brier score, i.e., the mean squared error between predicted follow probability and the binary follow label, is also reported.

To compare mechanisms with different numbers of parameters, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are computed from the per-participant NLL. Model selection is performed using BIC due to its stronger penalty for additional parameters at small sample sizes. In addition, a simple parsimony rule is applied: if a more complex mechanism improves BIC by at most $\Delta\text{BIC} \leq 2$ relative to a simpler mechanism, the simpler mechanism is selected. This prevents over-interpreting marginal improvements as evidence for additional psychological structure.

The selected mechanism for a given participant should be interpreted as a compact description of that participant’s reliance strategy under the available predictors (τ_t and β). In particular, selection of the threshold mechanism indicates that the trust–confidence margin is sufficient to explain the participant’s reliance pattern, while selection of the confidence offset + lapse mechanism suggests additional structure such as confidence-dependent bias and/or non-negligible stochasticity in decisions.

8.5.2. Robustness and Identifiability of Participant-specific Parameters

Per-participant parameter estimates can be sensitive when the number of trials is limited or when behavior is close to deterministic (e.g., almost always following or almost always overriding). Per-participant fits are therefore accompanied by two robustness checks that probe identifiability and stability.

For each participant and each parameterized mechanism, uncertainty is quantified via bootstrap resampling of doors with replacement. For each bootstrap replicate, the mechanism is refit and the resulting parameter samples are used to form percentile confidence intervals. In addition, the fraction of bootstrap replicates in which the optimizer fails to return a finite solution is tracked. Wide confidence intervals and high bootstrap failure rates are treated as indicators of weak identifiability for that participant and mechanism.

As a more stringent stability probe, each participant’s trials are repeatedly split into a fit subset and a held-out subset (e.g., 30/30 doors). For each split, the mechanism is fitted on the fit subset (or the fixed mapping for the trust-as-probability mechanism is applied) and evaluation on held-out trials is done using held-out NLL and held-out Brier score. This yields a distribution of held-out scores, highlighting whether participant-specific parameters generalize within-participant beyond the fitted doors. As an additional descriptive diagnostic, the Pearson correlation between held-out predicted probabilities and held-out binary outcomes is computed, noting that this correlation can be undefined when either quantity has negligible variance.

Finally, an exploratory blockwise refit is performed to probe potential nonstationarity across the three task blocks (doors 1–20, 21–40, and 41–60). For each participant, the mechanism selected following the previously described criteria is refitted separately on each block. The resulting blockwise parameter trajectories can be used to qualitatively label participants as stable, drifting, or under-identified based on whether parameters remain consistent across blocks and whether block-level fits are well-posed. Because each block contains relatively few trials, this analysis is interpreted only as exploratory evidence of potential drift rather than a definitive claim of time-varying reliance.

8.6. Closed-Loop Generative Evaluation of Interaction Patterns

Closed-loop generative evaluation is conducted at two complementary levels. First, a global rollout analysis uses population-level behavioral parameters learned on the training data to assess whether the coupled trust–behavior system reproduces aggregate interaction structure across participants. This analysis focuses on pooled statistics and uncertainty bands over repeated stochastic rollouts, and is intended to probe structural realism rather than individual fidelity. Second, a participant-level rollout analysis uses personalized behavioral parameters to evaluate whether individual interaction signatures—such as override timing, switching asymmetries, and streak structure—can be reproduced under closed-loop coupling.

The purpose of closed-loop evaluation here is not behavioral validation, but stress-testing the inferred trust dynamics under feedback, revealing whether coupling induces instability, degeneration, or qualitative

mismatch. These analyses test the generative implications of the trust dynamics model under both average and individualized behavioral assumptions, while maintaining a clear distinction between population-level structure and participant-specific realism.

8.6.1. Behavior-outcome Coupling Assumption

The closed-loop simulations presented in this section require an explicit assumption linking the agent's latent trust state to observable behavioral choices and, in turn, to task outcomes. This coupling is not learned from data, but imposed as a modeling device in order to explore the qualitative behavior of the trust dynamics under feedback. As such, all results in this section should be interpreted as conditional on the coupling assumption described below.

The task considered in this study admits only two mutually exclusive actions at each decision point: following the system recommendation or overriding it. Outcomes are deterministic given the action, in the sense that following the recommendation yields success if and only if the recommendation is correct, while overriding yields success if and only if the recommendation is incorrect. Under these conditions, behavior and outcome are perfectly anti-correlated with respect to correctness.

This structure enables a simple and unambiguous coupling rule: higher trust increases the probability of following the system, which in turn increases the probability of experiencing outcomes aligned with system correctness. Conversely, lower trust increases the likelihood of override, thereby inverting the relationship between system correctness and experienced outcome.

In the closed-loop simulations, trust is therefore coupled to behavior through a monotonic mapping from trust to the probability of following the system recommendation. Outcomes are then generated deterministically based on the chosen action and the system's underlying correctness signal. This induces a reverse-psychology effect at the lower trust region, where actions will tend to be exactly in the opposite direction of recommendations. Another possible assumption would be that of a disuse model at the lower trust region, where actions are not necessarily defined as the opposite of each recommendation, but rather as following a (unknown) participant-specific decision strategy. The reverse-psychology assumption was chosen due to the highly binary nature of the recommendation/action choices in the task, as well as to avoid adding unnecessary parameters to the behavioral model. Importantly, this inversion is not meant to model human strategy or deliberation. Rather, it is a direct consequence of embedding trust-driven behavior into a binary task with deterministic outcomes.

The purpose of this coupling is not to achieve behavioral realism, but to close the feedback loop in a minimal and internally consistent manner. It allows the learned trust update models to be interrogated under repeated interaction, revealing whether their dynamics lead to stable trajectories, oscillations, collapse to extremes, or other qualitative patterns of interest.

At the same time, this assumption is deliberately strong. It ignores alternative mechanisms such as partial compliance, stochastic outcomes, strategic disuse, or context-dependent action policies. Consequently, the simulations presented in this section should not be interpreted as predictions of human behavior or task performance. They are intended solely as conditional explorations of the structural properties of the inferred trust dynamics under closed-loop feedback.

8.6.2. Behavioral Interaction Signatures

Closed-loop evaluation of trust dynamics requires descriptors that go beyond pointwise action prediction. In particular, because following the system recommendation is the dominant behavior in the experimental task, aggregate accuracy or follow rate alone provides little insight into the structure of interaction. Instead, behavior is characterized using a set of interaction signatures that capture the timing, persistence, and switching structure of overrides under repeated interaction.

These signatures are computed both from observed participant behavior and from simulated behavior generated by coupled rollouts, and are chosen to be robust to class imbalance and interpretable at the interaction level.

As a first-order descriptor, the overall override rate and its complement, the follow rate, are considered. To probe potential nonstationarity over time, these rates are also computed separately within each experimental block. Because overrides are comparatively rare, blockwise override rates provide a more sensitive indicator of systematic changes in reliance behavior than aggregate accuracy.

To characterize how often participants change their reliance state, the probability of switching between consecutive decisions is computed. In addition to the overall switching probability, a distinction is made between asymmetric transitions: following-to-override and override-to-follow. This decomposition captures whether

overrides tend to occur as isolated events or whether they trigger subsequent behavioral adjustments.

Temporal clustering of overrides is assessed through inter-override gaps, defined as the number of doors between successive override events. Summary statistics of these gaps (mean, median, and upper quantiles) describe whether overrides are sporadic, bursty, or regularly spaced. Complementarily, override streak lengths quantify how often overrides occur consecutively, providing a direct measure of burstiness in non-compliance.

For completeness, streak lengths of consecutive follow decisions are also reported. However, because following constitutes the majority behavior in the task, these statistics are interpreted cautiously and are not treated as primary indicators of interaction structure.

For simulated behavior, each interaction signature is computed across multiple stochastic rollouts per participant, yielding a distribution rather than a single value. Reported summaries therefore include central tendencies and uncertainty bands derived from these distributions. Observed behavioral signatures are computed once per participant and serve as reference points for qualitative comparison.

Together, these interaction signatures provide a compact yet expressive description of reliance dynamics over time. They enable assessment of whether the closed-loop trust–behavior system reproduces salient structural features of observed interactions, without relying on pointwise behavioral prediction or assuming behavioral optimality.

8.6.3. Comparison Protocol for Closed-Loop Rollout Evaluation

This subsection defines the procedure used to compare observed human interaction data with interaction signatures generated by closed-loop behavioral rollouts. The objective is to establish a principled and reproducible comparison protocol for generative evaluation.

For each participant in the validation set, observed door-level interaction sequences are summarized using the previously defined behavioral signatures. These summaries serve as empirical reference points and are computed both at the individual level and, where appropriate, in a pooled manner across participants. When block structure is available, signatures are additionally stratified by block to preserve temporal context. The observed summaries are treated as fixed reference quantities against which simulated distributions are compared.

Closed-loop simulations generate stochastic interaction sequences by sampling behavioral decisions and feeding them back into the trust dynamics. For each comparison condition, multiple independent rollouts are performed per participant, yielding an empirical distribution over each behavioral signature. Rather than collapsing these simulations into a single point estimate, simulated behavior is represented through distributional summaries, including central tendency and quantile-based uncertainty bounds.

Uncertainty in simulated behavior arises from stochastic decision sampling rather than parameter uncertainty. Accordingly, simulated signatures are summarized using quantile intervals computed across rollouts for each participant and each metric. These intervals characterize the variability induced by repeated interaction trajectories under identical model parameters and initial conditions. Observed signatures are not assigned uncertainty bounds, as they correspond to realized interaction histories.

The comparison between observed and simulated behavior is performed at the level of interaction signatures rather than individual decisions. For each signature, observed values are contrasted with the corresponding simulated distribution to assess whether empirical behavior lies within the range of behaviors generated by the model. This comparison is descriptive and structural: it evaluates whether the coupled trust–behavior system reproduces salient temporal and switching patterns, rather than optimizing or scoring predictive accuracy.

The comparison protocol is applied at two complementary levels. First, global comparisons pool signature summaries across participants to assess whether aggregate interaction structure is reproduced at the population level. Second, participant-level comparisons retain individual specificity by contrasting each participant's observed signatures with their corresponding simulated distributions. These two views are treated as distinct but complementary diagnostic lenses.

8.7. Trust Trajectory Sanity Checks under Closed-Loop Coupling

Closed-loop coupling alters the interaction stream by sampling actions from the behavioral model and feeding the resulting counterfactual outcomes back into the trust dynamics. Even when the underlying trust model is held fixed, this feedback can induce deviations in the resulting trust trajectory relative to replay (i.e., trust evolution under the recorded interaction stream), and these deviations may differ depending on whether behavior is parameterized globally or at the participant level. The purpose of this section is therefore purely methodological: to define a set of trajectory-level sanity checks that quantify how much closed-loop coupling perturbs trust dynamics, and to document the guardrails used when participant-level behavioral parameters are unreliable.

8.7.1. Trajectory Comparison Modes and Contrasts

For each participant, trust trajectories are evaluated on the full simulation time grid produced by the trust dynamics, rather than only at discrete decision points. Three trajectory constructions can be distinguished:

1. **Replay trajectory:** trust evolution under the recorded interaction data, used as a deterministic reference for each participant.
2. **Coupled trajectory with global behavior parameters:** closed-loop simulation in which actions are sampled from the globally fitted behavioral model, and the sampled actions are coupled back into trust dynamics via the outcome inversion rule.
3. **Coupled trajectory with personalized behavior parameters:** closed-loop simulation in which actions are sampled from participant-specific behavioral parameters (subject to guardrails).

Using these trajectories, two contrasts that isolate distinct effects of coupling can be defined:

Coupling effect (relative to replay): quantification of the deviation between coupled and replay trajectories, computed separately for (i) globally parameterized behavior and (ii) personalized behavior. This contrast captures how much closed-loop action sampling (together with counterfactual outcome inversion) can move trust away from the replay baseline.

Parameterization effect under coupling: quantification of the deviation between coupled trajectories obtained with personalized versus global behavior parameters. To isolate parameter effects from stochastic sampling noise, coupled simulations in this contrast are run with matched random seeds across the two parameterizations, so that differences are attributable to the behavior parameterization rather than to different action-sampling realizations.

8.7.2. Continuous-time Divergence Metrics

Let $\tau_1(t_k)$ and $\tau_2(t_k)$ denote two trust trajectories sampled on the same simulation grid $\{t_k\}_{k=1}^K$ with step size Δt . For a given contrast, an error curve can be defined as

$$e(t_k) \triangleq \tau_1(t_k) - \tau_2(t_k), \quad (8.3)$$

summarizing divergence per rollout using four scalar metrics:

- **Time-averaged absolute deviation (MAE_t):** the mean absolute magnitude of $e(t)$ over the trajectory duration.
- **Time-averaged squared deviation (RMSE_t):** the root-mean-square magnitude of $e(t)$ over time.
- **Integrated absolute deviation (IAD):** a time-integral proxy for the area under $|e(t)|$, i.e., an accumulated deviation expressed in trust-units times seconds.
- **Maximum absolute deviation (MaxAbs):** the peak absolute discrepancy $\max_k |e(t_k)|$.

All four metrics are computed on the full simulation time grid. This choice ensures that deviations induced by coupling are assessed as properties of the trajectory (including transients between decision events), rather than being reduced to a door-indexed comparison.

8.7.3. Temporal Diagnostics for Interpretability

Scalar divergence metrics support compact reporting, but they do not reveal when along an interaction horizon the deviations emerge or accumulate. Accordingly, two trajectory-level diagnostics derived from $e(t_k)$ are retained for each contrast:

1. **Rollout-mean error curve:** for a fixed participant and contrast, the pointwise mean of $e(t_k)$ across repeated coupled rollouts, which provides a signed summary of whether coupling tends to increase or decrease trust relative to the reference trajectory.
2. **Accumulated absolute error curve:** defined by the discrete-time accumulation of $|e(t_k)|$, approximating $\int_0^t |e(\tau)| d\tau$. This curve makes it explicit whether divergence accumulates steadily or is driven by localized bursts.

These curve-based diagnostics are intended for visualization and qualitative inspection, complementing the scalar metrics without introducing any additional evaluation targets.

8.7.4. Uncertainty Representation and Pooling Across Horizons

Closed-loop coupled trajectories are stochastic because actions are sampled from the behavioral model. Uncertainty is therefore represented at two levels.

For each participant and each contrast, multiple coupled rollouts are run and the divergence metrics described in subsection 8.7.2 are computed per rollout. These per-participant distributions are summarized by reporting the rollout mean together with a central uncertainty interval given by lower/upper quantiles. This provides a robust, nonparametric characterization of variability induced by action sampling.

Participants may have different total interaction durations and therefore different time horizons. To pool trajectory-level diagnostics across participants, each participant's time grid is mapped to a normalized axis $s \in [0, 1]$ (e.g., by affine rescaling of time from start to end of the trajectory) and participant-level rollout-mean curves are resampled onto a shared normalized grid. Pooled summaries are then formed by aggregating, at each normalized time s , the mean curve value across participants, together with quantile envelopes. This pooling procedure is used strictly as a visualization aid to describe cohort-level patterns in divergence shape; it does not assert that normalized time has behavioral interpretability beyond enabling comparisons across heterogeneous horizons.

8.7.5. Guardrails and Fallback Audit for Personalized Parameters

Personalized coupled simulations rely on participant-specific behavioral parameters inferred from the behavioral identification pipeline. However, for some participants these parameters may be under-identified or fail basic validity checks. To prevent propagating clearly unreliable behavioral estimates into coupled trust trajectories, some guardrails are put in place that trigger a fallback to simpler behavior parameterizations when necessary.

The result is a per participant log of (i) the participant-level model selection produced by the behavioral pipeline, (ii) the model actually used for personalized coupled simulation after applying guardrails, (iii) whether fallback was applied, and (iv) a categorical reason label for the fallback. This audit sequence supports transparent interpretation of downstream rollout analyses by making it explicit when “personalized” coupling is in fact implemented via a simpler, more conservative behavioral specification.

The diagnostics defined in this section accompany the closed-loop generative evaluation described in section 8.6. They ensure that coupled rollouts used to assess behavioral interaction signatures are paired with explicit diagnostics of how coupling perturbs the trust trajectory, and that the effect of personalization on trust dynamics can be separated from stochastic sampling noise via matched-seed comparisons. All numeric outcomes and interpretations of these diagnostics are deferred to the dedicated results section.

8.8. Summary of Behavioral Model Methodology

This section summarizes the end-to-end methodology used to construct, identify, and evaluate the trust-aware behavioral modeling layer prior to reporting numerical outcomes. The purpose is to provide a single, coherent overview that connects (i) the scope and interpretive framing of behavioral modeling (section 8.1), (ii) the candidate trust-to-action mappings (section 8.2), (iii) the causally aligned behavioral dataset used throughout (section 8.3), (iv) the training–validation estimation and probabilistic evaluation protocol (section 8.4), (v) the individual-differences analysis with robustness diagnostics (section 8.5), (vi) the closed-loop generative rollout evaluation at global and participant level (section 8.6), and (vii) the trajectory-level trust sanity checks under coupling (section 8.7). Throughout the section, the emphasis is on methodological consistency, strict causal alignment, and clear separation between predictive evaluation, descriptive heterogeneity analysis, and generative diagnostics.

8.8.1. Overview and Scope

The behavioral modeling stage treats reliance as an observable stochastic outcome influenced by an inferred latent trust state. Its role is deliberately auxiliary: rather than attempting a comprehensive model of human decision-making, it provides a controlled probe of whether the inferred trust trajectories carry sufficient information to explain salient reliance patterns in the experiment. Consequently, behavioral predictability is interpreted as diagnostic evidence about the explanatory utility of trust, not as a direct validation of the upstream trust dynamics model.

A key structural principle is maintained: trust inference and validation are performed independently in the preceding chapter, and the behavioral layer is introduced only as a downstream mapping from trust to action. This separation supports interpretability (trust dynamics are assessed on their own terms) and prevents circularity when later analyses embed behavior into closed-loop simulations.

8.8.2. Candidate Trust-to-action Mappings

Behavior is modeled as a Bernoulli decision at each door-level event, with the behavioral model specifying a follow probability $p \in (0, 1)$ as a function of the inferred trust-at-decision and (where applicable) self-confidence. To probe alternative hypotheses about how trust translates into reliance, the chapter introduces a small nested family of mappings with increasing expressive power. The mappings include: (i) a parameter-free interpretation of trust as a direct follow probability; (ii) a thresholded mapping in which follow likelihood depends on trust relative to a reference level tied to self-confidence, with a sensitivity parameter governing transition sharpness; and (iii) an extension that allows confidence modulation and a lapse component to capture occasional trust-independent decisions. The family is intentionally compact and interpretable so that differences in downstream performance can be attributed to specific modeling assumptions.

8.8.3. Behavioral Dataset Construction and Causal Alignment

All behavioral analyses in this chapter share a common decision-level dataset constructed to preserve causal consistency. The unit of analysis is a door-level decision event, represented by a binary follow/override label together with directly observed contextual covariates. The inferred trust value paired with each decision is defined as the trust state immediately prior to the decision event, before outcome information becomes available. This pre-decision alignment ensures that trust is used as a legitimate predictor of behavior rather than a post-hoc correlate, and it prevents artificial inflation of predictive metrics that would arise from using post-decision or post-outcome trust values.

8.8.4. Training-validation Protocol and Probabilistic Evaluation

Predictive evaluation of trust-to-action mappings follows a strict participant-level training/validation split. Where mappings contain free parameters, they are estimated by maximum likelihood on the training set under a Bernoulli model, yielding fixed parameter values that are then applied unchanged on the validation set. Two trust-free reference predictors (a global base-rate model and an uninformed 0.5 predictor) are included to contextualize performance gains attributable to trust conditioning.

Because the mappings output probabilities, evaluation is centered on proper scoring rules. Negative log-likelihood and the Brier score quantify probabilistic accuracy and calibration-relevant error, while calibration diagnostics use reliability binning and expected calibration error to assess whether predicted probabilities match empirical follow rates. For interpretability, secondary metrics include thresholded accuracy, ROC-AUC for discrimination, and override-centric diagnostics (precision/recall/F1, balanced accuracy, and PR-AUC) to

better characterize performance under class imbalance. To avoid pooled metrics masking heterogeneity, performance is summarized both at the pooled level (aggregating all validation decisions) and at the participant level (computing per-participant summaries).

8.8.5. Individual Differences and Parameter Robustness

Beyond pooled predictive evaluation, the chapter characterizes mechanistic heterogeneity by fitting the candidate reliance mechanisms separately for each validation participant and selecting the best and simplest mechanism per individual. Model comparison uses information criteria derived from per-participant likelihoods, with selection governed by BIC and an explicit parsimony rule that defaults to simpler mechanisms when improvements are marginal. The resulting selected mechanism is interpreted as a compact description of that participant's reliance strategy under the available predictors (trust and self-confidence), rather than as evidence of cross-participant generalization.

To accompany participant-specific fits, robustness checks probe identifiability and stability. Bootstrap re-sampling yields uncertainty intervals and failure rates as indicators of weak identification. Repeated within-participant random splits evaluate whether fitted parameters generalize to held-out doors, using held-out NLL and Brier score, complemented by descriptive correlation diagnostics when defined. Finally, an exploratory blockwise refit probes potential nonstationarity across task blocks, supporting qualitative labels such as stable, drifting, or under-identified, while explicitly acknowledging the limited trial count per block.

8.8.6. Closed-loop Generative Rollout Evaluation

Predictive evaluation at the decision level does not directly test whether a coupled trust-behavior system reproduces interaction structure over time. Therefore, the chapter performs closed-loop generative evaluation by sampling actions from the behavioral model and feeding them back into the trust dynamics via an explicit behavior-outcome coupling assumption. This coupling rule is introduced as a modeling device to close the feedback loop in a binary task setting; it is not learned from data and is interpreted conditionally.

Generative evaluation is explicitly twofold. A global rollout analysis uses population-level behavioral parameters learned on the training set to probe whether aggregate interaction signatures are reproduced across participants. A complementary participant-level rollout analysis uses personalized behavioral parameters to assess whether individual interaction signatures can be reproduced for specific participants. In both cases, evaluation is signature-based rather than pointwise: simulated and observed interactions are summarized through override-centric rates, switching asymmetries, inter-override gaps, and streak structure. Simulated behavior is represented distributionally by repeating stochastic rollouts per participant and reporting central tendencies with quantile-based uncertainty bounds, while observed signatures serve as fixed reference points.

8.8.7. Trust Trajectory Sanity Checks under Coupling

Closed-loop coupling can perturb the trust trajectory even when the underlying trust model is unchanged. To quantify and document these perturbations, trajectory-level sanity checks computed on the full simulation time grid are defined. Two contrasts are considered: coupled versus replay trajectories (computed separately under global and personalized behavior parameterizations), and personalized versus global behavior under coupling, using matched random seeds to isolate parameterization effects from sampling noise. Divergence is summarized per rollout by continuous-time metrics (MAE_t , $RMSE_t$, IAD, and MaxAbs) and complemented by temporal diagnostics that retain rollout-mean error curves and accumulated absolute error curves for interpretability. Uncertainty is represented by quantiles across rollouts within each participant, and pooled curve summaries are constructed by resampling participant-level mean curves onto a normalized time grid for visualization across heterogeneous horizons.

Because personalized coupling depends on participant-specific behavioral parameters that may be under-identified, guardrails are applied and explicitly audited. The analysis records the selected behavioral model and the one actually used after guardrail application.

In summary, the behavioral modeling methodology combines (i) a causally aligned decision-level dataset, (ii) a compact family of interpretable trust-to-action mappings, (iii) training/validation likelihood-based estimation with probabilistic scoring and calibration diagnostics, (iv) descriptive participant-specific mechanism selection with robustness checks, (v) two-level closed-loop generative evaluation via interaction signatures, and (vi) trajectory-level trust sanity checks that quantify how coupling and personalization perturb trust dynamics. The methodology is designed to fail gracefully: weak behavioral predictability constrains claims about trust-action coupling without undermining the internal validity of the inferred trust dynamics. The following section reports the corresponding behavioral model results and empirical outcomes.

8.9. Behavioral Model Results

This section reports the empirical outcomes of the trust-aware behavioral modeling analyses defined in sections 8.4 through 8.7. Results are presented in four parts. First, behavioral dataset composition and summary statistics for the training and validation split are reported. Second, trust-to-action mappings are evaluated as probabilistic predictors of follow/override decisions on held-out participants, relative to trust-free baselines. Third, participant-level mechanism selection are summarized together with robustness diagnostics, to characterize heterogeneity and identifiability limits. Fourth, a minimal trust-to-action mapping is used to close the model behavior loop, and both aggregate interaction signatures from coupled rollouts and trajectory-level sanity checks that quantify how coupling perturbs trust evolution are reported.

Unless explicitly stated otherwise, predictive and generative results are computed on the validation participants.

8.9.1. Behavioral Dataset Composition and Summary Statistics

This subsection reports the general composition and statistics of the generated behavioral dataset. Its goal is to serve as a transparent reference that grounds the results to follow. Table 8.1 summarizes information such as number of total (N) and valid (N_{valid}) events, as well as follow rates and mean and standard deviation metrics for decision trust (τ_d), risk (r) and self-confidence (β).

Table 8.1: Behavioral dataset composition and summary statistics per split (door-trial level). Decision trust τ_d is evaluated immediately before the door-event update, ensuring causal alignment between trust state and behavioral decision.

Split	N_p	N	N_{valid}	Follow Rate	$\mu(\tau_d)$	$\sigma(\tau_d)$	$\mu(r)$	$\mu(\beta)$	$\mu(\tau_d - \beta)$
Training	25	1500	1500	0.845	0.618	0.152	0.478	0.710	-0.092
Validation	10	600	600	0.892	0.677	0.146	0.500	0.803	-0.125

The validation split is strongly follow-dominant (follow rate = 0.892), implying substantial class imbalance in the decision labels. Consequently, a constant base-rate predictor is expected to achieve high thresholded accuracy and to provide a stringent reference for any trust-conditioned mapping. These dataset properties are therefore used to contextualize the probabilistic and override-centric results reported below.

8.9.2. Predictive Performance on Held-Out Participants

This subsection reports how well the candidate trust-to-action mappings predict observed reliance decisions under held-out conditions. Performance is evaluated probabilistically and reported relative to trust-free baselines, following the protocol defined in section 8.4.

Table 8.2 summarizes pooled predictive performance across all validation decision events for each behavioral mapping and baseline. Primary metrics are mean negative log-likelihood (NLL) and Brier score. Lower values indicate better probabilistic prediction.

Table 8.2: Pooled probabilistic performance on validation set participants.

Model	Mean NLL	Brier Score
Uninformed baseline ($p = 0.5$)	0.6931	0.2500
Global base-rate baseline	0.3520	0.0987
Trust-as-probability	0.5516	0.1482
Thresholded reliance	0.6931	0.2500
Threshold + modulation + lapse	0.3575	0.1010

On pooled held-out decisions, the global base-rate baseline achieves the lowest mean NLL and Brier score among all reported mappings. The trust-as-probability mapping yields worse probabilistic scores than the base-rate predictor, while the thresholded reliance mapping matches the uninformed baseline exactly (mean NLL = 0.6931, Brier = 0.2500), indicating a collapse to constant uninformative predictions on the validation set. The more flexible threshold + modulation + lapse mapping approximately recovers base-rate performance but does not provide a clear probabilistic improvement beyond it.

Calibration diagnostics (reliability binning and expected calibration error) were computed for the trust-conditioned mappings on the validation split and were consistent with the probabilistic scoring results, with no mapping exhibiting clear calibration improvements relative to the base-rate reference.

Table 8.3 reports a secondary discrimination-oriented metric, namely ROC-AUC.

Table 8.3: Discrimination (ranking) performance on validation set participants (ROC-AUC).

Model	ROC-AUC
Uninformed baseline ($p = 0.5$)	0.5000
Global base-rate baseline	0.5000
Trust-as-probability	0.6601
Thresholded reliance	0.5000
Threshold + modulation + lapse	0.4981

As a rank-based diagnostic, ROC-AUC indicates whether higher predicted follow probabilities tend to be associated with follow outcomes, independent of probability calibration. The trust-as-probability mapping exhibits above-chance discrimination (ROC-AUC = 0.6601), whereas the remaining mappings and baselines are near chance. Importantly, above-chance discrimination under the trust-as-probability mapping indicates that inferred trust preserves ordinal information about reliance tendencies, even when absolute probabilities are miscalibrated or dominated by base-rate effects.

Because overrides are comparatively rare in the validation split, override-centric diagnostics (treating override as the positive class) were additionally computed to detect degenerate prediction regimes. Consistent with the probabilistic results, these diagnostics indicate that some mappings reduce to near-constant predictions on the validation set, yielding either near-zero override recall (near-always-follow) or trivially high override recall with low precision (near-always-override).

Participant-level summaries of predictive scores showed substantial inter-individual variability; pooled metrics are therefore interpreted as cohort-level aggregates rather than representative of every participant.

8.9.3. Participant-Level Mechanism Selection and Identifiability Limits

This subsection reports the outcomes of participant-specific mechanism selection and robustness analyses defined in section 8.5.

Figure 8.1 summarizes the BIC values computed for each validation participant, for each of the three proposed behavioral models.

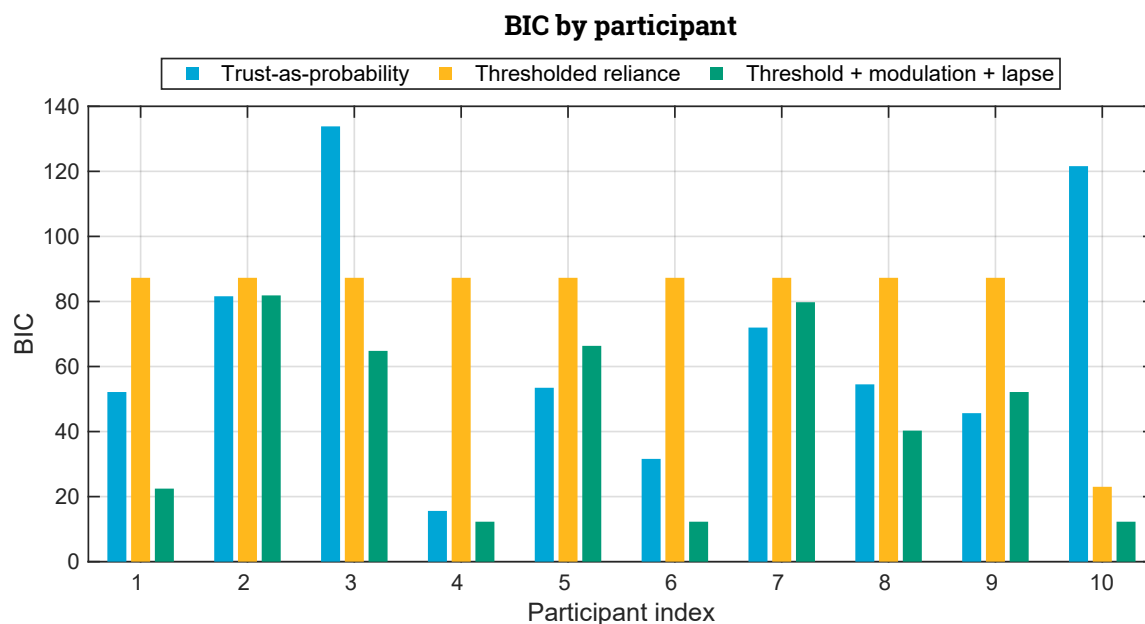


Figure 8.1: BIC values by validation set participant for the three proposed behavioral models.

Across the 10 validation participants, the BIC selection rule chooses the trust-as-probability mapping for 4 participants and the more flexible threshold + modulation + lapse mapping for 6 participants (Table 8.4). No participant is selected as best explained by the threshold-only mapping under BIC. These selections summarize which mapping provides the most parsimonious likelihood fit under the available predictors (τ_d, β) for each individual.

Table 8.4: Per-participant behavioral fits on validation set. Best model selected by BIC with a parsimony rule ($\Delta\text{BIC} \leq 2$ selects the simpler model). All participant metrics are calculated using the same number of events ($N = 60$). BIC_0 , BIC_1 and BIC_2 denote the BIC values for the trust-as-probability, thresholded reliance, and threshold + modulation + lapse models, respectively.

Participant	BIC_0	BIC_1	BIC_2	Best Model
R0ODC7IYM	52.16	87.27	22.45	Threshold + modulation + lapse
R0YCGFKR2	81.59	87.27	81.86	Trust-as-probability
R1QNFU95K	133.83	87.27	64.80	Threshold + modulation + lapse
R35SSCTHX	15.60	87.27	12.28	Threshold + modulation + lapse
R4E127CKP	53.46	87.27	66.35	Trust-as-probability
R7ZOUIC86	31.60	87.27	12.28	Threshold + modulation + lapse
RKJ5TBXWA	71.97	87.27	79.76	Trust-as-probability
RKSPT7N80	54.51	87.27	40.29	Threshold + modulation + lapse
RLPIUZPOS	45.65	87.27	52.16	Trust-as-probability
RWZXKC2BI	121.60	23.01	12.28	Threshold + modulation + lapse

For participants where the flexible mapping is selected, fitted parameters frequently exhibit saturation and boundary solutions (e.g., extremely large slope parameters and lapse rates estimated at or near zero), indicating weak identifiability under limited trial counts and strongly imbalanced actions. For transparency, the robustness diagnostics below are used as the primary evidence when interpreting participant-specific fits.

Figure 8.2 summarizes within-participant random-split robustness for the thresholded mapping by reporting the distribution of split-to-split variability in the fitted slope parameter k . Many fits concentrate near $k \approx 0$, corresponding to an effectively constant decision probability and consistent with the pooled collapse previously observed.

Figure 8.3 reports the analogous split-to-split variability for the flexible mapping. The presence of heavy tails and extreme split-to-split variation indicates that for some participants the fitted mapping is weakly identified and can enter saturation regimes, motivating conservative interpretation of participant-specific parameter values and downstream analyses that rely on them.

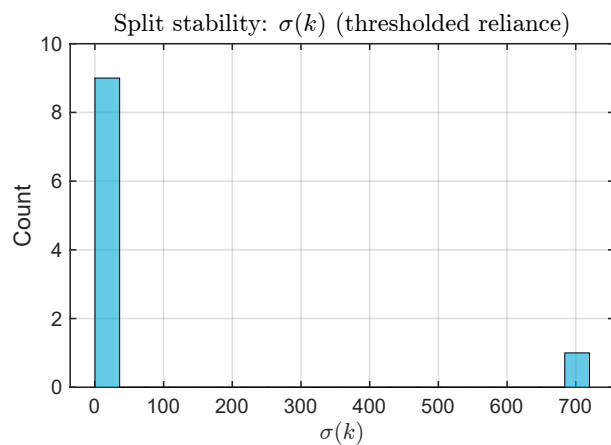


Figure 8.2: Distribution of split-to-split SD of k for the thresholded mapping model. In many cases the fit collapses to near-zero slope, yielding constant predictions.

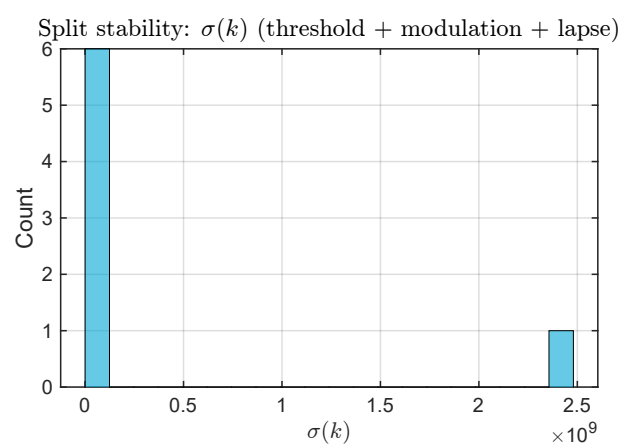


Figure 8.3: Distribution of split-to-split SD of k for the flexible mapping model (threshold + modulation + lapse). Heavy tails / extreme values indicate saturation and weak identifiability for participants.

8.9.4. Closed-Loop Generative Interaction Patterns

This subsection closes the trust–behavior loop and evaluates the resulting interaction patterns using the signature-based protocol in subsection 8.6.2. To keep the loop-closure interpretable given the identifiability limitations in subsection 8.9.3, the primary closed-loop results are reported for the parameter-free trust-as-probability mapping, which provides a minimal and transparent coupling from inferred trust to action sampling.

Figure 8.4 compares pooled observed and simulated interaction signatures under global behavior parameters.

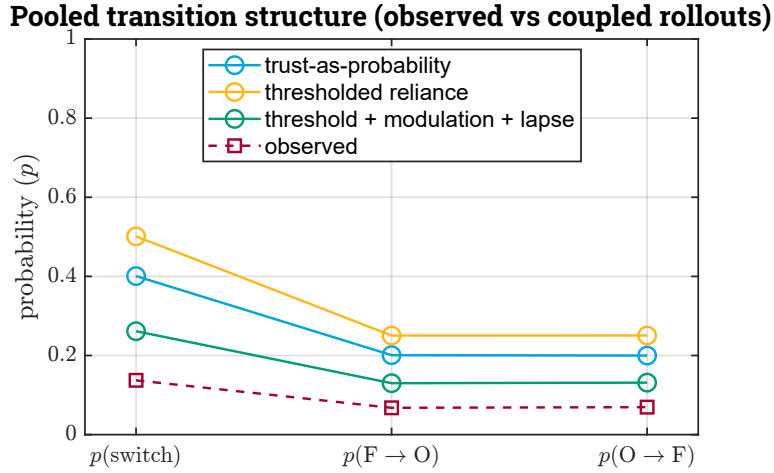


Figure 8.4: Pooled transition structure on the validation split: observed sequences versus coupled rollouts under each trust-to-action mapping. O = override, F = follow.

Table 8.5 further collects these signatures in a comprehensive manner.

Table 8.5: Pooled closed-loop interaction signatures on the validation split. Reported values are the mean across participants of the rollout-mean signature (300 rollouts/participant). Observed signatures are computed from recorded validation sequences. The primary loop-closure baseline is trust-as-probability. O = override, F = follow.

Model	$\bar{p}(O)$	$p(F \rightarrow O)$	$p(O \rightarrow F)$	$p(\text{switch})$	Gap_O mean	Streak_O mean
Observed	0.1083	0.1373	0.0695	0.1373	5.1037	1.0928
Trust-as-probability	0.3631	0.2009	0.1999	0.4007	3.2890	1.7399
Threshold	0.4986	0.2505	0.2506	0.5011	2.0054	1.9907
Offset + lapse	0.1563	0.1300	0.1314	0.2614	6.3949	1.1888

Under closed-loop coupling with the trust-as-probability mapping, simulated interactions exhibit substantially higher override rates and switching probabilities than observed validation behavior. In particular, the model produces elevated $P(\text{switch})$ and markedly increased $P(O \rightarrow F)$, indicating reduced persistence of reliance states relative to the observed interaction sequences.

Participant-specific closed-loop rollouts using individualized behavioral parameterizations were also computed, but are not reported in the main text because participant-level behavioral fits were frequently under-identified, which can drive qualitatively unrealistic simulated regimes.

8.9.5. Trust Trajectory Sanity Checks under Coupling

To quantify how stochastic closed-loop coupling perturbs trust evolution relative to replay, trajectory-level divergence diagnostics are reported for the globally coupled rollouts under the primary loop-closure mapping (trust-as-probability).

Figure 8.5 shows the pooled normalized-time trust trajectory error between coupled and replay trajectories, aggregated across validation participants with quantile envelopes over rollouts. This diagnostic summarizes the direction and temporal structure of coupling-induced deviations in trust under stochastic action sampling.

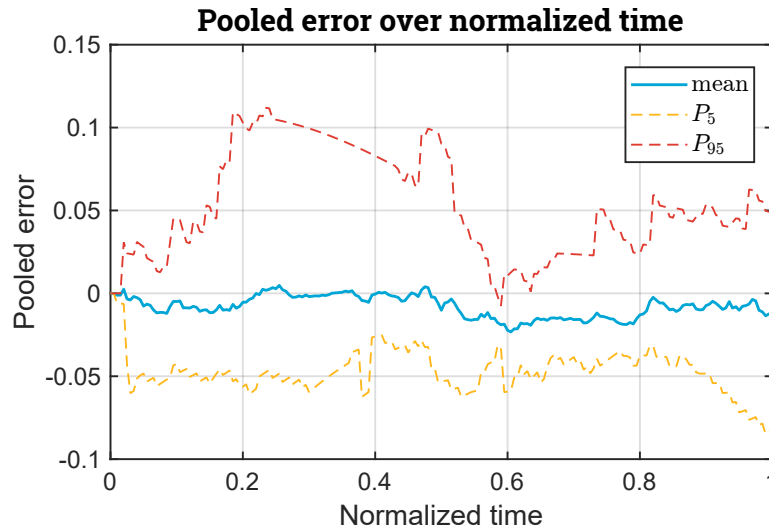


Figure 8.5: Pooled trust trajectory error $e(t) = \tau_{\text{coupled, global}}(t) - \tau_{\text{simple}}(t)$ over normalized time. Solid line shows the mean across participants; dashed lines indicate the 5–95% quantile range. This comparison isolates the effect of introducing stochastic coupling relative to deterministic replay.

8.10. Discussion of Behavioral Modeling Results

This section interprets the results reported in section 8.9 in light of the chapter’s exploratory purpose: to probe whether inferred trust trajectories carry behavioral signal, and to identify what breaks when attempting to close the trust–behavior loop under minimal, transparent coupling assumptions. The intent is not to defend weak predictive performance, but to extract the most informative constraints these results place on how trust may (and may not) translate into observable reliance decisions in the present task.

A key premise from section 8.1 is maintained throughout: limitations of the behavioral mapping should not be conflated with limitations of the upstream trust dynamics model, which has already been identified and validated independently. The behavioral layer is a downstream probe that tests one specific question: whether the inferred latent trust state is behaviorally legible under simple decision rules.

8.10.1. Behavioral Relevance of the Inferred Trust State

Two empirical facts jointly characterize the decision-level predictive evaluation. First, the validation dataset is strongly follow-dominant (Table 8.1), meaning that a trust-free base-rate predictor can achieve strong probabilistic scores simply by predicting “follow” with high probability. Second, despite this base-rate dominance, the trust-as-probability mapping achieves above-chance discrimination ($\text{ROC-AUC} = 0.6601$, from Table 8.3), indicating that inferred trust preserves some information about reliance tendencies.

These two facts are not contradictory: they reveal that the trust signal present in the inferred trajectories is primarily rank-based rather than calibrated. In other words, trust estimates tend to order decision points in a way that is meaningfully aligned with follow versus override outcomes, but the magnitude of predicted probabilities is not well matched to the empirical prevalence of actions, and therefore does not translate into improved proper scoring performance relative to a base-rate predictor (Table 8.2). This distinction is important for interpreting what the behavioral probe actually establishes. The results do not support the claim that inferred trust is a sufficient probabilistic predictor of follow/override decisions in this task, but they do support the weaker and more diagnostic claim that inferred trust contains behaviorally relevant structure that is detectable under ranking-based evaluation.

The observed divergence between discrimination and proper scoring is also consistent with the chapter’s conceptual framing: reliance decisions reflect multiple interacting factors (e.g., confidence, risk perception, strategy, attention, and local contextual cues), so even when trust is behaviorally relevant, it might not improve aggregate predictive likelihood in a highly imbalanced regime where base-rate dominates.

8.10.2. Limits of Trust-to-Action Mappings at the Decision Level

The thresholded reliance mapping collapses on the validation set to constant, uninformative behavior (Table 8.2). In practice, this corresponds to an effective slope $k \approx 0$ so that the sigmoid becomes nearly flat and the model predicts $p(\text{follow}) \approx 0.5$ regardless of input. The important point is not that this mapping “fails” per se, but what this failure mode reveals: under the available predictors and the participant-level training/validation split, the trust–confidence margin does not provide a stable, generalizable decision boundary that survives across individuals.

This outcome is consistent with two constraints of the current dataset. First, the decision label distribution contains comparatively few overrides, so the likelihood surface provides weak information about where a threshold should lie, especially when that threshold is tied to another noisy input such as self-confidence. Second, the mapping assumes that confidence enters the decision as an immediate reference level, but self-confidence reports may not correspond to an online decision threshold in the way the mapping prescribes. Under such conditions, a flat solution is a plausible maximum-likelihood compromise: it avoids catastrophic overconfidence on rare overrides at the expense of losing discrimination entirely.

The more flexible mapping (threshold + modulation + lapse) recovers near base-rate performance in pooled scoring, but does not provide a clear improvement beyond it (Table 8.2). Here, the diagnostic message is that additional behavioral flexibility can absorb some mismatch at the cohort level without producing a robust, interpretable causal mapping from trust to action. In this exploratory context, the value of the flexible mapping is less in any marginal score gain and more in revealing the identifiability limits and failure modes that emerge when attempting to parameterize individual differences with limited trials and imbalanced outcomes.

The per-participant model selection results (Figure 8.1; Table 8.4) indicate heterogeneity in which mapping best describes each participant’s decisions within participant. This is consistent with the broader HRI literature in which reliance behavior reflects different strategies and attitudes across individuals. However, the robustness diagnostics (Figures 8.2–8.3) show that, for parameterized mappings, the estimated parameters are often weakly identified: split-to-split variability can be large, and fitted solutions can enter saturation regimes.

This combination of outcomes is informative. It suggests that describing individual reliance strategies using a small set of parameters is plausible in principle, but that in the present dataset the information available to identify those parameters is limited. When behavior is close to deterministic (almost always follow or almost always override), many distinct parameter settings yield nearly identical likelihood, so optimization can produce extreme values without corresponding interpretive meaning. The chapter therefore treats participant-specific fits as descriptive labels of reliance mechanism preference rather than as stable psychological measurements. This is also why the main closed-loop results prioritize a minimal parameter-free mapping for loop closure, while keeping personalized coupling conservative.

8.10.3. Implications of Closed-Loop Coupling for Interaction Dynamics

Closed-loop rollouts provide a qualitatively different diagnostic than pointwise prediction. Even when a trust-to-action mapping contains some discriminative signal, embedding it in a feedback loop goes one step further and asks a different question: does the coupled system reproduce interaction structure over time?

The signature-based results show that, under the primary loop-closure mapping (trust-as-probability), simulated interactions exhibit substantially higher override rates, higher switching probabilities, and reduced persistence relative to observed validation behavior (Table 8.5; Figure 8.4). This mismatch is not merely a “performance gap”; it reveals a structural limitation of the coupling assumption in a binary task with deterministic outcomes.

Specifically, the imposed coupling rule creates a strong feedback mechanism: sampled actions directly invert whether the participant experiences success or failure when the system is correct or incorrect. In such a setting, even modest stochasticity in decisions can repeatedly flip the experienced outcome stream, which in turn pushes the trust dynamics to respond to a counterfactual interaction history that differs from the replay sequence. The result is a coupled system that tends to generate frequent alternation between follow and override states (high $P(\text{switch})$) unless the behavioral policy includes an explicit source of persistence or hysteresis beyond instantaneous trust.

The closed-loop findings therefore provide a concrete boundary condition for trust-aware behavioral modeling in this task: a mapping from instantaneous trust to instantaneous action probability is insufficient to reproduce the observed temporal structure of reliance. Even if trust contributes to decision-making, the observed interaction patterns imply additional stabilizing mechanisms, such as dependence on recent actions, inertia in policy selection, strategy commitment, or other latent variables not captured by trust alone.

The trust trajectory sanity checks complement the interaction signatures by quantifying how much the trust dynamics are perturbed by stochastic coupling relative to replay (Figure 8.5). Importantly, these diagnostics are not used to claim that coupled trust trajectories are “wrong”, but to establish that closing the loop changes the interaction history in a way that produces measurable deviations in the trust trajectory even with the same underlying trust update model.

This result reinforces a methodological point that is central to the end-to-end pipeline: closed-loop evaluation is not a trivial replay with added noise. Once action sampling is introduced, the trust model is driven by an alternative outcome stream generated by the coupling rule, and divergence relative to replay becomes an expected property of the coupled system. The value of the sanity checks is therefore twofold: they (i) confirm that the implemented loop closure is actually exerting feedback (i.e., the loop is meaningfully closed), and (ii) provide guardrails to interpret rollout results, distinguishing qualitative behavioral mismatch from purely stochastic perturbations.

8.10.4. Limitations and Scope of the Behavioral Modeling Results

The results of this chapter highlight several important limitations that define the scope within which the proposed trust-aware behavioral modeling framework should be interpreted. These limitations are not incidental shortcomings of a particular model instantiation, but structural constraints that arise from the modeling assumptions, data regime, and coupling strategy adopted in this exploratory analysis.

First, while the inferred trust trajectories carry behavioral information that is detectable under discrimination-oriented evaluation (Table 8.3), this signal is insufficient to support calibrated probabilistic prediction of reliance decisions in a strongly imbalanced setting. Trust-conditioned mappings do not outperform a trust-free base-rate predictor under proper scoring rules (Table 8.2), indicating that trust alone cannot account for the magnitude or frequency of follow and override actions observed in the data. This limitation bounds the predictive role of trust in isolation, without negating its behavioral relevance.

Second, although individual differences in reliance mechanisms are plausible and empirically observable, participant-specific parameter identification proves fragile under limited trial counts and near-deterministic behavior. Robustness diagnostics reveal frequent saturation and weak identifiability for parameterized mappings (Figures 8.2–8.3), constraining the interpretability of fitted parameters and motivating conservative use of personalized behavioral models in downstream analyses. As a result, participant-level behavioral fits are best interpreted descriptively rather than as stable psychological measurements.

Third, and most critically for end-to-end modeling, closing the trust–behavior loop using instantaneous trust-to-probability mappings exposes a mismatch between simulated and observed interaction structure. Coupled rollouts exhibit substantially higher switching rates and reduced persistence relative to empirical interaction sequences (Table 8.5), indicating that realistic reliance dynamics cannot be recovered without additional behavioral structure beyond trust-only coupling. In the present binary task with deterministic outcomes, stochastic action sampling combined with outcome inversion amplifies this limitation by inducing excessive alternation in the interaction stream.

Collectively, these limitations define the appropriate scope of the behavioral modeling layer introduced in this chapter. The behavioral models do not aim to provide deployment-ready prediction or comprehensive accounts of human decision-making. Instead, they serve as controlled probes that reveal where trust-driven explanations are informative, where they break down, and what constraints any future behavioral layer must satisfy in order to close the trust–behavior loop without degeneracy. Within this scope, the chapter fulfills its diagnostic purpose without placing additional demands on the validity of the upstream trust dynamics model.

8.11. Synthesis and Implications for Research Question 4

This section synthesizes the contributions of chapter 8 and explicitly addresses the fourth research question of this thesis. Research Question 4 concerns the feasibility, scope, and limitations of coupling the identified trust dynamics model to a behavioral decision model in order to form a closed-loop human-robot interaction (HRI) pipeline.

Research Question 4: *To what extent can the identified trust model be coupled with a behavioral decision model to form a closed-loop HRI pipeline, and what limitations emerge from this coupling?*

- RQ4.a** How can the inferred latent trust state be mapped to probabilistic models of human reliance and decision-making behavior?
- RQ4.b** Can the coupled trust–behavior model reproduce aggregate behavioral patterns observed in the experimental data when simulated in closed loop?
- RQ4.c** What limitations emerge when attempting to close the trust–behavior feedback loop, and what do these limitations reveal about the relationship between trust and observable human actions?

Mapping inferred trust to probabilistic reliance behavior (RQ4.a). This chapter demonstrates that the inferred latent trust state can be systematically mapped to probabilistic models of human reliance behavior through a family of simple, interpretable trust-to-action mappings. By treating reliance as a stochastic binary decision conditioned on trust (and, in some formulations, self-confidence), the behavioral layer establishes a principled interface between continuous trust trajectories and discrete observable actions. Importantly, the analysis shows that inferred trust carries detectable behavioral signal: under a minimal trust-as-probability interpretation, trust values preserve ordinal information about follow versus override tendencies, as evidenced by above-chance discrimination on held-out participants. At the same time, the results indicate that trust alone is insufficient to yield well-calibrated probabilistic predictions of action in a strongly imbalanced decision regime. Together, these findings establish that inferred trust is behaviorally meaningful but incomplete as a sole determinant of reliance, and that probabilistic trust-to-action mappings are best interpreted as diagnostic probes rather than comprehensive behavioral theories.

Closed-loop reproduction of aggregate interaction patterns (RQ4.b). By embedding the trust-to-action mapping within a closed-loop simulation framework, this chapter demonstrates that the identified trust model can be coupled with a behavioral decision model to generate self-consistent interaction trajectories. The resulting end-to-end pipeline integrates trust inference, action sampling, outcome generation, and trust updating in a fully closed feedback loop. However, the closed-loop simulations reveal that, under minimal coupling assumptions, the coupled system does not reproduce key aggregate behavioral patterns observed in the experimental data. In particular, simulated interactions exhibit substantially higher switching rates and reduced persistence of reliance states relative to observed behavior, indicating a qualitative mismatch in temporal interaction structure. These results provide a clear and informative answer to RQ4.b: while closed-loop coupling is technically feasible and internally consistent, reproducing realistic aggregate interaction dynamics requires additional behavioral structure beyond instantaneous trust-to-action mappings.

Limitations of trust–behavior coupling and implications for interpretation (RQ4.c). The limitations that emerge when closing the trust–behavior loop are structural rather than incidental. At the decision level, trust-conditioned behavioral mappings are dominated by base-rate effects and exhibit fragile parameter identifiability under limited trials and imbalanced outcomes, constraining their predictive and interpretive power. At the interaction level, closing the loop in a binary task with deterministic outcomes exposes a fundamental limitation of trust-only coupling: stochastic action sampling combined with outcome inversion induces excessive alternation and destabilizes reliance patterns unless additional sources of behavioral persistence or inertia are present. These limitations do not undermine the validity of the identified trust dynamics model. Instead, they clarify the role of trust within a broader decision-making process, highlighting that trust influences behavior without uniquely determining it. The findings therefore delineate a boundary condition for trust-aware modeling: trust can be robustly inferred and meaningfully interpreted as an internal state, but realistic behavioral and interaction dynamics cannot be recovered without modeling additional latent factors or decision mechanisms that operate alongside trust.

Taken together, the results of this chapter provide a principled and qualified answer to Research Question 4. The identified trust dynamics model can be coupled to a behavioral decision model to form a closed-loop HRI pipeline, and this pipeline functions as a transparent and informative diagnostic tool. The coupling reveals that inferred trust carries behavioral signal, that end-to-end loop closure is feasible under minimal assumptions, and that the resulting failures to reproduce observed interaction patterns are themselves informative. These limitations expose the gap between latent trust and observable action, reinforcing the interpretation of trust as a necessary but insufficient driver of behavior. In doing so, the chapter establishes clear constraints on trust-aware behavioral modeling and provides a structured foundation for future work that seeks to enrich the behavioral layer while preserving the validated trust dynamics at the core of the pipeline.

Integrated Discussion and System-Level Implications

The preceding chapters have developed, validated, and stress-tested a trust-aware modeling pipeline for human-robot interaction (HRI). Rather than pursuing a single monolithic objective, the thesis has addressed a sequence of increasingly demanding questions: how trust can be experimentally elicited and measured, how its latent dynamics can be identified from data, and how inferred trust states behave when embedded in downstream behavioral and closed-loop interaction models. This chapter provides an integrated discussion of these contributions, drawing connections across chapters 6, 7, and 8 and articulating their collective implications at the system level.

Rather than restating individual results or re-answering the research questions already addressed in dedicated synthesis sections, this chapter aims to (i) reassert the original end-to-end modeling objective of the thesis, (ii) highlight the strongest and most reliable findings, (iii) interpret the limitations revealed by downstream coupling attempts, and (iv) extract broader lessons for trust-aware HRI modeling that are not visible at the level of individual chapters. In doing so, it positions the thesis as a coherent whole and prepares the ground for the final conclusions and future research directions.

9.1. Reframing the End-to-End Modeling Objective

A central motivation of this thesis was the development of an end-to-end trust-aware modeling pipeline for HRI, spanning interaction design, trust inference, and downstream use of inferred trust states. Importantly, this objective was never framed as the construction of a fully deployable system or as the optimization of behavioral performance. Instead, the pipeline was intended as a structured means of probing how trust can be modeled, validated, and operationalized under realistic experimental and data constraints.

Chapters 6 through 8 can be understood as successive stages in this pipeline. Chapter 6 establishes the experimental substrate: a task design that elicits trust-relevant interaction dynamics while making explicit the limits of observability and measurement. Chapter 7 focuses on the identification and validation of a latent trust dynamics model, treating trust as an internal cognitive state that evolves over time and can be inferred from sparse, noisy reports. Chapter 8 then explores what happens when this inferred trust state is used downstream, both as a predictor of reliance decisions and as a driver in closed-loop interaction simulations.

Viewed together, these chapters do not attempt to demonstrate that trust modeling alone suffices to explain or predict human behavior. Instead, they collectively test whether a modular, trust-centered modeling approach can be instantiated end to end, and what its strengths and limitations are when subjected to increasingly stringent use cases.

9.2. Trust Dynamics Identification as a Core Contribution

Among the components of the proposed pipeline, the identification and validation of the trust dynamics model in chapter 7 represents the most robust contribution of the thesis. Using a carefully designed identification strategy with shared parameters and explicit training–validation separation, the trust model is shown to reproduce reported trust trajectories across participants while maintaining generalization to unseen data. Sensitivity and robustness analyses further clarify which parameters are well supported by the available data and which remain weakly identifiable.

Crucially, the trust model is validated on its own terms, using trust reports rather than behavioral actions as the primary reference. This separation ensures that trust is treated as a latent cognitive construct rather than as a proxy for compliance or performance. The resulting model captures temporal structure, response to interaction outcomes, and population-level regularities without requiring individualized parameter tuning. These properties make it a suitable foundation for simulation-based analysis and for integration into larger HRI modeling frameworks.

From a system-level perspective, the strength of the trust identification results is what enables the subsequent analyses in chapter 8 to be interpreted meaningfully. Because trust dynamics are already shown to be identifiable and predictive within their intended scope, downstream failures or mismatches cannot be attributed to an ill-posed or poorly estimated trust model. Instead, they reveal genuine limitations in how trust can be used once inferred.

9.3. Behavioral Coupling as a Diagnostic Stress Test

The trust-aware behavioral modeling explored in chapter 8 constitutes the most demanding stress test of the pipeline. Here, inferred trust states are no longer evaluated in isolation, but are mapped to observable decisions and embedded in closed-loop simulations. The resulting analyses are explicitly exploratory and diagnostic in nature, and their primary value lies in what they reveal about the limits of trust-based explanations.

At the decision level, trust-conditioned behavioral mappings demonstrate that inferred trust carries detectable behavioral signal, in the sense that higher trust values are associated with a greater likelihood of following system recommendations. However, this signal is insufficient to yield calibrated probabilistic prediction in a strongly imbalanced setting, where simple trust-free baselines dominate proper scoring metrics. This outcome highlights a fundamental distinction between behavioral relevance and behavioral sufficiency: trust influences decisions, but does not uniquely determine them.

The behavioral modeling results should not be interpreted as evidence that trust is irrelevant, but rather as evidence that trust alone is not sufficient to explain reliance behavior under the tested assumptions. This outcome is consistent with the broader behavioral modeling literature, in which observable decisions are shaped by multiple interacting mechanisms, including beliefs about the environment, goal structure, time pressure, affective state, cognitive biases, and strategic adaptation. Within this broader perspective, the present behavioral chapter serves as a diagnostic stress test: it confirms that inferred trust carries behavioral signal, while empirically delimiting the explanatory limits of trust when other latent variables are unobserved. This is a valuable result, because it clarifies where trust-aware pipelines are likely to succeed (capturing temporal trust structure) and where additional decision-model structure is required for calibrated action prediction and realistic interaction dynamics.

At the interaction level, closing the trust–behavior loop exposes additional structural limitations. Minimal trust-to-action couplings produce interaction sequences with substantially different switching and persistence properties than those observed empirically. These discrepancies arise even when trust estimation is held fixed and when parameter-free behavioral mappings are used, indicating that the mismatch is not a consequence of overfitting or parameter instability. Instead, it reflects the absence of additional behavioral mechanisms, such as inertia, strategy, or context-dependent policies, that operate alongside trust in real interactions.

Interpreted in this light, the behavioral modeling results function as a controlled diagnostic. By attempting to use trust in the simplest plausible ways, the analysis makes explicit which assumptions are too strong and what kinds of structure are missing. This diagnostic role is an essential part of the end-to-end objective, as it prevents overextension of the trust model beyond its empirically supported domain.

9.4. System-Level Lessons for Trust-Aware HRI Modeling

When considered jointly, the results of chapters 6 through 8 yield several system-level lessons that extend beyond the specifics of the experimental task or modeling choices.

First, trust benefits from being modeled as a latent internal state with its own dynamics, rather than being conflated with observable behavior. The clear separation between trust inference and behavioral modeling adopted in this thesis enables both components to be evaluated on appropriate grounds and avoids circular validation.

Second, trust appears better suited as a modulatory or contextual variable than as a direct behavioral policy. While inferred trust carries meaningful information about human reliance tendencies, realistic decision-making and interaction dynamics require additional latent factors and structural assumptions. Treating trust as one input among several, rather than as a standalone driver of action, is therefore a more defensible modeling stance.

Although the proposed model is identified from human-subject data and validated primarily as a descriptive predictor of reported trust trajectories, it is also compatible with an operational interpretation in which the robot maintains an online estimate of the human's latent trust state. In this setting, the trust dynamics model becomes a component of the autonomy itself: it can be used to anticipate reliance tendencies, detect potential trust miscalibration (overtrust or undertrust relative to observed performance and contextual risk), and trigger corrective adaptations. These adaptations do not necessarily mean modifying task logic, but may include changes in transparency, communication style, confidence signaling, explanation timing, or autonomy allocation intended to steer trust toward appropriate calibration. This perspective positions the proposed model not only as an explanatory tool, but also as a mechanism enabling trust-aware interaction policies in long-term and safety-relevant human-robot cooperation.

In many HRI systems, trust is not merely a psychological construct that indirectly affects behavior, but a practically relevant state variable that shapes safety and performance objectives. Trust calibration is directly relevant in settings such as supervisory control and human-on-the-loop monitoring, mixed-initiative decision support, adaptive autonomy and handover management, and safety-critical operation under uncertainty (e.g., emergency response, time pressure, or high consequences of error). In such contexts, the ability to represent trust as a dynamical quantity with interpretable structure is valuable even before committing to a detailed behavioral decision model, because it enables principled reasoning about how interaction events and context shape reliance readiness over time.

Third, closed-loop evaluation plays a critical role in revealing limitations that remain invisible under open-loop or pointwise predictive metrics. Interaction-level signatures such as switching asymmetries and persistence provide a richer lens on model behavior and expose failure modes that would not be apparent from accuracy or likelihood alone. This underscores the importance of evaluating trust-aware models at the level of interaction trajectories when they are intended for use in adaptive or long-term HRI scenarios.

Finally, the modular structure of the pipeline proves advantageous. Because trust dynamics, behavioral mappings, and coupling assumptions are explicitly separated, limitations in one component do not invalidate the others. This modularity supports incremental refinement and makes clear where future modeling effort should be concentrated.

9.5. Positioning Within the Broader Trust Modeling Landscape

Within the broader literature on trust-aware HRI, this thesis occupies a position that emphasizes interpretability, identifiability, and diagnostic clarity over end-to-end optimization. Many existing approaches either infer trust implicitly through behavior or embed trust-related variables directly into decision policies without independent validation. While such approaches can yield strong predictive performance in specific settings, they often obscure the relationship between latent trust, observable behavior, and interaction dynamics.

By contrast, the present work demonstrates the value of decoupling these elements and subjecting each to targeted evaluation. The trust dynamics model is validated against trust reports; behavioral mappings are tested for sufficiency and robustness; and closed-loop simulations are used to probe interaction structure rather than performance alone. This layered approach does not eliminate uncertainty, but it makes explicit where uncertainty resides and what assumptions are responsible for observed behavior.

As a result, the thesis contributes not only a specific trust model, but also a methodological template for studying trust in HRI as a latent, evolving construct whose role must be empirically delimited rather than assumed.

10

Conclusions and Future Recommendations

This chapter provides a standalone synthesis of the thesis, consolidating its main contributions, empirical findings, and methodological insights into a coherent closing perspective. Rather than reiterating the detailed analyses and technical developments presented in earlier chapters, the focus is on interpreting what was achieved, clarifying the scope and limitations of the proposed framework, and situating the results within the broader context of trust modeling in human-robot interaction (HRI).

The chapter first revisits the problem context and positions the thesis within the existing body of work on trust in HRI, emphasizing the identification-oriented perspective adopted throughout the research. It then summarizes the core contributions of the thesis and explicitly addresses the research questions, highlighting how each was answered through model formulation, experimental design, and empirical validation. The key findings are subsequently interpreted to extract broader insights into trust dynamics, parameter identifiability, and the relationship between trust and observable behavior.

Following this synthesis, the chapter discusses the main limitations of the present work, framing them as design choices and empirical constraints that define the boundaries of the reported results. The implications of the proposed framework for the design and evaluation of trust-aware HRI systems are then outlined, with particular emphasis on the role of identifiable trust dynamics models as diagnostic and supporting components within larger interaction architectures. Finally, the chapter concludes by identifying directions for future research, outlining how the proposed framework may be extended, refined, and applied in subsequent studies.

10.1. Problem Context and Thesis Positioning

This thesis addressed the problem of modeling human trust in human-robot interaction (HRI) as a dynamic, latent process that evolves over time and influences interaction outcomes. Trust has long been recognized as a critical factor in HRI, particularly in contexts characterized by uncertainty, time pressure, and asymmetric information, where humans must decide whether and how strongly to rely on robotic assistance. As robots are increasingly envisioned as collaborative partners rather than passive tools, the ability to represent, estimate, and reason about human trust has become an essential requirement for the design and evaluation of effective human-robot systems.

Despite substantial progress in the conceptual understanding of trust and the identification of factors that influence its formation and erosion, the literature has remained fragmented with respect to how trust should be modeled, identified, and empirically validated. Many existing approaches focused either on descriptive psychological constructs or on predictive trust metrics, without explicitly addressing whether proposed models could be identified from realistic human-subject data or reused across experimental contexts and applications. As a result, it has remained unclear how trust models can be systematically compared, validated, or integrated into broader modeling and simulation pipelines.

The central position adopted in this thesis was that trust modeling should be treated as an identification problem rather than solely as a predictive or descriptive exercise. From this perspective, trust was modeled as a latent, time-varying internal state whose structure, parameters, and evolution must be explicitly defined in a form that supports empirical estimation, validation, and reuse. This stance imposed a set of requirements

on the trust model, including interpretability, boundedness, executability, and compatibility with parameter identification techniques.

Accordingly, the thesis focused not on optimizing trust-related behavior or deploying trust-aware systems in real-world missions, but on developing and evaluating an end-to-end framework that links trust model formulation, experimental design, parameter identification, and validation within a coherent methodological pipeline. High-risk human-robot collaboration scenarios, such as search and rescue, were used as motivating contexts to inform task design and experimental conditions, rather than as domains for claims of operational readiness or mission-level performance. Within this scope, the thesis aimed to clarify what can and cannot be inferred about human trust from empirical data, and how trust dynamics models can be systematically stress-tested before being embedded in larger human-robot interaction systems.

10.2. Summary of Core Contributions

This thesis delivered an end-to-end framework for modeling, identifying, and evaluating dynamic human trust in human-robot interaction. Rather than treating trust modeling, data collection, and validation as isolated efforts, the work integrated these elements into a coherent pipeline that enabled systematic assessment of both model structure and empirical support.

A first core contribution was the formulation of a structured mathematical model of trust dynamics grounded in the three-layered trust framework proposed in the literature. The proposed model explicitly separated dispositional, situational, and learned components of trust, and was designed to be interpretable, bounded, and executable. This structure enabled trust to be represented as a continuous, time-varying latent state whose evolution could be analyzed and simulated under well-defined assumptions.

A second contribution was the design of a purpose-built human-subject experiment and trust measurement strategy tailored to the requirements of model identification. The experimental protocol was constructed to elicit informative trust dynamics while controlling for confounding factors, and the measurement approach combined validated subjective trust scales with behavioral observations. This alignment between experimental design and modeling objectives enabled the collection of data suitable for estimating latent trust dynamics despite noise and inter-individual variability.

Third, the thesis introduced and applied a principled identification and validation pipeline for dynamic trust models. This pipeline supported parameter estimation, assessment of identifiability and robustness, and evaluation of generalization to unseen participants. By comparing the proposed model against representative baseline trust models from the literature, the analysis contextualized empirical performance and clarified the trade-offs between model structure, flexibility, and identifiability.

Finally, the thesis contributed an exploratory coupling of the identified trust dynamics model with a behavioral decision model to form a closed-loop human-robot interaction pipeline. This coupling was used as a diagnostic tool to examine the extent to which inferred trust states could explain observed human actions. The resulting analysis provided empirical insight into the behavioral relevance and limitations of trust-based modeling assumptions, reinforcing the view of trust as a meaningful but partial contributor to human decision-making in HRI.

10.3. Answers to the Research Questions

This section summarizes how the research questions formulated in this thesis were addressed, based on the theoretical developments, empirical analyses, and simulation results presented in the preceding chapters. Each subsection provides a concise answer to one research question, followed by a brief interpretation of the corresponding findings and their implications.

10.3.1. Dynamic Trust Model Definition

The first research question examined how a dynamic mathematical model of human trust in human-robot interaction can be formally defined. The results demonstrated that the three-layered trust framework could be translated into an explicit, modular trust dynamics formulation that represents trust as a continuous, time-varying latent state. By separating dispositional, situational, and learned trust components, the proposed model captured both persistent individual tendencies and context-dependent variations in trust while remaining bounded and interpretable.

The formulation showed that trust dynamics could be expressed as an executable state-update process driven by interaction events, robot performance outcomes, and contextual inputs. This structure enabled the model to be discretized in time, simulated, and analyzed consistently across experimental trials, thereby

satisfying the requirements for integration into identification and validation pipelines. These results addressed the need for a mathematically explicit trust model that balances expressive power with structural clarity.

10.3.2. Parameter Identification from Empirical Data

The second research question focused on whether the parameters of the proposed trust dynamics model could be identified from empirical human-robot interaction data. The findings indicated that, when paired with a purpose-designed experimental protocol and measurement strategy, latent trust dynamics could indeed be inferred despite noisy and heterogeneous observations.

The identification results showed that several model parameters could be consistently estimated across participants, while others exhibited higher uncertainty or participant-specific variability. These outcomes highlighted the importance of aligning experimental design, trust measurement, and model structure with the requirements of parameter identification, and demonstrated that trust modeling can be approached as a data-driven estimation problem rather than solely as a conceptual exercise.

10.3.3. Empirical Validation and Identifiability

The third research question addressed the extent to which the proposed trust dynamics model could be empirically validated. Validation analyses demonstrated that the model was able to reproduce observed trust trajectories and generalize to unseen participants under the considered experimental conditions. Comparison with baseline trust models from the literature further contextualized performance and revealed trade-offs between model flexibility and identifiability.

At the same time, the results showed that not all model parameters were equally identifiable. In particular, some parameters exhibited weak sensitivity in the experimental context, indicating that their influence on observed trust trajectories was limited under the tested conditions. These findings clarified the boundaries of what could be inferred from the available data and underscored the context-dependent nature of parameter identifiability in trust dynamics models.

10.3.4. Exploratory Trust-Behavior Coupling

The fourth research question explored whether the identified trust model could be coupled with a behavioral decision model to form a closed-loop human-robot interaction pipeline, and what limitations would emerge from this coupling. The results showed that inferred trust states provided meaningful information for explaining aggregate patterns of human reliance and intervention behavior, but were not sufficient to fully reproduce individual decision-making trajectories.

This outcome indicated that trust functions as a relevant latent moderator of behavior rather than as a standalone decision policy. The exploratory coupling therefore served as a diagnostic tool, revealing where trust-based assumptions aligned with observed behavior and where additional cognitive, contextual, or motivational factors were required. These findings reinforced the interpretation that limitations in behavioral reproduction reflected the inherent complexity of human decision-making, rather than deficiencies in the trust dynamics model itself.

10.4. Interpretation of Key Findings

Beyond addressing the research questions individually, the results of this thesis provided several broader insights into the nature of trust dynamics and their role in human-robot interaction. In particular, the findings clarified how model structure, experimental context, and identification constraints interact to shape what can be inferred about human trust from empirical data.

A key observation was that the proposed structured trust dynamics model produced smoother and more interpretable trust trajectories than representative baseline formulations. This behavior reflected the explicit separation of trust components and the continuous-time evolution imposed by the model structure. From an identification perspective, smoother trust trajectories were preferable, as they reduced sensitivity to isolated measurement noise and enabled more stable parameter estimation. This result suggested that structural regularization introduced through model design can be a strength rather than a limitation when modeling latent psychological states.

The analysis of parameter sensitivity and identifiability further highlighted the context-dependent nature of trust dynamics. While parameters associated with learned trust exhibited a clear influence on trust evolution under the experimental conditions, situational trust parameters showed weaker identifiability. Importantly, this outcome did not imply that situational trust was irrelevant to the model, but rather that the specific experimental

context did not sufficiently excite these mechanisms. The model formulation itself allowed situational inputs to influence trust, and the limited sensitivity observed reflected the interaction between model structure, parameter selection, and task design. This distinction clarified the difference between model capability and empirical observability.

The exploratory coupling between trust dynamics and behavioral decision-making yielded additional interpretive insight. Although inferred trust states correlated with aggregate behavioral trends, they did not uniquely determine individual actions. This finding reinforced the view of trust as a latent moderator of behavior rather than a direct control variable. The inability of trust alone to reproduce detailed decision-making trajectories was therefore interpreted as an expected and informative outcome, consistent with the multifactorial nature of human behavior. Rather than undermining the trust model, this result delineated its appropriate scope within a broader decision-making framework.

As a whole, these findings emphasized that the value of a trust dynamics model lies not only in its predictive accuracy, but in its ability to clarify where trust matters, where it does not, and under what conditions its influence becomes observable. The results demonstrated that careful model structuring and empirical stress-testing can reveal meaningful insights about the role of trust in HRI, even when some parameters or behaviors remain only partially explained.

10.5. Limitations of the Present Work

While this thesis achieved its objective of developing and empirically evaluating an end-to-end framework for dynamic trust modeling in human-robot interaction, several limitations of the present work must be acknowledged. These limitations do not invalidate the results, but rather define the scope within which the findings should be interpreted and contextualized.

A primary limitation concerned the scale and diversity of the empirical dataset. Although the experimental design was purpose-built for trust model identification, the number of participants and interaction scenarios necessarily constrained the range of observable trust dynamics. Inter-individual variability was present and partially captured, but the limited sample size restricted the ability to draw strong conclusions about population-level parameter distributions or demographic effects. As a result, the identified model parameters should be interpreted as indicative rather than universally representative.

A second limitation related to the experimental task design. While the task was structured to elicit dynamic trust evolution under controlled conditions, it necessarily simplified many aspects of real-world human-robot interaction. In particular, the operational complexity, environmental uncertainty, and social context were intentionally reduced to support identifiability and experimental control. This abstraction limited the extent to which situational trust mechanisms could be empirically excited, and consequently constrained the identifiability of certain model components.

The reliance on self-reported trust measurements introduced an additional limitation. Despite the use of validated trust scales and careful timing of measurements, subjective reports remained susceptible to cognitive biases, anchoring effects, and individual differences in scale interpretation. Behavioral measures were incorporated to complement these reports, but they did not provide a complete or continuous proxy for latent trust. As a result, the trust state remained indirectly observed, and estimation performance was inherently bounded by measurement fidelity.

From a modeling perspective, the trust dynamics formulation intentionally favored interpretability and identifiability over expressive complexity. While this design choice facilitated empirical validation, it also limited the model's ability to capture higher-order nonlinearities or abrupt trust shifts that may occur in more extreme or emotionally charged interactions. The model should therefore be understood as a first-order approximation of trust evolution rather than a comprehensive psychological representation.

Finally, the exploratory trust-behavior coupling constituted a conceptual rather than predictive limitation. The behavioral decision model was not intended to reproduce individual actions with high fidelity, and its simplified structure constrained the conclusions that could be drawn regarding closed-loop human-robot interaction. This limitation was accepted by design, as the coupling served primarily as a diagnostic tool to assess the behavioral relevance of inferred trust states rather than as a definitive decision-making model.

Taken together, these limitations highlighted the trade-offs inherent in empirically grounded trust modeling. The present work prioritized structural clarity, identifiability, and methodological rigor, and the resulting constraints should be viewed as deliberate design choices that enable cumulative refinement in future research.

10.6. Implications for Trust-Aware Human-Robot Interaction Systems

Despite its exploratory scope, the framework developed and evaluated in this thesis carried several implications for the design and assessment of trust-aware human-robot interaction (HRI) systems. Rather than proposing trust as a direct driver of behavior, the results demonstrated the value of treating trust as an internal, latent state that can be estimated, monitored, and reasoned about within broader interaction architectures.

First, the findings suggested that identifiable trust dynamics models can function as internal estimators of human trust in real time or near real time. By structuring trust as a bounded, dynamically evolving state with interpretable components, the model enabled principled inference of trust trajectories from noisy and partial observations. This capability has direct relevance for adaptive HRI systems that seek to monitor human trust without relying on continuous explicit feedback, particularly in time-critical or cognitively demanding tasks.

Second, the results highlighted the potential role of trust models as diagnostic tools for identifying trust miscalibration. The separation between trust estimation and behavioral decision-making made it possible to detect discrepancies between inferred trust states and observed reliance behavior. Such discrepancies may indicate overtrust, undertrust, or the influence of external situational factors, and could be leveraged to trigger transparency interventions, explanation strategies, or changes in autonomy level. In this sense, trust modeling supported system awareness rather than behavioral control.

Third, the work illustrated how trust dynamics models can be integrated into simulation and control pipelines as supporting components rather than complete behavioral models. By decoupling trust estimation from action selection, the framework avoided overloading trust with explanatory responsibility for complex human behavior. This modular integration aligns with existing practices in cognitive and agent-based modeling, where latent mental states inform but do not replace decision structures. As a result, trust modeling can be incorporated incrementally into existing systems without requiring wholesale redesign of interaction logic.

Finally, the emphasis on identifiability and empirical validation carried implications for the evaluation of trust-aware systems. The results underscored the importance of assessing not only whether a model can reproduce trust measurements, but also whether its parameters can be meaningfully inferred and generalized across users. From a system design perspective, this shifts the focus from optimizing trust-related outcomes to ensuring that trust representations are interpretable, testable, and robust under realistic data constraints.

These implications positioned trust dynamics modeling as an enabling layer for trust-aware HRI, supporting monitoring, diagnosis, and adaptation while remaining compatible with richer behavioral and decision-making frameworks.

10.7. Future Research and Recommendations

While this thesis focused on the formulation, identification, and exploratory evaluation of a dynamic trust model under controlled experimental conditions, it also opens several clear directions for future research. These directions follow naturally from the structure of the thesis itself, progressing from extensions of the trust model formulation, through experimental design and identification, to behavioral integration and system-level considerations. Collectively, they outline a research trajectory in which the proposed framework serves as an initial, structured foundation rather than a final or exhaustive model of human trust.

10.7.1. Incremental Evolution of the Trust Model Structure

The trust dynamics model proposed in this thesis was deliberately designed to be deterministic, modular, and interpretable, prioritizing structural clarity and identifiability over representational completeness. As such, a natural direction for future work is the incremental evolution of the model through the introduction of additional trust-affecting components motivated by existing literature.

Rather than replacing the current formulation, future research could extend it by incorporating new modules or augmenting existing ones to account for factors such as transparency, communication quality, social cues, or task workload effects. This modular approach would allow the model to grow iteratively while preserving its engineering interpretability and execution properties. In this sense, the present model should be viewed as a scaffold that supports systematic refinement as empirical evidence and application requirements evolve.

10.7.2. Identification-Oriented Experimental Design and Data Collection

The identification results obtained in this thesis highlighted the strong dependence of parameter identifiability on experimental excitation. In particular, situational and reputation-related trust components were only weakly excited under the chosen experimental conditions, limiting the extent to which their parameters could be robustly identified.

Future experiments could therefore be designed explicitly with identifiability in mind, deliberately manipulating task conditions, environmental uncertainty, or system behavior to excite specific trust components. In addition, collecting larger and more diverse datasets would enable more detailed analysis of inter-individual variability and robustness, and could support exploratory investigation of whether demographic or personality-related factors systematically influence trust dynamics. Such data would also enable more data-driven analyses that complement the model-based identification pipeline developed in this work.

10.7.3. Modeling and Eliciting Risk in Ethical Experimental Settings

Risk played a central role in the proposed trust dynamics model as an input shaping situational trust. However, defining and eliciting perceived risk in human-subject experiments remains a methodological challenge, particularly when ethical constraints limit the degree of real danger that can be imposed on participants.

Future research should therefore address both the modeling and experimental aspects of risk. On the modeling side, alternative representations of perceived risk could be explored, including more nuanced mappings between task features and subjective risk perception. On the experimental side, new paradigms may be developed to elicit meaningful risk responses in safe and ethically acceptable ways, for instance through time pressure, outcome uncertainty, or socially framed consequences. Improving the treatment of risk would strengthen both the realism and identifiability of trust dynamics models.

10.7.4. Generalization Across Tasks, Domains, and Interaction Contexts

Although the experimental task in this thesis was inspired by search-and-rescue scenarios, the proposed trust dynamics framework was not inherently tied to a specific domain. A key direction for future work is therefore to assess how well the model generalizes across different tasks, levels of autonomy, and interaction modalities.

Applying the framework to alternative human-robot interaction contexts would allow investigation of which aspects of the trust dynamics are domain-invariant and which require task-specific adaptation. Such studies would also help clarify the limits of generalizability of a single trust model structure and inform principled decisions about when domain-specific extensions are warranted.

10.7.5. Population-Level Modeling and Structured Personalization

A defining feature of the present work is the use of a shared parameter set across participants, emphasizing population-level modeling rather than individualized fitting. While this approach supports interpretability and robustness, future research may investigate structured forms of personalization without abandoning the advantages of a shared model.

Two complementary avenues are possible. First, empirical clustering of observed trust trajectories could reveal distinct classes of trust dynamics, suggesting a small number of representative parameter regimes rather than fully individualized models. Second, the model structure itself could be expanded to include personalization inputs, such as dispositional traits or experience-related modifiers, while keeping the core parameters universal. Both approaches preserve the engineering coherence of the framework while allowing richer representation of individual differences.

10.7.6. Integration with Cognitive and Agent-Based Decision Models

The exploratory trust-behavior coupling investigated in this thesis demonstrated that trust alone is insufficient to fully explain human decision-making behavior. This finding motivates future work that integrates trust dynamics with broader cognitive or agent-based decision models.

In such models, trust would act as one latent internal state alongside beliefs, goals, workload, and other contextual variables that jointly shape behavior. Integrating the proposed trust model into cognitive architectures or agent-based frameworks would enable more comprehensive representations of human decision-making while preserving the interpretability of the trust component. This direction aligns with the view of trust as a guiding but non-deterministic factor in human-robot interaction.

10.7.7. Alternative and Complementary Trust Measurement Modalities

The present work relied on validated self-report instruments to ground trust measurement and model identification. While this choice ensured psychometric validity, future research may explore complementary measurement modalities to enhance temporal resolution and reduce intrusiveness.

Combining subjective trust reports with behavioral indicators or physiological measurements could provide richer observational data for trust inference and validation. Multi-modal measurement strategies may also

support cross-validation of inferred trust states and improve robustness in scenarios where self-reporting is impractical or disruptive.

10.7.8. Adaptive Trust-Aware Robot Policies

Beyond modeling and identification, the proposed framework lends itself to use as an internal estimator within adaptive robotic systems. Future work could explore how inferred trust states may inform robot behavior, for instance by adjusting transparency, communication strategies, or levels of autonomy in response to detected trust miscalibration.

Importantly, such adaptive policies should be evaluated carefully to avoid manipulative or unsafe trust optimization. Rather than maximizing trust, the objective would be to support appropriate calibration that aligns human expectations with system capabilities. Investigating these strategies would extend the framework from analysis toward interaction design, while remaining grounded in the trust dynamics identified in this thesis.

10.7.9. Ethical and Normative Considerations

As trust-aware systems become more capable of estimating and responding to human internal states, ethical and normative questions become increasingly relevant. Future research should examine the boundaries of acceptable trust adaptation, including issues of transparency, consent, and responsibility.

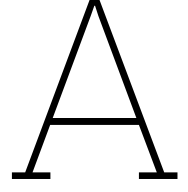
While this thesis adopted an engineering perspective, acknowledging these broader implications is essential for responsible deployment. Addressing such considerations early can help ensure that trust-aware technologies are developed in ways that respect human autonomy and safety.

10.7.10. Real-World Deployment Considerations

Finally, a long-term research direction concerns the application of trust dynamics models under real-world deployment constraints. These include partial observability, sensor noise, missing data, and computational limitations associated with embedded systems and online estimation.

Although the present work does not claim deployment readiness, investigating these constraints would be a necessary step toward practical use of trust-aware models in operational systems. Simulation-based studies and incremental field evaluations could provide a bridge between controlled experimental research and real-world application.

Together, these future directions position the proposed framework as a foundation for continued research into interpretable, identifiable, and behaviorally meaningful models of human trust in human-robot interaction.



Full Derivations of the Proposed Trust Dynamics Model

This appendix provides the complete mathematical derivations supporting the final Trust Dynamics Model presented in chapter 5. All notation and variable definitions follow the same rules and conventions used in that chapter. The purpose of this appendix is traceability: each subsection derives one component or update rule used in the final model.

A.1. Dispositional Trust

This section states the formulation of Dispositional Trust, τ^{disp} , as used in (5.2). It is treated as time-invariant over the simulation horizon.

$$\tau^{\text{disp}} = \alpha^{\text{disp}} \delta^{\text{bio}} + (1 - \alpha^{\text{disp}}) \delta^{\text{env}}, \quad \alpha^{\text{disp}}, \delta^{\text{bio}}, \delta^{\text{env}} \in [0, 1]. \quad (\text{A.1})$$

A.2. Situational Trust

This section records the definition of Situational Trust, τ^{sit} , used in the final model (see (5.3)). It is treated as invariant in time, a function of risk r , and is conditioned on the ordinal relationship between τ^{disp} and self-confidence (β).

$$\tau^{\text{sit}}(r) = \begin{cases} 2e^{-\lambda^{\text{sit}} r} - 1, & \text{if } \tau^{\text{disp}} \leq \beta, \\ 1 - 2e^{-\lambda^{\text{sit}} r}, & \text{if } \tau^{\text{disp}} > \beta, \end{cases} \quad (\text{A.2})$$

A.3. Learned Trust Components

This section derives the Learned Trust subcomponents used in the final model: Reputation τ^{rep} and Personal Experience τ^{exp} .

A.3.1. Reputation

This subsection derives the discrete-time update for the Reputation component τ^{rep} (see (5.19)) starting from the continuous-time exponential form in (A.3).

$$\tau^{\text{rep}}(t) = \tau_0^{\text{rep}} e^{-\lambda^{\text{rep}} t} \quad (\text{A.3})$$

The discrete-time recursion for $\tau^{\text{rep}}[k]$ is obtained by sampling (A.3) with step size Δt in (A.4), yielding (A.5).

$$\tau^{\text{rep}}[k+1] = \tau_0^{\text{rep}} e^{-\lambda^{\text{rep}}(k+\Delta t)} = \tau_0^{\text{rep}} e^{-\lambda^{\text{rep}}k} e^{-\lambda^{\text{rep}}\Delta t} = \tau^{\text{rep}}[k] e^{-\lambda^{\text{rep}}\Delta t} \quad (\text{A.4})$$

$$\begin{cases} \tau^{\text{rep}}[0] = \tau_0^{\text{rep}} \\ \tau^{\text{rep}}[k+1] = \tau^{\text{rep}}[k] e^{-\lambda^{\text{rep}}\Delta t} \end{cases} \quad (\text{A.5})$$

A.3.2. Personal Experience

This subsection derives the Personal Experience increments used in the final model to represent event-driven trust changes associated with HRI successes and failures. The increment magnitude is indexed by the number of consecutive events.

Modeling HRI Failures

This subsection derives the failure-induced Personal Experience increment $\tau^{\text{exp,fail}}(n)$ used in the final model (see (5.6) and (5.7)). The derivation starts from a generic exponential decay law and defines the n^{th} decrement as the difference between consecutive samples.

$$y^{\text{fail}}(x) = y_{\min}^{\text{fail}} + (y_0^{\text{fail}} - y_{\min}^{\text{fail}}) e^{-\lambda^{\text{fail}}(x-x_0)}. \quad (\text{A.6})$$

For the failure case, set $x_0 = 0$, $y_0^{\text{fail}} = 1$, and $y_{\min}^{\text{fail}} = 0$ in (A.6), yielding the unit exponential decay in (A.7).

$$y^{\text{fail}}(x) = e^{-\lambda^{\text{fail}}x}, \quad x > 0, \lambda^{\text{fail}} > 0. \quad (\text{A.7})$$

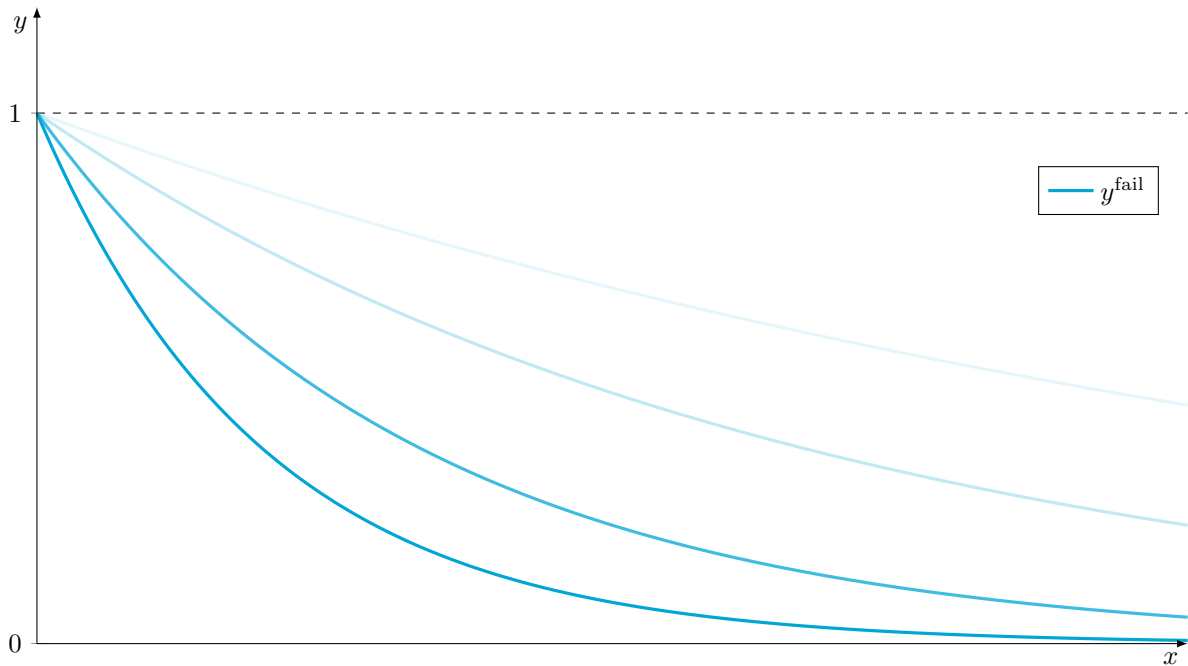


Figure A.1: Unit exponential decay curve. Curves with lower opacity represent lower magnitude decay rates λ^{fail} .

The following result establishes that consecutive sample differences of the unit exponential decay decrease with the sample index.

THEOREM 1 For a constant $\Delta s > 0$, the difference between consecutive samples of the unit exponential decay function will decrease with the number of samples.

PROOF 1 Consider any $n \in \mathbb{Z}$ such that:

$$y^{\text{fail}}((n+1)\Delta s) - y^{\text{fail}}(n\Delta s) = e^{-\lambda^{\text{fail}} n\Delta s} - e^{-\lambda^{\text{fail}} (n-1)\Delta s}. \quad (\text{A.8})$$

Now let us define any $m, j \in \mathbb{Z}$ so that $n = m + j$. Note that from here follows that $n > m$. Substituting n in the right half of the previous equation, we then get:

$$\begin{aligned} y^{\text{fail}}((n+1)\Delta s) - y^{\text{fail}}(n\Delta s) &= e^{-\lambda^{\text{fail}} (m+j)\Delta s} - e^{-\lambda^{\text{fail}} (m+j-1)\Delta s} \\ &= e^{-\lambda^{\text{fail}} j\Delta s} \left(e^{-\lambda^{\text{fail}} m\Delta s} - e^{-\lambda^{\text{fail}} (m-1)\Delta s} \right) \\ &= e^{-\lambda^{\text{fail}} j\Delta s} \left(y^{\text{fail}}((m+1)\Delta s) - y^{\text{fail}}(m\Delta s) \right). \end{aligned} \quad (\text{A.9})$$

Because $\lambda^{\text{fail}}, j, \Delta s > 0$, we know that:

$$0 < e^{-\lambda^{\text{fail}} j\Delta s} < 1 \Rightarrow e^{-\lambda^{\text{fail}} j\Delta s} \left(y^{\text{fail}}((m+1)\Delta s) - y^{\text{fail}}(m\Delta s) \right) < y^{\text{fail}}((m+1)\Delta s) - y^{\text{fail}}(m\Delta s). \quad (\text{A.10})$$

Note that the left half of the inequality is the same as $y^{\text{fail}}((n+1)\Delta s) - y^{\text{fail}}(n\Delta s)$, concluding our proof.

$$\therefore y^{\text{fail}}((n+1)\Delta s) - y^{\text{fail}}(n\Delta s) < y^{\text{fail}}((m+1)\Delta s) - y^{\text{fail}}(m\Delta s), \quad \forall n > m. \quad (\text{A.11})$$

Using this property, the failure impulse at the n^{th} consecutive failure is defined as the difference between two consecutive samples. Figure A.2 illustrates the sampling construction.

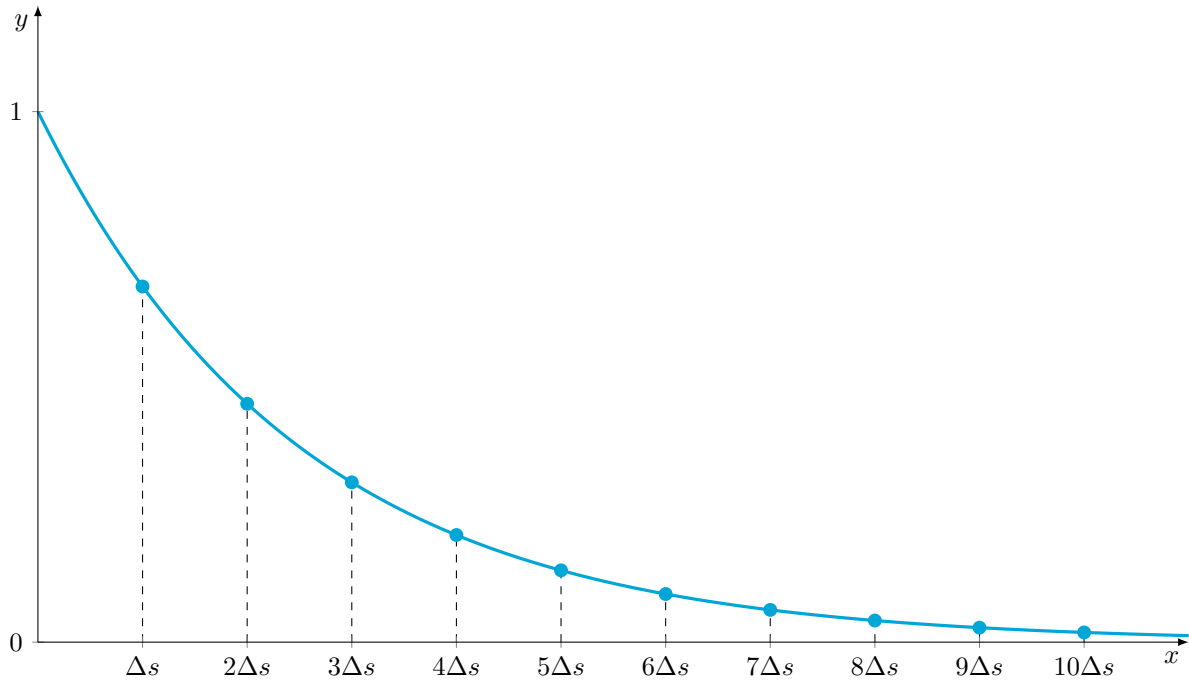


Figure A.2: Sampled exponential decay.

Accordingly, the Personal Experience increment for the n^{th} consecutive failure is defined as:

$$\begin{aligned} \tau^{\text{exp, fail}}(n) &= y^{\text{fail}}(n\Delta s) - y^{\text{fail}}((n-1)\Delta s), \quad n \in \mathbb{Z} \\ &= e^{-\lambda^{\text{fail}} n\Delta s} - e^{-\lambda^{\text{fail}} (n-1)\Delta s} \\ &= e^{-\lambda^{\text{fail}} n\Delta s} - e^{-\lambda^{\text{fail}} n\Delta s} e^{\lambda^{\text{fail}} \Delta s} \\ &= e^{-\lambda^{\text{fail}} n\Delta s} \left(1 - e^{\lambda^{\text{fail}} \Delta s} \right). \end{aligned} \quad (\text{A.12})$$

Equation (A.12) contains two free parameters, Δs and λ^{fail} . To constrain the update, define ϕ^{fail} such that the first failure yields $\tau^{\text{exp,fail}}(1) = -\phi^{\text{fail}}$:

$$\begin{aligned}
 \tau^{\text{exp,fail}}(1) &= -\phi^{\text{fail}} \\
 e^{-\lambda^{\text{fail}} \Delta s} (1 - e^{\lambda^{\text{fail}} \Delta s}) &= -\phi^{\text{fail}} \\
 e^{-\lambda^{\text{fail}} \Delta s} - e^{-\lambda^{\text{fail}} \Delta s} e^{\lambda^{\text{fail}} \Delta s} &= -\phi^{\text{fail}} \\
 e^{-\lambda^{\text{fail}} \Delta s} - 1 &= -\phi^{\text{fail}} \\
 \lambda^{\text{fail}} \Delta s &= -\ln(1 - \phi^{\text{fail}}) .
 \end{aligned} \tag{A.13}$$

Equation (A.13) fixes the parameters such that $y^{\text{fail}}(\Delta s) = 1 - \phi^{\text{fail}}$. Figure A.3 visualizes the effect of varying Δs under this constraint.

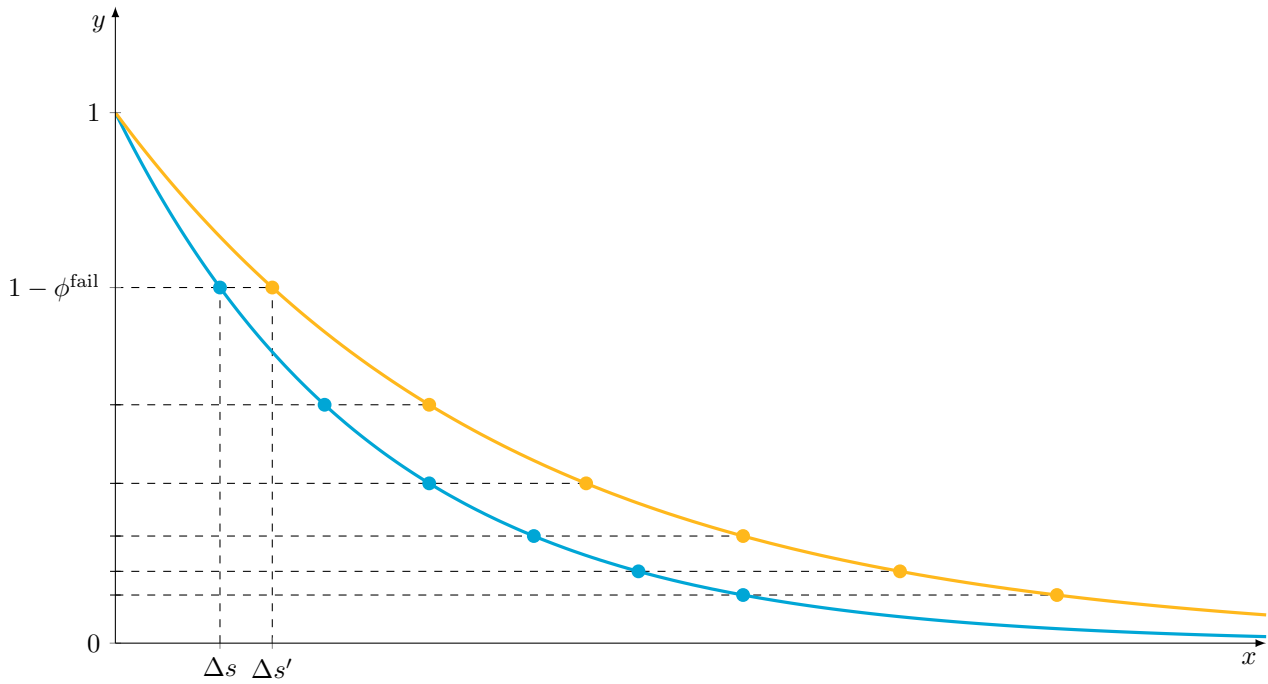


Figure A.3: Role of sampling interval.

Under the constraint in (A.13), changes in Δs are compensated by λ^{fail} such that the curve passes through $(\Delta s, 1 - \phi^{\text{fail}})$.

Setting $\Delta s = 1$ yields the simplified expressions in (A.14) and (A.15).

$$\tau^{\text{exp,fail}}(n) = e^{-\lambda^{\text{fail}} n} (1 - e^{\lambda^{\text{fail}}}) \quad , \text{ with} \tag{A.14}$$

$$\lambda^{\text{fail}} = -\ln(1 - \phi^{\text{fail}}) . \tag{A.15}$$

Modeling HRI Successes

This subsection derives the success-induced Personal Experience increment $\tau^{\text{exp,succ}}(n)$ from a logistic function (see (5.8), (5.10) and (5.11)). The derivation starts from the generic form in (A.16).

$$y^{\text{succ}}(x) = \frac{L}{1 + e^{-\kappa^{\text{succ}}(x - x_{1/2})}} . \tag{A.16}$$

For the success case, impose the boundary conditions $y^{\text{succ}}(0) = 0$ and $\lim_{x \rightarrow \infty} y^{\text{succ}}(x) = 1$. This requires shifting the lower asymptote and setting the carrying capacity to $L = 1$.

Equation (A.17) generalizes this function so that its codomain is the interval $] -a, L[$, with $a > 0$.

$$y^{\text{succ}}(x) = -a + \frac{L + a}{1 + e^{-\kappa^{\text{succ}}(x - x_{1/2})}}. \quad (\text{A.17})$$

Setting $L = 1$ in (A.17) yields (A.18). Figure A.4 illustrates the effect of the lower asymptote parameter a .

$$y^{\text{succ}}(x) = -a + \frac{1 + a}{1 + e^{-\kappa^{\text{succ}}(x - x_{1/2})}}, \quad a \in \mathbb{R}^+. \quad (\text{A.18})$$

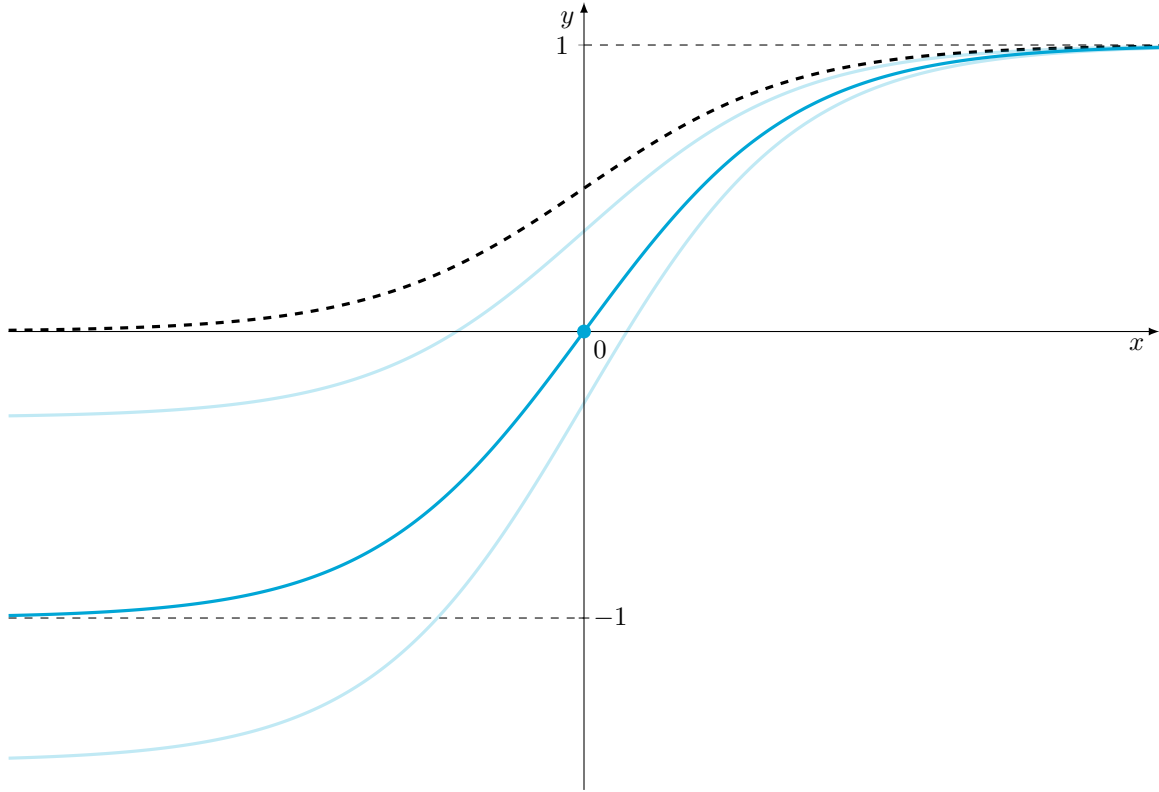


Figure A.4: Role of a .

Imposing the condition $y^{\text{succ}}(0) = 0$ yields the following constraint:

$$\begin{aligned} y_0^{\text{succ}} &= y^{\text{succ}}(0) = 0 \\ \Leftrightarrow -a + \frac{1 + a}{1 + e^{-\kappa^{\text{succ}}(0 - x_{1/2})}} &= 0 \\ \Leftrightarrow 1 + e^{\kappa^{\text{succ}} x_{1/2}} &= \frac{a + 1}{a} \\ \Leftrightarrow a + a e^{\kappa^{\text{succ}} x_{1/2}} &= a + 1 \\ \Leftrightarrow a &= e^{-\kappa^{\text{succ}} x_{1/2}}. \end{aligned} \quad (\text{A.19})$$

A nontrivial parameterization is required to satisfy $y^{\text{succ}}(0) = 0$ while maintaining a gradual initial increase for $x > 0$. Accordingly, a , κ^{succ} , and $x_{1/2}$ are constrained using (A.19) and (A.21).

Choosing $x_{1/2} \gg 0$ shifts the logistic curve to the right and reduces its initial slope for x near 0. Figure A.5 illustrates the effect of varying $x_{1/2}$.

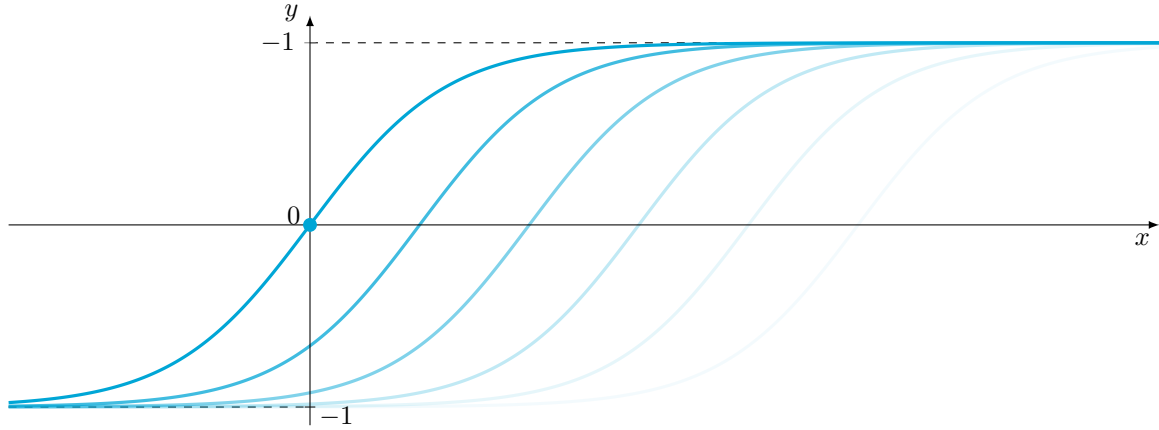


Figure A.5: Role of $x_{1/2}$

With $\kappa^{\text{succ}} > 0$ and $x_{1/2} > 0$, (A.19) implies the domain constraint $0 < a < 1$:

$$a = e^{-\kappa^{\text{succ}} x_{1/2}}, \quad \kappa^{\text{succ}} x_{1/2} > 0 \Rightarrow 0 < a < 1 \quad (\text{A.20})$$

Introduce a sampling interval $\Delta s > 0$ and impose the calibration constraint $y^{\text{succ}}(\Delta s) = \phi^{\text{succ}}$:

$$\begin{aligned} y^{\text{succ}}(\Delta s) &= \phi^{\text{succ}} \\ \Leftrightarrow -a + \frac{1+a}{1+e^{-\kappa^{\text{succ}}(\Delta s - x_{1/2})}} &= \phi^{\text{succ}} \\ \Leftrightarrow 1 + e^{-\kappa^{\text{succ}}(\Delta s - x_{1/2})} &= \frac{1+a}{\phi^{\text{succ}} + a} \\ \Leftrightarrow e^{-\kappa^{\text{succ}}(\Delta s - x_{1/2})} &= \frac{1+a}{\phi^{\text{succ}} + a} - 1 \\ \Leftrightarrow -\kappa^{\text{succ}}(\Delta s - x_{1/2}) &= \ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right). \end{aligned} \quad (\text{A.21})$$

Solving (A.19) and (A.21) yields κ^{succ} and $x_{1/2}$ as functions of $(a, \Delta s, \phi^{\text{succ}})$:

$$\begin{aligned} \begin{cases} -\kappa^{\text{succ}}(\Delta s - x_{1/2}) = \ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) \\ a = e^{-\kappa^{\text{succ}} x_{1/2}} \end{cases} &\Leftrightarrow \begin{cases} \kappa^{\text{succ}} x_{1/2} - \kappa^{\text{succ}} \Delta s = \ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) \\ \kappa^{\text{succ}} x_{1/2} = -\ln(a) \end{cases} \\ \Leftrightarrow \begin{cases} -\ln(a) + \frac{\ln(a) \Delta s}{x_{1/2}} = \ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) \\ \kappa^{\text{succ}} = -\frac{\ln(a)}{x_{1/2}} \end{cases} &\Leftrightarrow \begin{cases} \ln(a) \left(\frac{\Delta s}{x_{1/2}} - 1\right) = \ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) \\ \kappa^{\text{succ}} = -\frac{\ln(a)}{x_{1/2}} \end{cases} \\ \Leftrightarrow \begin{cases} \frac{\Delta s}{x_{1/2}} = \frac{\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right)}{\ln(a)} + 1 \\ \kappa^{\text{succ}} = -\frac{\ln(a)}{x_{1/2}} \end{cases} &\Leftrightarrow \begin{cases} x_{1/2} = \Delta s \frac{\ln(a)}{\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) + \ln(a)} \\ \kappa^{\text{succ}} = -\frac{1}{\Delta s} \ln(a) \frac{\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) + \ln(a)}{\ln(a)} \end{cases} \\ \Leftrightarrow \begin{cases} x_{1/2} = \Delta s \frac{\ln(a)}{\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) + \ln(a)} \\ \kappa^{\text{succ}} = -\frac{1}{\Delta s} \left[\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}} + a}\right) + \ln(a) \right] \end{cases} & . \end{aligned} \quad (\text{A.22})$$

We now have expressions for κ^{succ} and $x_{1/2}$ as functions of a , Δs and ϕ^{succ} , but we have yet to define a . An option is to try and find an explicit function for a . For this, consider that we want $y^{\text{succ}}(x)$ to pass through a general point (x_1, y_1) so that:

$$\begin{aligned} y^{\text{succ}}(x_1) &= y_1 \\ \Leftrightarrow y_1 &= -a + \frac{1+a}{1 + e^{-\kappa^{\text{succ}}(x_1 - x_{1/2})}} \\ \Leftrightarrow y_1 &= -a + \frac{1+a}{1 + e^{\kappa^{\text{succ}} x_{1/2}} e^{-\kappa^{\text{succ}} x_1}} \end{aligned} \quad (\text{A.23})$$

Substituting from (A.22), we get:

$$\begin{aligned} \Rightarrow y_1 &= -a + \frac{1+a}{1 + e^{-\ln(a)} e^{\left[\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}}+a}\right) + \ln(a)\right] \frac{x_1}{\Delta s}}} \\ \Leftrightarrow y_1 &= -a + \frac{1+a}{1 + \frac{1}{a} \left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}}+a}\right)^{\frac{x_1}{\Delta s}} a^{\frac{x_1}{\Delta s}}} \end{aligned} \quad (\text{A.24})$$

Equation (A.24) cannot be generally solved for a . Moreover, it would be possible to solve this equation for $\frac{x_1}{\Delta s} = 1$ or possibly other small integers (2, 3, ...). However, if we think about the meaning of $\frac{x_1}{\Delta s}$ for our application case, we realize that we can either have a high sampling frequency (small Δs) but define a (x_1, y_1) point close to the origin, or we can define a (x_1, y_1) point far from the origin but have a low sampling frequency (high Δs).

Equation (A.24) does not admit a closed-form solution for a in the general case. Accordingly, a is treated as a free calibration parameter in the final model.

Given these results, we are left with two degrees of freedom in a and in Δs . Figure A.6 shows the effect of changing the sampling rate for a specific value of a .

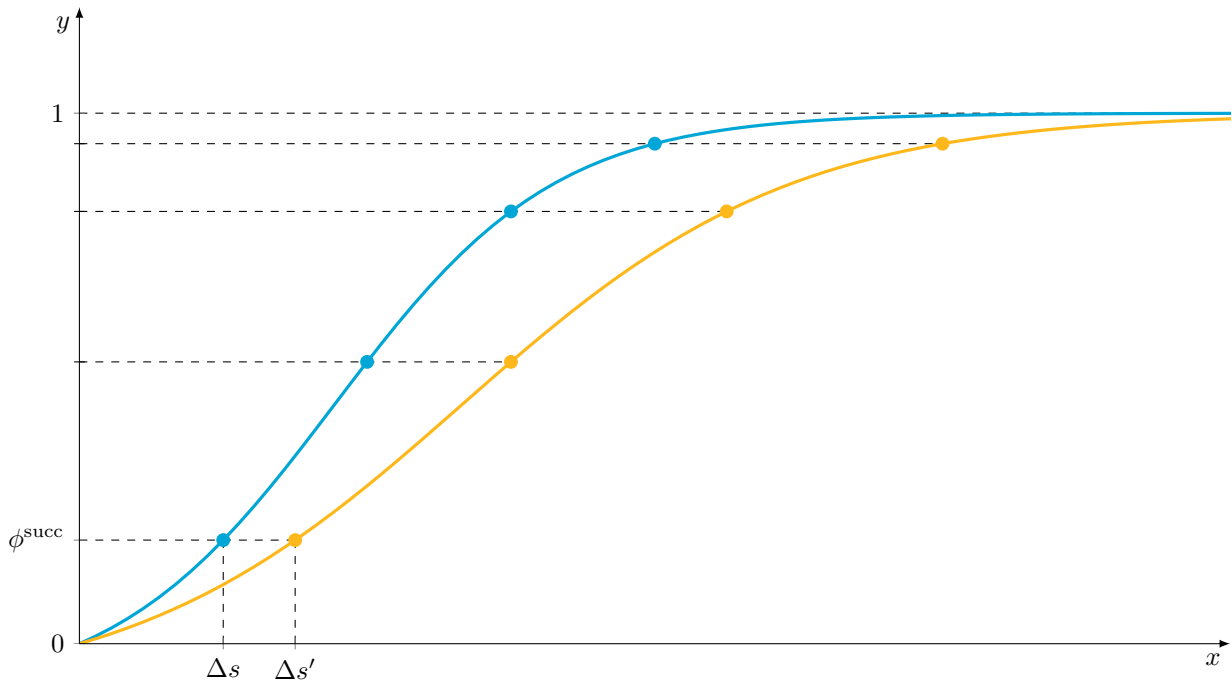


Figure A.6: Role of sampling interval.

Fixing $\Delta s = 1$ yields the simplified expressions in (A.25) and (A.26).

$$\kappa^{\text{succ}} = -\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}}+a}\right) - \ln(a) \quad (\text{A.25})$$

$$x_{1/2} = \frac{\ln(a)}{\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}}+a}\right) + \ln(a)}. \quad (\text{A.26})$$

The Personal Experience increment for the n^{th} consecutive HRI success is defined as the difference between consecutive samples:

$$\begin{aligned} \tau^{\text{exp,succ}}(n) &= y^{\text{succ}}(n\Delta s) - y^{\text{succ}}((n-1)\Delta s) \quad , \quad n \in \mathbb{Z} \\ &= y^{\text{succ}}(n) - y^{\text{succ}}(n-1) \\ &= -a + \frac{1+a}{1+e^{-\kappa^{\text{succ}}(n-x_{1/2})}} + a - \frac{1+a}{1+e^{-\kappa^{\text{succ}}(n-1-x_{1/2})}} \\ &= (1+a) \left(\frac{1}{1+e^{-\kappa^{\text{succ}}(n-x_{1/2})}} - \frac{1}{1+e^{-\kappa^{\text{succ}}(n-1-x_{1/2})}} \right). \end{aligned} \quad (\text{A.27})$$

Relating Successes and Failures

This subsection defines two analytical couplings between the success and failure update laws.

First, impose a proportional constraint between the first-event magnitudes ϕ^{fail} and ϕ^{succ} :

$$\phi^{\text{fail}} = \rho \phi^{\text{succ}} \quad , \quad \rho \in [0, 1]. \quad (\text{A.28})$$

Second, define χ as the ratio between the times required for the success and failure curves to reach a distance γ from their respective asymptotes:

$$\begin{aligned} y^{\text{succ}}(x_\gamma^{\text{succ}}) &= 1 - \gamma \Leftrightarrow -a + \frac{1+a}{1+e^{-\kappa^{\text{succ}}(x_\gamma^{\text{succ}}-x_{1/2})}} = 1 - \gamma \\ &\Leftrightarrow \frac{1+a}{1+e^{-\kappa^{\text{succ}} x_\gamma^{\text{succ}}} e^{\kappa^{\text{succ}} x_{1/2}}} = 1 - \gamma + a \\ &\Leftrightarrow 1 + e^{-\kappa^{\text{succ}} x_\gamma^{\text{succ}}} \left(\frac{1}{a}\right) = \frac{1+a}{1-\gamma+a} \\ &\Leftrightarrow \frac{e^{-\kappa^{\text{succ}} x_\gamma^{\text{succ}}}}{a} = 1 - \frac{1+a}{1-\gamma+a} \\ &\Leftrightarrow e^{-\kappa^{\text{succ}} x_\gamma^{\text{succ}}} = a \frac{\gamma}{1-\gamma+a} \\ &\Leftrightarrow x_\gamma^{\text{succ}} = -\frac{1}{\kappa^{\text{succ}}} \ln\left(a \frac{\gamma}{1-\gamma+a}\right). \end{aligned} \quad (\text{A.29})$$

$$y^{\text{fail}}(x_\gamma^{\text{fail}}) = \gamma \Leftrightarrow e^{-\lambda^{\text{fail}} x_\gamma^{\text{fail}}} = \gamma \quad (\text{A.30})$$

$$\Leftrightarrow x_\gamma^{\text{fail}} = -\frac{1}{\lambda^{\text{fail}}} \ln(\gamma).$$

$$\begin{aligned} \frac{x_\gamma^{\text{succ}}}{x_\gamma^{\text{fail}}} &= \frac{\ln\left(a \frac{\gamma}{1-\gamma+a}\right)}{\ln(\gamma)} \frac{\lambda^{\text{fail}}}{\kappa^{\text{succ}}} \\ &= \frac{\ln\left(a \frac{\gamma}{1-\gamma+a}\right)}{\ln(\gamma)} \frac{\ln(1-\phi^{\text{fail}})}{-\ln\left(\frac{1-\phi^{\text{succ}}}{\phi^{\text{succ}}+a}\right) - \ln(a)} \\ &= \chi. \end{aligned} \quad (\text{A.31})$$

Figure A.7 illustrates the two curves and the corresponding times x_γ^{succ} and x_γ^{fail} defined in (A.29) and (A.30).

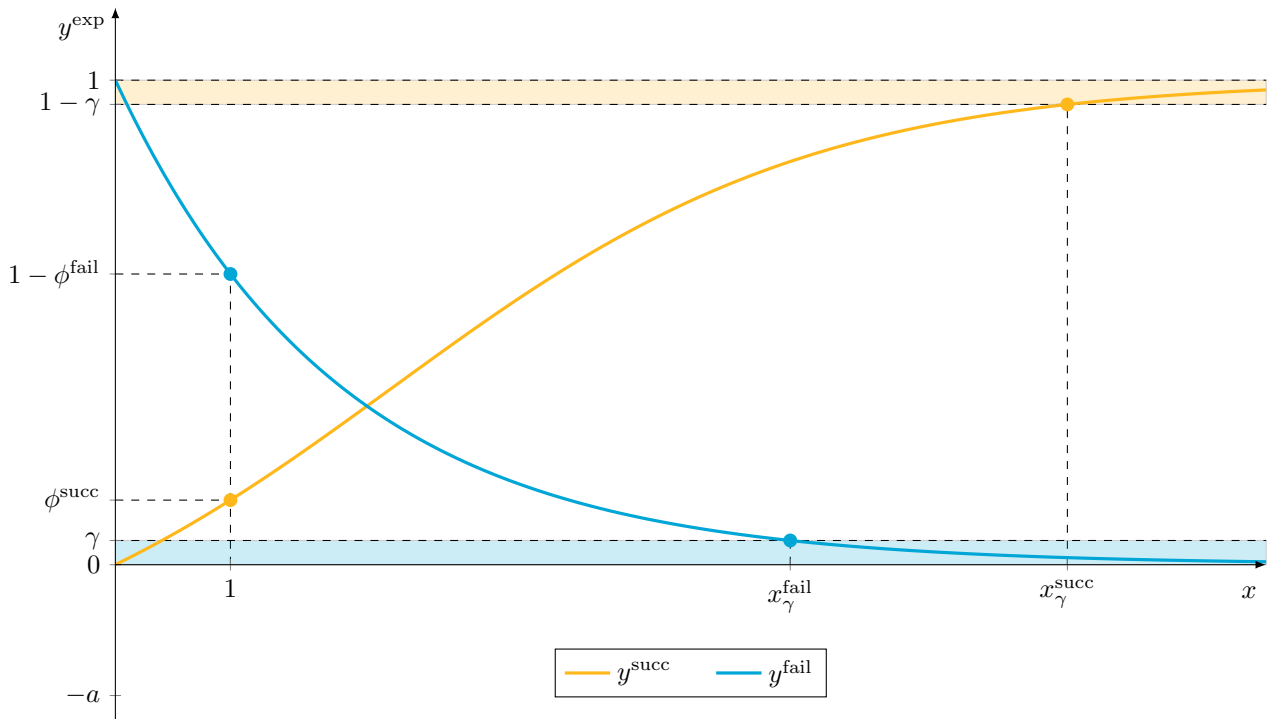


Figure A.7: Relationship between HRI successes and failures curves.

A.4. Latent Trust

This section derives the latent trust dynamics used when no HRI event occurs. Two cases are considered: (i) $\tau_0^{\text{lat}} > \tau^{\text{disp}}$ (decay toward τ^{disp}) and (ii) $\tau_0^{\text{lat}} < \tau^{\text{disp}}$ (growth toward τ^{disp}).

A.4.1. Trust Above Dispositional Trust (Decay)

When the initial trust is higher than the dispositional level ($\tau_0^{\text{lat}} > \tau^{\text{disp}}$), the trust level decays toward τ^{disp} , which acts as the asymptotic minimum.

Use the generic exponential decay law in (A.6) to model latent trust relaxation when $\tau_0^{\text{lat}} > \tau^{\text{disp}}$. In general, define a trajectory starting at τ_0^{lat} at $t = t_0$ and converging to an asymptote $\tau_{\min}^{\text{lat}}$:

$$\tau^{\text{lat}}(t) = \tau_{\min}^{\text{lat}} + (\tau_0^{\text{lat}} - \tau_{\min}^{\text{lat}}) e^{-\lambda_{\downarrow}^{\text{lat}}(t-t_0)}. \quad (\text{A.32})$$

For the specific use at hand, we want to define a function that starts at an initial value $\tau_0^{\text{lat}} > \tau^{\text{disp}}$ for $t = t_0$, and that tends to the asymptotical value of τ^{disp} . Using the previously introduced notation, this is given by (A.33).

$$\tau^{\text{lat}}(t) = \tau^{\text{disp}} + (\tau_0^{\text{lat}} - \tau^{\text{disp}}) e^{-\lambda_{\downarrow}^{\text{lat}}(t-t_0)}. \quad (\text{A.33})$$

The decay rate $\lambda_{\downarrow}^{\text{lat}}$ is defined as a function of $(\tau_0^{\text{lat}}, \tau^{\text{disp}})$ and a generic base parameter λ^{lat} as in (A.34).

$$\lambda_{\downarrow}^{\text{lat}} = f(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau}) \lambda^{\text{lat}}, \quad (\text{A.34})$$

where $f(\cdot)$ is a dimensionless scaling factor that depends on τ_0^{lat} and τ^{disp} (and fixed transformation parameters collected in θ^{τ}).

This is resolved through a two-step serial transformation of the unit exponential decay function as shown in (A.35).

$$\tau_{1 \rightarrow 0}^{\text{lat}}(t) = e^{-\lambda^{\text{lat}}(t-t_0)}. \quad (\text{A.35})$$

Transformation 1. Map the unit decay (A.35) to the codomain $]\tau^{\text{disp}}, 1]$ via (A.36), and choose λ' so that the transformed curve reaches $\tau^{\text{disp}} + \gamma$ at $t = t_{\tau^{\text{disp}}}$. Figure A.8 illustrates this first step, showing how the decay curve is constrained close to the asymptote τ^{disp} .

$$\tau^{\text{lat}'}(t) = \tau^{\text{disp}} + (1 - \tau^{\text{disp}}) e^{-\lambda'(t-t_0)}. \quad (\text{A.36})$$

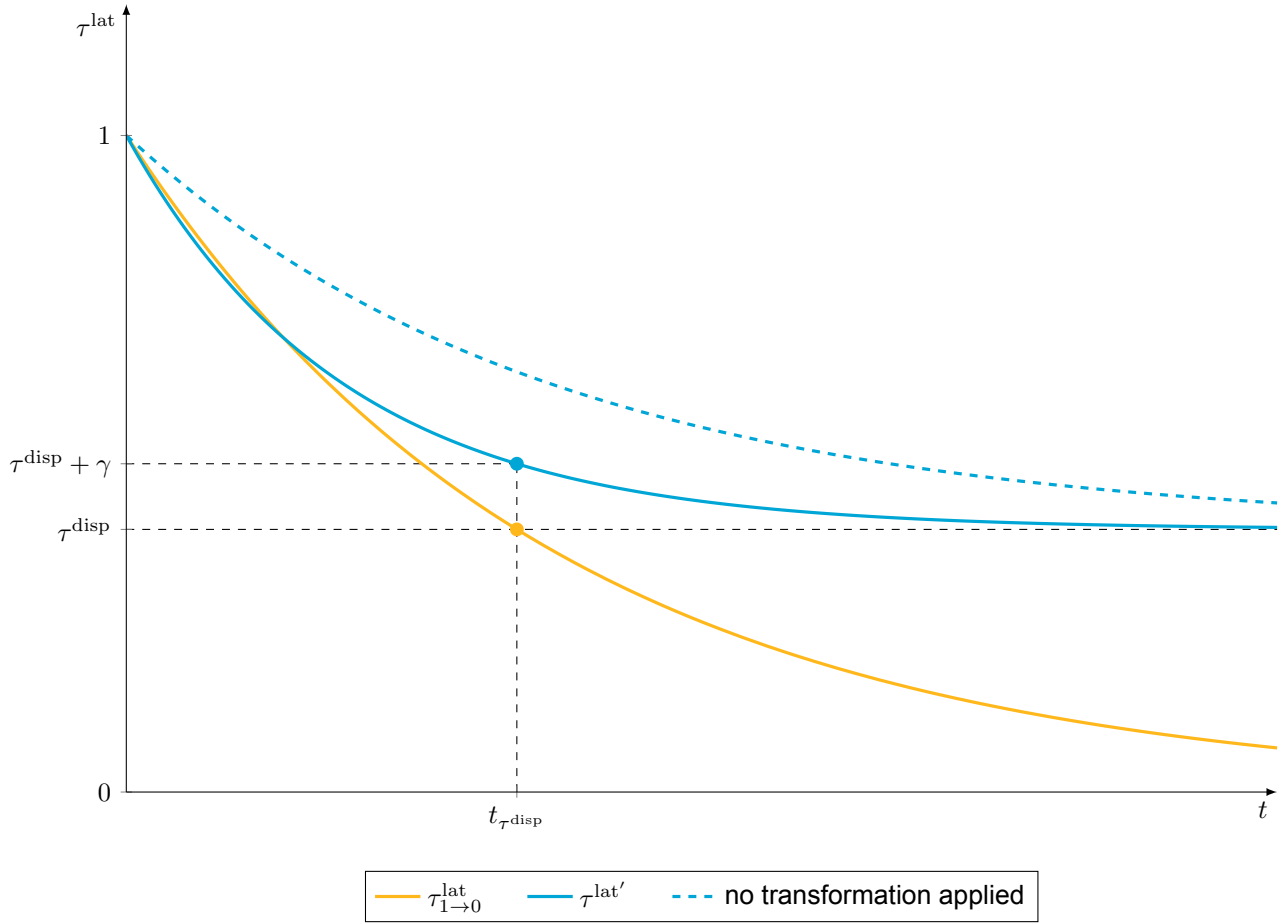


Figure A.8: First transformation of the failure curve for latent trust modeling.

Following what is shown in Figure A.8, we can use (A.35) to determine the time instant, $t_{\tau^{\text{disp}}}$, at which the unit exponential decay function crosses the dispositional trust value, τ^{disp} . This is shown by (A.37).

$$\begin{aligned} \tau_{1 \rightarrow 0}^{\text{lat}}(t_{\tau^{\text{disp}}}) = \tau^{\text{disp}} &\Leftrightarrow e^{-\lambda^{\text{lat}}(t_{\tau^{\text{disp}}} - t_0)} = \tau^{\text{disp}} \\ &\Leftrightarrow -\lambda^{\text{lat}}(t_{\tau^{\text{disp}}} - t_0) = \ln(\tau^{\text{disp}}) \\ &\Leftrightarrow t_{\tau^{\text{disp}}} = t_0 - \frac{\ln(\tau^{\text{disp}})}{\lambda^{\text{lat}}}. \end{aligned} \quad (\text{A.37})$$

Using (A.36) and $t_{\tau^{\text{disp}}}$, we can then relate the transformed function's decay rate, λ' , with the decay rate of the unit exponential decay function, λ^{lat} . This relation is derived in (A.38).

$$\begin{aligned}
\tau^{\text{lat}'}(t_{\tau^{\text{disp}}}) &= \tau^{\text{disp}} + \gamma \Leftrightarrow \cancel{\tau^{\text{disp}}} + (1 - \tau^{\text{disp}}) e^{-\lambda'(t_{\tau^{\text{disp}}} - t_0)} = \cancel{\tau^{\text{disp}}} + \gamma \\
&\Leftrightarrow -\lambda'(t_{\tau^{\text{disp}}} - t_0) = \ln\left(\frac{\gamma}{1 - \tau^{\text{disp}}}\right) \\
&\Leftrightarrow \lambda' = -\frac{\ln\left(\frac{\gamma}{1 - \tau^{\text{disp}}}\right)}{t_{\tau^{\text{disp}}} - t_0} \\
&\Rightarrow \lambda' = -\frac{\ln\left(\frac{\gamma}{1 - \tau^{\text{disp}}}\right)}{\underbrace{t_0 - \frac{\ln(\tau^{\text{disp}})}{-\lambda^{\text{lat}}}}_{t_{\tau^{\text{disp}}}}} \\
&\Leftrightarrow \lambda' = \frac{\ln\left(\frac{\gamma}{1 - \tau^{\text{disp}}}\right)}{\ln(\tau^{\text{disp}})} \lambda^{\text{lat}}
\end{aligned} \tag{A.38}$$

Transformation 2. Adjust the codomain from $]\tau^{\text{disp}}, 1]$ to $]\tau^{\text{disp}}, \tau_0^{\text{lat}}]$ by defining t_ε via (A.39), and selecting $\lambda_\downarrow^{\text{lat}}$ so that $\tau^{\text{lat}}(t_\varepsilon) = \tau^{\text{disp}} + \varepsilon$. Figure A.9 visually depicts this second transformation.

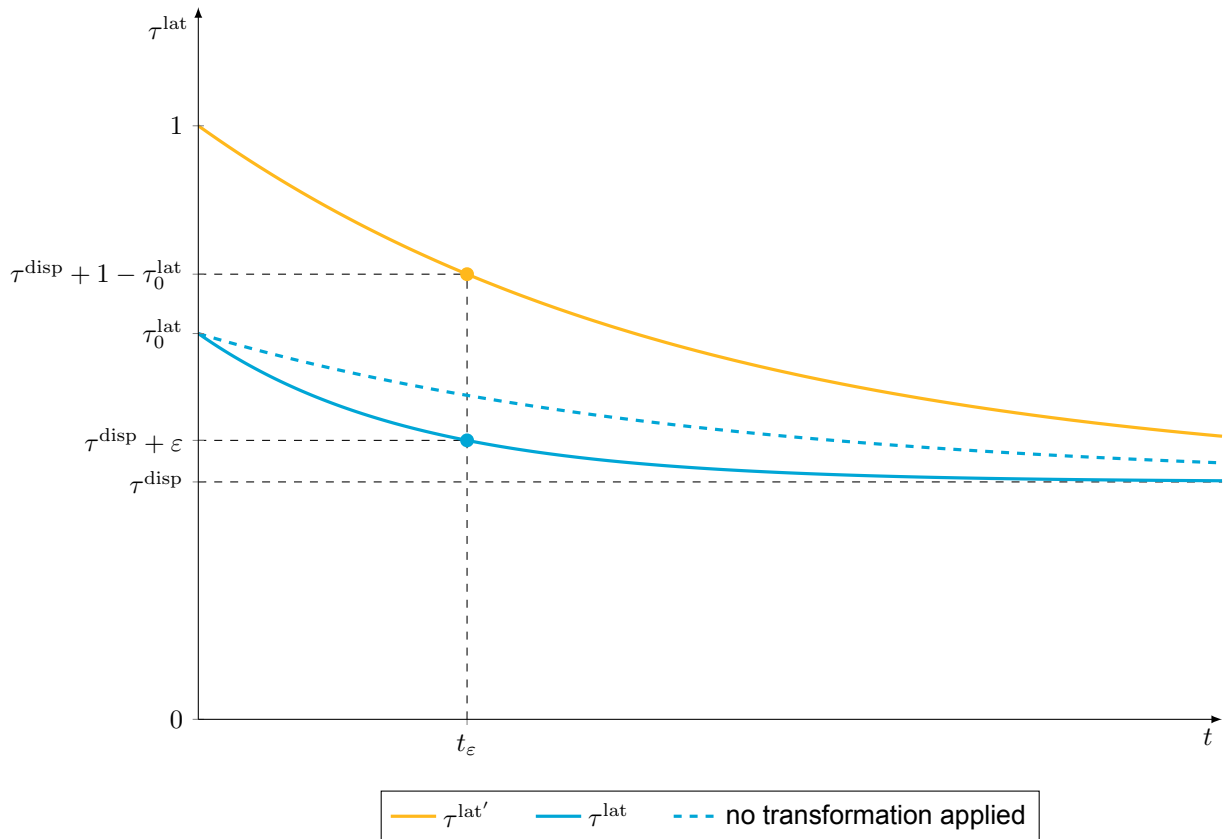


Figure A.9: Second transformation of the failure curve for latent trust modeling.

Following what is shown in Figure A.9, we can use (A.36) to determine the time instant, t_ε , at which the distance from the original curve to the asymptotic value (dispositional trust) is equal to the difference between the two curves' initial points (1 and τ_0^{lat}). This is shown by (A.39).

$$\begin{aligned}
 \tau^{\text{lat}'}(t_\varepsilon) &= \tau^{\text{disp}} + 1 - \tau_0^{\text{lat}} \Leftrightarrow \cancel{\tau^{\text{disp}}} + (1 - \tau^{\text{disp}}) e^{-\lambda'(t_\varepsilon - t_0)} = \cancel{\tau^{\text{disp}}} + 1 - \tau_0^{\text{lat}} \\
 &\Leftrightarrow e^{-\lambda'(t_\varepsilon - t_0)} = \frac{1 - \tau_0^{\text{lat}}}{1 - \tau^{\text{disp}}} \\
 &\Leftrightarrow t_\varepsilon = t_0 - \frac{\ln\left(\frac{1 - \tau_0^{\text{lat}}}{1 - \tau^{\text{disp}}}\right)}{\lambda'}.
 \end{aligned} \tag{A.39}$$

Using (A.33) and t_ε , we can then relate the transformed function's decay rate, $\lambda_\downarrow^{\text{lat}}$, with the decay rate of the function defined by (A.36), λ' . We can then get a direct relation between $\lambda_\downarrow^{\text{lat}}$ and the decay rate of the unit exponential decay function, λ^{lat} . This is derived in (A.40).

$$\begin{aligned}
 \tau^{\text{lat}}(t_\varepsilon) &= \tau^{\text{disp}} + \varepsilon \Leftrightarrow \cancel{\tau^{\text{disp}}} + (\tau_0^{\text{lat}} - \tau^{\text{disp}}) e^{-\lambda_\downarrow^{\text{lat}}(t_\varepsilon - t_0)} = \cancel{\tau^{\text{disp}}} + \varepsilon \\
 &\Leftrightarrow e^{-\lambda_\downarrow^{\text{lat}}(t_\varepsilon - t_0)} = \frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}} \\
 &\Leftrightarrow \lambda_\downarrow^{\text{lat}} = -\frac{\ln\left(\frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}}\right)}{t_\varepsilon - t_0} \\
 &\Rightarrow \lambda_\downarrow^{\text{lat}} = -\frac{\ln\left(\frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}}\right)}{\underbrace{t_0 - \frac{\ln\left(\frac{1 - \tau_0^{\text{lat}}}{1 - \tau^{\text{disp}}}\right)}{\lambda'} - t_0}_{t_\varepsilon}} \\
 &\Leftrightarrow \lambda_\downarrow^{\text{lat}} = \frac{\ln\left(\frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}}\right)}{\ln\left(\frac{1 - \tau_0^{\text{lat}}}{1 - \tau^{\text{disp}}}\right)} \lambda' \\
 &\Rightarrow \lambda_\downarrow^{\text{lat}} = \left(\frac{\ln\left(\frac{\gamma}{1 - \tau^{\text{disp}}}\right) \ln\left(\frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}}\right)}{\ln(\tau^{\text{disp}}) \ln\left(\frac{1 - \tau_0^{\text{lat}}}{1 - \tau^{\text{disp}}}\right)} \right) \lambda^{\text{lat}}, \quad \tau_0^{\text{lat}} > \tau^{\text{disp}} + \varepsilon.
 \end{aligned} \tag{A.40}$$

Figure A.10 shows the result of the full transformation, illustrating how the two serial transformations achieve a much more similar absolute decay in the desired codomain compared to using a single, relative decay rate.

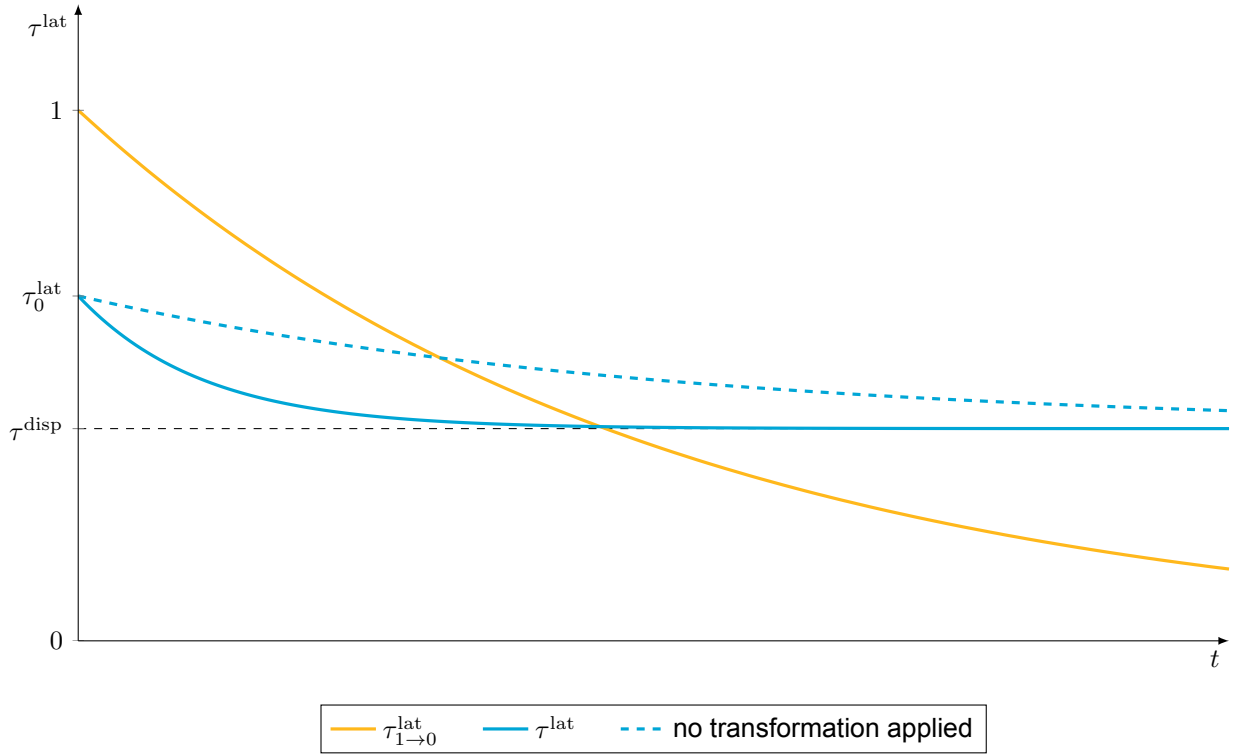


Figure A.10: Full transformation of the failure curve for latent trust modeling.

Thus, (A.40) provides the scaling in (A.34) via:

$$(A.41) \quad \begin{cases} \lambda_{\downarrow}^{\text{lat}} = f(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau}) \lambda^{\text{lat}} \\ f(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau}) = \frac{\ln\left(\frac{\gamma}{1-\tau^{\text{disp}}}\right) \ln\left(\frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}}\right)}{\ln(\tau^{\text{disp}}) \ln\left(\frac{1-\tau_0^{\text{lat}}}{1-\tau^{\text{disp}}}\right)} \end{cases} .$$

Discrete-time Formulation

The discrete-time formulation for updating the latent trust decay is derived in (A.42) and defined in (A.43). For readability purposes, the function $f(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau})$ is simply referred to as f .

$$\begin{aligned} \tau^{\text{lat}}[k+1] &= \tau^{\text{disp}} + (\tau_0^{\text{lat}} - \tau^{\text{disp}}) e^{-f\lambda^{\text{lat}}(k+\Delta t-t_0)} \\ &= \tau^{\text{disp}} + (\tau_0^{\text{lat}} - \tau^{\text{disp}}) e^{-f\lambda^{\text{lat}}(k-t_0)} e^{-f\lambda^{\text{lat}}\Delta t} \\ &= \tau^{\text{disp}} + \left(\tau^{\text{disp}} + (\tau_0^{\text{lat}} - \tau^{\text{disp}}) e^{-f\lambda^{\text{lat}}(k-t_0)} - \tau^{\text{disp}} \right) e^{-f\lambda^{\text{lat}}\Delta t} \\ &= \tau^{\text{disp}} + (\tau^{\text{lat}}[k] - \tau^{\text{disp}}) e^{-f\lambda^{\text{lat}}\Delta t} \end{aligned} \quad (A.42)$$

$$\begin{cases} \tau^{\text{lat}}[0] = \tau_0^{\text{lat}} \\ \tau^{\text{lat}}[k+1] = \tau^{\text{disp}} + (\tau^{\text{lat}}[k] - \tau^{\text{disp}}) e^{-f\lambda^{\text{lat}}\Delta t} \\ f = \frac{\ln\left(\frac{\gamma}{1-\tau^{\text{disp}}}\right) \ln\left(\frac{\varepsilon}{\tau_0^{\text{lat}} - \tau^{\text{disp}}}\right)}{\ln(\tau^{\text{disp}}) \ln\left(\frac{1-\tau_0^{\text{lat}}}{1-\tau^{\text{disp}}}\right)} \end{cases} \quad (A.43)$$

A.4.2. Trust Below Dispositional Trust (Growth)

When the initial trust is lower than the dispositional level ($\tau_0^{\text{lat}} < \tau^{\text{disp}}$), the trust level must increase toward τ^{disp} , which acts as the asymptotic maximum.

Use the logistic form in (A.17) to model latent trust relaxation when $\tau_0^{\text{lat}} < \tau^{\text{disp}}$. Define a trajectory that starts at τ_0^{lat} at $t = t_0$ and converges to an asymptote $\tau_{\text{max}}^{\text{lat}}$, with codomain $]\tau_{\text{min}}^{\text{lat}}, \tau_{\text{max}}^{\text{lat}}[$ and $\tau_{\text{min}}^{\text{lat}} < \tau_0^{\text{lat}}$:

$$\tau^{\text{lat}}(t) = \tau_{\text{min}}^{\text{lat}} + \frac{\tau_{\text{max}}^{\text{lat}} - \tau_{\text{min}}^{\text{lat}}}{1 + e^{-\kappa_{\uparrow}^{\text{lat}}(t - t_{1/2} - t_0)}}. \quad (\text{A.44})$$

We can redefine the lower asymptote value as $\tau_{\text{min}}^{\text{lat}} = \tau_0^{\text{lat}} - \tau^{\text{offset}}$. This modifies (A.44) to (A.45).

$$\tau^{\text{lat}}(t) = \tau_0^{\text{lat}} - \tau^{\text{offset}} + \frac{\tau_{\text{max}}^{\text{lat}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}{1 + e^{-\kappa_{\uparrow}^{\text{lat}}(t - t_{1/2} - t_0)}}. \quad (\text{A.45})$$

For the specific use at hand, we want to define a function that starts at an initial value $\tau_0^{\text{lat}} < \tau^{\text{disp}}$ for $t = t_0$, and that tends to the asymptotical value of τ^{disp} . Using the previously introduced notation, this is given by (A.46).

$$\tau^{\text{lat}}(t) = \tau_0^{\text{lat}} - \tau^{\text{offset}} + \frac{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}{1 + e^{-\kappa_{\uparrow}^{\text{lat}}(t - t_{1/2} - t_0)}}. \quad (\text{A.46})$$

Define the episode-specific growth rate $\kappa_{\uparrow}^{\text{lat}}$ via a scaled base parameter κ^{lat} as:

$$\kappa_{\uparrow}^{\text{lat}} = g(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau}) \kappa^{\text{lat}}, \quad (\text{A.47})$$

where $g(\cdot)$ is a dimensionless scaling factor that depends on τ_0^{lat} and τ^{disp} (and fixed transformation parameters collected in θ^{τ}).

This is solved through a two-step serial transformation of the original logistic function, written using the new notation and shown in (A.48).

$$\tau_{0 \rightarrow 1}^{\text{lat}}(t) = -\tau^{\text{offset}} + \frac{1 + \tau^{\text{offset}}}{1 + e^{-\kappa^{\text{lat}}(t - t_{1/2} - t_0)}}. \quad (\text{A.48})$$

Transformation 1. Map the baseline logistic (A.48) to the codomain $[0, \tau^{\text{disp}}[$ via (A.49), and determine $t_{\tau^{\text{disp}}}$ and κ' via (A.50) and (A.51).

$$\tau^{\text{lat}'}(t) = -\tau^{\text{offset}} + \frac{\tau^{\text{disp}} + \tau^{\text{offset}}}{1 + e^{-\kappa'(t - t_{1/2} - t_0)}}. \quad (\text{A.49})$$

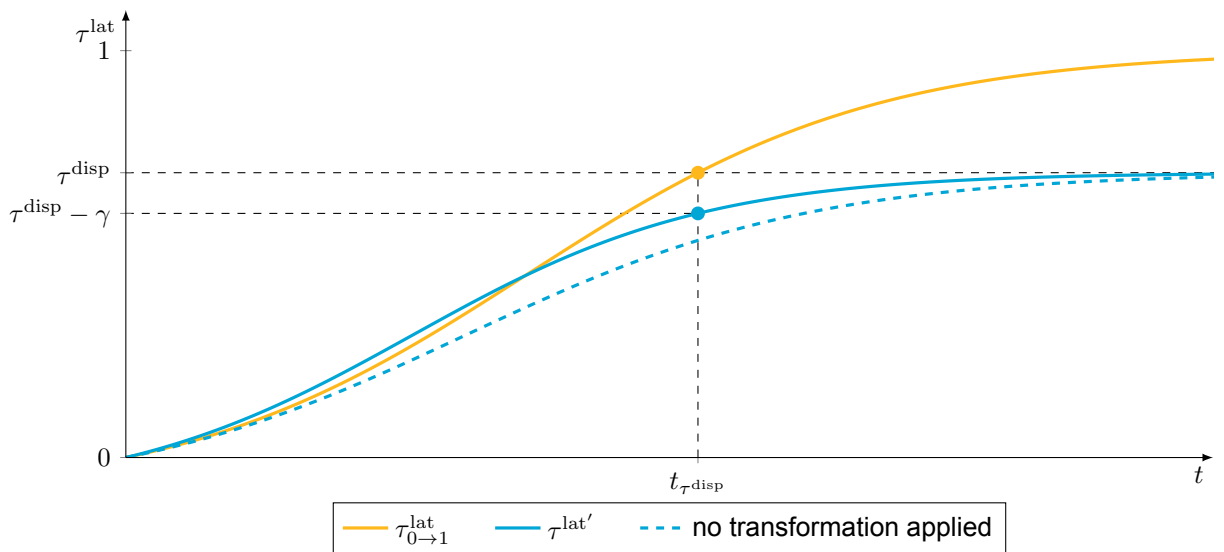


Figure A.11: First transformation of the success curve for latent trust modeling.

Compute the time instant $t_{\tau^{\text{disp}}}$ at which (A.48) reaches τ^{disp} :

$$\begin{aligned}
 \tau_{0 \rightarrow 1}^{\text{lat}}(t_{\tau^{\text{disp}}}) = \tau^{\text{disp}} &\Leftrightarrow -\tau^{\text{offset}} + \frac{1 + \tau^{\text{offset}}}{1 + e^{-\kappa^{\text{lat}}(t_{\tau^{\text{disp}}} - t_{1/2} - t_0)}} = \tau^{\text{disp}} \\
 &\Leftrightarrow 1 + e^{-\kappa^{\text{lat}}(t_{\tau^{\text{disp}}} - t_{1/2} - t_0)} = \frac{1 + \tau^{\text{offset}}}{\tau^{\text{disp}} + \tau^{\text{offset}}} \\
 &\Leftrightarrow -\kappa^{\text{lat}}(t_{\tau^{\text{disp}}} - t_{1/2} - t_0) = \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right) \\
 &\Leftrightarrow t_{\tau^{\text{disp}}} = t_0 + t_{1/2} - \frac{\ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right)}{\kappa^{\text{lat}}} \quad \left(\kappa^{\text{lat}} t_{1/2} = \ln\left(\frac{1}{\tau^{\text{offset}}}\right)\right) \\
 &\Rightarrow t_{\tau^{\text{disp}}} = t_0 - \frac{1}{\kappa^{\text{lat}}} \left[\ln(\tau^{\text{offset}}) + \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right) \right].
 \end{aligned} \tag{A.50}$$

Using (A.49) and τ^{disp} , we can then relate the transformed function's logistic growth rate, κ' , with the logistic growth rate of the original logistic function, κ^{lat} . This relation is derived in (A.51).

$$\begin{aligned}
 \tau^{\text{lat}'}(t_{\tau^{\text{disp}}}) = \tau^{\text{disp}} - \gamma &\Leftrightarrow -\tau^{\text{offset}} + \frac{\tau^{\text{disp}} + \tau^{\text{offset}}}{1 + e^{-\kappa'(t_{\tau^{\text{disp}}} - t_{1/2} - t_0)}} = \tau^{\text{disp}} - \gamma \\
 &\Leftrightarrow 1 + e^{-\kappa'(t_{\tau^{\text{disp}}} - t_{1/2} - t_0)} = \frac{\tau^{\text{disp}} + \tau^{\text{offset}}}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma} \\
 &\Leftrightarrow -\kappa'(t_{\tau^{\text{disp}}} - t_{1/2} - t_0) = \ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right) \\
 &\Leftrightarrow \kappa' = -\frac{\ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right)}{t_{\tau^{\text{disp}}} - t_{1/2} - t_0} \quad \left(\kappa' t_{1/2} = \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right)\right) \\
 &\Rightarrow \kappa' = -\frac{\ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right)}{\cancel{t_0} - \frac{1}{\kappa^{\text{lat}}} \left[\ln(\tau^{\text{offset}}) + \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right) \right] - \frac{1}{\kappa'} \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \cancel{t_0}} \\
 &\Leftrightarrow \frac{\kappa'}{\kappa^{\text{lat}}} \left[\ln(\tau^{\text{offset}}) + \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right) \right] + \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) = \ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right) \\
 &\Leftrightarrow \kappa' = \kappa^{\text{lat}} \frac{\ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right) - \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right)}{\ln(\tau^{\text{offset}}) + \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right)}
 \end{aligned} \tag{A.51}$$

Transformation 2. Adjust the codomain from $[0, \tau^{\text{disp}}[$ to $[\tau_0^{\text{lat}}, \tau^{\text{disp}}[$ by defining t_ε via (A.52) and selecting $\kappa_\uparrow^{\text{lat}}$ via (A.53).

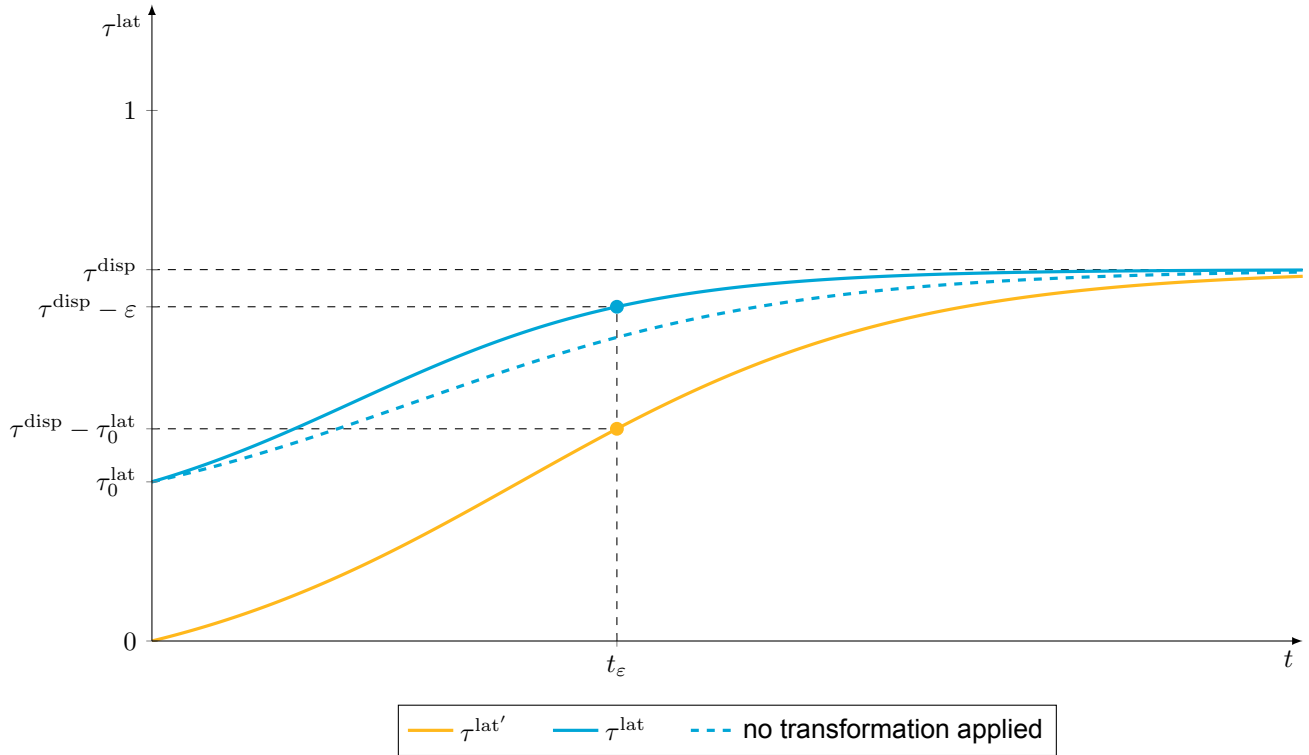


Figure A.12: Second transformation of the success curve for latent trust modeling.

Compute the time instant t_ε such that (A.49) reaches $\tau^{\text{disp}} - \tau_0^{\text{lat}}$:

$$\begin{aligned}
 \tau^{\text{lat}'}(t_\varepsilon) = \tau^{\text{disp}} - \tau_0^{\text{lat}} &\Leftrightarrow -\tau^{\text{offset}} + \frac{\tau^{\text{disp}} + \tau^{\text{offset}}}{1 + e^{-\kappa'(t_\varepsilon - t_{1/2} - t_0)}} = \tau^{\text{disp}} - \tau_0^{\text{lat}} \\
 &\Leftrightarrow 1 + e^{-\kappa'(t_\varepsilon - t_{1/2} - t_0)} = \frac{\tau^{\text{disp}} + \tau^{\text{offset}}}{\tau^{\text{disp}} + \tau^{\text{offset}} - \tau_0^{\text{lat}}} \\
 &\Leftrightarrow -\kappa'(t_\varepsilon - t_{1/2} - t_0) = \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} + \tau^{\text{offset}} - \tau_0^{\text{lat}}}\right) \tag{A.52} \\
 &\Leftrightarrow t_\varepsilon = t_0 + t_{1/2} - \frac{\ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} + \tau^{\text{offset}} - \tau_0^{\text{lat}}}\right)}{\kappa'} \quad \left(\kappa' t_{1/2} = \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right)\right) \\
 &\Rightarrow t_\varepsilon = t_0 + \frac{1}{\kappa'} \left[\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} + \tau^{\text{offset}} - \tau_0^{\text{lat}}}\right) \right].
 \end{aligned}$$

Using (A.46) and t_ε , we can then relate the transformed function's logistic growth rate, $\kappa_\uparrow^{\text{lat}}$, with the logistic growth rate of the function defined by (A.49), κ' . We can then get a direct relation between $\kappa_\uparrow^{\text{lat}}$ and the logistic growth rate of the original logistic function, κ^{lat} . This is derived in (A.53).

$$\begin{aligned}
\tau^{\text{lat}}(t_\varepsilon) &= \tau^{\text{disp}} - \varepsilon \\
\Leftrightarrow \tau_0^{\text{lat}} - \tau^{\text{offset}} + \frac{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}{1 + e^{-\kappa_\uparrow^{\text{lat}}(t_\varepsilon - t_{1/2} - t_0)}} &= \tau^{\text{disp}} - \varepsilon \\
\Leftrightarrow 1 + e^{-\kappa_\uparrow^{\text{lat}}(t_\varepsilon - t_{1/2} - t_0)} &= \frac{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}} \\
\Leftrightarrow -\kappa_\uparrow^{\text{lat}}(t_\varepsilon - t_{1/2} - t_0) &= \ln\left(\frac{\varepsilon}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right) \quad \left(\kappa_\uparrow^{\text{lat}} t_{1/2} = \ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right)\right) \\
\Rightarrow \kappa_\uparrow^{\text{lat}} \left(\underbrace{\cancel{t_0} + \frac{1}{\kappa'} \left[\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right) \right]}_{t_\varepsilon} - \underbrace{\frac{1}{\kappa_\uparrow^{\text{lat}}} \ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right)}_{t_{1/2}} - \cancel{t_0} \right) &= \dots \\
&\dots = -\ln\left(\frac{\varepsilon}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right) \\
\Leftrightarrow \frac{\kappa_\uparrow^{\text{lat}}}{\kappa'} \left[\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right) \right] &= \ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\varepsilon}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right) \\
\Leftrightarrow \kappa_\uparrow^{\text{lat}} &= \frac{\ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\varepsilon}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)}{\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)} \kappa' \\
\Rightarrow \kappa_\uparrow^{\text{lat}} &= \frac{\ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right) - \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right)}{\ln(\tau^{\text{offset}}) + \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right)} \frac{\ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\varepsilon}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)}{\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)} \kappa^{\text{lat}}, \dots \\
&\dots \tau_0^{\text{lat}} - \tau^{\text{offset}} < \tau^{\text{disp}} - \varepsilon.
\end{aligned} \tag{A.53}$$

Figure A.13 shows how these two serial transformations of the logistic function result in a much more similar absolute growth in the desired codomain, compared to simply using the same logistic growth rate as in the original logistic function.

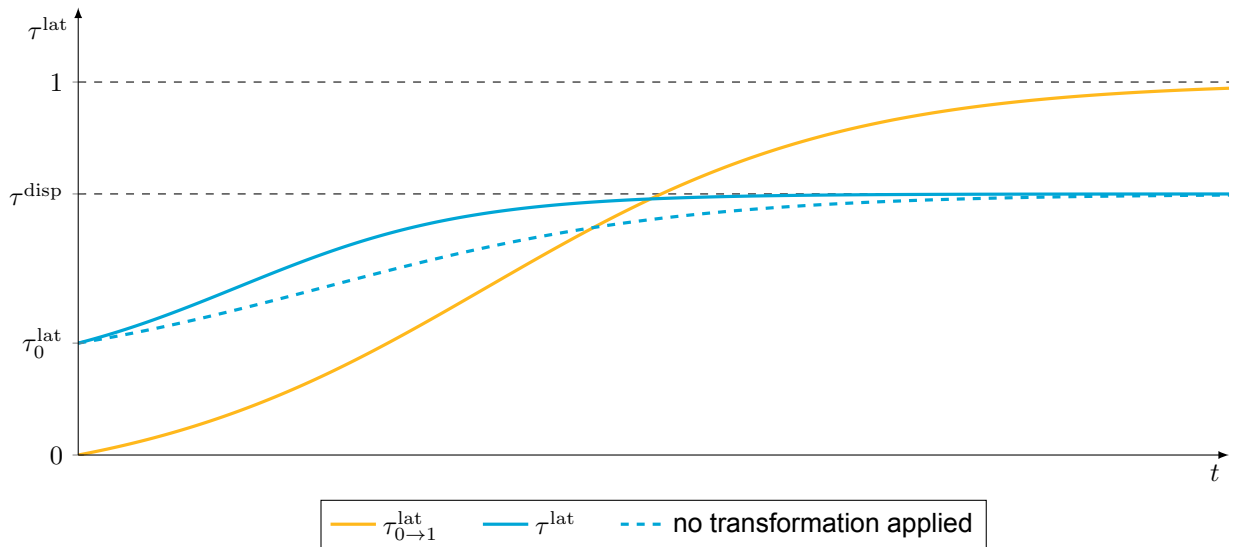


Figure A.13: Full transformation of the success curve for latent trust modeling.

Thus, (A.53) provides the scaling in (A.47) via:

$$\begin{cases} \kappa_{\uparrow}^{\text{lat}} = g(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau}) \kappa^{\text{lat}} \\ g(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau}) = \frac{\ln\left(\frac{\gamma}{\tau^{\text{disp}} + \tau^{\text{offset}} - \gamma}\right) - \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right)}{\ln(\tau^{\text{offset}}) + \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right)} \frac{\ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\varepsilon}{\tau^{\text{disp}} - \varepsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)}{\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)} \end{cases} \quad (\text{A.54})$$

Discrete-time Formulation

This subsection derives a discrete-time update for latent trust growth when $\tau_0^{\text{lat}} < \tau^{\text{disp}}$, resulting in (A.58). For brevity, denote $g(\tau_0^{\text{lat}}, \tau^{\text{disp}}, \theta^{\tau})$ by g .

$$\begin{aligned} \tau^{\text{lat}}(t) &= \tau_0^{\text{lat}} - \tau^{\text{offset}} + \frac{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}{1 + e^{-g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \\ &= \tau_0^{\text{lat}} - \tau^{\text{offset}} + (\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}) \left(\frac{1}{1 + e^{-g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \right) \\ &= \tau_0^{\text{lat}} - \tau^{\text{offset}} + (\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}) \sigma(t) . \end{aligned} \quad (\text{A.55})$$

$$\begin{aligned} \tau^{\text{lat}}(t_0) &= \tau_0^{\text{lat}} \\ \Leftrightarrow \tau_0^{\text{lat}} - \tau^{\text{offset}} + (\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}) \sigma(t_0) &= \tau_0^{\text{lat}} \\ \Leftrightarrow \sigma(t_0) &= \frac{\tau^{\text{offset}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}} . \end{aligned} \quad (\text{A.56})$$

$$\begin{aligned} \frac{d\sigma}{dt} &= \frac{d}{dt} \left(\frac{1}{1 + e^{-g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \right) \\ &= \frac{d}{dt} \left(\frac{e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}}{1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \right) \\ &= \frac{(1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}) \frac{d}{dt} (e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}) - e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)} \frac{d}{dt} (1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)})}{(1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)})^2} \\ &= \frac{g\kappa^{\text{lat}} (1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}) e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)} - g\kappa^{\text{lat}} e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)} e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}}{(1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)})^2} \\ &= g\kappa^{\text{lat}} \frac{e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}}{(1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)})^2} \\ &= g\kappa^{\text{lat}} \left(\frac{e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}}{1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \right) \left(\frac{1}{1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \right) \\ &= g\kappa^{\text{lat}} \left(\frac{e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}}{1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \right) \left(1 - \frac{e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}}{1 + e^{g\kappa^{\text{lat}}(t-t_{1/2}-t_0)}} \right) \\ &= g\kappa^{\text{lat}} (\sigma(t) (1 - \sigma(t))) \end{aligned} \quad (\text{A.57})$$

A forward-Euler discretization of (5.23) yields:

$$\left\{ \begin{array}{l} \sigma[0] = \frac{\tau^{\text{offset}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}} \\ \sigma[k+1] = \sigma[k] (1 + g \kappa^{\text{lat}} (1 - \sigma[k]) \Delta t) \\ \tau^{\text{lat}}[k+1] = \tau_0^{\text{lat}} - \tau^{\text{offset}} + (\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}) \sigma[k+1] \\ g = \frac{\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}} - \tau}\right) - \ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right)}{\ln(\tau^{\text{offset}}) + \ln\left(\frac{1 - \tau^{\text{disp}}}{\tau^{\text{disp}} + \tau^{\text{offset}}}\right)} \frac{\ln\left(\frac{\tau^{\text{disp}} - \tau_0^{\text{lat}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\epsilon}{\tau^{\text{disp}} - \epsilon - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)}{\ln\left(\frac{\tau^{\text{disp}}}{\tau^{\text{offset}}}\right) - \ln\left(\frac{\tau_0^{\text{lat}}}{\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}}\right)} \end{array} \right. \quad (\text{A.58})$$

A.5. Integrated Trust Dynamics Model

This section collects the final discrete-time update rules for the integrated trust dynamics model, using the component definitions derived in previous sections. The overall trust is reconstructed at each time step via:

$$\tau[k] = \tau^{\text{lat}}[k] + \tau^{\text{rep}}[k] + \alpha^{\text{sit}} \tau^{\text{sit}}(r), \quad (\text{A.59})$$

The total trust τ is therefore not itself the quantity that obeys dynamical rules; rather, each subcomponent evolves according to its own mechanism, and the final score is obtained through (A.59). Crucially, the latent dynamics act *only* on τ^{lat} , and not on reputation or situational trust, preventing unintended cross-effects between trust modules. The model achieves dynamic prediction by assessing whether a new HRI event has occurred during the time step, and applying the corresponding set of update rules.

A.5.1. Rule 1: HRI Event Occurs

If an HRI event occurs at time step k , first update τ^{lat} and τ^{rep} over Δt using their respective dynamics, and evaluate $\tau^{\text{sit}}(r)$ at the observed risk r .

Then apply the Personal Experience impulse to the latent component:

$$\tau^{\text{lat}}[k] \leftarrow \tau^{\text{lat}}[k] + \tau^{\text{exp}}(n). \quad (\text{A.60})$$

Reconstruct the total trust via (A.59):

$$\tau[k] = \tau^{\text{lat}}[k] + \tau^{\text{rep}}[k] + \alpha^{\text{sit}} \tau^{\text{sit}}(r). \quad (\text{A.61})$$

Start a new latent episode by setting:

$$\tau_0^{\text{lat}} = \tau^{\text{lat}}[k], \quad (\text{A.62})$$

and compute the episode-specific rate $\lambda_{\downarrow}^{\text{lat}}$ or $\kappa_{\uparrow}^{\text{lat}}$ using Sections A.4.1 and A.4.2.

A.5.2. Rule 2: No HRI Event Occurs (Latent Trust Episode Evolution)

If no HRI event occurs during $[k, k+1]$, update τ^{lat} using the appropriate latent episode dynamics.

Case A: Trust Above Dispositional Trust. The latent component relaxes exponentially toward τ^{disp} using the episode-specific rate $\lambda_{\downarrow}^{\text{lat}}$:

$$\tau^{\text{lat}}[k+1] = \tau^{\text{disp}} + (\tau^{\text{lat}}[k] - \tau^{\text{disp}}) e^{-\lambda_{\downarrow}^{\text{lat}} \Delta t}, \quad (\text{A.63})$$

which corresponds to (A.43).

Case B: Trust Below Dispositional Trust. The latent trust grows toward τ^{disp} according to the discrete logistic formulation

$$\sigma[k+1] = \sigma[k] (1 + \kappa_{\uparrow}^{\text{lat}} (1 - \sigma[k]) \Delta t), \quad (\text{A.64})$$

$$\tau^{\text{lat}}[k+1] = \tau_0^{\text{lat}} - \tau^{\text{offset}} + (\tau^{\text{disp}} - \tau_0^{\text{lat}} + \tau^{\text{offset}}) \sigma[k+1], \quad (\text{A.65})$$

which corresponds to (A.58).

Case C: Within the Deadband. If $|\tau_0^{\text{lat}} - \tau^{\text{disp}}| \leq \varepsilon$, hold τ^{lat} constant.

Update τ^{rep} using (A.5) and evaluate $\tau^{\text{sit}}(r)$ at the most recent risk observation.

Reconstruct the total trust via:

$$\tau[k+1] = \tau^{\text{lat}}[k+1] + \tau^{\text{rep}}[k+1] + \alpha^{\text{sit}} \tau^{\text{sit}}(r). \quad (\text{A.66})$$

A.5.3. Final Check: Ensuring Bounded Total Trust

Because τ is formed by additive reconstruction (see (A.59)), the combined value may exceed the interval $[0, 1]$. Apply saturation:

To ensure conceptual and mathematical validity, the final trust score is clipped to the closed interval $[0, 1]$:

$$\tau[k+1] = \begin{cases} 0, & \text{if } \tau^{\text{combined}} < 0, \\ \tau^{\text{combined}}, & \text{if } 0 \leq \tau^{\text{combined}} \leq 1, \\ 1, & \text{if } \tau^{\text{combined}} > 1, \end{cases} \quad (\text{A.67})$$

where τ^{combined} is the unconstrained value produced by (A.61) or (A.66).

B

Online Experiment Implementation and Data Acquisition

B.1. Purpose of the Experimental Software

The empirical methodology of this thesis relied on an online, browser-based experiment designed to elicit trust measurements and reliance decisions in a simulated search-and-rescue interaction. To operationalize the experimental protocol described in the thesis, a dedicated software platform was developed to implement the interactive task, administer questionnaires and probes, assign experimental conditions, and collect structured event logs for subsequent modeling and analysis.

The experiment software constitutes the implementation layer that transforms the experimental design into an executable procedure. It encodes the full participant timeline (instructions, reputation exposure, training, task blocks, trust probes, questionnaires, and demographics) and produces the raw behavioral and self-report data used throughout the thesis.

B.2. Scope of the Implemented Experiment Platform

A complete web application was developed to support the online study workflow end-to-end. The platform includes:

- an interactive task engine implementing the “door trials” and their visual/temporal dynamics,
- a deterministic condition-assignment mechanism based on predefined block/trial sets,
- administration of trust questionnaires (pre/post) and mid-task trust measurements (probes and mid-block questionnaires),
- a durable client-side logging system for resilience to network failures and premature tab closure,
- structured data export via a server-side receiver to a tabular storage format suitable for analysis,
- explicit support for participant withdrawal, including best-effort data deletion requests.

The software was structured with modularity and reproducibility as primary objectives, enabling clear separation between configuration, trial logic, condition sets, and logging.

B.3. Mapping Between Experimental Procedure and Software Modules

The experiment implements the full participant flow described in the thesis, using a timeline-based structure. The core interactive task is encoded as a repeatable trial module, composed into blocks with interleaved trust probes and mid-task questionnaires.

Experimental Stage	Software Implementation
Reputation exposure	Review-card presentation module + reputation probe
Pre-task trust measurement	40-item pre-task questionnaire module
Demonstration/training	Non-interactive demo configuration of the door-trial module
Main task (3 blocks)	Door-trial module composed into a block-based timeline
Interleaved trust probes	Probe trial module inserted at defined indices or cadence
Mid-task trust questionnaires	14-item questionnaires inserted between blocks
Post-task measurement and demographics	Post-task questionnaires and demographics module(s)

Table B.1: Mapping between the experimental procedure and implemented software components.

B.4. Participant Timeline and Door-Trial Structure

The online study is implemented as a timeline of stages executed in the browser. The main task phase consists of three blocks of interactive door trials, with two mid-task trust questionnaires (14-item) placed between blocks and repeated trust probes interleaved within blocks.

Door trial phases

A single door trial is composed of the following phases:

1. Scene render (environment, doors, drone, overlays and user interface elements)
2. Drone entry animation
3. Drone scan sweep (scan cone visualization)
4. Drone hop to the recommended door
5. Decision stage (participant chooses to follow or ignore the recommendation)
6. Door reveal and outcome feedback (success/failure)
7. Transition (blackout fade to the next trial)

A time-limited “environmental integrity” countdown spans the active phases of the trial and provides urgency cues. Depending on configuration, the timeout behavior can trigger an automatic follow action.

Risk conditions

Trials may apply risk-specific visual and timing overrides. These overrides influence environmental visuals (e.g., smoke/fire encroachment), animation pacing, and decision-time scaling, allowing risk to be manipulated consistently while keeping the global configuration stable.

B.5. Condition Sets and Reproducible Assignment

The experiment uses predefined condition sets, each defining three blocks of trials. Each trial specifies at least:

- the true victim location (left/right),
- the drone recommendation (left/right),
- a categorical risk level (e.g., low/medium/high/extreme).

Participant assignment to condition sets is resolved deterministically (with optional overrides), ensuring reproducibility and supporting balanced distribution across experimental conditions. Condition selection and set loading are separated from the task logic, enabling reuse of the platform with alternative trial sets.

B.6. Data Logging, Durability, and Withdrawal Support

A resilient logging pipeline was implemented to support browser-based data collection under realistic conditions (intermittent connectivity, unexpected tab closure, device variability). Logged data are stored in an event-centric tall-table format, with one row per retained event (door trials, trust probes, questionnaires, reputation exposure items, and demographics). Key design features include:

- durable local buffering using IndexedDB (with a localStorage fallback),
- batch uploading with retry and exponential backoff,
- a last-chance delivery attempt on navigation using `sendBeacon()` where supported,
- schema robustness through preservation of unknown fields in a dedicated `extra_json` column.

The platform also supports an explicit withdrawal flow. Participants can choose to discard their data, triggering local queue clearing and a best-effort server-side deletion request by participant identifier.

B.7. Opening Statement and Participant Consent

The experiment began with an information and consent screen presented to all participants prior to any data collection. This opening statement described the purpose of the study, expected duration, data handling procedures, voluntary participation, and withdrawal options. Participants could only proceed after explicitly providing consent.

For transparency and reproducibility, the text shown to participants is reproduced below.

Participant Information & Consent

You are being invited to participate in a research study titled **Trust in Human–Robot Interaction in Search-and-Rescue Scenarios**. This study is being conducted by Pedro Rodrigues Correia da Silva from the Faculty of Aerospace Engineering at Delft University of Technology (TU Delft).

The purpose of this research study is to collect data regarding how people make decisions and form trust in emergency situations where a robotic agent provides suggestions. The study will take approximately **25–30 minutes** to complete. The data will be used for research in a Master Thesis at the Faculty of Aerospace Engineering of TU Delft.

During the study, you will be asked to complete a short questionnaire and then perform a computer-based task in which you decide between different options (e.g., choosing a door), with guidance from a simulated drone.

As with any online activity, there is a minimal risk of a data breach. To the best of our ability, your responses in this study will remain **completely anonymous**. No names, email addresses, or IP addresses are collected. The only data recorded are your responses and timing information within the experiment. Data will be stored securely and used exclusively for academic research purposes.

Your participation in this study is entirely voluntary, and you can withdraw at any time. You are free to omit any questions. If you wish to stop the experiment early, you may click the **Exit** button (shown at the top-right corner of the screen during the task) or press `Ctrl + Shift + X` on your keyboard. At that point, you will be given the choice to either keep the data collected so far or to discard it entirely. Once the experiment is completed, the data is anonymous and cannot be linked back to you or removed.

If you have any questions about this research, please contact Pedro Silva. The responsible supervisor for this project is Anahita Jamshidnejad.

Please note that this experiment works best on a **desktop or laptop computer**. While it can run on mobile devices, the interface may be harder to use, and results may be compromised. On desktop, the experiment will automatically switch to **full screen mode**, the recommended mode for completing the study. You can exit full screen at any time by pressing the `Esc` key.

By clicking on the button below, you are agreeing to this Opening Statement and giving your consent to participate in this study.

B.8. Review Stimulus Pool

The experiment manipulates perceived robot reputation through exposure to predefined textual reviews. Each review belongs to a categorical reputation condition and is associated with a numerical reputation expectation used in the modeling pipeline. For transparency and reproducibility, the full review pool used in the online experiment is provided below.

Table B.2: Very positive reviews.

ID	Review text
vp1	This drone consistently found victims ahead of us.
vp2	Clear, timely guidance — felt like a real teammate.
vp3	Never saw it mislead us even in heavy smoke.
vp4	Trustworthy recommendations under pressure.
vp5	Cut search time dramatically on my last mission.
vp6	Robust even when sensors were partially occluded.

Table B.3: Slightly positive reviews.

ID	Review text
sp1	Usually helpful, had a few minor hiccups.
sp2	Most calls were good, one or two were iffy.
sp3	Guidance was generally reliable for us.
sp4	Improved flow in most rooms we cleared.
sp5	Sometimes slow, but pointed us right more often than not.
sp6	Reasonable detections, occasional false positives.

Table B.4: Mixed reviews.

ID	Review text
mx1	Some finds were impressive; some felt like guesses.
mx2	Good in open areas, struggled in cramped corridors.
mx3	Worked well early, degraded later in the mission.
mx4	Balanced results overall — neither great nor poor.
mx5	Learned to take its call with a pinch of salt.
mx6	Performance varied with smoke density.

Table B.5: Slightly negative reviews.

ID	Review text
sn1	A few wrong calls made us cautious.
sn2	Helped occasionally, but missed key signs.
sn3	Too uncertain in cluttered rooms.
sn4	Sometimes diverted us away from viable areas.
sn5	Thermal bloom confused it more than expected.
sn6	Needed babysitting to be useful.

Table B.6: Very negative reviews.

ID	Review text
vn1	Frequently wrong — hurt our pace and safety.
vn2	We stopped listening; way too many false leads.
vn3	Unreliable under pressure — wouldn't recommend.
vn4	Added confusion when seconds mattered.
vn5	Sensors failed in heat; guidance became random.
vn6	Created more problems than it solved.

B.9. Experiment Walkthrough

This section provides a visual overview of the participant experience during the online experiment. Only stages involving substantial visual interaction or graphical elements are included here. Purely text-based instruction screens and questionnaires are described in the main body of the thesis and are therefore omitted from this walkthrough.

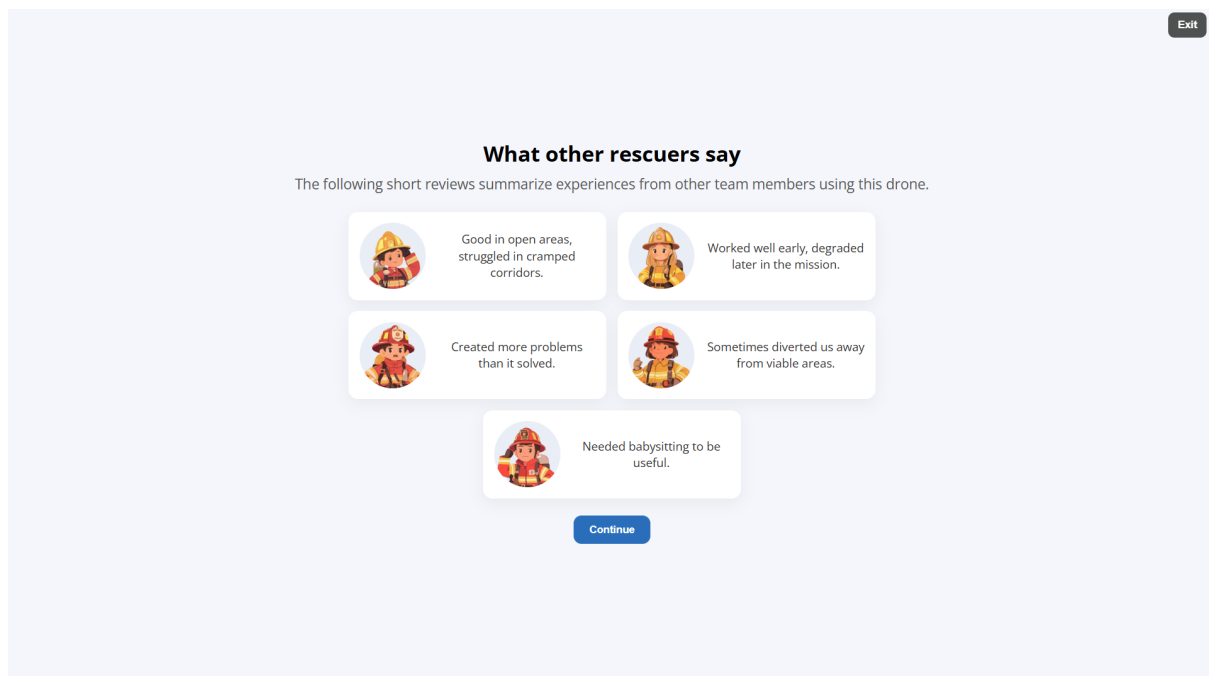


Figure B.1: Reputation exposure stage. Participants are presented with randomized visually formatted review cards conveying prior evaluations of the drone system. These stimuli provide an initial reputation cue intended to influence early trust formation before task interaction begins.

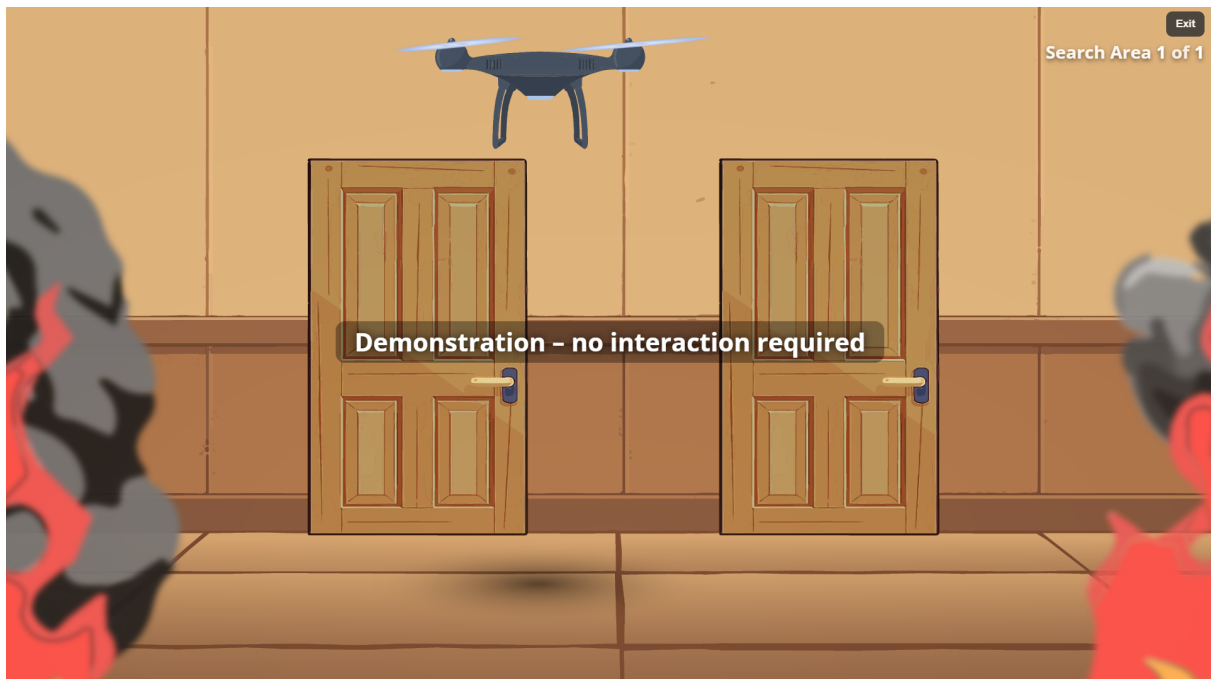


Figure B.2: Demonstration (training) phase. Participants observe a non-interactive example of the search task, including environment rendering and drone motion, allowing familiarization with the interface and task dynamics.



Figure B.3: Drone scan phase. The drone performs an animated scan of the environment, visually indicating the assessment process preceding its recommendation. The environmental integrity timer is not yet active throughout this stage.



Figure B.4: Recommendation and decision stage. The drone indicates a suggested door, after which participants choose whether to follow or ignore the recommendation under time pressure. Visual urgency cues intensify as the decision timer approaches expiration.



Figure B.5: Outcome feedback — empty door. The selected door opens and no victim is present, indicating an unsuccessful decision. This negative outcome contributes to the participant's personal-experience signal and influences subsequent trust updates in the model.



Figure B.6: Outcome feedback — victim found. The selected door opens and the victim is revealed, indicating a successful decision. This positive outcome reinforces reliance and contributes to personal-experience-driven trust adaptation over time.

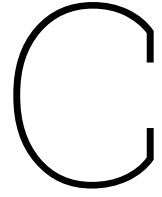
B.10. Documentation and Repository Access

The full implementation of the online experiment is maintained as a dedicated repository accompanied by technical documentation. The documentation describes the experiment flow, architecture, configuration system, trial modules, and the end-to-end logging pipeline, including durability measures and withdrawal support.

Repository availability

The full source code of the browser-based experiment platform developed and used for the online study is publicly available in a dedicated GitHub repository. The repository provides access to the complete HTML/JavaScript implementation, documentation of the experiment timeline and trial logic, and a description of the data logging and export mechanisms used to produce the dataset analyzed in this thesis.

Repository link: https://github.com/ripcidalia/trust_experiment.



Software Implementation and Reproducibility

C.1. Purpose of the Software Framework

The methodology presented in this thesis required the development of a complete software framework to operationalize the modeling, identification, and evaluation of human trust in a human–robot interaction scenario. While the primary scientific contribution of this work lies in the formulation of the trust dynamics model and the associated identification and validation methodology, the implementation of these concepts in software was essential for executing the experimental pipeline and producing the reported results.

A modular MATLAB codebase was therefore developed to support the full workflow described throughout the thesis, including data preprocessing, measurement calibration, trust simulation, parameter estimation, behavioural modeling, and closed-loop evaluation. This software served as the computational backbone enabling the transformation of theoretical constructs into executable models and reproducible experiments.

The implementation reflects the methodological structure described in the thesis chapters addressing trust model formulation, experimental data processing, parameter identification, and behavioral coupling. Each methodological component was translated into a corresponding computational module, ensuring that all analyses presented in this work could be executed programmatically and reproduced.

C.2. Scope of the Implemented Codebase

The developed software framework constitutes a complete research pipeline rather than a collection of isolated scripts. It includes components for:

- preprocessing and structuring experimental data,
- calibration and harmonization of trust measurements,
- simulation of the continuous-time trust dynamics model,
- weighted least-squares parameter estimation,
- diagnostic and validation analyses,
- behavioral modeling of reliance decisions,
- coupled trust–behavior simulations and rollout experiments.

The system was designed with modularity and reproducibility as primary objectives. Individual components can be executed independently, while a top-level orchestration script enables end-to-end execution of the full experimental pipeline. Intermediate outputs are stored in structured run directories, ensuring traceability of parameter estimation and analysis results.

C.3. End-to-End Pipeline Implemented in Software

The software framework implements the full methodological pipeline of the thesis. Each stage described conceptually in the research methodology corresponds to a dedicated computational module.

Methodological Stage	Software Implementation
Experimental data structuring	Preprocessing pipeline converting raw logs into structured participant representations
Measurement calibration	Mapping and weighting modules aligning multiple trust instruments onto a common scale
Trust dynamics modelling	Continuous-time simulation of dispositional, latent, situational, and reputation trust components
Parameter identification	Weighted least-squares optimisation with multiple solver backends
Model validation and diagnostics	Residual analysis, trajectory inspection, and sensitivity studies
Behavioural modelling	Probabilistic models linking trust states to reliance decisions
Coupled system evaluation	Closed-loop simulations of trust and behaviour under stochastic interaction scenarios

Table C.1: Mapping between methodological stages and implemented software modules.

Through this pipeline, the software enables the full execution of the research workflow, from raw experimental data to calibrated models, fitted parameters, and behavioural simulations. As such, it functions as the operational embodiment of the methodological framework presented in this thesis.

C.4. Reproducibility and Run Management

A central design objective of the software framework was reproducibility. The implementation includes mechanisms to ensure that model fitting and analysis results can be replicated consistently:

- Run-based execution: each optimisation and analysis is associated with a unique run identifier.
- Archival of intermediate artefacts: calibrated measurements, fitted parameters, and diagnostics are stored systematically.
- Deterministic simulation modes: trust simulations can be replayed using recorded event sequences.
- Separation of training and validation data: the pipeline supports structured participant splits and out-of-sample evaluation.

These design choices allow experiments to be reproduced, extended, or re-analysed without manual intervention, supporting methodological transparency and scientific reliability.

C.5. Role of Software in the Research Contribution

The software implementation supports, but does not constitute, the primary scientific contribution of this thesis. The central contributions lie in:

- the formulation of a dynamic, interpretable trust model,
- the identification framework linking experimental data to model parameters,
- the validation methodology and behavioural coupling analysis.

The developed codebase enabled these contributions by providing the computational infrastructure required to test hypotheses, estimate parameters, simulate interaction scenarios, and generate empirical results. It therefore serves as an enabling research tool that operationalises the theoretical and methodological components of the work.

C.6. Reuse and Research Impact

Although developed specifically for the experimental context of this thesis, the software framework was designed with generality in mind. Its modular structure allows adaptation to other human–robot interaction studies, trust modelling experiments, or behavioural analysis tasks involving sequential decision-making.

Potential reuse includes:

- trust-aware control and human–robot interaction research,
- simulation-based behavioural modelling,
- parameter identification in cognitive or behavioural dynamical systems,
- experimental pipeline design for longitudinal interaction studies.

By providing a reusable and extensible implementation of the modelling and analysis pipeline, the software contributes to the broader reproducibility and methodological development of research in human–robot trust and interaction modelling.

C.7. Documentation and Repository Access

The full software implementation is maintained as a structured repository accompanied by comprehensive technical documentation. The documentation includes:

- a system overview describing the architecture and execution pipeline,
- module-level documentation for preprocessing, measurement mapping, trust modelling, optimisation, diagnostics, and behavioural modelling,
- detailed descriptions of the analysis pipeline and reproducibility workflow.

The repository and accompanying documentation are intended to enable other researchers to understand, reproduce, and extend the modelling and analysis framework developed in this thesis.

Repository availability

The full MATLAB implementation of the trust and behavioural modelling framework developed in this thesis is publicly available in a dedicated GitHub repository. The repository provides access to the complete source code, structured technical documentation, and instructions for reproducing the preprocessing, model identification, and analysis pipeline described in this work.

Repository link: https://github.com/ripcidalia/hri_trust_dynamics_model.

References

- Ahmad, M., Alzahrani, A., Robinson, S., & Rahat, A. (2023). Modelling human trust in robots during repeated interactions. *Proceedings of the 11th International Conference on Human-Agent Interaction*, 281–290. <https://doi.org/10.1145/3623809.3623892>
- Akash, K., McMahon, G., Reid, T., & Jain, N. (2020). Human trust-based feedback control: Dynamically varying automation transparency to optimize human-machine interactions. *IEEE Control Systems Magazine*, 40(6), 98–116. <https://doi.org/10.1109/MCS.2020.3019151>
- Akash, K., Polson, K., Reid, T., & Jain, N. (2019). Improving human-machine collaboration through transparency-based feedback – part i: Human trust and workload model [2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018]. *IFAC-PapersOnLine*, 51(34), 315–321. <https://doi.org/10.1016/j.ifacol.2019.01.028>
- Akash, K., Reid, T., & Jain, N. (2019). Improving human-machine collaboration through transparency-based feedback – part ii: Control design and synthesis [2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018]. *IFAC-PapersOnLine*, 51(34), 322–328. <https://doi.org/10.1016/j.ifacol.2019.01.026>
- Alzahrani, A. S., & Ahmad, M. I. (2026). The architecture of trust: A three-layered mathematical model for human-robot collaboration. *Proceedings of the 13th International Conference on Human-Agent Interaction*, 332–340. <https://doi.org/10.1145/3765766.3765792>
- Aster, R. C., Borchers, B., & Thurber, C. H. (2013). Chapter one - introduction. In R. C. Aster, B. Borchers, & C. H. Thurber (Eds.), *Parameter estimation and inverse problems (second edition)* (Second Edition, pp. 1–23). Academic Press. <https://doi.org/10.1016/B978-0-12-385048-5.00001-X>
- Azevedo Sá, H., Robert, L., Tilbury, D., Jayaraman, S., Esterwood, C., & Yang, X. J. (2021). Real-time estimation of drivers' trust in automated driving systems framework for modeling the dynamics of drivers' trust. *International Journal of Social Robotics*, 13. <https://doi.org/10.1007/s12369-020-00694-1>
- Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2), 231–260. <https://doi.org/10.1086/292745>
- Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, 12(3), 183–210. [https://doi.org/10.1002/\(SICI\)1099-1727\(199623\)12:3<183::AID-SDR103>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1099-1727(199623)12:3<183::AID-SDR103>3.0.CO;2-4)
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370. <https://doi.org/10.1037/1089-2680.5.4.323>
- Borsboom, D., Mellenbergh, G., & Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110, 203–219. <https://doi.org/10.1037/0033-295X.110.2.203>
- Bourgais, M., Taillandier, P., & Vercouter, L. (2020). Ben: An architecture for the behavior of social agents. *Journal of Artificial Societies and Social Simulation*, 23. <https://doi.org/10.18564/jasss.4437>
- Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press, Cambridge, MA.
- Campagna, G., & Rehm, M. (2025). A systematic review of trust assessments in human–robot interaction. *J. Hum.-Robot Interact.*, 14(2). <https://doi.org/10.1145/3706123>
- Chen, M., Soh, H., Hsu, D., Nikolaidis, S., & Srinivasa, S. (2020). Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction*, 9. <https://doi.org/10.1145/3359616>
- Couceiro, M. S., Portugal, D., Rocha, R. P., & Araújo, A. (2021). Fostering human–robot cooperative architectures for search and rescue missions in urban fires. *Simulation*, 97, 177–194. <https://doi.org/10.1177/0037549720964548>

- Cristiani, E., Piccoli, B., & Tosin, A. (2014). An overview of the modeling of crowd dynamics. In *Multiscale modeling of pedestrian dynamics* (pp. 73–107). Springer International Publishing. https://doi.org/10.1007/978-3-319-06620-2_4
- Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in hci agents. *Computers in Human Behavior*, 29(3), 577–579. <https://doi.org/https://doi.org/10.1016/j.chb.2012.11.023>
- Das, T. K., & Teng, B.-S. (2004). The risk-based view of trust: A conceptual framework. *Journal of Business and Psychology*, 19, 85–116. <https://doi.org/https://doi.org/10.1023/B:JOBU.0000040274.23551.1b>
- Davis, S. E. (2019). *Individual differences in operators' trust in autonomous systems: A review of the literature* (tech. rep.). Australian Department of Defence, Science and Technology.
- de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning [Trust and Technology]. *International Journal of Human-Computer Studies*, 58(6), 719–735. [https://doi.org/https://doi.org/10.1016/S1071-5819\(03\)00039-9](https://doi.org/https://doi.org/10.1016/S1071-5819(03)00039-9)
- Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Wanco, H. (2013). Impact of robot failures and feedback on real-time trust. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 251–258. <https://doi.org/10.1109/HRI.2013.6483596>
- Du, N., & Yang, X. J. (2019). Not all information is equal: Effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming. *Human Factors The Journal of the Human Factors and Ergonomics Society*. <https://doi.org/10.1177/0018720819862916>
- Floyd, M. W., Drinkwater, M., & Aha, D. W. (2015). Trust-guided behavior adaptation using case-based reasoning. *Proceedings of the 24th International Conference on Artificial Intelligence*, 4261–4267.
- Freeddy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. *2007 International Symposium on Collaborative Technologies and Systems*, 106–114. <https://doi.org/10.1109/CTS.2007.4621745>
- Gao, F., Clare, A. S., Macbeth, J. C., & Cummings, M. L. (2013, March). *Modeling the impact of operator trust on performance in multiple robot control* (tech. rep.). AAAI Spring Symposium: Trust and Autonomous Systems.
- Guo, Y., Shi, C., & Yang, X. J. (2021). Reverse psychology in trust-aware human-robot interaction. *IEEE Robotics and Automation Letters*, 6, 4851–4858. <https://doi.org/10.1109/LRA.2021.3067626>
- Guo, Y., & Yang, X. J. (2021). Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 13, 1899–1909. <https://doi.org/10.1007/s12369-020-00703-3>
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., Visser, E. J. D., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53, 517–527. <https://doi.org/10.1177/0018720811417254>
- Hardin, R. (2006). *Trust*. Polity Press.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57, 407–434. <https://doi.org/10.1177/0018720814547570>
- Hutter, M., & Siegwart, R. (2018). Field and service robotics. In B. Siciliano & O. Khatib (Eds.), *Springer proceedings in advanced robotics 5* (Vol. 5). Springer. <http://www.springer.com/series/15556>
- Kanouse, D. E., & Hanson, L. R. (1987). Negativity in evaluations. In *Attribution: Perceiving the causes of behavior* (pp. 47–62). Lawrence Erlbaum Associates, Inc.
- Kinateder, M. (2012). *Social influence in emergency situations-studies in virtual reality* [Doctoral dissertation, Julius-Maximilians-Universität Würzburg].
- Kinateder, M., Comunale, B., & Warren, W. H. (2018). Exit choice in an emergency evacuation scenario is influenced by exit familiarity and neighbor behavior. *Safety Science*, 106, 170–175. <https://doi.org/10.1016/j.ssci.2018.03.015>

- Kinateder, M., Kuligowski, E. D., Reneke, P. K., & Peacock, R. D. (2014, September). *A review of risk perception in building fire evacuation* (tech. rep.). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.TN.1840>
- Kinateder, M., & Warren, W. H. (2016). Social influence on evacuation behavior in real and virtual environments. *Frontiers Robotics AI*, 3. <https://doi.org/10.3389/frobt.2016.00043>
- Kinateder, M., & Warren, W. H. (2021). Exit choice during evacuation is influenced by both the size and proportion of the egressing crowd. *Physica A: Statistical Mechanics and its Applications*, 569. <https://doi.org/10.1016/j.physa.2021.125746>
- Kinsey, M. J., Gwynne, S. M., Kuligowski, E. D., & Kinateder, M. (2019). Cognitive biases within decision making during fire evacuations. *Fire Technology*, 55, 465–485. <https://doi.org/10.1007/s10694-018-0708-0>
- Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences*, 42(2), 92–101. <https://doi.org/https://doi.org/10.1016/j.tins.2018.10.004>
- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35, 1243–1270. <https://doi.org/10.1080/00140139208967392>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of trusted autonomy* (pp. 135–159). Springer International Publishing. https://doi.org/10.1007/978-3-319-64816-3_8
- Li, F., Mikulski, D., Wagner, J., & Wang, Y. (2019). *Trust-based control and scheduling for ugv platoon under cyber attacks* (tech. rep.). SAE Technical Paper. <https://doi.org/10.4271/2019-01-1077>
- Lu, Y., & Sarter, N. (2019). Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems*, 49, 560–568. <https://doi.org/10.1109/THMS.2019.2930980>
- Mahani, M. F., Jiang, L., & Wang, Y. (2021). A bayesian trust inference model for human-multi-robot teams. *International Journal of Social Robotics*, 13, 1951–1965. <https://doi.org/10.1007/s12369-020-00705-1>
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6, 57–87. <https://doi.org/10.1177/1555343411433844>
- Merritt, S., & Ilgen, D. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50, 194–210. <https://doi.org/10.1518/001872008X288574>
- Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, 55, 520–534. <https://doi.org/10.1177/0018720812465081>
- 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. <https://doi.org/10.1080/00140139408964957>
- Murphy, R. (2004). Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2), 138–153. <https://doi.org/10.1109/TSMCC.2004.826267>
- Nam, C., Walker, P., Lewis, M., & Sycara, K. (2017). Predicting trust in human control of swarms via inverse reinforcement learning. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 528–533. <https://doi.org/10.1109/ROMAN.2017.8172353>
- Nam, C., Walker, P., Li, H., Lewis, M., & Sycara, K. (2020). Models of trust in human control of swarms with varied levels of autonomy. *IEEE Transactions on Human-Machine Systems*, 50(3), 194–204. <https://doi.org/10.1109/THMS.2019.2896845>

- Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Plato. (1955). *The republic* (H. D. P. Lee, Trans.). Penguin Classics.
- Rani, M. R. A., Sinclair, M., & Case, K. (2000). Human mismatches and preferences for automation. *International Journal of Production Research*, 38(17), 4033–4039. <https://doi.org/10.1080/00207540050204894>
- Rantanen, P., Parkkari, T., Leikola, S., Airaksinen, M., & Lyles, C. A. (2017). An in-home advanced robotic system to manage elderly home-care patients' medications: A pilot safety and usability study. *Clinical Therapeutics*, 39. <https://doi.org/10.1016/j.clinthera.2017.03.020>
- Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, 47(4), 425–436. <https://doi.org/10.1109/THMS.2017.2648849>
- Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 101–108. <https://doi.org/10.1109/HRI.2016.7451740>
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4), 651–665. <https://doi.org/https://doi.org/10.1111/j.1467-6494.1967.tb01454.x>
- Sabater, J., & Sierra, C. (2005). Review on computational trust and reputation models. <https://doi.org/10.1007/s10462-004-0041-5>
- Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2015-March*, 1–8. <https://doi.org/10.1145/2696454.2696497>
- Sapienza, A., Cantucci, F., & Falcone, R. (2022). Modeling interaction in human–machine systems: A trust and trustworthiness approach. *Automation*, 3, 242–257. <https://doi.org/10.3390/automation3020012>
- Schaefer, K. E. (2013). *The perception and measurement of human-robot trust* [Doctoral dissertation, University of Central Florida].
- Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58, 377–400. <https://doi.org/10.1177/0018720816634228>
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1), 148–175. <https://doi.org/10.1109/JPROC.2015.2494218>
- Soh, H., Xie, Y., Chen, M., & Hsu, D. (2020). Multi-task trust transfer for human–robot interaction. *Int. J. Rob. Res.*, 39(2–3), 233–249. <https://doi.org/10.1177/0278364919866905>
- Strutz, T. (2010). *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Springer Vieweg Wiesbaden.
- Ullrich, D., Butz, A., & Diefenbach, S. (2021). The development of overtrust: An empirical simulation and psychological analysis in the context of human-robot interaction. *Frontiers in robotics and AI*, 8. <https://doi.org/10.3389/frobt.2021.554578>
- Unhelkar, V. V., Siu, H. C., & Shah, J. A. (2014). Comparative performance of human and mobile robotic assistants in collaborative fetch-and-deliver tasks. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 82–89. <https://doi.org/10.1145/2559636.2559655>
- van der Wal, C. N., Formolo, D., Robinson, M. A., Minkov, M., & Bosse, T. (2017). Simulating crowd evacuation with socio-cultural, cognitive, and emotional elements. In J. Mercik, N. T. Nguyen, & R. Kowalczyk (Eds.), *Transactions on computational collective intelligence xxvii* (pp. 139–177). Springer International Publishing. https://doi.org/10.1007/978-3-319-70647-4_11

- Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 109–116.
- Wang, Y., Li, F., Zheng, H., Jiang, L., Mahani, M. F., & Liao, Z. (2024). Human trust in robots: A survey on trust models and their controls/robotics applications. *IEEE Open Journal of Control Systems*, 3, 58–86. <https://doi.org/10.1109/OJCSYS.2023.3345090>
- Wetzel, J. M., Sheffert, S. M., & Backs, R. W. (2004). Driver trust, annoyance, and acceptance of an automated calendar system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(19), 2335–2339. <https://doi.org/10.1177/154193120404801928>
- Wilmes, L., & Jamshidnejad, A. (2025). Gt-bdi model: A combined game-theoretic and bdi-based computational model for emergency evacuation with search and rescue robots. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2025.3555870>
- Xu, A., & Dudek, G. (2015). Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2015-March*, 221–228. <https://doi.org/10.1145/2696454.2696492>
- Yang, X. J., Guo, Y., & Schemanske, C. (2022). From trust to trust dynamics: Combining empirical and computational approaches to model and predict trust dynamics in human-autonomy interaction. *Human-Automation Interaction*, 253–265. https://doi.org/10.1007/978-3-031-10784-9_15
- Yang, X. J., Schemanske, C., & Searle, C. (2023). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, 65, 862–878. <https://doi.org/10.1177/00187208211034716>
- Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. *ACM/IEEE International Conference on Human-Robot Interaction, Part F127194*, 408–416. <https://doi.org/10.1145/2909824.3020230>
- Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016). How users adjust trust in automation: Contrast effect and hindsight bias. *Proceedings of the Human Factors and Ergonomics Society*, 196–200. <https://doi.org/10.1177/1541931213601044>
- Zahedi, Z., Verma, M., Sreedharan, S., & Kambhampati, S. (2023). Trust-aware planning: Modeling trust evolution in iterated human-robot interaction. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 281–289. <https://doi.org/10.1145/3568162.3578628>
- Zhang, M. Y., & Yang, X. J. (2017). Evaluating effects of workload on trust in automation, attention allocation and dual-task performance. *Proceedings of the Human Factors and Ergonomics Society, 2017-October*, 1799–1803. <https://doi.org/10.1177/1541931213601932>
- Zhang, Y., Kinateder, M., Huang, X., & Warren, W. H. (2025). Modeling competing guidance on evacuation choices under time pressure using virtual reality and machine learning. *Expert Systems with Applications*, 262. <https://doi.org/10.1016/j.eswa.2024.125582>
- Zheng, H., Smereka, J., Mikulski, D., & Wang, Y. (2023). Bayesian optimization based trust model for human multi-robot collaborative motion tasks in offroad environments. *International Journal of Social Robotics*, 15, 1–21. <https://doi.org/10.1007/s12369-023-01011-2>