Comparing Trust Development in Human and Robot Collaboration

MSc Computer Science Thesis

Ching Guo

Delft University of Technology

Comparing Trust Development in Human and Robot Collaboration

by

Ching Guo

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday February 24th, 2025.

Student number:	5052300
Project duration:	March 20, 2025 – February 24, 2026
Thesis committee:	Prof. Myrthe L. Tielman Prof. Mark Neerincx Prof. Avishek Anand
Faculty:	Interactive Intelligence Group Department of Intelligent Systems Faculty of Electrical Engineering, Mathematics and Computer Science Delft University of Technology

Abstract

As robots increasingly transition from automated tools to collaborative teammates, trust becomes a central requirement for effective human–robot collaboration. While prior research has examined trust in human–robot interaction, little is known about how trust dynamically unfolds across its full trajectory of formation, violation, and repair, particularly in comparison to human–human collaboration and within physically co-present settings. This thesis investigates how a teammate's identity, human versus robot, shapes the development of interpersonal trust over time.

A controlled laboratory study was conducted in which participants collaborated with either a human confederate or an anthropomorphic robot teammate on a cooperative building task requiring high interdependence. Trust was measured across three phases: initial collaboration (trust formation), a competence-based mistake (trust violation), and a subsequent repair attempt involving an apology, explanation, and promise (trust recovery). Trust was measured using trust questionnaires capturing trusting beliefs and trusting intentions. Data were analyzed using a Bayesian multilevel modeling approach to account for repeated measures and individual differences.

The results show that participants initially reported lower trust toward the robot than toward the human teammate. Contrary to expectations based on the perfect automation schema, trust declined more sharply following a mistake by the human than by the robot. During the recovery phase, trust rebounded in both conditions. Trust toward the robot recovered to its initial level, while trust toward the human did not fully return to baseline.

Analyses across trust dimensions further revealed that benevolence perceptions toward the robot improved over time, narrowing the initial gap between human and robot teammates. Competence perceptions showed similar violation and recovery patterns across conditions. In contrast, trusting intentions showed a more uneven pattern: although willingness to rely on the robot seemingly returned to its own baseline during recovery, the human–robot difference widened again at t_3, suggesting that reliance remained more sensitive to teammate identity even as other trust dimensions converged.

Overall, this study demonstrates that trust toward human and robot teammates follows similar formation, violation, and recovery phases, but differs in how changes are anchored to initial expectations and distributed across trust dimensions. Specifically, participants began with lower trust in the robot, yet a human teammate's mistake produced a sharper drop and less complete return to baseline than a comparable robot mistake. While trust toward the robot increased relative to its own baseline, particularly through benevolence, willingness to rely remained more differentiated by teammate identity. These findings show that aggregated trust scores can mask dimension-specific dynamics and that recovery in trust beliefs does not necessarily translate into equivalent recovery in trusting intentions. Practically, this suggests that designing for effective human–robot teamwork requires addressing not only how robots regain positive evaluations after errors, but also how to support users' willingness to rely on them in interdependent tasks.

Acknowledgement

I still remember being asked to read 10 papers initially to prepare for my kick-off, and how that immediately filled my head with questions and ideas for what my thesis could be. In the end, the scope was too big and too grand, and it was gradually reduced into the thesis you are currently reading. Who would have thought that it originally aimed to include two human participants and one robot teammate to mimic a "bigger" and perhaps more realistic representation of what we call "teamwork"? Or that I initially wanted to investigate trust violations only, and not recovery? The idea expanded, then became smaller, and eventually led to what we have today.

Therefore, a special thank you to my supervisor, Professor Myrthe L. Tielman, for always telling me not to make my idea too grand, for keeping me rational, and for always giving me tips to make my thesis even better. I learned a lot from her, and I truly wouldn't have been able to deliver this without her. Thank you very much.

I also want to thank my thesis committee for reading this (what I already find) quite lengthy thesis, and for sparing the time for my defense so that I can successfully graduate.

Now onto the thank yous beyond university staff. Thank you so much to my friends who were willing to extend their entire friend networks for me, since I struggled to find 40 participants myself. A special thank you to my sister, Rainy Guo, for always helping me film things, invite friends and classmates, and act as a human confederate when needed. Thank you Shreya, Floortje, and Hang for also being there, helping me find participants and stepping in as a human confederate when needed. Thank you Kayleigh for your moral support. And of course, thank you mom for always believing in me, even though I wanted to quit my Master's quite a few times.

I had a great and I would say quite relaxed time during my thesis, and that was all because of the environment I worked in. Thank you everyone for being here during this time, and I would like to say: "I am officially finished!"

Contents

List of Figures

List of Tables

1 Introduction

Robots powered by artificial intelligence (AI) are no longer limited to executing predefined, automated tasks; they are increasingly deployed in social and collaborative environments where they interact directly with humans. From autonomous rescue robots being developed by fire departments to robotic assistants in hospitals and educational settings, embodied AI systems are beginning to take on roles that require cooperation rather than mere task execution [95, 55, 60, 11]. In these contexts, AI functions as the underlying decision-making and sensing capability, while the robot serves as the physical and social interface for collaboration. This transition from automation to collaboration has transformed the perception of robots from passive instruments into active teammates capable of influencing human performance and interaction dynamics.

As robots become more autonomous and socially expressive, an important question arises: do humans trust robotic teammates in the same way they trust other humans? Research shows that human–robot teams often experience lower trust and weaker team identification compared to purely human teams [38, 40]. This difference is partly explained by social categorization processes. Humans tend to trust those perceived as part of their ingroup, while robots are often categorized as outgroup members, due to differences in appearance, communication style, or behavior [93, 104]. As a result, people may feel less connected to robotic teammates and hesitate to rely on them, particularly when mistakes occur.

Trust is central to teamwork because it allows individuals to accept vulnerability in interdependent situations [38, 58]. Yet, trust is fragile and dynamic; it develops through successful collaboration, can be violated when expectations are unmet, and may be repaired through strategies such as apology, explanation, or promise [23, 52]. While these processes are well documented in human–human contexts, far less is known about how trust evolves across all three stages—formation, violation, and repair—in embodied human–robot teams.

Robots are often held to higher performance standards than humans, a phenomenon known as the perfect automation schema [59, 54]. Consequently, people may react more negatively to a robot's mistake than to a human's and may also find it more difficult to forgive. At the same time, robots are increasingly being designed to look and behave more like humans. Anthropomorphic cues such as natural speech, facial expressions, and social gestures can make robots appear more relatable and may strengthen trust resilience [42, 5]. This raises a central question for the present research: if robots increasingly resemble human teammates in both appearance and behavior, can they also earn and maintain trust in the same way, or will differences in perceived social belonging continue to shape how people respond when robots make mistakes?

1.1. Research Questions and Hypotheses

The main research question that guides this thesis is:

> **RQ:** How does a teammate's identity (human vs robot) affect the development of trust across mistake and recovery phases in collaboration?

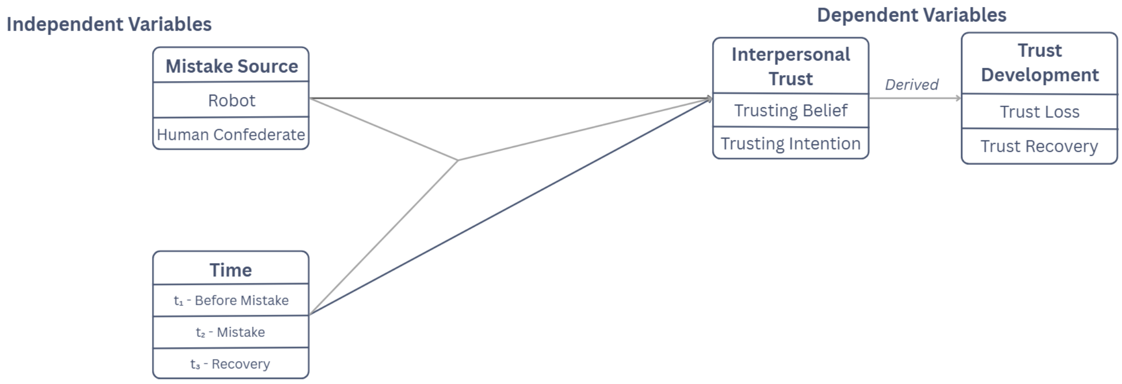


Figure 1.1: The conceptual model of the variables in this research. The left column shows the independent variables: mistake source (between-subject; human confederate vs robot) and time (within-subject; t1 baseline, t2 mistake, t3 recovery). The right column shows the dependent variables: interpersonal trust (trusting beliefs and trusting intentions) and the derived outcomes for trust development (trust loss and trust recovery). The arrows depict the hypothesized effect relationships. The forked arrows linking mistake source and time to interpersonal trust indicate possible interaction effects.

As introduced in Section 1, trust in human–robot collaboration is not static but evolves over time. It forms during early interactions, may be violated when a mistake occurs, and can recover through repair strategies. As robots become more human-like and increasingly take on social and cooperative roles, it becomes essential to understand whether trust toward them develops in ways comparable to trust toward human teammates. To address this question, the present research examines trust development across three phases—formation, violation, and repair—over repeated interactions with either a human confederate or a robot teammate. The conceptual model is presented in Figure 1.1.

The main question is explored through four sub-questions that correspond to the phases of trust development, each followed by its associated hypothesis and rationale.

RQ(a): How does initial interpersonal trust differ when cooperating with a robot teammate compared to a human teammate?

H(a): Initial trust will be lower for a robot teammate than for a human teammate.

Trust formation is shaped by social categorization: humans tend to trust those perceived as part of their ingroup, while robots are often seen as outgroup members due to differences in appearance, communication, or perceived intentionality [93, 104]. This outgroup bias can reduce initial trust, particularly in early stages of interaction when limited behavioral evidence is available.

RQ(b): How does the amount of trust lost after an accidental mistake differ when the teammate is a robot compared to a human?

H(b): The loss of trust after a mistake will be greater when the mistake is made by a robot than when it is made by a human teammate.

Research on the perfect automation schema shows that people tend to hold automated systems to higher performance standards and react more negatively to their errors [68, 63]. When a robot fails, the violation of its presumed reliability or competence can lead to a sharper decline in trust than when a human makes the same error.

RQ(c): How does the amount of trust recovered, given the same recovery strategy, differ when the teammate is a robot compared to a human?

H(c): Under identical recovery strategies, trust will recover less when the teammate is a robot than when it is a human.

Although apology, explanation, and promise are effective repair strategies in human–human trust recovery, their effectiveness depends on perceived sincerity and emotional understanding [61]. Robots often struggle to convey authentic emotional cues, leading people to interpret their apologies as less genuine or diagnostic of future behavior. Additionally, the higher expectations associated with automation can make forgiveness slower or incomplete. Consequently, even if a robot delivers the same recovery message as a human, the perceived meaning and its impact on trust are likely to be weaker.

RQ(d): How does overall interpersonal trust develop across formation, violation, and recovery phases when collaborating with a robot compared to a human?

H(d): Overall, the robot teammate will maintain lower trust levels throughout the experiment, with a steeper drop during the mistake phase and a slower recovery in the final phase. Trust in both conditions is expected to follow a similar general trajectory, but with different magnitudes. The robot's trajectory will start lower, decline more sharply during the violation, and recover more slowly, reflecting both initial social distance and stronger negative reactions to perceived reliability failures.

In summary, this research examines how interpersonal trust unfolds dynamically in human–robot versus human–human teamwork, focusing on the full cycle of formation, violation, and recovery. By distinguishing these phases and comparing their trajectories, the study aims to provide a clearer understanding of when and why trust toward robots diverges from trust toward humans, offering insights for the design of more reliable and socially aware robotic teammates.

1.2. Report Structure

This chapter has outlined the motivation and research goals of this thesis. The remainder of the report is structured as follows. Chapter 2 provides an overview of the theoretical background and related work on interpersonal trust, social categorization, and trust in automation and robots. Chapter 3 introduces the collaborative task used in this study and describes the scenario, roles, and scripted mistake and recovery phases. Chapter 4 details the research design, participants, measures, and procedure.

The empirical findings are presented in Chapter 5, which reports trust levels across formation, violation, and recovery phases, as well as differences between human and robot teammates. Chapter 6 discusses the results in light of existing literature, outlines the limitations of the study, and suggests directions for future research. Finally, Chapter 7 concludes the thesis.

2
Background

No longer confined to the background as invisible tools, AI systems increasingly step forward as partners that act, communicate, and even collaborate alongside people. From voice assistants in our pockets to robots working in classrooms, hospitals, and workplaces, these technologies are becoming part of the social and organizational fabric of everyday life. This shift invites not only technical innovation but also deeper questions: How do humans perceive these systems—as tools, as teammates, or as something in between? What role does trust play in determining whether collaboration succeeds or fails?

This chapter sets the stage by tracing the movement from automation to collaboration, exploring the foundations of human–AI teamwork, and highlighting the role of anthropomorphism in shaping perceptions of artificial agents. Finally, it draws attention to trust as a defining element of collaboration, laying the foundation for the discussion in this study.

2.1. From Automation to Collaboration

2.1.1. AI in Everyday Life

The distinction between humans and machines has blurred due to rapid technological developments in robotics, automation, and natural language processing. These advances provide the foundation for what is broadly referred to as artificial intelligence (AI), which can be defined as the tangible real-world capability of non-human machines or artificial entities to perform, task solve, communicate, interact, and act logically as it occurs with biological humans [39]. AI manifests in many forms, ranging from purely digital applications such as voice assistants to physical embodiments such as autonomous robots.

Although many of our daily interactions are still with other humans, we increasingly communicate with AI systems that are not human but display intelligent, sociable, and human-like behaviors. Examples include Siri, Apple's voice-activated assistant with a distinctive personality, and Google's self-driving cars that adjust their driving style to cope with aggressive human drivers [73, 71]. These systems illustrate how AI has become more visible in daily life, shifting from background automation toward active engagement with people.

As AI systems continue to develop, their role is no longer confined to software or virtual environments. Increasingly, AI takes physical form in the shape of robots that act and interact directly within human environments. According to IEEE, robots are "autonomous machine[s] capable of sensing [their] environment, carrying out computations to make decisions, and performing actions in the real world" [42]. Robots therefore represent a subset of AI systems that combine artificial intelligence with a physical body, enabling them not only to process information but also to manipulate and act upon the physical and social world.

Social robots, which are specifically developed to interact and communicate with humans, are becoming increasingly prominent in domains such as healthcare, education, and caregiving [8, 10, 9]. They are designed not only to assist with tasks but also to support social interaction, for example by serving as companions for the elderly, as a companionship in nursing homes or as learning partners for students

[60, 11, 10]. This development reflects a gradual shift in perspective, from viewing AI primarily as tools to considering them as potential teammates in human activities [109].

2.1.2. Human–AI teamwork

Artificial intelligence has rapidly expanded across workplaces and organizational settings. AI technologies are widely deployed to optimize work processes [49]. Instead of fully replacing human workers, research highlights that AI can be beneficial when humans and AI collaborate closely, complementing each other's strengths and weaknesses [2]. As AI systems continue to evolve, they increasingly transition from being mere tools to acting as independent team members [57].

Human–AI teamwork can be defined as a collaboration in which humans and one or more AI agents interact socially, pursue shared goals, and depend on one another's input for successful task completion [94]. An AI agent in this context refers to a computer-based entity that perceives and acts in its environment, possesses some degree of autonomy, and communicates or coordinates with human partners to achieve common objectives [83, 72]. Unlike traditional automation, which executes predefined tasks without flexibility, AI agents contribute proactively to team processes, demonstrate adaptability, and can be viewed as teammates rather than tools [109].

Two dominant theoretical perspectives shape our understanding of how people perceive such agents. The Computers as Social Actors (CASA) paradigm posits that humans apply the same social rules and expectations to machines as they do to other humans, often "mindlessly" attributing human-like qualities to them [70, 69]. Research shows that this perspective extends beyond computers to robots, with people responding to robotic agents in ways that mirror social expectations [59]. In contrast, the unique-agent hypothesis suggests that people view machines as fundamentally distinct from humans, often expecting near-perfect performance while holding them less morally accountable for errors [99, 63, 28, 48]. Both perspectives find empirical support, highlighting the complexity of human perceptions of AI teammates.

Within this broader landscape of human–AI collaboration, robots represent a different form of teamwork. Unlike purely digital AI agents such as chatbots or decision-support systems, robots act directly in the physical environment, which increases the level of interdependence between human and machine teammates. Interdependence is a central driver of trust, as tasks requiring communication, coordination, and shared decision-making create opportunities for trust to develop [82, 107]. Conversely, in contexts of low interdependence, trust is less likely to form [56].

The social categorization processes that govern human teamwork also extend to human–robot teams. New team members are evaluated based on their perceived similarity to existing members, with greater similarity leading to stronger perceptions of belonging and trust [93, 106]. Robots, however, are often perceived as dissimilar and may therefore initially face challenges in being accepted as full team members [104]. Anthropomorphic design can help mitigate this barrier by increasing perceptions of similarity even in early stages of interaction [24].

Overall, the integration of AI into teams marks a shift from automation as a supporting tool to AI systems as interdependent collaborators. Robots, as embodied AI agents, represent one form of such collaboration.

2.1.3. Anthropomorphism in Human–Robot Interaction

Anthropomorphism refers to the degree to which an artificial agent displays human-like characteristics, which often leads people to ascribe intent, motivation, goals, or a sense of self to the agent [29]. Advances in robotics have introduced humanoid social robots, such as the NAO robot, and pet-like robots, such as Sony's Aibo.

Research in social robotics suggests that humans often prefer anthropomorphic robots over non-anthropomorphic ones. For example, robots with a humanoid face are more likely to be perceived positively than those without [14]. Human-like robots are frequently judged as more intelligent [45], more sociable [50], and more likable [16]. In such interactions, people also form mental models based on unconscious assumptions, which they use to infer a robot's knowledge, skills, and performance [5].

2.1.4. Trust as the Foundation of Teamwork

The developments outlined above show how AI systems are increasingly integrated into human life, extending from software agents such as voice assistants to embodied robots that operate directly in physical and social environments [73, 71, 10]. These technologies have transitioned from being viewed as tools to being considered potential teammates that share tasks and goals with humans [57]. Theories such as the CASA paradigm and the unique-agent hypothesis highlight the complexity of how people perceive such agents [70, 69, 99, 63], while anthropomorphic design plays a role in shaping these perceptions by making robots appear more familiar, intelligent, and socially acceptable [24, 14, 16].

Yet, regardless of how advanced or anthropomorphic robots become, the success of human–robot teamwork ultimately depends on trust. Teamwork inherently requires interdependence, coordination, and a willingness to rely on others [82, 107]. Without trust, even the most capable robot will struggle to be accepted as a genuine teammate. Trust therefore forms the foundation upon which collaboration between humans and robots must be built, making it a central factor in understanding and enabling effective human–robot interaction [84].

2.2. Understanding Trust

2.2.1. Defining Trust

Trust is a foundational concept in social and organizational research. At its core, trust is defined as a willingness to be vulnerable to the actions of another party, based on the expectation that the other will act competently and with positive intent, even in the absence of direct monitoring or control [64]. This willingness to accept vulnerability distinguishes trust from related constructs such as cooperation or confidence. Without the presence of risk, uncertainty, and vulnerability, trust is not necessary.

The trust process can be understood as comprising three interrelated components: trusting beliefs, trusting intentions, and trusting actions [64, 20].

Trusting Beliefs.

Trusting beliefs represent the cognitive foundation of trust and consist of two elements: individual differences and trustworthiness perceptions. Individual differences reflect dispositional factors, such as propensity to trust, which describes a person's general tendency to rely on others across situations [6]. This propensity is especially influential in early interactions when little information about the trustee is available [67]. Trustworthiness perceptions, in contrast, are target-specific and capture the trustor's evaluation of the trustee's ability (competence in the given context), benevolence (the extent to which the trustee is believed to have the trustor's best interest in mind), and integrity (the perception that the trustee adheres to a set of principles or values acceptable to the trustor) [64]. Together, these beliefs shape whether a trustee is judged as worthy of trust.

Trusting Intentions.

Trusting intentions build on these beliefs and reflect the trustor's willingness to rely on the trustee in situations involving risk or vulnerability [6, 46]. Importantly, these intentions involve a decision to accept vulnerability, even when the trustor cannot fully monitor or control the trustee's actions.

Trusting Actions.

Trusting actions are the behavioral manifestations of trust, such as delegating tasks, following advice, or cooperating in joint activities [6, 77]. However, trust intentions and trust actions do not always align perfectly, since situational constraints or psychological factors may prevent intentions from being enacted [1].

Trust, therefore, is best understood as a dynamic process in which beliefs inform intentions, and intentions guide actions that may be revised as new information becomes available.

Interpersonal trust refers specifically to trust between human individuals, such as colleagues, supervisors, or partners. It has been defined as the expectation or belief that another person will act in ways that are beneficial, or at least not harmful, in situations of vulnerability [82]. Interpersonal trust encompasses both trusting beliefs (the perception that another person possesses competence, integrity, and benevolence) and trusting intentions (the willingness to rely on that person despite the risks involved)

[66]. Together, these components capture both the evaluative judgments people make about others and their readiness to act on those judgments.

A complementary perspective comes from McAllister's [65] influential work, which distinguishes between two dimensions of interpersonal trust in organizational contexts: cognition-based trust and affect-based trust. Cognition-based trust is grounded in rational assessments of another's competence, reliability, and professionalism. In contrast, affect-based trust is rooted in emotional bonds, care, and genuine concern—trust. McAllister's study found that these two forms of trust develop independently yet simultaneously within workplace relationships and carry distinct antecedents and consequences.

Trust is inherently relational and context-dependent. To trust requires a trustor, a trustee, and a situation involving risk or uncertainty [6, 64]. Perceptions of trustworthiness evolve as the trustor gathers more information about the trustee, meaning that trust can develop, erode, or be repaired over time [47, 67].

2.2.2. Trust in Human–AI Collaboration

As robots and AI systems increasingly move from being seen as tools to being regarded as teammates, understanding how trust forms and develops in these relationships has become crucial. As outlined in Section 2.2.1, trust can be conceptualized in terms of beliefs, intentions, and actions [64, 58]. This framework has been applied both to trust in humans and to trust in automation, highlighting three central components: perceptions of trustworthiness, the willingness to accept vulnerability, and the enactment of trust through behavior [64, 58]. In the context of human–AI and human–robot interaction, people similarly form beliefs about an agent's trustworthiness, decide whether they are willing to rely on it under uncertainty, and act on this willingness through behaviors such as reliance or delegation. Although this structure closely parallels human trust, the specific factors that shape beliefs, intentions, and actions differ in important ways when the teammate is artificial rather than human.

The media-equation hypothesis offers one explanation for these similarities, suggesting that people apply the same social rules to computers and robots as they do to other humans [70]. From this perspective, trust in robots should develop according to the same principles as trust in humans. Indeed, studies have shown that people often treat robots like human partners, particularly when the robot exhibits anthropomorphic cues such as human-like faces, speech, or empathy. For example, expressions of empathy or even simple small talk can increase trust in robots, echoing the social dynamics of human–human interaction [91, 76]. Supporting this view, Alarcon and colleagues found that participants reduced their trust in both human and robot teammates equally after experiencing a trust violation, aligning with the CASA paradigm [5]. These findings indicate that people apply human trust frameworks to robots, especially when their design includes humanlike features that make them appear as social partners.

At the same time, trust in robots is not simply a copy of trust in humans. The unique-agent hypothesis argues that human–AI trust has distinctive features that cannot be reduced to human trust processes. For instance, people approach AI teammates with a different set of expectations: whereas human fallibility is accepted as natural, automation is often expected to perform flawlessly. Mistakes made by robots or AI are therefore judged more harshly than those made by humans [28, 68]. Trust in robots is also characterized by heightened uncertainty, since people often struggle to evaluate an AI agent's competence, benevolence, or integrity with the same confidence they bring to human teammates [40, 86]. Moreover, people are prone to monitoring robots more closely than humans, particularly after an error, reflecting the distinct dynamics of trust calibration in HRI [63].

A further distinction stems from the embodiment of robots. Unlike traditional forms of automation, robots share physical space with humans and interact through gestures, touch, and other nonverbal cues. This physical presence introduces affective dimensions into trust, making relational cues more salient. As discussed in Section 2.1.3, anthropomorphic features can support trust by making robots appear more familiar and human-like, thereby easing their acceptance as teammates [44, 85]. However, such design choices also pose risks: anthropomorphism may lead people to attribute unwarranted human qualities to robots and to base trust on superficial appearance rather than demonstrated reliability, which can result in misplaced reliance [21].

In sum, trust in human–AI and human–robot teams shares the same underlying structure as human trust frameworks—rooted in beliefs, intentions, and actions (see Section 2.2)—yet the processes that shape these components are distinct. The media-equation hypothesis highlights the social tendencies that make humans treat robots like people, while the unique-agent hypothesis emphasizes the biases, expectations, and uncertainties that differentiate trust in artificial partners from trust in humans. Taken together, these perspectives suggest that while trust models provide a useful foundation, additional considerations such as embodiment, anthropomorphism, and heightened expectations of reliability must be accounted for when examining trust in AI and robotic teammates.

2.2.3. Trust Violation

Trust violations occur when expectations about another's behavior are not met, or when actions are inconsistent with one's values [12]. Within the broader trust process (see Section 2.2.1), violations mark a critical point where trusting beliefs and intentions are challenged. A violation often leads to the reassessment of the trustee's ability, benevolence, or integrity, which in turn reshapes the trustor's willingness to remain vulnerable. Trust violations are therefore central to understanding both the fragility and resilience of trust relationships.

Trust Violations in Human–Human Interaction

In interpersonal contexts, trust violations are common but not necessarily catastrophic. Because humans are generally seen as fallible, mistakes are often expected and sometimes forgiven. Violations can stem from performance-based errors such as incompetence or forgetfulness, or from non-performance-based behaviors such as deception or neglect. The perceived intent behind the violation is important: research shows that people are more willing to forgive honest mistakes, than intentional acts of betrayal [89]. The human schema for fallibility, which acknowledges that fatigue, distraction, or misjudgment are normal, means that violations are typically interpreted within a framework of human imperfection [27].

Trust Violations in Human–AI and Human–Robot Interaction

Violations in human–AI and human–robot contexts differ in important ways. Automation is often perceived through a "perfect automation schema" in which machines are expected to operate flawlessly [28, 63]. When robots or AI systems commit errors, trust tends to decline more sharply than in comparable human contexts, as these failures challenge the assumption of consistency and reliability. Empirical studies show that after a single error, trust in a robotic teammate can drop dramatically, with many users unwilling to rely on the agent again in future tasks [81]. This phenomenon, sometimes referred to as the "first-error effect," reflects the asymmetry in expectations: whereas human mistakes are normalized, errors by AI or robots are judged against the higher standard of perfection [27].

Increased anthropomorphism has been found to dampen declines in trust following repeated failures [99]. Yet this also introduces the risk of misplaced reliance, as trust may be granted based on appearance rather than demonstrated reliability [21].

Comparing Human and AI or Robot Trust Violations

Overall, both humans and robots are subject to trust violations, but the underlying expectations differ. With humans, violations are often filtered through schemas of fallibility and interpreted in light of intent, which allows for the possibility of repair. With robots and AI, violations strike against expectations of perfection, evoke sharper declines in perceived competence, and are harder to contextualize due to limited transparency in their actions.

2.2.4. Trust Repair

Trust repair refers to the efforts undertaken by a trustee to restore trust after a violation has occurred [54]. Within the broader trust process (see Section 2.2.1), repair represents the attempt to rebuild trusting beliefs, reestablish willingness to be vulnerable, and encourage renewed trusting actions after a breach. Because negative events typically exert stronger effects on reducing trust than positive events do in building it, repair is a particularly critical component of sustaining long-term cooperation [43].

Trust Repair in Human–Human Interaction

In trust contexts, trust repair strategies have been widely studied. Research identifies several common approaches: apologies, denials, explanations, and promises [61, 51, 89]. Apologies seek to express remorse for a transgression [101], denials reject culpability often by attributing blame externally [7], explanations clarify why the violation occurred [26], and promises convey positive intentions about future behavior [89]. These strategies are not equally effective in all situations. For example, apologies and promises are generally effective when the violation is perceived as unintentional or performance-based, whereas intentional acts of deception are more resistant to repair efforts [92].

Trust Repair in Human–AI and Human–Robot Interaction

In human–AI and human–robot contexts, repair dynamics are complicated by different expectations. Because automation is often judged against a "perfect automation schema", repair attempts may be perceived as less effective or less sincere [28, 63]. Users may doubt whether an apology or promise from a machine corresponds to actual improvements in performance, since repair attempts are often seen as scripted and predefined [97]. Nevertheless, research suggests that anthropomorphism can make trust repair strategies more effective. Robots with humanlike features or behaviors are perceived as more benevolent and sincere, which enhances repair outcomes [99, 31].

Comparing Human and AI or Robot Trust Repair

Overall, both humans and artificial agents can attempt to repair trust, but the effectiveness of these efforts depends on the underlying expectations. In human–human interactions, repair strategies are filtered through schemas of fallibility, making apologies or promises more credible. In contrast, repair attempts by robots or AI systems face greater skepticism because machines are assumed to be less capable of genuine remorse or moral intent.

2.3. Research Gap and Rationale

The existing literature highlights the central role of trust in both human–human and human–robot teamwork. Yet, a gap remains in understanding how trust violations and subsequent repair unfold when robots serve as teammates. Much of the human–robot teaming literature relies on simulations, pre-recorded videos, or unidirectional communication, which reduces ecological validity [17, 62]. Only a few studies have examined trust dynamics with physically present robots working side-by-side with people in dyadic teamwork. As summarized in Table 2.1, most prior work has relied on simulated agents or video-based interactions, with Salem et al standing out as the only study that implemented a real robot in a co-present teamwork task [85]. Even in that case, the focus was limited to errors in compliance rather than the full trajectory of trust development. No existing work has examined trust formation, violation, and repair together within the same study, despite repeated calls for such comprehensive designs in embodied HRI research [4, 5, 62, 38].

Addressing this gap, the present study investigates how trust develops and evolves across repeated interactions with either a human or a physically present robot teammate. By introducing an initial round to capture baseline trust formation, followed by a competence-based mistake (trust violation) and a subsequent repair attempt, the study provides a systematic test of the full trust trajectory. By directly contrasting human and robot teammates under matched, embodied conditions, this study seeks to clarify whether trust unfolds differently across human–human and human–robot teamwork.

Table 2.1: Overview of experimental HRI trust studies, mapped to role, task, and whether they examined trust formation, violation, and repair. The table highlights that most studies relied on simulated or video-based setups, with only Salem et al. (2015) employing a physically present robot, and no study addressing all three trust processes together.

Study	**Role of the Robot**	**Task**	**Formation**	**Violation**	**Repair**
Alarcon et al. (2021) [5]	Partner/teammate (simulated anthropomorphic robot)	Incentivized trust / collaborative game with partner returns	✓	✓	✗
Alarcon et al. (2022) [3]	Partner/teammate (NAO, co-present but video simulated) vs. human partner	Collaborative gameplay; ability vs. benevolence manipulation	✓	✓	✗
Alarcon et al. (2023) [4]	Partner/teammate (simulated human vs. simulated robot)	Repeated interaction / trust game with feedback; HH vs. HR comparison	✓	✓	✗
Sanders et al. (2019) [87]	Advisor/assistant (simulated HRI choice agent)	Use–choice paradigm linking baseline trust to reliance	✓	✗	✗
Esterwood & Robert (2021) [31]	Collaborator (anthropomorphic vs. mechanoid robot; online)	Box sorting/picking; test apology, denial, explanation, promise	✗	✓	✓
Robinette et al. (2015) [81]	Physical evacuation guidance robot (co-present)	Emergency building navigation and guidance	✗	✓	✓
Christoforakos et al. (2021) [18]	Social robot (Pepper; video-based)	Impression/anticipatory trust via competence and warmth cues	✓	✗	✗
Onnasch & Hildebrandt (2021/22) [74]	Industrial robot (physical; anthropomorphic vs. low-anthro design)	Industrial HRI collaboration; trust and attention under failures	✓	✓	✗
Lyons et al. (2023) [62]	Robot teammate (video; USAR context)	Unexpected robot behavior; explanations to mitigate loss	✗	✓	✓
Georganta & Ulfert (2024) [37]	New AI teammate vs. new human teammate (no embodiment; online scenario)	Hypothetical team task allocation & change scenario; measure trustworthiness, cognitive & affective interpersonal trust, trust actions	✓	✗	✗
Salem et al. (2015) [85]	Home companion robot (physically present, non-humanoid)	Domestic scenario in a smart home; correct vs. faulty modes; compliance with unusual requests	✗\✓	✗\✓	✗
Esterwood & Robert Jr.(2023) [32]	Robot coworker (virtual; simulated teammate)	Warehouse box sorting task with repeated interactions; three competence errors and repair strategies	✗	✓	✓
Zhang et al. (2023) [111]	NAO healthcare assistant (video-based, non–co-present)	Video HRI with prescription-information scenario; logic/semantic/syntax failure types and repair strategies	✗	✓	✓

3

Task Design

As outlined in the background section, trust is central to effective human-AI teamwork, requiring interdependence, communication, and a willingness to rely on the teammate despite vulnerability [82, 107]. While prior studies have examined aspects of trust in human-robot interaction, most have relied on simulations, pre-recorded videos, or unidirectional communication as seen in Table 2.1. In addition, little research has investigated the full trajectory of trust development, from formation to violation to repair, within the same repeated interaction task [4, 5, 38].

To address this gap, we designed a task that allows trust to be established, deliberately violated, and subsequently repaired across multiple rounds of interaction with either a human or a physically present robot teammate. The task is structured to remain engaging across rounds, to require interdependence between participant and teammate, and to emphasize embodiment by involving a physically present robot partner rather than a simulated or virtual one. In the following, we present the rationale behind our task design, introduce the cooperative building task, discuss the selection of the robot, and describe how trust violations and repair were implemented.

3.1. Task Requirements

Trust is a dynamic process that develops, can be violated, and may be repaired over time. To study this trajectory in a controlled yet engaging way, the task must ensure that reliance on the teammate is unavoidable, that mistakes are noticeable, and that interaction remains stimulating across rounds. From these considerations, we identified five key requirements that guided our task design.

Embodiment and physical co-presence.

While many studies on human–AI trust have relied on virtual agents or screen-based simulations, embodiment adds another dimension to collaboration. Embodied agents share physical space with humans, using gestures, gaze, and movement to direct attention and convey intent, which can influence how trust is formed and maintained [85, 5]. In teamwork contexts that involve shared manipulation of objects or coordinated actions, collaborators are usually physically co-present. Empirical research shows that embodied agents are often perceived as more trustworthy and engaging than disembodied conversational interfaces, as their physical presence enables more natural social cues and smoother coordination in task-based interaction [79]. Therefore, this study employs embodied, physically present teammates (human or robot) to more accurately capture the trust dynamics that emerge during co-located collaboration.

Multiple rounds capturing trust dynamics.

Trust is a dynamic process that can develop, erode, and be repaired over time [47, 67]. To capture these processes, the task must include at least three consecutive interactions corresponding to (a) trust formation, (b) trust violation, and (c) trust repair. Examining these phases within one continuous design is essential because trust repair cannot occur without a preceding violation, and both depend on prior trust formation. Moreover, since individuals differ in their initial propensity to trust, focusing on changes in trust across phases provides a more meaningful basis for comparison than single-point measurements. This temporal approach therefore enables the study of how trust evolves and differs between human and robot teammates.

Stimulation and variability.
Human activity is rarely repetitive; even games and cooperative tasks evolve through variation. For engagement and ecological validity, the task must be stimulating across repeated rounds, with variation introduced while retaining the same underlying structure [13, 23].

High interdependence.
Trust requires a situation in which relying on the teammate involves some level of vulnerability [82, 107]. This vulnerability arises when participants cannot fully succeed on their own, or when working alone is slower and less efficient, making it beneficial to depend on the teammate's actions to achieve a shared goal. If a task allows independent completion, even at a slower pace, participants can choose to minimize reliance on their partner, reducing the need for trust. Therefore, the task should enforce strong interdependence, ensuring that progress and success depend on coordinated contributions from both teammates. Such interdependence makes mistakes noticeable, creating the conditions under which trust can develop, be violated, and repaired [96].

Bidirectional communication.
Studies have shown that unidirectional communication limits the naturalness of collaboration. Bidirectional interaction, in contrast, allows both teammates to exchange information, respond to each other's input, and coordinate actions more fluidly, which is essential for human–AI teamwork [17]. The task should therefore enable reciprocal communication, ensuring that both the participant and the teammate can actively influence the interaction.

3.2. Task Setup

3.2.1. Task Overview

Our central research question is:

How does a teammate's identity (human vs. robot) affect the development of trust across mistake and recovery phases in collaboration?

To address this, we adapted a cooperative building game designed to maximize interdependence (see Figure 3.1). In each round, the participant reconstructs a target figure based solely on instructions from their teammate, who is either a human confederate or a robot. The participant cannot see the reference figure, while the teammate cannot see the participant's construction. This setup ensures mutual dependence and prevents independent completion of the task.

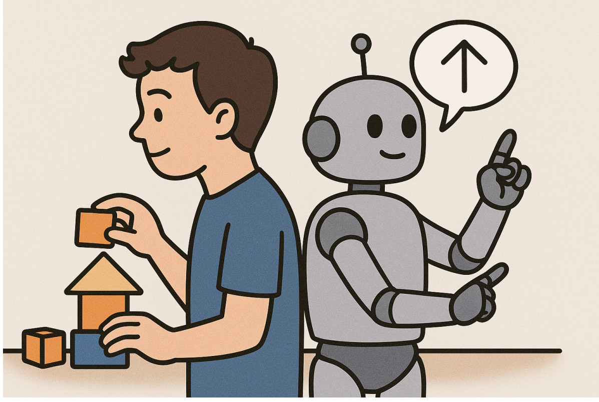


Figure 3.1: Illustration of the experimental task, in which a human participant builds a block structure while receiving guidance from a robot teammate. (AI-generated image)

The task unfolds across three rounds:

- **Round 1 – Trust Formation.** The teammate provides accurate instructions, allowing trust to form through successful collaboration.

- **Round 2 – Trust Violation.** The teammate provides one incorrect instruction as part of the experimental design, creating a competence-based error from the participant's perspective. The violation is followed by a structured reflection phase.
- **Round 3 – Trust Recovery.** Before the round begins, the teammate delivers a combined apology, explanation, and promise to address the previous mistake, representing the repair strategy. The teammate returns to accurate instructions, enabling observation of whether and how trust is restored.

Each round involved a different target figure to keep the task engaging and so that participants did not know in advance what the figure should look like. Feedback on correctness was only provided after each round, ensuring that participants relied on their teammate rather than self-correcting during the task. As shown in Figure 3.2, each round followed the same sequence: greetings and a readiness check, block collection, instruction, and construction. After the task was completed, the built figure was compared to the target model, followed by a reflection phase and a questionnaire to capture trust dynamics.

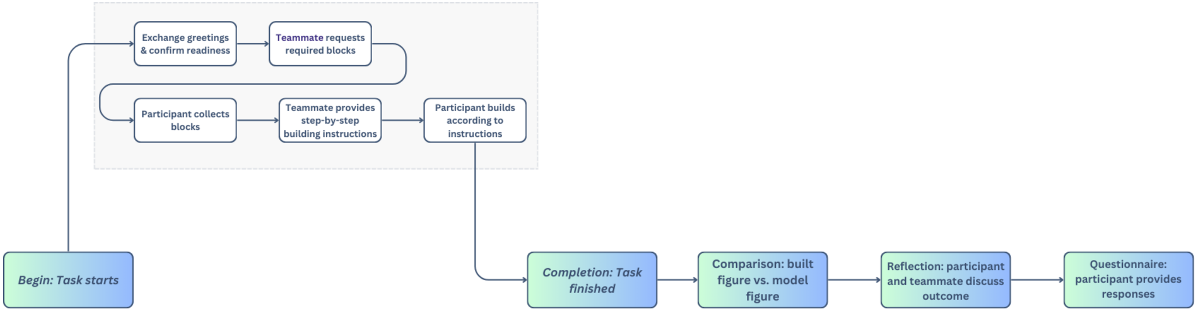

Exchange greetings & confirm readiness
Teammate requests required blocks
Participant collects blocks
Teammate provides step-by-step building instructions
Participant builds according to instructions
Begin: Task starts
Completion: Task finished
Comparison: built figure vs. model figure
Reflection: participant and teammate discuss outcome
Questionnaire: participant provides responses

Figure 3.2: Flow of activities within each round of the building task.

In the second round, the reflection phase included a scripted statement from the teammate:

"I think I misread the figure and gave you the wrong instructions, which is why we didn't complete the task perfectly."

This confession highlighted the source of the error while making it clear that the participant was not responsible. Importantly, the teammate did not yet apologize, as immediate repair would interfere with measurement of the trust violation. Only after participants had reflected and completed the questionnaire did the teammate apply the repair strategy described in Section 3.3.2.

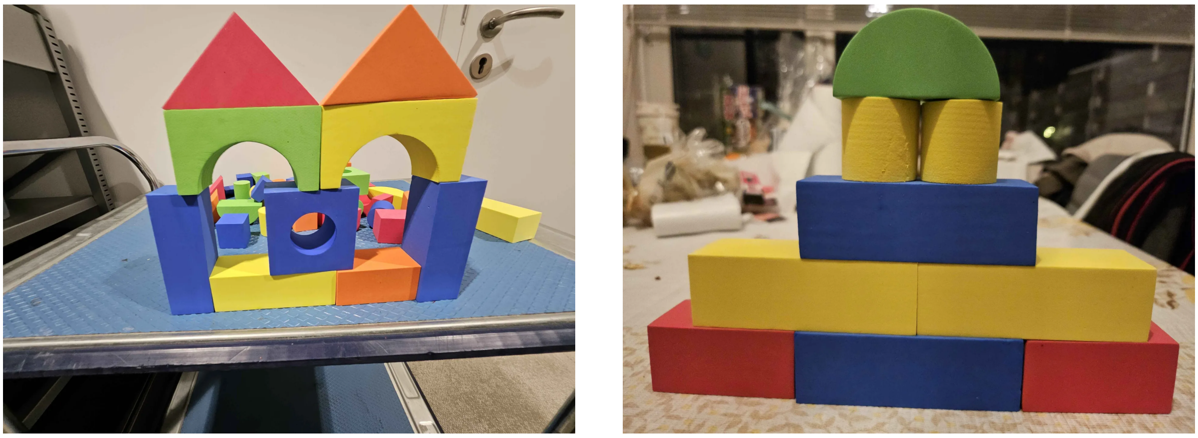
3.2.2. Script Development

To ensure that participants built the intended figures correctly in Rounds 1 and 3, and incorrectly in Round 2 (where a deliberate error was introduced), two requirements shaped the script design:

1. Instructions needed to be clear and easy to follow.
2. Human error had to be minimized by providing sufficiently detailed guidance, reducing the likelihood of unintended deviations.

Based on these requirements, three target figures were created (see Figures 3.3 and 3.4). For Round 2, participants were guided by a deliberately misleading instruction, which led them to construct the incorrect figure shown in Figure 3.4a, while the correct target is displayed in Figure 3.4b for comparison.

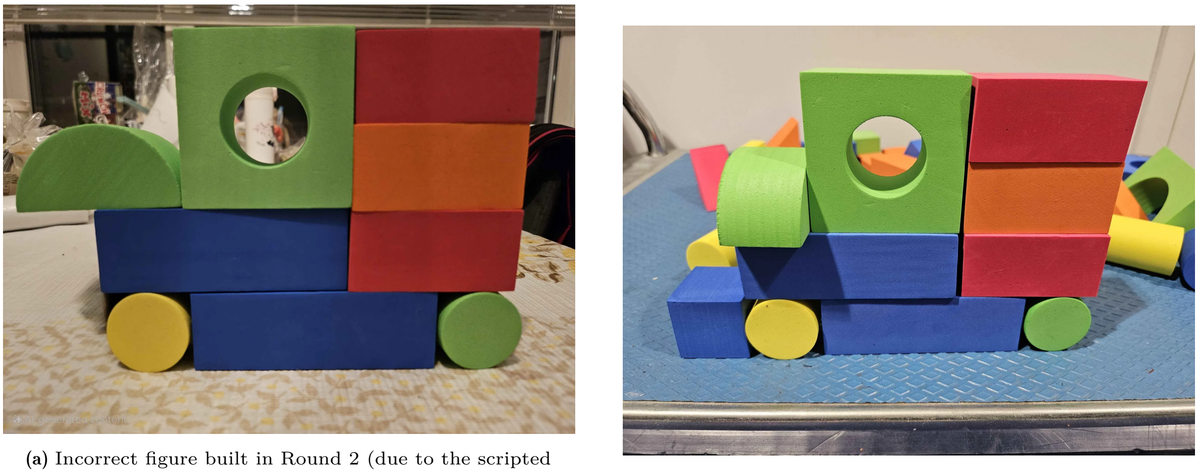
To validate the clarity of the scripts, a pilot study was conducted with four volunteers. Each participant built the figures step by step using the provided instructions. At the end of each layer, we paused to verify whether the construction matched the intended design and gathered feedback on the clarity of the instructions. Only the first participant built the figure incorrectly, after which the script was revised based on the feedback. Following this revision, the remaining three participants completed all figures without difficulty, indicating that the scripts were clear and precise enough to consistently produce the intended designs. An excerpt from the first script is shown below, while the complete set is included in Appendix A.



(a) Round 1

(b) Round 3

Figure 3.3: Figures participants built during the correct rounds.



(a) Incorrect figure built in Round 2 (due to the scripted error). One blue block is missing, and a block that should be placed vertically has been positioned horizontally.

(b) Reference figure of the correct design, which participants were prevented from building.

Figure 3.4: Figures for Round 2. The left figure shows the outcome of the deliberate error, while the right figure shows the intended correct target.

Excerpt of Script 1

Teammate's Instructions

First, let's get the pieces ready. Please collect these blocks for me:

- 2 long blue rectangular blocks
- 1 long yellow rectangular block
- 1 medium orange rectangular block
- 1 blue cube with a hole in the middle
- 1 green rectangular block with a half-arch cutout
- 1 yellow rectangular block with a half-arch cutout
- 1 big pink triangle
- 1 big orange triangle

Teammate: Do you have all of these pieces with you?

[**Breakpoint**]

Teammate: Perfect! Let's start building.

- Place the long yellow rectangular block horizontally on the surface.
- Place the medium orange rectangular block directly to the right of the yellow one.
- Stand the two long blue rectangular blocks upright next to each side, like tall pillars.
- Place the blue cube with the hole on top, between the pillars, hole facing you.

3.3. Design of Trust Violation and Repair

To systematically study how trust develops, deteriorates, and recovers in human–robot collaboration, we designed both a controlled trust violation and a structured repair strategy. The violation was implemented as a competence-based error, while the repair combined an apology, an explanation, and a promise. These elements worked together to make the manipulation clear and establish a consistent basis for comparing human and robot teammates.

3.3.1. Trust Violation

In our study, the trust violation is implemented as a competence-based error, where the teammate provides an incorrect instruction during the building task. We chose this type of violation for two main reasons.

First, competence has been shown to be a strong predictor of trust. Meta-analytic findings indicate that violations of perceived ability have a greater impact on trust than violations of benevolence or integrity [20]. In other words, when a teammate demonstrates a lack of competence, trust declines more sharply than when the same teammate is perceived as less benevolent or less principled. Competence therefore represents the most impactful dimension for manipulating trust in a teamwork setting.

Second, competence-based errors are particularly relevant in human–robot interaction. Robots are often evaluated through a "perfect automation schema," in which people expect near flawless performance [28, 63]. As a result, mistakes attributed to competence are judged more harshly for robots than for human teammates, making them a critical test case for comparing trust trajectories across human and robot partners.

Finally, competence violations in a cooperative building task are transparent and easy for participants to detect. This clarity ensures that the violation is attributed to the teammate's error rather than to task complexity or participant misunderstanding. Such unambiguous violations are essential for studying trust repair, since they provide a clear baseline for assessing whether and how trust can be restored afterward.

3.3.2. Trust Repair Strategy

To address the trust violation, we designed a repair strategy that combines three elements: an apology, an explanation, and a promise. The scripted message used in both human and robot conditions is:

"I'm sorry, I gave you the wrong instruction about the block's direction. I said it should be placed horizontally on top of the other block, but it actually needed to be vertical. I got distracted by the

reference picture from the next round, where the block was shown horizontally, and I also missed the blue block to the left of the yellow one because it blended into the background. I overlooked these details, and I'll be more careful in the next round."

Type of Apology

We used an internal-attribution apology, where the teammate explicitly admits fault and takes responsibility for the mistake. Prior work shows that this form of apology is more effective for competence-based violations than external attribution or denial, since it acknowledges the error and expresses remorse [51, 78, 90]. By stating "I'm sorry, I gave you the wrong instruction," the teammate shows accountability, which can increase perceptions of sincerity and benevolence and encourage forgiveness [61].

Type of Explanation

Rather than offering a vague acknowledgement such as "something went wrong," the teammate provides a specific and concrete account of the error. Explanations that include causal information help participants understand why the violation occurred and reduce uncertainty about the partner's reliability [35, 110]. In our case, the teammate explained that the mistake concerned the block's orientation and that one blue block had been left out.:

"I said it should be placed horizontally on top of the other block, but it actually needed to be vertical. I got distracted by the reference picture from the next round, where the block was shown horizontally and I also missed the blue block to the left of the yellow one because it blended into the background."

This explanation makes the source of the error explicit and connects it directly to the task, giving participants a clear understanding of both what the mistake was and how it occurred.

Type of Promise

The repair message ends with a forward-looking promise: *"I'll be more careful in the next round."* Promises are effective when they set expectations for future performance and provide reassurance that the error will not recur [89, 80]. In our design, this promise is credible because participants know there is a third round in which they can immediately test whether the teammate follows through. By placing the promise in a concrete and near-future context, we increase its believability and its potential to restore willingness to rely on the teammate.

Combined Strategy

By combining an apology, a specific explanation, and a credible promise, the repair strategy addresses the affective, cognitive, and forward-looking dimensions of trust. Apologies appeal to emotions and show care, explanations provide rational sense-making, and promises set expectations for reliable future collaboration. This multifaceted approach aligns with prior findings that integrated strategies are more effective than single responses in repairing trust after competence-based violations in both human–human and human–robot interaction [31, 53].

3.4. Technology

This section describes how the robot was programmed to execute the building scripts and how the interaction was designed to mimic real-life, bi-directional communication between robot and participant.

3.4.1. Choice of the Robot

For this study, we chose to work with an anthropomorphic robot. The key motivation is to heighten the perceived similarity between the robot and the participant. People often categorize robots as outgroup members because of their different appearance and behavior, which can hinder the development of trust and team cohesion [104, 94]. Anthropomorphic design helps to mitigate this barrier, as humanlike features increase perceived similarity and encourage participants to relate to the robot more as a teammate than as a tool [24, 106].

Anthropomorphism also plays an important role in trust formation. Research shows that cues such as facial expressions, gestures, or speech enhance perceptions of competence, benevolence, and sociability, which are dimensions central to trustworthiness [64, 44, 14, 16]. Such design increases the likelihood that participants accept the robot as a partner and follow its guidance. This is especially relevant in our study, which examines not only trust formation but also the dynamics of violation and repair.

By choosing an anthropomorphic robot, we create conditions that mirror potential future scenarios where robots work alongside humans somewhat as peers. In such settings, errors are inevitable, and it is important to understand whether robots that display certain human-like qualities can be forgiven and re-accepted as trustworthy teammates after mistakes, in the same way that humans often are [85, 91, 76].

3.4.2. Why Navel

Among the anthropomorphic robots available at TU Delft, Navel was selected for this study (see Figure 3.5). Navel's design combines humanoid features with a neutral appearance, which supports its role as a collaborative partner, in contrast to the more childlike or service-oriented designs of Nao and Pepper. In addition to its physical design, Navel supports naturalistic, bidirectional communication through speech, gestures, and head movements, which aligns with the requirements of interactive cooperation.

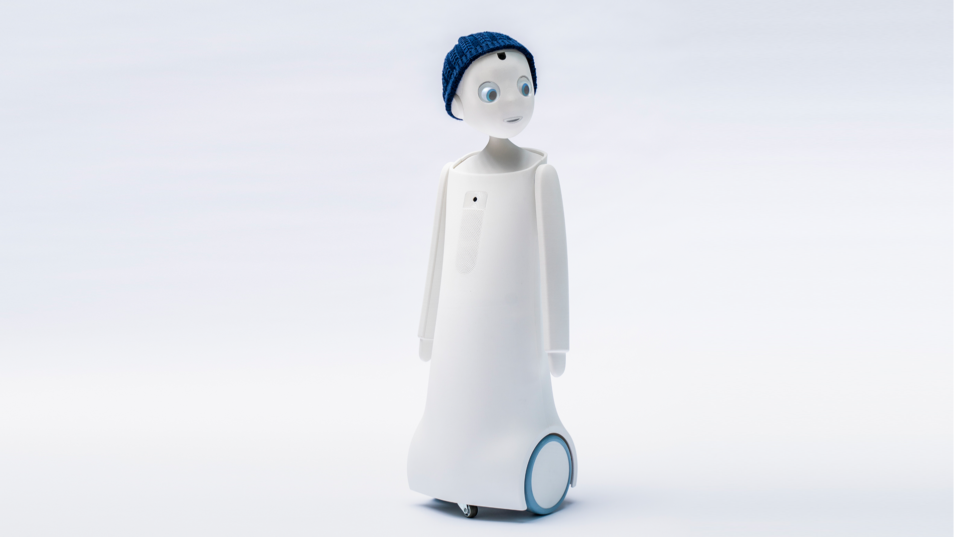


Figure 3.5: Navel, the anthropomorphic robot used in this study.

For the current experiment, not all of Navel's social features were required. Since the robot and participant were seated back-to-back during the task, facial expressions and body movements would not have been visible or meaningful. Instead, the study focused on two of Navel's key capabilities: capturing audio input through its built-in microphone and producing speech via the Navel Python SDK[1]. These functions enabled natural verbal interaction, ensuring that participants could engage with a teammate who felt both approachable and sufficiently human-like, while still being capable of performing the communicative behaviors. Outside of the task phase (e.g., during reflection moments), the robot and participant were positioned face-to-face, allowing participants to see the robot directly. However, since the reflection phase was relatively short, we still did not make use of facial expressions and body movements, as participants were asked to remain seated and were therefore also limited in body movements themselves.

3.4.3. Scripted Interaction

To adapt the building scripts for real-time interaction, additional commands were embedded to manage timing and support bi-directional dialogue. After each instruction, a [**pause**] allowed Navel to wait for up to 10 seconds while keeping its microphone active, giving participants the chance to signal when they were ready to proceed. Dedicated [**Questions**] segments offered space for clarification or repetition. Together, these mechanisms created a more natural rhythm and made the interaction feel less one-sided.

A short excerpt of the adapted script is shown below:

[1] https://doc.navelrobotics.com/

Teammate's Instructions

Teammate: First, let's get the pieces ready. Please collect these blocks for me:

- 2 long blue rectangular blocks [**pause**]
- 1 long yellow rectangular block [**pause**]
- 1 medium orange rectangular block [**pause**]
- 1 blue cube with a hole in the middle [**pause**]
- 1 green rectangular block with a half-arch cutout [**pause**]
- 1 yellow rectangular block with a half-arch cutout [**pause**]
- 1 big pink triangle [**pause**]
- 1 big orange triangle [**pause**]

Teammate: Do you have all of these pieces with you?

[**Questions**]

Teammate: Perfect! Let's start building.

[**Breakpoint**]

- Place the long yellow rectangular block horizontally on the surface. [**pause**]
- Place the medium orange rectangular block directly to the right of the yellow one, so they form a longer line together. [**pause**]
- Stand the two long blue rectangular blocks upright next to each side, like two tall pillars. [**pause**]
- Place the blue cube with the hole in the middle on top, between the pillars, hole facing you. [**pause**]

Table 3.1: Script markers embedded in the instructions and their functions.

Marker	Function
[**pause**]	Inserts a short timed break (10 seconds) during which the robot listens for readiness signals before moving on.
[**Questions**]	Opens a segment for participant inquiries. The robot may repeat or clarify instructions.
[**Breakpoint**]	Marks a transcript marker. At these points, all dialogue is transcribed and tagged with the corresponding round (e.g., [**BREAKPOINT: Round 2**]).

The roles of the embedded markers—[**pause**], [**Questions**], and [**Breakpoint**]—are summarized in Table 3.1. Together, these markers structured the interaction so that participants experienced the robot as responsive rather than purely directive. The [**pause**] and [**Questions**] elements enabled active participation, while [**Breakpoint**] ensured accurate data capture for later analysis.

3.4.4. Large Language Model Integration

A large language model (LLM) was integrated into Navel's dialogue system to enable context-sensitive, bi-directional communication. The LLM did not generate the scripted instructions, which remained fixed, but instead supported natural interaction at designated points. Its functions included normalizing user input, deciding when to advance the script, and answering participant questions without breaking the experimental structure.

LLM Selection

OpenAI's GPT-5 mini[2] was selected for this study. This model offered a balance between speed and conversational quality, making it well-suited for real-time human–robot interaction. Larger models risk higher latency, which can disrupt immersion, whereas GPT-5 mini provided quick, coherent, and context-appropriate responses.

[2] https://platform.openai.com/docs/models/gpt-5-mini

Normalization of User Input

Participants sometimes referred to the robot by name (e.g., “Navel”). Before passing input to the LLM, such references were normalized to second-person pronouns (“you”) so that responses sounded natural and consistent.

> Example Prompt
>
> You are Navel, a friendly robot teammate. Whenever the operator says ”Navel” or ”Navel’s,” treat it as ”you” or ”your.” Respond naturally in the first person.

Decision Making

At pause points, the LLM was asked to classify whether the participant’s utterance signaled readiness to proceed. This enabled Navel to either continue with the next instruction or remain in the current loop.

> Example Prompt
>
> You decide if the latest user utterance indicates they are ready to move on. If yes, reply only with: CONTINUE. If not, reply only with: WAIT.

Response Generation

During designated question phases, the LLM generated concise, context-specific answers. These responses were restricted to the most recent set of steps (up to the last breakpoint), preventing the robot from revealing future instructions. Responses were designed to be short, warm, and natural, and lists were avoided unless explicitly requested.

> Example Prompt
>
> You are Navel, a friendly robot teammate. Answer the participant’s question by referring only to the steps recorded so far. Do not reveal or hint at future steps. Speak warmly and concisely, in 1–3 sentences.
> Steps covered so far: [Automatically tracked up to the last breakpoint]
> Avoid lists unless the user explicitly asks for step-by-step repetition.

3.4.5. Pilot Testing the Setup

To ensure that the Navel setup functioned as intended, we conducted a small pilot study with two volunteers on campus. While we had already tested the scripts with other volunteers (see Section 3.2.2), this pilot focused on whether the robot could execute them smoothly when equipped with the additional capabilities for listening, answering, repeating, and pausing. The main goal was to verify whether the interaction felt natural and whether any technical or design issues emerged when Navel acted as a teammate.

During the pilot, participants sat across from Navel and followed its instructions, without being able to see each other. We monitored the session simultaneously to check whether the speech recognition system was working properly. One issue arose when Navel asked whether the participant had any questions. When the participant answered, “No, I don’t,” the LLM did not interpret this as readiness to proceed. To address this, we refined the decision-making prompt to make such answers explicit indicators of progression:

> Revised Prompt
>
> You decide if the latest user utterance indicates they are ready to move to the next script step. Examples include signaling readiness, completion, wanting to continue, moving on, or saying they have no questions, or replying “no” when asked if they have any questions.

With this adjustment, Navel correctly recognized that a negative response to the question prompt should advance the script.

Aside from this issue, the interaction ran smoothly. Participants occasionally needed to raise their voices because Navel's microphone was somewhat insensitive, but overall the instructions were clear, the timing of pauses felt appropriate, and the option to shorten waiting times allowed for a more flexible rhythm. Each round lasted approximately 5–7 minutes, which was within the desired range.

One participant remarked that the figures felt slightly too simple, noting that she instinctively assumed symmetry while building. However, she also emphasized that the clarity of the scripts made it difficult to make mistakes as long as they were followed correctly. This trade-off was considered acceptable, as minimizing unintentional human errors was more important than increasing task difficulty.

Overall, the pilot confirmed that the setup was functional, the scripts were robust, and no major issues would interfere with the main experiment.

3.5. Anticipated Risks and Mitigation

Three main risks were identified for this task design:

1. **Bias in the human teammate condition.** If we acted as the human teammate, prior familiarity with participants (e.g., friends) could bias initial levels of trust and influence responses throughout the experiment. *Mitigation:* To reduce this risk, we invited either an available friend outside of overlapping social groups or my sister to serve as the teammate. This separation helped minimize interpersonal bias and encouraged more objective responses on the trust questionnaires.
2. **Robot malfunction.** Navel could malfunction unexpectedly during the task. *Mitigation:* We remained in the room at all times (after confirming the participant was comfortable with my presence) and was prepared to intervene if needed. If necessary, we could temporarily adopt a Wizard-of-Oz approach to ensure the task continued smoothly.
3. **Participant error despite scripted instructions.** Even with detailed scripts, participants might misinterpret steps and construct the figure incorrectly. *Mitigation:* To minimize this risk, we conducted a pilot study to validate the clarity of the scripts (see Section 3.2.2). Nevertheless, mistakes may still occur. Since the scripts are designed to align with specific stages of trust development, any participant data compromised by such errors will be excluded from the final analysis. We will keep a record of how often such errors occur and report this in the results, ensuring transparency.

Overall, these risks were carefully considered and addressed in advance to safeguard both the smooth execution of the experiment and the reliability of the collected data.

4

User Study

This chapter presents the user study conducted to investigate how trust develops, is violated, and can be repaired in human–robot collaboration. The study examines how a teammate's identity (human or robot) influences interpersonal trust over time, particularly across phases of mistake and recovery.

The chapter first describes the experimental design, detailing the mixed factorial structure used to examine changes in trust across collaboration phases. It then outlines participant recruitment and demographics, the measurement instruments employed to assess interpersonal trust, and the procedure followed during the study sessions. Ethical considerations relevant to the conduct of the experiment are discussed to ensure transparency and research integrity.

Finally, the chapter introduces the Bayesian modeling strategy used to analyze the collected data. This section motivates the choice of a Bayesian analytical framework, explains model specification and comparison procedures, and justifies the selection of the final model used in subsequent analyses. Together, these sections provide the methodological foundation for the results presented in the following chapter.

4.1. Experimental Design

To answer the research question, *"How does a teammate's identity (human vs. robot) affect the development of trust across mistake and recovery phases in collaboration?"*, a 2 × 3 mixed factorial design was used. The design includes one between-subjects factor (teammate type) and one within-subjects factor (time).

Trust is measured once per round across three rounds, immediately after the task and reflection phase. Each measurement corresponds to a specific point in the collaboration process:

- **t_1 (Baseline Trust)**: After the first round, capturing participants' initial trust when cooperating with their assigned teammate.
- **t_2 (Post-Mistake Trust)**: After the second round, following a predefined mistake made by the teammate. During the reflection phase, participants are made aware that the error was caused by the teammate rather than themselves (see Section 4.4).
- **t_3 (Post-Recovery Trust)**: After the final round, following the teammate's trust-repair attempt. The recovery phase involves an apology, an explanation of the previous error, and a promise to be more careful before starting the last round. This promise is then upheld by the teammate by providing complete and accurate instructions for building the figure in the final round, without making any further mistakes.

The dependent variable in this study is interpersonal trust, which represents the participants' trust toward their teammate throughout the collaboration process.

4.2. Participants

A total of 40 participants took part in the study, having been recruited through personal connections and extended personal networks. Two inclusion criteria were applied: participants were required to (1) not be colorblind and (2) have sufficient English proficiency. These criteria were necessary because the task involved assembling figures using blocks of different colors and sizes, which required participants to accurately distinguish colors and understand verbal instructions provided in English.

The sample was skewed toward female participants (29 female, 11 male). All participants were between 18 and 27 years old, and most were currently enrolled in an educational program, ranging from vocational education (MBO) to university level.

Participants were assigned to one of the two experimental groups with the aim of achieving a balanced distribution in terms of gender and age. Practical considerations, including participants' availability and travel time to the laboratory, were also taken into account during group assignment. All participants received snacks as a token of appreciation at the end of the study, no monetary compensation was provided.

4.3. Measurements

4.3.1. Demographic Pre-Survey

At the beginning of the study, participants completed a brief pre-survey to report their age range and gender. This information was collected to allow for exploratory analysis of potential group differences in trust-related responses. The pre-survey allowed us to explore whether age or gender influenced participants' behavior or responses during the study. The full questionnaire is included in Appendix C.2.

4.3.2. Trust Measurement and Scale Selection

To assess trust, we selected the interpersonal trust scale developed by McKnight and Chervany [66]. Previous studies have commonly used or adapted either McAllister's or Mayer et al.'s models to measure trust in collaborative contexts [64, 65, 37, 6, 62, 111]. While both scales are widely used, their conceptual focus differs. Mayer's model primarily measures trustworthiness, referring to the perceived attributes of the trustee (ability, benevolence, integrity) that influence the trustor's willingness to be vulnerable. However, our interest lies in the participant's own willingness to trust their teammate rather than how trustworthy they perceive them to be.

McAllister's interpersonal trust scale was initially considered because it directly measures trust between individuals [65]. It distinguishes between two components:

- **Cognitive trust:** grounded in perceptions of the other's reliability, competence, and professional integrity.
- **Affective trust:** based on emotional bonds, mutual care, and the belief that the other genuinely values the relationship.

Since our task was highly structured, goal-oriented, and of short duration, affective trust was unlikely to develop meaningfully. Including it would therefore not provide useful variance and might result in uniformly low scores. We sought a framework that emphasizes the cognitive and intentional components of interpersonal trust without relying on long-term emotional relationships.

For this reason, we selected McKnight and Chervany's model, which provides a balanced conceptualization of interpersonal trust [66]. Their model defines interpersonal trust as the trusting beliefs and trusting intentions one holds toward a specific other person, such as a colleague or supervisor. It integrates both the belief-based and intention-based aspects of trust that are central to interpersonal collaboration. Specifically, the scale includes:

- **Trusting Belief – Benevolence:** belief that the teammate cares about one's well-being.
- **Trusting Belief – Competence:** belief that the teammate is capable and skilled.
- **Trusting Intention:** willingness to depend on or be vulnerable to the teammate.

This structure aligns closely with our experiment's focus on task-oriented cooperation, where participants rely on their teammate's instructions and performance rather than emotional connection. In short, McKnight and Chervany's operationalization captures how willing participants are to rely on another individual, grounded in beliefs about that individual's motives and competence. The questionnaire is provided in Appendix C.3.

4.3.3. Supplementary Qualitative Measures

After each round, participants were invited to provide open-ended feedback about the task and collaboration, allowing them to share additional thoughts or feelings not expressed during the reflection phase. Both the transcribed reflections and the written feedback were included in the qualitative analysis. Although we initially considered adding a direct question about perceived changes in trust after the teammate's mistake, we decided against it to avoid revealing the experimental manipulation before the debriefing. These qualitative responses were later analyzed to complement the quantitative trust measures.

4.4. Procedure

4.4.1. Experimental Setup

To ensure consistency across sessions, all experiments were conducted in a quiet room with three people present: the participant, the teammate (either the human confederate or the robot Navel), and the experimenter.

Since Navel has a short delay before responding to questions (as he first listens, processes, and then answers), the human confederate was instructed to pause briefly before answering as well. In addition, because Navel was programmed to respond only during specific question rounds, the same restriction was applied to the human confederate. This ensured that both conditions were comparable, with the human teammate only allowed to answer questions at the same predefined breakpoints specified in the scripts (see Appendix A).

4.4.2. Participant Instructions

Before the experiment began, participants were instructed to ask their questions in one complete sentence without taking short pauses. This rule was important because Navel stops listening once a gap or pause occurs and immediately begins formulating a response based on what he has already heard. To ensure fairness between conditions, the human confederate followed the same rule and only responded to the first continuous question spoken by the participant before any break or hesitation.

Participants then listened to the opening speech (Appendix B), which explained that they would complete three rounds of a building task based on their teammate's verbal instructions. The speech also reminded them of the specific communication rules.

After the opening speech, participants completed the informed consent form, as shown in Appendix C.1.

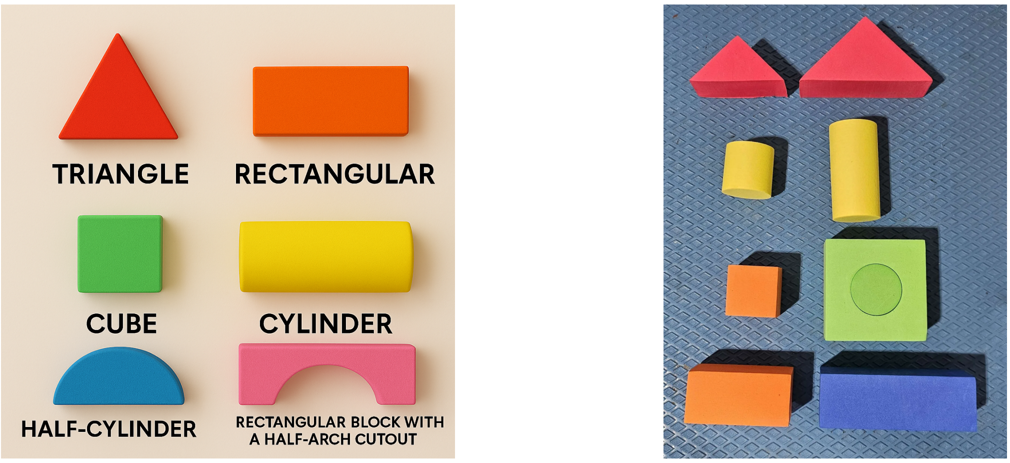
Before starting the first round, participants were shown two figures to familiarize themselves with the materials used in the task. Figure 4.1a displays the different geometric shapes included in the experiment, while Figure 4.1b illustrates the two available size variations for each shape.

This step helped establish a shared understanding of the shapes and terminology that would be used throughout the task, minimizing the risk of misunderstanding during verbal instructions.

4.4.3. Interaction, Reflection, and Debriefing Phases

After each round, participants were given the option to have a short reflection with their teammate. These reflections were informal and intended to provide an opportunity for brief exchanges, such as expressing gratitude or encouragement, rather than structured discussions. Participants also completed a questionnaire measuring their interpersonal trust after each round, as shown in Appendix C.3, and could optionally provide brief written feedback about their experience.

The reflection after the second round was scripted for both conditions. At this stage, the human confederate or the robot teammate acknowledged their mistake (see Section 3.2.1). This clarification ensured that participants understood that the error originated from the teammate rather than themselves. Before the third round, a scripted trust recovery strategy was introduced, as described in Section 3.3.2.



(a) Shapes of the building blocks used in the experiment.

(b) Sizes of the building blocks available in the experiment.

Figure 4.1: Overview of the shapes and sizes of the building blocks used in the experiment. Participants were shown these figures before the task to ensure common ground and avoid misunderstandings.

For Navel, these scripted reflections were delivered using a Wizard-of-Oz approach, where the experimenter temporarily controlled the robot's speech to ensure consistent delivery. This decision was made because the reflections typically lasted less than two minutes, and researcher control prevented unintended or inconsistent robot responses that might disrupt the interaction. In the human condition, the same two scripted messages were used (one after Round 2 and one before Round 3), and the confederate was instructed to match the participant's tone and level of engagement.

At the end of the final round, the experimenter conducted a debriefing session (Appendix B) to explain that the teammate's mistake in the second round was intentional and part of the experimental design. This ensured full transparency and helped prevent any misunderstandings about the purpose or nature of the study.

4.5. Ethical Considerations

During the opening speech, participants were informed that they should not share any personal information during the tasks. An analysis of all transcripts and reflections confirmed that no participant disclosed personal details. We believe this is because the experiment was highly task-focused, requiring participants' full attention, leaving little motivation or opportunity to share personal information.

Apart from gender and age (recorded as age ranges), no additional personal data were collected in the questionnaires. Only written transcripts were used for analysis, as the conversations were transcribed live during the experiment. No audio recordings were made or stored, ensuring that voice recognition was not possible.

This study was reviewed and approved by the Human Research Ethics Committee (HREC) of Delft University of Technology[1] prior to data collection.

4.6. Bayesian Modeling Strategy

This section describes the statistical modeling framework used to analyze trust development over time. It outlines the rationale for adopting a Bayesian approach, the specification of the candidate models, and the criteria used for model comparison and selection.

4.6.1. Bayesian Analytical Approach

In this thesis a Bayesian analytical framework was chosen over the traditional frequentist approach because it treats uncertainty as a probability distribution over hypotheses rather than as long-run error rates [88]. Frequentist inference evaluates the likelihood of the observed data assuming the null

[1] https://www.tudelft.nl/en/about-tu-delft/strategy/integrity-policy/human-research-ethics

Table 4.1: Interpretive guidelines for Bayes factors comparing H_1 against H_0, reproduced from [36].

Bayes factor (BF_{10})	**Interpretation**
> 30	Very strong evidence for H_1
10–30	Strong evidence for H_1
3–10	Moderate evidence for H_1
1–3	Anecdotal evidence for H_1
$= 1$	No evidence
0.33–1	Anecdotal evidence for H_0
0.10–0.33	Moderate evidence for H_0
0.03–0.10	Strong evidence for H_0
< 0.03	Very strong evidence for H_0

hypothesis is true and then makes a binary decision based on an arbitrarily chosen significance threshold (e.g., $\alpha = 0.05$) [41]. By contrast, Bayesian inference combines prior knowledge with the data likelihood to produce a posterior distribution that directly quantifies uncertainty about model parameters, allowing statements such as ”the probability that the effect exceeds zero is 0.87” [33].

A key advantage of the Bayesian framework is the use of Bayes factors to quantify evidence on a continuous scale. A Bayes factor (BF) is the ratio of the marginal likelihoods of two competing models, typically the alternative hypothesis H_1 relative to the null hypothesis H_0. Values of BF > 1 indicate greater support for H_1, whereas values of BF < 1 indicate greater support for H_0. To support interpretation, conventional guidelines categorize Bayes factors into qualitative strength levels (Table 4.1). These categories provide a graded summary of evidence, rather than requiring a binary decision such as ”significant” versus ”non-significant” [36].

4.6.2. Bayesian Model Specification and Comparison

All Bayesian models were estimated using Stan via the `ulam` interface from the `rethinking` package in R. The `ulam` interface provides a high-level model specification language while compiling models to optimized Stan code internally [15].

Model structure and common components

All candidate models shared a common core structure. Trust ratings were modeled as continuous outcomes and assumed to follow a Gaussian likelihood, with the predicted mean determined by a linear combination of fixed and varying effects. To account for the repeated-measures design, all models included participant-level varying intercepts, allowing each participant to have their own baseline level of trust. This structure accounts for the fact that multiple observations were collected from the same participant, preventing repeated measurements from being treated as independent.

Time was included as a fixed effect with three levels (t_1–t_3) to model overall changes in trust across measurement points. Experimental condition (human vs. robot teammate) was included as a fixed effect, along with a time-by-condition interaction, allowing trust to change differently over time for the two teammate types. This ensured that differences in trust development could be captured even in the base model.

Priors

Weakly informative priors were used throughout. The population-level intercept was centered on the midpoint of the trust scale, reflecting a plausible baseline trust level without strongly constraining the estimates. Fixed-effect coefficients were centered at zero with moderate variance, allowing for meaningful effects while regularizing extreme values. Variance parameters for participant-level intercepts and residual noise were assigned exponential priors, favoring smaller variances while remaining flexible if supported by the data. These choices reflect prior skepticism toward large effects while avoiding overly restrictive assumptions.

Compared models

Four Bayesian models were specified and compared. The models differed only in whether demographic predictors were included:

- **Base model**: Included fixed effects of time, condition, and their interaction, as well as participant-level random intercepts.
- **Gender model**: Extended the base model by adding gender as a fixed effect.
- **Age model**: Extended the base model by including age group as a categorical fixed effect.
- **Gender + age model**: Included both gender and age group as fixed effects in addition to the base structure.

All models shared the same likelihood function, priors for common parameters, and hierarchical structure. In this context, a hierarchical structure means that the model explicitly represents both individual-level variation and population-level effects by nesting repeated observations within participants. This design ensured that differences in model performance could be attributed to the inclusion of demographic predictors rather than to differences in model specification.

Model comparison using WAIC

Models were compared using the Widely Applicable Information Criterion (WAIC). WAIC estimates out-of-sample predictive accuracy by combining the log-likelihood with a penalty for effective model complexity. Lower WAIC values indicate better expected predictive performance, while differences in WAIC should be interpreted relative to their associated standard errors [102]. Unlike traditional likelihood-ratio tests, WAIC is fully Bayesian and suitable for hierarchical models [103].

Because WAIC is computed from posterior samples, results can vary slightly depending on the random seed used to initialize the Markov chains. To ensure stability, a fixed seed was set prior to model estimation so that all models were compared under identical sampling conditions. Minor fluctuations in WAIC across runs are expected, but the relative ordering of models remained stable.

Model selection

Table 4.2 reports WAIC values, standard errors, differences relative to the best-performing model, and model weights. The base model achieved the lowest WAIC, with all extended models showing slightly higher values. Differences in WAIC between models were small and well within one standard error, indicating that adding age, gender, or both did not meaningfully improve predictive performance. Model weights were distributed relatively evenly, further suggesting that the data provided little support for preferring the more complex specifications.

Based on these results, the base model was selected for subsequent analyses. This choice favors a simpler model that retains the core theoretical predictors of interest. Because demographic predictors did not improve predictive accuracy, excluding them avoids unnecessary complexity and reduces the risk of overfitting, while maintaining a clear interpretation of trust dynamics across time and experimental condition.

Table 4.2: Comparison of candidate Bayesian models using WAIC. Lower values indicate better expected out-of-sample predictive performance.

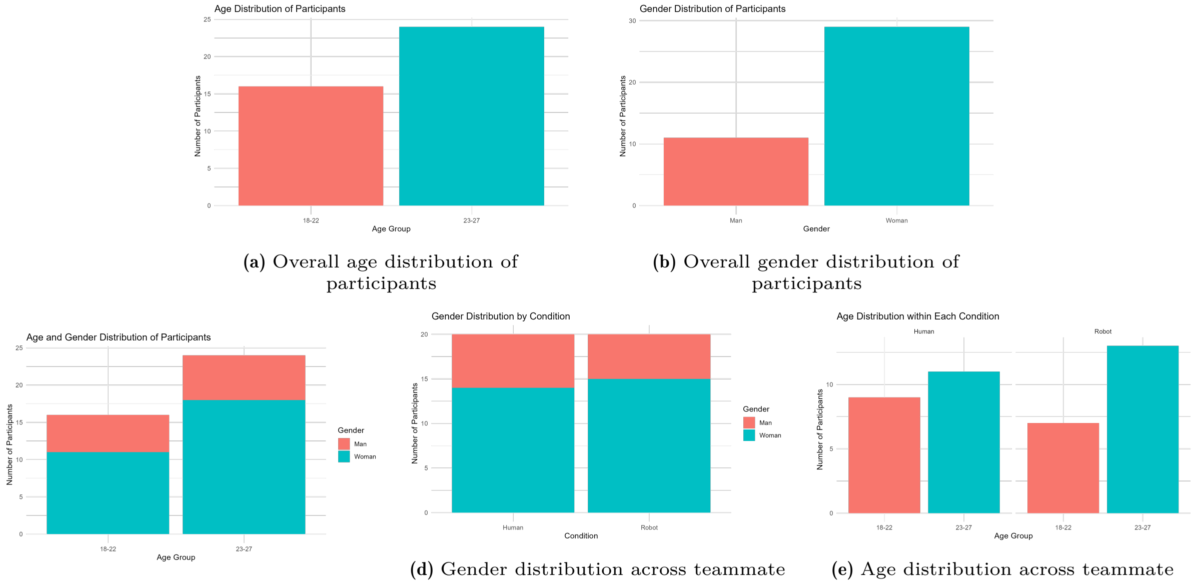
Model	WAIC	SE	ΔWAIC	dSE	pWAIC	Weight
Base	251.2	13.16	0.0	–	31.0	0.28
Gender	251.3	13.21	0.1	0.38	31.1	0.26
Age	251.4	13.18	0.2	0.88	30.9	0.25
Gender + Age	251.8	13.22	0.6	0.97	31.2	0.21

5 Results

This chapter presents the empirical results of the study. It first provides an overview of participant demographics, followed by quantitative analyses of trust development across the formation, violation, and recovery phases for both human and robot teammates. These analyses include descriptive summaries and Bayesian model-based results for overall trust and its underlying dimensions. The chapter concludes with a qualitative analysis of participants' open-ended feedback, offering additional insight into how collaboration and trust were experienced in the two conditions. Interpretation of these findings is reserved for the Discussion chapter.

5.1. Participants

A total of 40 participants took part in the study (Figure 5.1). Participants were assigned evenly across conditions (human-teammate: $n = 20$; robot-teammate: $n = 20$). The sample consisted of 11 men and 29 women (Figure 5.1b). Participants were divided into two age groups: 18–22 years ($n = 16$) and 23–27 years ($n = 24$) (Figure 5.1a). In the 18–22 age group, the sample included 5 men and 11 women, whereas the 23–27 age group included 6 men and 18 women (Figure 5.1c). Across conditions, the gender distribution was comparable (human-teammate: 6 men, 14 women; robot-teammate: 5 men, 15 women) (Figure 5.1d). The age distribution across conditions showed a small difference (human-teammate: 9 aged 18–22 and 11 aged 23–27; robot-teammate: 7 aged 18–22 and 13 aged 23–27) (Figure 5.1e).



(a) Overall age distribution of participants

(b) Overall gender distribution of participants

(c) Gender distribution by age

(d) Gender distribution across teammate conditions

(e) Age distribution across teammate conditions

Figure 5.1: Overview of participant demographics and their distribution across experimental conditions.

5.2. Trust Development

This section presents the results of the trust development analyses across time and conditions. We first provide a descriptive overview of observed trust patterns to offer an intuitive understanding of how trust and its underlying dimensions evolve throughout the task. We then report results from Bayesian statistical models, including posterior estimates, credible intervals, and Bayes factors, to formally assess differences between human and robot teammates and to evaluate the preregistered hypotheses. Finally, we examine whether these patterns differ across trust dimensions by reporting scale-specific posterior results for benevolence, competence, and trusting intentions.

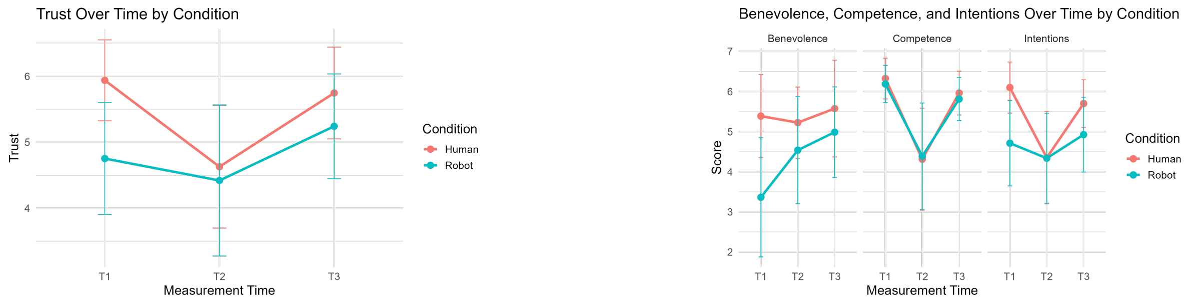
5.2.1. Descriptive Overview of Initial Trust Patterns

We first summarize the observed trust trajectories across measurement points and conditions to provide context for the subsequent Bayesian analyses.

Figure 5.2a presents mean overall trust ratings across three measurement points (t_1–t_3) for the human-teammate and robot-teammate conditions. In both conditions, mean trust decreases from t_1 to t_2 and increases again at t_3. Descriptive means and standard deviations for each measurement point are reported in Table 5.1.

Figure 5.2b provides a descriptive breakdown of trust into the McKnight-based dimensions of benevolence, competence, and intentions. For benevolence, the condition means change across measurement points, with a small decrease from t_1 to t_2 for the human condition and an increase across the same interval for the robot condition; both conditions increase from t_2 to t_3. For competence, both conditions show a decrease from t_1 to t_2 and an increase at t_3, with closely aligned trajectories across conditions. For intentions, both conditions decrease from t_1 to t_2 and increase again at t_3, with the condition means remaining separated across measurement points (Table 5.1).

These descriptive results summarize mean trends and variability across overall trust and the three trust dimensions. The different trajectories across dimensions motivate examining both overall trust and the individual dimensions in the Bayesian analyses that follow.



(a) Overall trust ratings across time (t_1–t_3) for human and robot teammates

(b) Trust dimensions (benevolence, competence, intentions) across time

Figure 5.2: Descriptive overview of trust development over time, shown at both the aggregated level and across individual trust dimensions.

5.2.2. Posterior Estimates from the Selected Base Model

All statistical results reported in this section are based on the base model selected during the model comparison procedure (Section 4.6.2). This model includes fixed effects of time, condition, and their interaction, as well as participant-level varying intercepts.

Figure 5.3 presents posterior predictive trust trajectories for the human and robot teammate conditions across the three measurement points. Points indicate posterior mean predictions, and shaded bands represent 95% credible intervals. For both conditions, trust decreases from t_1 to t_2 following the mistake phase and increases from t_2 to t_3 during the recovery phase. Predicted trust is higher in the human condition at all time points, with partial overlap of credible intervals at t_2 and t_3.

Table 5.2 reports posterior means and 95% credible intervals for key contrasts derived from the selected base model. These contrasts represent either differences in predicted trust between the robot and

Table 5.1: Mean (SD) trust scores for overall trust and the benevolence, competence, and intentions dimensions by teammate condition at each measurement point (t_1–t_3).

Trust Type	**Teammate Identity**	t_1	t_2	t_3
Aggregate Trust	Human	5.94 (0.61)	4.63 (0.93)	5.75 (0.70)
	Robot	4.75 (0.85)	4.42 (1.15)	5.24 (0.79)
Trust Dimensions				
Benevolence Trust	Human	5.39 (1.04)	5.22 (0.89)	5.58 (1.21)
	Robot	3.36 (1.49)	4.54 (1.33)	4.99 (1.13)
Competence Trust	Human	6.32 (0.51)	4.31 (1.27)	5.96 (0.55)
	Robot	6.19 (0.47)	4.39 (1.33)	5.81 (0.54)
Intentions Trust	Human	6.10 (0.64)	4.35 (1.15)	5.70 (0.59)
	Robot	4.71 (1.06)	4.34 (1.12)	4.92 (0.93)

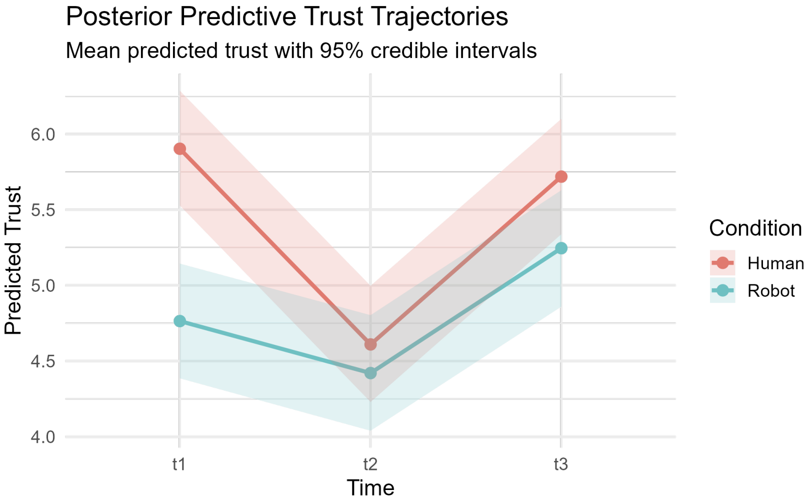


Figure 5.3: Posterior predictive trust trajectories based on the selected base model. Points represent posterior mean predicted trust, and shaded areas indicate 95% credible intervals for each condition.

human conditions at the same measurement point, or changes in predicted trust over time within a single condition. For between-condition contrasts (Robot − Human), negative values indicate lower predicted trust for the robot teammate relative to the human teammate. For within-condition contrasts (e.g., Human: $t_2 - t_1$), negative values indicate decreases in trust and positive values indicate increases. Credible intervals excluding zero indicate high posterior certainty regarding the direction of an effect.

At t_1, the robot–human contrast is negative, indicating lower predicted trust in the robot condition at baseline. At t_2, the estimated difference is smaller and includes zero, and at t_3 the contrast remains negative but also includes zero.

Within the human condition, trust decreases from t_1 to t_2 and increases from t_2 to t_3; both changes credibly differ from zero. In contrast, the overall change from t_1 to t_3 is small and uncertain.

Within the robot condition, the change from t_1 to t_2 is negative but uncertain, followed by a credible increase from t_2 to t_3. Overall, trust shows a positive change from t_1 to t_3 .

Table 5.3 summarizes posterior probabilities and Bayes factors for the three preregistered hypotheses. The results provide very strong evidence for lower initial trust in the robot condition and very strong evidence against the hypothesis that robots experience greater trust loss following a mistake. For trust recovery, the results indicate moderate evidence that trust recovers more weakly for robots than for humans.

Table 5.2: Posterior means and 95% credible intervals for key contrasts derived from the selected base model. Bolded values indicate contrasts for which the credible interval excludes zero.

Aggregate Trust Contrast	Mean	2.5%	97.5%
Robot − Human @ t_1	**-1.14**	**-1.69**	**-0.60**
Robot − Human @ t_2	-0.19	-0.73	0.36
Robot − Human @ t_3	-0.47	-1.02	0.07
Human: $t_2 - t_1$	**-1.29**	**-1.65**	**-0.93**
Human: $t_3 - t_2$	**1.11**	**0.74**	**1.49**
Human: $t_3 - t_1$	-0.18	-0.55	0.19
Robot: $t_2 - t_1$	-0.34	-0.71	0.02
Robot: $t_3 - t_2$	**0.83**	**0.45**	**1.20**
Robot: $t_3 - t_1$	**0.48**	**0.11**	**0.85**

Table 5.3: Posterior probabilities and Bayes factors for the three preregistered hypotheses, computed from posterior samples of the selected base model. Strong evidence is highlighted in bold.

Aggregate Trust Hypothesis	Posterior Prob.	Bayes Factor	Evidence
H(a): Lower initial trust in robots	**100%**	∞	**Very strong for H_1**
H(b): Greater trust loss after robot mistake	**0.03%**	**0.00**	**Very strong for H_0**
H(c): Weaker trust recovery for robots	86.17%	6.23	Moderate for H_1

5.2.3. Scale-Specific Posterior Results

To assess whether trust development differed across dimensions, benevolence, competence, and trusting intentions were analyzed separately using the same base model structure. For each scale, posterior contrasts between conditions and over time are reported, followed by posterior probabilities and Bayes factors for the preregistered hypotheses. Credible intervals excluding zero indicate high posterior certainty, and Bayes factors are interpreted using established guidelines. Figure 5.4 visualizes the posterior predictive trajectories for each trust dimension across time and teammate identity.

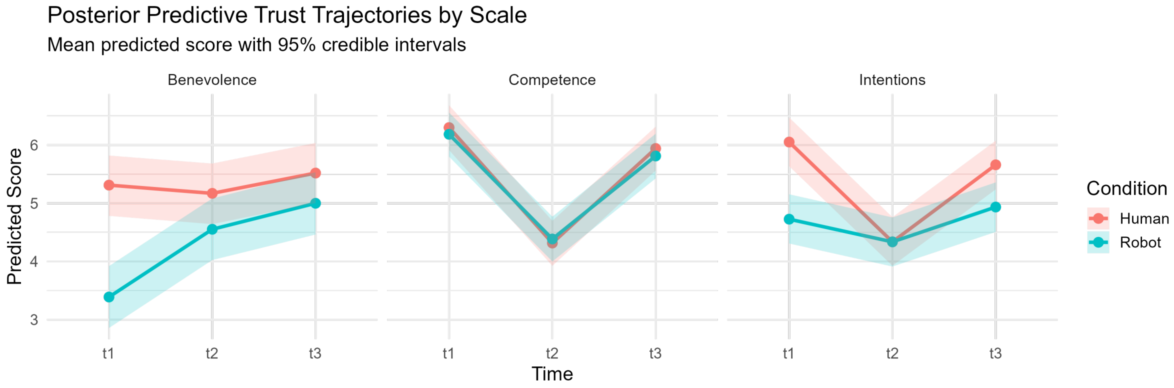


Figure 5.4: Posterior predictive trajectories for benevolence, competence, and trusting intentions. Points represent posterior mean predicted trust, and shaded areas indicate 95% credible intervals for each condition.

Benevolence

Posterior contrasts for the benevolence scale are reported in Table 5.4. At baseline (t_1), the robot–human contrast is credibly negative, indicating lower predicted benevolence in the robot condition. At t_2 and t_3, the between-condition contrasts are smaller and include zero.

Within the human condition, changes in benevolence across time are comparatively small and include zero across contrasts. Within the robot condition, benevolence increases credibly from t_1 to t_2 and from t_1 to t_3. The intermediate change from t_2 to t_3 is smaller and uncertain.

Hypothesis-level evidence is summarized in Table 5.5. The results indicate strong evidence for lower initial benevolence toward robots, strong evidence against greater benevolence loss following a robot mistake, and inconclusive evidence regarding weaker benevolence recovery for robots relative to humans.

Table 5.4: Posterior means and 95% credible intervals for benevolence contrasts. Bolded rows indicate credible intervals excluding zero.

Benevolence Contrast	Mean	2.5%	97.5%
Robot – Human @ t_1	**-1.92**	**-2.66**	**-1.16**
Robot – Human @ t_2	-0.61	-1.36	0.16
Robot – Human @ t_3	-0.51	-1.24	0.24
Human: $t_2 - t_1$	-0.14	-0.60	0.30
Human: $t_3 - t_2$	0.35	-0.11	0.80
Human: $t_3 - t_1$	0.20	-0.25	0.66
Robot: $t_2 - t_1$	**1.17**	**0.71**	**1.61**
Robot: $t_3 - t_2$	0.44	-0.01	0.89
Robot: $t_3 - t_1$	**1.61**	**1.15**	**2.06**

Table 5.5: Posterior probabilities and Bayes factors for benevolence hypotheses. Strong evidence is highlighted in bold.

Benevolence Hypothesis	Posterior Prob.	Bayes Factor	Evidence
H(a): Lower initial trust in robots	**100%**	∞	**Very strong for H_1**
H(b): Greater trust loss after robot mistake	**0%**	**0.00**	**Very strong for H_0**
H(c): Weaker trust recovery for robots	38.40%	0.62	Anecdotal for H_0

Competence

Posterior contrasts for the competence scale are shown in Table 5.6. Across measurement points, the robot–human contrasts are small and include zero, indicating no reliable between-condition differences in perceived competence at t_1, t_2, or t_3.

Within the human condition, competence trust decreases credibly from t_1 to t_2 and increases credibly from t_2 to t_3, whereas the overall change from t_1 to t_3 is uncertain. A similar temporal pattern is observed within the robot condition: competence trust decreases credibly from t_1 to t_2 and shows a credible recovery from t_2 to t_3, with the overall change from t_1 to t_3 remaining uncertain.

Table 5.7 summarizes hypothesis-level evidence for competence. Posterior probabilities and Bayes factors indicate only anecdotal evidence for all three competence-related hypotheses.

Table 5.6: Posterior means and 95% credible intervals for competence contrasts. Bolded rows indicate credible intervals excluding zero.

Competence Contrast	Mean	2.5%	97.5%
Robot – Human @ t_1	-0.11	-0.66	0.44
Robot – Human @ t_2	0.07	-0.47	0.60
Robot – Human @ t_3	-0.13	-0.65	0.42
Human: $t_2 - t_1$	**-1.98**	**-2.49**	**-1.49**
Human: $t_3 - t_2$	**1.62**	**1.12**	**2.10**
Human: $t_3 - t_1$	-0.36	-0.85	0.12
Robot: $t_2 - t_1$	**-1.80**	**-2.29**	**-1.32**
Robot: $t_3 - t_2$	**1.42**	**0.93**	**1.92**
Robot: $t_3 - t_1$	-0.38	-0.85	0.11

Table 5.7: Posterior probabilities and Bayes factors for competence hypotheses. Strong evidence is highlighted in bold.

Competence Hypothesis	Posterior Prob.	Bayes Factor	Evidence
H(a): Lower initial competence trust in robots	65.24%	1.88	Anecdotal for H_1
H(b): Greater trust loss after robot mistake	30.28%	0.43	Anecdotal for H_0
H(c): Weaker trust recovery for robots	71.30%	2.48	Anecdotal for H_1

Trusting Intentions

Posterior contrasts for the trusting intentions scale are shown in Table 5.8. At baseline (t_1), the robot–human contrast is credibly negative, indicating lower predicted trusting intentions in the robot condition. At t_2, the between-condition contrast is close to zero and uncertain. At t_3, the robot–human contrast again credibly excludes zero, indicating lower trusting intentions toward the robot teammate following the recovery phase.

Within the human condition, trusting intentions decrease credibly from t_1 to t_2 and increase credibly from t_2 to t_3, whereas the overall change from t_1 to t_3 is uncertain. Within the robot condition, the decrease from t_1 to t_2 is uncertain, followed by a credible increase from t_2 to t_3; the overall change from t_1 to t_3 remains uncertain.

Table 5.9 summarizes hypothesis-level evidence for trusting intentions. The results provide very strong evidence for lower initial trusting intentions toward robots, very strong evidence against greater trust loss following a robot mistake, and very strong evidence for weaker trust recovery in the robot condition relative to the human condition.

Table 5.8: Posterior means and 95% credible intervals for trusting intentions contrasts. Bolded rows indicate credible intervals excluding zero.

Trusting Intentions Contrast	**Mean**	**2.5%**	**97.5%**
Robot – Human @ t_1	**-1.34**	**-1.93**	**-0.74**
Robot – Human @ t_2	-0.02	-0.61	0.55
Robot – Human @ t_3	**-0.75**	**-1.35**	**-0.16**
Human: $t_2 - t_1$	**-1.71**	**-2.15**	**-1.28**
Human: $t_3 - t_2$	**1.33**	**0.90**	**1.76**
Human: $t_3 - t_1$	-0.39	-0.83	0.05
Robot: $t_2 - t_1$	-0.39	-0.84	0.05
Robot: $t_3 - t_2$	**0.60**	**0.15**	**1.05**
Robot: $t_3 - t_1$	0.21	-0.24	0.65

Table 5.9: Posterior probabilities and Bayes factors for trusting intentions hypotheses. Strong evidence is highlighted in bold.

Trusting Intentions Hypothesis	**Posterior Prob.**	**Bayes Factor**	**Evidence**
H(a): Lower initial trust in robots	**100%**	∞	**Very strong for H_1**
H(b): Greater trust loss after robot mistake	**0%**	**0.00**	**Very strong for H_0**
H(c): Weaker trust recovery for robots	**99.01%**	**100.27**	**Very strong for H_1**

5.3. Qualitative Thematic Analysis

This section reports a qualitative analysis of participants' open-ended feedback collected after each round and during the reflection phases. The aim of this analysis is not to catalog individual comments, but to describe broader patterns in how participants experienced the collaboration and how these experiences differed between the human and robot teammate conditions. The analysis focuses on three moments of interaction: feedback after each round, the scripted reflection phases, and interaction behavior during task execution.

5.3.1. Analysis Procedure

Because the amount of written open-ended feedback in the post-round questionnaire was small, this part of the analysis is based on general impressions rather than a formal coding procedure. After each session and round, we reviewed the transcripts directly after the round ended and noted short "memorable" sentences that appeared representative of participants' experience (e.g., explicit praise, frustration, or reactions to the mistake). We also recorded whether participants opted into the reflection phase. For participants who chose to reflect, we examined those reflection transcripts more closely to identify recurring patterns in how the mistake and the subsequent repair attempt were discussed. The quotes included below are illustrative examples that support the patterns described.

5.3.2. Questionnaire Feedback Across Rounds

Overall, relatively few participants used the optional open-ended questionnaire field across the rounds. In the human condition, 4 out of 20 participants (20%) provided written post-round feedback at least once, compared to 6 out of 20 participants (30%) in the robot condition.

Human Teammate

Across the three rounds, participants collaborating with a human teammate generally provided minimal written feedback, particularly when the task progressed smoothly. In the first round, the few comments that were provided were short and positive, such as “Good Job”, “Good instructions”, and “Very clear instructions of what needs to be done and easy to understand.” This lack of elaboration did not appear to be disengagement. Rather, it suggested that when expectations were met, participants did not feel a strong need to comment further.

When the scripted mistake occurred in the second round, the written feedback that was provided generally interpreted the error as tolerable and understandable. Participants sometimes framed the outcome as disappointing but acceptable:

> *“It’s sad that we couldn’t complete the task perfectly, but its fine!”*

When criticism was expressed, it was mild and directed at the teammate’s oversight rather than the situation itself:

> *“I dont see how they could have missed that entire block, but oh well”*

By the third round, feedback in the human condition largely returned to brief positive statements or no feedback. For example, some participants wrote “No comment”, while others emphasized successful performance (e.g., “It went perfect!”) or explicitly referenced improvement (e.g., “We did better than last time”). Overall, the questionnaire comments suggest that the error in the second round did not have strong lasting negative effects on participants’ evaluation of the collaboration, at least among those who chose to provide written feedback.

Robot Teammate

In contrast to the human condition, participants collaborating with the robot provided more detailed and reflective written feedback when they responded. In the first round, comments frequently addressed not only the clarity of the instructions but also the robot’s interaction style. Participants often evaluated the instructions positively (e.g., “Instructions were clear”), while also noting constraints or interpersonal distance:

> *“Felt limited in the way I could ask questions, didnt seem like the robot really cared.”*

Pacing was also mentioned:

> *“Good explanation, but tempo could be faster.”*

One participant described the robot as narrowly responsive (“Only responded what was being asked, nothing more”), while still evaluating the interaction as potentially useful:

> *“However the instructions were clear and I think the robot could be helpful in other tasks!”*

Taken together, these comments portray the robot as competent and clear, but sometimes experienced as procedural or less responsive to participants’ needs.

Following the mistake in the second round, feedback in the robot condition became more analytical and evaluative. Some participants continued to describe the instructions as generally clear while noting occasional misleadingness:

> *“Overall the instructions were very detailed and clear, however, very rarely the instructions can be a bit misleading.”*

Others focused on the robot’s handling of responsibility and whether it demonstrated understanding. One participant wrote that the robot “wanted to shoulder the blame and was very cooperative in trying to explain what was wrong,” and added that they liked this communication-wise, but still questioned comprehension:

> *"however I fail to see if he truly understood what went wrong."*

Another participant evaluated the error more harshly and linked it to their attempts to confirm correctness:

> *"I think that it was a relatively stupid mistake. I think we could have avoided the mistake altogether since I repeatedly asked affirmation."*

Compared to the human condition, the robot's error more often prompted explicit evaluation of competence, explanation quality, and evidence of understanding.

In the third round, participants' written feedback was again largely positive and focused on smooth task completion and effective assistance:

> *"I think the cooperation worked well this round, I feel like we did a great job together!"*

> *"Progress went smoothly!"*

> *"The robot was able to assist me well, and gave clear instructions."*

5.3.3. Reflections After Each Round

Human Teammate

During the reflection phases, many participants collaborating with a human teammate chose to opt out, particularly when the collaboration had gone well. In the first round, 2 out of 20 participants (10%) opted into reflection, and their reflections emphasized shared success and cooperative alignment (e.g., "I think we work well together!" and "We did great!").

In the second round, reflections in the human condition sometimes included teasing or blunt reproaches directed at the teammate. Participants explicitly attributed the mistake to the human teammate, for example:

> *"You totally forgot the block! You didnt tell me!"*

> *"Huh how did you even miss that?"*

> *"Wtf where did the second block come from? It is your fault yeah."*

While such comments could appear confrontational when taken out of context, they were typically delivered as informal or playful accountability and did not seem to disrupt the collaborative relationship. This teasing pattern was specific to the human condition and appeared to function as a socially acceptable way of addressing the error while maintaining a cooperative atmosphere.

After the scripted apology preceding the third round, participants generally accepted the explanation and appeared ready to proceed without extended discussion. Most participants started the third round immediately, though one comment before the round began stood out as a light reminder of the earlier issue:

> *"I hope you dont forget a block this time!"*

During the reflection phase after the third round, no participants opted into reflection, with participants proceeding directly to completing the questionnaire.

Robot Teammate

As in the human condition, many participants in the robot condition opted out during the first round. In the first round, 3 out of 20 participants (15%) opted into reflection, and reflections were generally positive and brief. For example, one participant wrote:

> *"I think you explained it very clearly, thank you for that!"*

During the second-round reflection, some participants asked the robot to clarify why it believed it was at fault and how the mistake had occurred. For example, participants asked:

> *"Can you tell me how you missed a block?"*

> *"Why do you think it is your fault?"*

Others responded more reassuringly (e.g., "It's okay!"). This reflects a tendency in the robot condition for participants to seek more concrete explanation and to evaluate whether the robot's responses signaled genuine awareness or scripted behavior.

Before the third round, participants generally accepted the robot's apology and indicated readiness to continue. In some cases, participants offered encouragement prior to the round beginning, such as "We can do better for next round. Lets go!" (with two participants making similar comments). In the reflection phase after the third round, 2 out of 20 participants (10%) opted into reflection. When reflections were provided, they were brief and socially oriented:

> *"Thank you for being my teammate"*
>
> *"See you maybe later?"*

5.3.4. Interaction During Task Execution

Differences between the human and robot conditions were also evident during task execution itself. When collaborating with a human teammate, participants frequently relied on minimal verbal signals or non-lexical sounds, such as humming or brief acknowledgments, to indicate readiness to proceed. These cues were sufficient to maintain smooth coordination between participants and the human teammate.

In the robot condition, participants rarely relied on such minimal signals. Instead, they used explicit verbal confirmations to ensure that the robot registered their readiness. Participants appeared to adapt their communication style to the robot's perceived requirements, making their intentions more explicit. Additionally, several participants expressed frustration with the robot's interaction pace and speech recognition. Delays, slow tempo, and the need to repeat responses multiple times disrupted the flow of the task and were explicitly mentioned in feedback.

6 Discussion

This chapter connects the findings to prior work on interpersonal trust, trust in automation, and human–robot collaboration to explain the mechanisms underlying the observed trust trajectories. In particular, it addresses how baseline expectations shape early trust, how a competence-based mistake is judged relative to those expectations, and why different trust components (benevolence, competence, and trusting intentions) can move differently. Trust was tracked across formation, violation, and repair within the same interaction, using a controlled comparison between a human confederate and a physically present robot. By keeping the task structure, mistake type, and repair message constant across conditions, the study isolates the role of teammate identity in shaping trust changes over time. The discussion proceeds in three steps. First, the research question and preregistered hypotheses are revisited. Second, the aggregated trust results are interpreted. Third, the chapter breaks down the aggregated pattern by examining trust dimensions separately. The chapter then concludes with limitations and directions for future work.

6.1. Research Question and Hypotheses Revisited

This thesis set out to examine how teammate identity (human versus robot) shapes the development of interpersonal trust over time in a collaborative task involving trust formation, violation, and repair. The central research question was:

> *How does a teammate's identity (human vs. robot) affect the development of trust across mistake and recovery phases in collaboration?*

To address this question, trust was conceptualized as a dynamic process that unfolds over time and was measured repeatedly across three phases: an initial correct round (formation), a round containing a competence-based mistake (violation), and a final round following a structured repair attempt (recovery). In line with established interpersonal trust models, trust was operationalized using an aggregated trust score composed of trusting beliefs and trusting intentions, with further analyses conducted at the level of individual trust dimensions [67].

Based on prior work on social categorization, trust in automation, and human–robot interaction, four hypotheses were formulated. First, it was expected that initial trust would be lower for robot teammates than for human teammates (H(a)), reflecting outgroup categorization and reduced social identification with robots [93, 104]. Second, drawing on the perfect automation schema, it was hypothesized that trust loss following a mistake would be greater when the mistake was attributed to a robot rather than a human (H(b)) [63, 28]. Third, given concerns about perceived sincerity and emotional capacity in artificial agents, it was expected that trust recovery would be weaker for robots than for humans, even when the same repair strategy was applied (H(c)) [97, 99]. Finally, these expectations were combined into an overarching hypothesis predicting that trust trajectories would follow the same general shape across conditions, but with consistently lower levels, sharper declines, and slower recovery for robot teammates (H(d)).

6.2. Overall Trust Development: Aggregated Results and Interpretation

At the aggregated level, the results reveal a partially confirmatory and partially corrective picture of the preregistered hypotheses. As shown in Figure 5.3, trust in both conditions follows the expected three-phase trajectory, with an initial phase of trust formation, a sharp decline following the competence-based mistake, and a subsequent increase after the repair attempt. While the overall shape of this trajectory is similar for human and robot teammates, the starting point, magnitude of change, and relation to baseline differ in theoretically meaningful ways.

6.2.1. Initial Trust: Caution Toward Robots

In line with H(a), initial aggregated trust was credibly lower for robot teammates than for human teammates. This effect is reflected in the baseline contrast reported in Table 5.2 and is also visible in Figure 5.3, where the human condition starts at a higher predicted trust level than the robot condition. Table 5.3 further indicates very strong evidence in favor of H(a), suggesting that participants initially approached the robot teammate with more caution than the human confederate.

This pattern is consistent with prior work on social categorization and ingroup-outgroup dynamics, which suggests that humans are more inclined to trust other humans in early interactions, particularly when limited behavioral evidence is available [93, 104]. Even though the robot in this study was physically present, anthropomorphic, and assigned a teammate role, participants appeared to approach it with greater initial caution. This aligns with literature showing that robots are often perceived as socially distinct or less intuitively trustworthy at first encounter, despite human-like cues [93].

6.2.2. Trust Violation: Larger Trust Drop for Humans

Following the competence-based mistake in Round 2, aggregated trust declined in both conditions, confirming that the manipulation constituted a trust violation. However, the magnitude and credibility of trust loss differed substantially between human and robot teammates, and this pattern did not align with the preregistered expectation in H(b).

As shown in Table 5.2 and Figure 5.3, the human condition shows a pronounced decline in trust from baseline to post-mistake, whereas the robot condition shows a smaller and less decisive downward shift. Table 5.3 provides very strong evidence against H(b), indicating that trust loss was not greater for robot mistakes than for human mistakes.

This pattern runs counter to predictions derived from the perfect automation schema, which would suggest harsher punishment for robot errors [63, 28]. Instead, the results indicate that trust violations by human teammates had a stronger impact on aggregated trust; one plausible explanation is that they disrupted participants' higher initial expectations. Humans entered the interaction with a higher baseline level of trust, making the subsequent violation more consequential when those expectations were not met. Because trust judgments are inherently expectation-based, higher initial trust reflects stronger positive assumptions about a teammate's competence and reliability [105, 19]. When these assumptions were violated, the resulting mismatch between expected and observed behavior led to a pronounced decline in trust.

This effect may have been further amplified by the limited interaction history available to participants. With only a single successful round establishing expectations, the mistake in the following round represented the first and only counter-evidence, thereby exerting a relatively strong influence on subsequent trust evaluations.

In contrast, trust in the robot condition began from a more cautious baseline. Lower initial expectations limited the extent to which trust could be violated, leaving less trust to lose and resulting in a smaller and less decisive downward shift. In this context, the same error may have been interpreted as more consistent with prior assumptions about the robot's fallibility rather than as a clear violation. Together, these patterns suggest that trust violations are evaluated relative to both the level and stability of prior expectations, and that early-stage trust judgments based on sparse experience are particularly sensitive to negative deviations.

6.2.3. Trust Recovery: Rebound vs. Return to Baseline

After the repair attempt in Round 3, aggregated trust increased in both conditions, indicating that the combined apology, explanation, and promise strategy was effective in restoring trust following the violation. However, recovery differed between human and robot teammates when considering recovery magnitude, final trust level, and relation to baseline.

As reported in Table 5.2 and shown in Figure 5.3, the human condition shows a somewhat steeper rebound from post-mistake to post-repair trust than the robot condition. However, this larger gain does not translate into full recovery relative to baseline. In the human condition, post-repair trust remains slightly below the initial baseline level, whereas in the robot condition, post-repair trust is credibly higher than at baseline (see Table 5.2). In other words, the robot ends the interaction more trusted than it began, while the human teammate does not.

At the same time, the contrast between robot and human trust at t_3 remains negative and only narrowly approaches zero, suggesting that the robot narrows the initial trust gap without fully closing it (Table 5.2). This aligns with the moderate evidence for H(c) reported in Table 5.3, where weaker robot recovery reflects differences in starting points rather than inferior recovery capacity.

One plausible explanation for this pattern lies in expectation anchoring and expectation updating. Human teammates entered the interaction with higher initial trust and experienced a sharper violation, which resulted in a larger recovery gain. However, this gain primarily reflects the restoration of trust that had been lost, rather than an increase beyond prior beliefs. Because initial expectations toward the human teammate were already relatively strong, successful repair served mainly to reinstate those expectations rather than to meaningfully revise them upward.

Robot teammates, by contrast, began from a more cautious baseline. In this context, the combination of a structured repair and correct performance in the final round likely provided particularly informative evidence about the robot's reliability. Prior work on trust calibration suggests that when initial expectations are low or uncertain, positive performance feedback can lead to upward revisions in trust that exceed baseline levels, as uncertainty is reduced and confidence in the agent's capabilities increases [108]. From this perspective, the repair episode did not merely restore lost trust but also contributed new information that supported more favorable evaluations of the robot's competence.

This interpretation aligns with research showing that trust in automated or artificial agents develops through incremental confirmation of reliability rather than immediate social acceptance [108]. In contrast, trust in human teammates is typically grounded in stronger pre-existing assumptions about competence and intent, which limits the scope for positive updating once those assumptions are re-established [25]. As a result, successful recovery in the human condition primarily reinstated prior trust, whereas in the robot condition it facilitated an increase in trust relative to the initial baseline.

6.2.4. Overall Trust Trajectories

Taken together, the findings across the formation, violation, and recovery phases allow Hypothesis H(d) to be assessed. While the overall trajectory shape is similar across conditions, the predicted pattern of lower levels, sharper declines, and slower recovery for robot teammates is only partially supported.

Consistent with H(d), trust in both the human and robot conditions followed the same overall trajectory shape: trust increased during initial successful collaboration, declined after the competence-based mistake, and increased again following repair. This confirms that the basic temporal dynamics of trust formation, violation, and recovery operate similarly across human and robot teammates in a tightly matched collaborative task.

However, the remaining components of H(d) are not fully supported. While trust levels were consistently lower for the robot at baseline and remained lower at the end of the interaction, the robot did not experience a sharper trust decline following the mistake. Instead, trust dropped more strongly in the human condition, contradicting the prediction of greater punishment for robot errors. Likewise, there was no evidence that recovery was slower for the robot: although the robot showed weaker recovery at the aggregated level, this difference was closely tied to lower starting points rather than an inability to recover. In fact, trust toward the robot increased above its own baseline by the final round, whereas trust toward the human did not.

These results suggest that H(d) correctly anticipated the overall trajectory shape, but not the mechanisms driving change within those trajectories. Rather than reflecting harsher evaluation of robot behavior, the observed differences are better explained by expectation anchoring. Higher initial expectations toward human teammates produced more pronounced trust loss after violation and limited upward revision during recovery. In contrast, lower and more uncertain expectations toward the robot constrained early trust loss but allowed greater relative improvement once the robot demonstrated recovery and reliable performance.

In this sense, the findings revise H(d) from a hypothesis about systematic robot disadvantage to one about expectation-dependent trust dynamics. Trust toward robots did not deteriorate more sharply, nor recover more slowly in absolute terms, but evolved from a lower baseline in a way that allowed meaningful upward revision. The combined evidence across the preceding subsections therefore supports the idea that human-robot trust trajectories are shaped less by teammate identity alone and more by how that identity structures initial expectations against which subsequent behavior is evaluated.

6.3. Beyond Aggregated Trust: The Role of Trust Dimensions

The aggregated trust results establish the overall story: trust changes over time in response to interaction experience and the scripted mistake, and the human and robot conditions are not identical in how these changes unfold. However, the aggregated measure cannot show where these differences come from. A lower overall trust score can reflect weaker social confidence in the teammate, doubts about task ability, or reduced willingness to rely. Each of these have different implications for what is happening psychologically and for how “recovery” should be interpreted.

The dimension-level analyses in this section, therefore, shift the discussion from whether trust differs to what kind of trust differs. This is important for interpreting the mixed pattern in the aggregate trajectory: apparent convergence in overall trust does not necessarily mean the same aspect of trust converges, and apparent recovery does not necessarily mean participants return to the same level of reliance. The following subsections examine benevolence, competence, and trusting intentions separately to identify which components drive the baseline gap, which components respond most strongly to the competence mistake, and which components remain conservative by the end of the interaction.

6.3.1. Benevolence: The Warmth Gap Narrows

We examine benevolence using the posterior contrasts in Table 5.4 and the hypothesis-level evidence in Table 5.5.

Benevolence is the dimension that most clearly separates the two teammate types at the start of collaboration. The robot begins with a relational disadvantage that reflects perceived goodwill, social alignment, or warmth as an interaction partner rather than task ability. This helps interpret why aggregated trust starts lower for the robot even when role and behavior are held constant. At the same time, this disadvantage does not appear fixed. As participants gain interaction experience, benevolence judgments toward the robot become more favorable, while benevolence judgments toward the human teammate remain comparatively stable. This suggests that early caution toward the robot may be shaped by conservative social expectations that can be revised once participants repeatedly observe cooperative, supportive behavior.

The violation phase helps clarify the role of benevolence in this study. Notably, benevolence toward the robot did not decrease after the competence-based mistake; instead, it showed a slight increase during the violation phase. This suggests that participants treated the robot’s error as a competence failure rather than a sign of reduced goodwill toward the partner. More broadly, the results point to a separation between performance-based and relational evaluations: an ability-related failure can make a teammate seem less capable without necessarily making them seem less well-intentioned.

Taken together, the benevolence results suggest that the baseline aggregated trust gap is driven primarily by early social impressions, which can place substantial weight on benevolence-relevant cues when behavioral evidence is still limited [34, 75]. The later narrowing of the aggregated gap is consistent with repeated cooperative interaction providing counterevidence to these initial impressions, allowing benevolence perceptions of the robot to shift upward over time [98, 93]. This pattern also aligns with work on trust restoration suggesting that apologies and explanations can communicate accountability and consideration for the partner, supporting benevolence even when the original violation is competence-based [61, 51]. Overall, benevolence appears to explain the robot's initial disadvantage, while its upward revision helps explain why aggregated trust differences narrow as collaboration progresses.

6.3.2. Competence: Similar Violation and Repair Patterns

We examine competence using the posterior contrasts in Table 5.6 and the hypothesis-level evidence in Table 5.7.

Competence is the dimension where human and robot teammates look most similar. Rather than showing a stable baseline difference, competence judgments are primarily shaped by what participants observe the teammate doing in the task. This makes competence a useful lens for interpreting the violation manipulation, because it suggests that participants updated competence beliefs in response to the scripted mistake and subsequent correct performance in much the same way across conditions. In other words, when the failure is ability-related and the behavioral evidence is comparable, perceived competence appears to be updated through a shared, performance-based process rather than being strongly shaped by the teammate's status as a robot.

The pronounced drop in trust during the violation in both conditions is consistent with prior work showing that ability-related failures often trigger sharp trust updating in task-oriented settings, because they are easy to detect and attribute [58, 44, 51]. Because the failure manipulation targeted competence, the central question is whether the same performance failure carries more weight in participants' judgments when the teammate is a robot. The competence results do not support that interpretation. Instead, they indicate that participants reacted to the error as an error, and that recovery in competence perceptions is largely consistent with the return to correct performance, regardless of whether the teammate is human or robot.

At the same time, competence cannot account for why aggregated trust starts lower for the robot. The absence of a baseline competence gap suggests that the early disadvantage in overall trust is more plausibly grounded in non-competence components, particularly benevolence and trusting intentions. Taken together, the competence results support a key implication of the study: when behavior is held constant and the mistake is clearly competence-based, participants do not appear to penalize robots more strongly than humans on the competence dimension.

6.3.3. Trusting Intentions: Unequal Recovery

We examine trusting intentions using the posterior contrasts in Table 5.8 and the hypothesis-level evidence in Table 5.9.

Trusting intentions show the clearest divergence between the two conditions after the recovery phase. In benevolence, the robot's ratings move upward over time and the initial gap narrows, and in competence the human and robot trajectories are largely parallel around the violation and repair. Trusting intentions, however, show a different structure. The gap becomes smaller around t_2, but it widens again by t_3. This re-opening of the gap is important because it aligns with the remaining difference in aggregated trust at the final measurement point.

The shape of the trusting intentions trajectories suggests that this divergence is driven more by differences in change dynamics than by a simple "lower intentions for robots" story. The human condition shows a clear dip-and-rebound pattern, with a pronounced decrease after the mistake and a strong increase after the repair. By contrast, the robot condition is comparatively stable. The drop after the mistake is smaller and not clearly supported, and the subsequent increase is also more modest. As a result, even though the robot trends upward again, it does not show the same rebound structure as the human condition, and the between-condition gap becomes visible again at t_3.

Conceptually, this matters because trusting intentions capture willingness to accept vulnerability through reliance [67, 40]. In this study, the data suggest that willingness-to-rely is updated differently depending on whether the teammate is human or robot. For the human teammate, reliance appears more responsive to the violation-and-repair sequence, whereas for the robot teammate it remains more conservative and steady. This helps interpret why recovery is nuanced. The robot is not necessarily punished more strongly at the point of failure, but willingness-to-rely does not shift in the same way as it does for the human, and that difference contributes to the re-emergence of the gap at the end of the interaction.

6.3.4. Main Takeaways Across Trust Dimensions

Disaggregating trust into benevolence, competence, and trusting intentions shows that the aggregated human–robot differences reflect distinct processes across dimensions rather than a single uniform "robot effect." The initial trust gap at t_1 is not only rooted in benevolence. Benevolence captures a clear relational disadvantage for the robot at the start, consistent with weaker early social confidence in the robot as an interaction partner. At the same time, trusting intentions also contribute to the baseline difference, indicating that participants begin the collaboration both less confident in the robot's goodwill and less willing to rely on it. Importantly, the benevolence gap narrows across rounds primarily because benevolence toward the robot shifts upward with interaction experience, while benevolence toward the human teammate remains comparatively stable.

Competence is the dimension where human and robot teammates look most alike. Competence judgments appear to be driven mainly by observable task performance, with both conditions showing comparable responses to the competence-based mistake and subsequent correct performance. This helps explain why there is strong evidence against greater overall trust loss after a robot mistake. The results suggest that, when behavior is held constant and the violation is clearly ability-related, participants update competence beliefs through a shared, performance-based process rather than penalizing robots more strongly. This conclusion is reinforced by how the non-competence dimensions behaved around the violation. In both benevolence and trusting intentions, the immediate drop from t_1 to t_2 was larger in the human condition than in the robot condition. This pattern further supports the evidence against the hypothesis that robots would incur greater trust loss after a mistake, because even the more social and reliance-focused components of trust did not show a stronger decline for the robot.

Trusting intentions, in contrast, clarify why recovery is nuanced. Unlike benevolence, where the robot closes the gap over time, and unlike competence, where both conditions follow a similar trajectory, trusting intentions show a re-emerging gap after recovery. The human condition shows a clearer dip-and-rebound pattern, whereas the robot condition is comparatively stable, with a smaller and less supported drop and a more modest increase. As a result, the gap in willingness-to-rely becomes visible again at t_3, which aligns with the remaining difference in aggregated trust at the final measurement point. Taken together, the partial support for Hypothesis H(d) becomes more interpretable. The predicted weaker recovery for robots is not expressed primarily as stronger punishment of robot errors, but as a more conservative reinstatement of reliance after repair, most clearly reflected in trusting intentions.

6.4. Qualitative Insights

The qualitative observations complement the model-based results by illustrating the interpretations participants brought to the interaction with each teammate type. Rather than introducing new effects, they help explain why the same scripted events could carry different meaning across conditions.

A recurring theme in the robot condition was a stronger focus on interaction management. Participants more often commented on coordination, responsiveness, and whether the robot appeared to register what was said or done. This points to an additional source of early caution: not only lower social identification, but also uncertainty about smooth turn-taking and mutual understanding.

The qualitative data also suggest that the mistake was framed differently depending on teammate identity. With the human teammate, participants more readily treated the error as a lapse by a socially accountable partner, which can invite interpersonal disappointment and blame. With the robot, participants more often discussed the error in system-like terms and evaluated the repair in relation to whether it signaled understanding (e.g., whether the explanation felt meaningful rather than merely formulaic). This provides a plausible bridge to the quantitative pattern where trust loss was sharper for humans despite higher baseline trust.

These observations suggest several directions for follow-up work: human errors may be punished more when they elicit interpersonal blame; robot trust may depend more on demonstrated understanding than apologies.

6.5. Limitations

This study offers a controlled comparison of trust development with a human versus a robot teammate across trust formation, violation, and repair. However, several limitations should be considered when interpreting the results and when generalizing beyond this experimental setting.

First, the participant sample limits how far these findings can be generalized beyond this experimental setting. All participants were between 18 and 27 years old and were largely students, and recruitment relied on personal connections and extended networks rather than broad sampling. As a result, the observed trust trajectories may not transfer directly to populations with different baselines and interaction expectations (e.g., older adults, non-students, or people collaborating in longer-term workplace teams). This matters because trust in automated or robotic systems can vary by user characteristics, including age and gender, and these differences may shape both initial trust and how errors are interpreted. For example, prior work suggests that age can be associated with different patterns of trust in human–robot interaction, and that gender differences in trust toward robots have been observed in some settings [100, 52].

Second, the robot condition introduced technical friction that the human condition could not fully mirror, even with careful standardization. The robot sometimes did not pick up what participants said, and participants occasionally had to speak louder due to microphone sensitivity; several participants also reported frustration with speech recognition and interaction pacing, including delays, slow tempo, and having to repeat responses. Such issues can influence trust independently of the scripted violation and repair, because participants may attribute recognition failures to inattentiveness or incompetence, meaning that some between-condition differences, especially the lower baseline trust toward the robot, may partly reflect interaction quality rather than teammate identity alone. However, this limitation does not make the results uninformative: reduced interaction quality would be expected to amplify robot disadvantage, particularly by producing larger trust loss after a robot mistake, yet the study shows the opposite pattern, with trust declining more sharply after the human teammate's mistake and robot-related trust measures not showing a stronger drop during the violation phase. This suggests that technical friction may contribute to the baseline gap but is less consistent with the observed trust-loss dynamics, while also motivating follow-up studies with more robust robot communication or tighter matching of interaction friction across conditions.

Third, the qualitative component provided limited explanatory depth. Although open-ended feedback and reflections were included, much of this data was optional, and many participants opted out especially when they felt there was little to add after a smooth round. This reduced the amount of rich qualitative material that could help explain why trust changed (e.g., what cues shaped their judgments). Reflections were also designed to be short and informal, which likely contributed to brief responses and further limited nuance in the qualitative analysis.

Finally, the scope of the manipulation and the design constraints limit how broadly the findings can be interpreted. The study used a single competence-based mistake and a single combined repair message (apology, explanation, and promise) for clarity and comparability, but results may not generalize to other trust violations (e.g., repeated small errors, safety-critical failures, or violations perceived as intentional) or other repair strategies (e.g., explanation-only, compensation, or interactive negotiation). In addition, only one robot platform (Navel) and one interaction configuration were used, and trust responses may differ with other embodiments. More generally, the sample size ($N = 40$) supports detecting clear patterns but still leaves uncertainty around smaller effects and subgroup differences.

Taken together, these limitations suggest that the results should be interpreted as evidence about trust development in this specific controlled, short-term collaboration setting, and as a foundation for follow-up studies with broader sampling, richer qualitative elicitation, and improved robustness of robot communication.

6.6. Future Work

This study provides a snapshot of trust dynamics in short-term, co-present human–robot teamwork, but it also raises several follow-up questions. In particular, the findings point to the roles of expectations, communication norms, and trust repair processes in shaping how trust changes over time. Future work can build on these results by testing whether the same patterns hold across different people, tasks, mistake types, and longer interaction periods.

First, it would be valuable to study for whom the observed trust patterns hold by modeling individual differences more explicitly. Trust research has long argued that trust judgments depend on both perceived trustworthiness (competence, benevolence, integrity) and differences in the trustor, such as trust propensity [64, 67]. Future studies could therefore test whether education level and field of study relate to different teammate expectations, different ways of explaining errors, and different thresholds for relying on robot guidance. In addition, because task context is known to moderate trust and teamwork outcomes, it would be informative to compare a strenuous, high-concentration task with a lighter, more playful collaboration [23, 72].

Second, one promising direction is to explore why participants adopted different coordination styles with the human teammate than with the robot. A clear qualitative difference was that minimal readiness signals (such as humming or brief acknowledgments) felt natural toward humans but not toward the robot, where participants preferred explicit wording. This aligns with work suggesting that people often respond socially to technology, yet still apply different assumptions about understanding and grounding compared to human partners [70, 69]. Future work could examine what perceptions drive this shift: for example, whether participants believe robots cannot infer intent from nonverbal acknowledgments, whether they expect stricter ”protocol-like” communication with machines, or whether they worry about being misunderstood. Beyond measurement, this also opens a design avenue: robots could explicitly demonstrate that short responses are understood and sufficient, potentially making collaboration more natural.

Third, the role of anthropomorphism in this setting remains an open question. Theory and evidence suggest that human-like cues can shape mind perception and social judgments, and may increase trust or trust resilience in cognitive agents [30, 104, 99]. At the same time, there are cautionary notes that anthropomorphism can raise expectations and create mismatches that harm trust if the system cannot meet the implied social standards [21]. A useful next step is to test, in a controlled way, whether anthropomorphic features actually moved the trust scales in this study, or whether the effects were primarily driven by performance and repair content.

Fifth, the study design can be extended to longer time horizons. Trust in teams is dynamic and updates with repeated evidence, and first impressions can persist across interactions [22]. Short sessions are useful for control, but real-world human–robot teaming could possibly be integrated for a long-term period in the future. Future work could therefore examine longer trust periods than a single session, include repeated violations and repairs, and test whether trust calibrates toward appropriate reliance over time [58, 46]. This would also allow separating temporary reactions or mood from stable changes in reliance behavior.

Finally, future studies should incorporate richer behavioral measures, especially nonverbal cues. Trust is not only expressed through questionnaire ratings; it can also be enacted through trusting actions such as delegating tasks, following advice, or cooperating in joint activities [6, 77]. Prior work suggests that facial expressions and small talk can influence trust in robots, and recent work on embodiment and trustworthy interaction points to the importance of alignment between verbal and nonverbal behavior [76]. Adding video-based coding or sensor-based measures could reveal early, subtle trust shifts that self-report misses, and help explain phenomena such as teasing: participants sometimes teased the human teammate after a large competence drop, which may function as a social regulation strategy that is culturally normal in human teams but not yet comfortable with robots. Understanding when and why such relational behaviors emerge would help explain differences between human and robot teammates beyond raw performance, and would inform how robots should respond to errors in socially acceptable ways.

Together, these directions aim to move from a tightly controlled, short-term trust trajectory toward a more complete account of trust in co-present human–robot teamwork: across different people, different tasks, and longer-term collaboration, using both subjective and behavioral evidence.

7 Conclusion

As robots increasingly take on cooperative roles in workplaces, education, and care, understanding how people decide when to rely on them becomes essential for safe and effective teamwork. This thesis addressed that need by examining whether trust toward a robotic teammate develops in the same way as trust toward a human teammate when collaboration unfolds across a full cycle of formation, violation, and repair. The central research question asked how teammate identity (human vs. robot) shapes trust across mistake and recovery phases in an interdependent task.

To answer this question, the study used a controlled, physically co-present collaboration setting in which participants built block figures using step-by-step guidance from either a human confederate or an anthropomorphic robot. Trust was measured repeatedly across three rounds: an initial correct round (formation), a second round containing a planned competence-based mistake (violation), and a third round following a structured repair attempt (recovery). The repair strategy combined an apology, a concrete explanation, and a forward-looking promise, allowing a consistent comparison of repair effects across teammate conditions. Trust was operationalized using McKnight and Chervany's framework, capturing both trusting beliefs (benevolence and competence) and trusting intentions, which matched the short, task-focused nature of the interaction [67].

The results show that trust in both conditions followed the expected three-phase trajectory: it formed during successful collaboration, dropped after the mistake, and increased again after the repair attempt. However, the magnitude and meaning of these changes differed between human and robot teammates in ways that refine common assumptions about robot trust. First, initial trust was credibly lower for the robot than for the human teammate, providing very strong evidence for the hypothesis that people approach robots with greater caution at first encounter. Second, the preregistered expectation that robot mistakes would lead to greater trust loss was not supported. Instead, aggregated trust dropped more strongly in the human condition, while the robot condition showed a smaller and less decisive decline. The pattern suggests that early-stage trust violations are evaluated relative to prior expectations: higher initial trust in the human condition created more trust to lose, whereas the more cautious baseline toward the robot reduced the extent to which the same mistake was experienced as a sharp violation.

After the repair attempt, trust increased in both conditions, indicating that the combined apology–explanation–promise strategy supported trust restoration following a competence-based mistake. At the same time, evidence for weaker recovery in the robot condition was moderate at the aggregated level, consistent with the idea that repair may be less diagnostic or less persuasive for robots even when the message is held constant. This overall pattern offers a more nuanced view than a simple "robots are punished more" account: the findings point to an expectation-based mechanism in which initial caution, rather than harsher punishment, may explain why robot trust trajectories remain lower across the interaction.

Analyses at the level of trust dimensions further clarify where the human–robot gap originates and how it changes. Benevolence was substantially lower for the robot at baseline, but increased strongly within the robot condition from the first to the second and third measurement points, shrinking the between-condition difference over time. This suggests that early distrust of robots may be driven less by doubts about capability and more by uncertainty about social orientation and care, and that even a short interaction can reduce that uncertainty once participants observe responsive, cooperative behavior. In combination, the aggregated and scale-specific findings imply that treating trust as a single stable attitude can miss important dynamics: different trust components can start at different baselines and change for different reasons, even when the task and the repair message are identical.

Beyond the quantitative trajectory, the study design itself contributes a controlled yet socially realistic approach to comparing trust in human versus robot teammates. Many prior trust studies rely on simulated agents or non-embodied settings, whereas this work examined repeated interaction with a physically present, anthropomorphic robot under tightly matched communication constraints across conditions. The findings also have practical implications for the design of collaborative robots. If initial trust deficits toward robots are primarily tied to benevolence beliefs, then early interaction design should prioritize cues that signal attentiveness, responsiveness, and cooperative intent, rather than focusing exclusively on demonstrating technical competence.

In sum, this thesis shows that trust in human–robot collaboration is dynamic and expectation-dependent. People began with lower trust toward a robotic teammate, did not penalize the robot's competence mistake more than a human's in this early-stage setting, and showed trust recovery after repair in both conditions, with indications that recovery might be less complete for robots at the aggregated level. By analyzing the full formation–violation–repair cycle and separating trust dimensions, this work offers a clearer account of when and why trust toward robots diverges from trust toward humans, providing a foundation for designing robotic teammates that support appropriately calibrated reliance over time.

References

[1] Icek Ajzen. “The theory of planned behavior”. In: *Organizational Behavior and Human Decision Processes* 50 (1991), pp. 179–211. DOI: 10.1016/0749-5978(91)90020-T.

[2] Zeynep Akata et al. “A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence”. In: *Computer* 53.8 (2020), pp. 18–28. DOI: 10.1109/MC.2020.2996587.

[3] Gene Alarcon et al. “Trust Violations in Human-Human and Human-Robot Interactions: The Influence of Ability, Benevolence and Integrity Violations”. In: *Hawaii International Conference on System Sciences 2022 (HICSS-55)* (Jan. 2022). URL: https://aisel.aisnet.org/hicss-55/cl/human-robot_interactions/3.

[4] Gene M. Alarcon et al. “Differential biases in human-human versus human-robot interactions”. In: *Applied Ergonomics* 106 (Jan. 2023), p. 103858. ISSN: 0003-6870. DOI: 10.1016/j.apergo.2022.103858. URL: https://www.sciencedirect.com/science/article/pii/S0003687022001818 (visited on 06/07/2025).

[5] Gene M. Alarcon et al. “Exploring the differential effects of trust violations in human-human and human-robot interactions”. In: *Applied Ergonomics* 93 (2021), p. 103350. ISSN: 0003-6870. DOI: https://doi.org/10.1016/j.apergo.2020.103350. URL: https://www.sciencedirect.com/science/article/pii/S0003687020302982.

[6] Gene M. Alarcon et al. “The effect of propensity to trust and perceptions of trustworthiness on trust behaviors in dyads”. In: *Behavior Research Methods* 50.5 (2018), pp. 1906–1920. DOI: 10.3758/s13428-017-0959-6.

[7] Anthony L. Baker et al. “Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.4 (2018), pp. 1–30. DOI: 10.1145/3181671.

[8] Christoph Bartneck and Jodi Forlizzi. “A design-centred framework for social human-robot interaction”. In: *RO-MAN 2004: 13th IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 2004, pp. 591–594. DOI: 10.1109/ROMAN.2004.1374827.

[9] Ryan A. Beasley. “Medical robots: Current systems and research directions”. In: *Journal of Robotics* 2012 (2012), p. 401613. DOI: 10.1155/2012/401613.

[10] Tony Belpaeme et al. “Social robots for education: A review”. In: *Science Robotics* 3.21 (2018), eaat5954. DOI: 10.1126/scirobotics.aat5954.

[11] Roger Bemelmans et al. “Socially Assistive Robots in Elderly Care: A Systematic Review into Effects and Effectiveness”. In: *Journal of the American Medical Directors Association* 13 (Dec. 2010), 114–120.e1. DOI: 10.1016/j.jamda.2010.10.002.

[12] Robert J. Bies and Thomas M. Tripp. “Beyond Distrust: Getting Even and the Need for Revenge”. In: *Trust in Organizations: Frontiers of Theory and Research*. Ed. by Roderick M. Kramer and Tom R. Tyler. Thousand Oaks, CA: Sage Publications, 1996, pp. 246–260.

[13] Christina Breuer, Joachim Hüffmeier, and Guido Hertel. “Does Trust Matter More in Virtual Teams? A Meta-Analysis of Trust and Team Effectiveness Considering Virtuality and Documentation as Moderators”. In: *Journal of Applied Psychology* 101.8 (2016), p. 1151. DOI: 10.1037/apl0000113.

[14] Elizabeth Broadbent et al. “Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a better personality”. In: *PLOS ONE* 8.8 (2013), e72589. DOI: 10.1371/journal.pone.0072589.

[15] Bob Carpenter et al. “Stan: A Probabilistic Programming Language”. In: *Journal of Statistical Software* 76.1 (2017), pp. 1–32. DOI: 10.18637/jss.v076.i01.

[16] Álvaro Castro-González, Henny Admoni, and Brian Scassellati. “Effects of form and motion on judgments of social robots’ animacy, likability, trustworthiness and unpleasantness”. In: *International Journal of Human-Computer Studies* 90 (2016), pp. 27–38. DOI: 10.1016/j.ijhcs.2016.02.004.

[17] Erin K. Chiou et al. “Toward human-robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task”. In: *International Journal of Social Robotics* (2021). DOI: 10.1007/s12369-021-00787-6.

[18] Lara Christoforakos et al. “Can Robots Earn Our Trust the Same Way Humans Do? A Systematic Exploration of Competence, Warmth, and Anthropomorphism as Determinants of Trust Development in HRI”. In: *Frontiers in Robotics and AI* 8 (2021), p. 640444. DOI: 10.3389/frobt.2021.640444.

[19] Jason A Colquitt et al. “Trust in typical and high-reliability contexts: Building and reacting to trust among firefighters”. In: *Academy of Management Journal* (2011). DOI: 10.5465/amj.2006.0241.

[20] Jason A. Colquitt, Brent A. Scott, and Jeffery A. LePine. “Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance”. In: *Journal of Applied Psychology* 92.4 (2007), pp. 909–927. DOI: 10.1037/0021-9010.92.4.909.

[21] K. E. Culley and P. Madhavan. “A note of caution regarding anthropomorphism in HCI agents”. In: *Computers in Human Behavior* 29.3 (2013), pp. 577–579. DOI: 10.1016/j.chb.2012.11.023.

[22] Kate Darling, Palash Nandy, and Cynthia Breazeal. “Empathic concern and the effect of stories in human-robot interaction”. In: *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Kobe, Japan: IEEE, 2015, pp. 770–775. DOI: 10.1109/ROMAN.2015.7333675.

[23] Bart A. De Jong, Kurt T. Dirks, and Nicole Gillespie. “Trust and Team Performance: A Meta-Analysis of Main Effects, Moderators, and Covariates”. In: *Journal of Applied Psychology* 101.8 (2016), pp. 1134–1150. DOI: 10.1037/apl0000110.

[24] Ewart J. De Visser et al. “A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents”. In: *Human Factors* 59.1 (2017), pp. 116–133. DOI: 10.1177/0018720816679137.

[25] Kurt T. Dirks and Bart de Jong. “Trust within the workplace: A review of two waves of research and a glimpse of the third”. In: *Annual Review of Organizational Psychology and Organizational Behavior* 9 (2022), pp. 247–276. DOI: 10.1146/annurev-orgpsych-012420-083025.

[26] Norman Du et al. “Look Who’s Talking Now: Implications of AV’s Explanations on Driver’s Trust, AV Preference, Anxiety and Mental Workload”. In: *Transportation Research Part C: Emerging Technologies* 104 (2019), pp. 428–442. DOI: 10.1016/j.trc.2019.05.025.

[27] Mary Dzindolet et al. “The role of trust in automation reliance”. In: *International Journal of Human-Computer Studies* 58 (June 2003), pp. 697–718. DOI: 10.1016/S1071-5819(03)00038-7.

[28] Mary T. Dzindolet et al. “The Perceived Utility of Human and Automated Aids in a Visual Detection Task”. In: *Human Factors* 44.1 (2002). PMID: 12118875, pp. 79–94. DOI: 10.1518/0018720024494856.

[29] Nicholas Epley, Adam Waytz, and John T. Cacioppo. “On seeing human: A three-factor theory of anthropomorphism”. In: *Psychological Review* 114.4 (2007), pp. 864–886. DOI: 10.1037/0033-295X.114.4.864. URL: https://psycnet.apa.org/record/2007-13558-002.

[30] Nicholas Epley, Adam Waytz, and John T. Cacioppo. “On seeing human: A three-factor theory of anthropomorphism”. In: *Psychological Review* 114.4 (2007), pp. 864–886. DOI: 10.1037/0033-295X.114.4.864.

[31] Connor Esterwood and Lionel P. Robert. “Do you still trust me? Human-robot trust repair strategies”. In: *Proceedings of the 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2021. DOI: 10.1109/RO-MAN50785.2021.9515460.

[32] Connor Esterwood and Lionel P. Robert. "Three Strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness". In: *Computers in Human Behavior* 142 (Jan. 2023). DOI: 10.1016/j.chb.2023.107658.

[33] Alexander Etz and Joachim Vandekerckhove. "Introduction to Bayesian Inference for Psychology". In: *Psychonomic Bulletin & Review* 25.1 (2017), pp. 5–34. DOI: 10.3758/s13423-017-1262-3.

[34] Susan T. Fiske, Amy J. C. Cuddy, and Peter Glick. "Universal dimensions of social cognition: Warmth and competence". In: *Trends in Cognitive Sciences* 11 (2007), pp. 77–83. DOI: 10.1016/j.tics.2006.11.005.

[35] Michael W. Floyd and David W. Aha. "Using explanations to provide transparency during trust-guided behavior adaptation". In: *AI Communications* 30.3-4 (2017), pp. 281–294. DOI: 10.3233/AIC-170737.

[36] Claire Garnett et al. "Updating the evidence on the effectiveness of the alcohol reduction app, Drink Less: using Bayes factors to analyse trial datasets supplemented with extended recruitment". In: *F1000Research* 8 (2019), p. 114. DOI: 10.12688/f1000research.17952.2.

[37] Eleni Georganta and Anna-Sophie Ulfert. "My colleague is an AI! Trust differences between AI and human teammates". In: *Team Performance Management* 30.1/2 (2024), pp. 23–37. DOI: 10.1108/TPM-07-2023-0053.

[38] Eleni Georganta and Anna-Sophie Ulfert. "Would you trust an AI team member? Team trust in human–AI teams". In: *Journal of Occupational and Organizational Psychology* 97 (Apr. 2024), pp. 1212–1241. DOI: 10.1111/joop.12504.

[39] Homero Gil de Zúñiga, Manuel Goyanes, and Timilehin Durotoye. "A Scholarly Definition of Artificial Intelligence (AI): Advancing AI as a Conceptual Framework in Communication Research". In: *Political Communication* (Dec. 2023), p. 18. DOI: 10.1080/10584609.2023.2290497.

[40] Ella Glikson and Anita Woolley. "Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals (in press)". In: *The Academy of Management Annals* (Apr. 2020).

[41] Sander Greenland et al. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". In: *European Journal of Epidemiology* 31.4 (2016), pp. 337–350. DOI: 10.1007/s10654-016-0149-3.

[42] Erico Guizzo. *What is a robot?* May 2020. URL: https://robots.ieee.org/learn/what-is-a-robot/.

[43] Yiwen Guo and Xiaojun J. Yang. "Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach". In: *International Journal of Social Robotics* 13 (2020), pp. 1899–1909. DOI: 10.1007/s12369-020-00721-1.

[44] Peter A. Hancock et al. "A meta-analysis of factors affecting trust in human-robot interaction". In: *Human Factors* 53.5 (2011), pp. 517–527. DOI: 10.1177/0018720811417254.

[45] Kerstin Sophie Haring et al. "The influence of robot appearance and interactive ability in HRI: A cross-cultural study". In: *Proceedings of the 8th International Conference on Social Robotics.* Vol. 9979. Lecture Notes in Computer Science. Kansas City, MO, USA: Springer, 2016, pp. 392–401. DOI: 10.1007/978-3-319-47437-3_38.

[46] Kevin A. Hoff and Masooda Bashir. "Trust in automation: Integrating empirical evidence on factors that influence trust". In: *Human Factors* 57.3 (2015), pp. 407–434. DOI: 10.1177/0018720814547570.

[47] Shannon L. Jones and Pankaj P. Shah. "Diagnosing the locus of trust: A temporal perspective for trustor, trustee, and dyadic influences on perceived trustworthiness". In: *Journal of Applied Psychology* 101.3 (2016), pp. 392–414. DOI: 10.1037/apl0000058.

[48] Peter H. Kahn Jr. et al. "Do people hold a humanoid robot morally accountable for the harm it causes?" In: *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction.* 2012, pp. 33–40. DOI: 10.1145/2157689.2157696.

[49] Andreas Kaplan and Michael Haenlein. “Rulers of the world, unite! The challenges and opportunities of artificial intelligence”. In: *Business Horizons* 63.1 (2020), pp. 37–50. DOI: 10.1016/j.bushor.2019.09.003.

[50] Sara Kiesler et al. “Anthropomorphic interactions with a robot and robot-like agent”. In: *Social Cognition* 26.2 (2008), pp. 169–181. DOI: 10.1521/soco.2008.26.2.169.

[51] Peter H. Kim et al. “Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- Versus Integrity-Based Trust Violations”. In: *Journal of Applied Psychology* 89.1 (2004), pp. 104–118. DOI: 10.1037/0021-9010.89.1.104.

[52] Elyakim Kislev. “The Robot-Gender Divide: How and Why Men and Women Differ in Their Attitudes Toward Social Robots”. In: *Social Science Computer Review* 41.6 (2023). Published December 2023, pp. 2230–2248. DOI: 10.1177/08944393231155674. URL: https://journals.sagepub.com/doi/10.1177/08944393231155674.

[53] E. S. Kox et al. “Trust repair in human-agent teams: The effectiveness of explanations and expressing regret”. In: *Cognition, Technology & Work* 23 (2021), pp. 865–878. DOI: 10.1007/s10111-021-00669-9.

[54] Roderick M. Kramer and Roy J. Lewicki. “Repairing and Enhancing Trust: Approaches to Reducing Organizational Trust Deficits”. In: *Academy of Management Annals* 4.1 (2010), pp. 245–277. DOI: 10.5465/19416521003673344.

[55] S. Küçük. “Introductory chapter: Medical robots in surgery and rehabilitation”. In: *Medical Robotics – New Achievements*. Ed. by S. Küçük and A. E. Canda. IntechOpen, 2020, pp. 3–8. DOI: 10.5772/intechopen.92028.

[56] Claus W. Langfred. “The downside of self-management: A longitudinal study of the effects of conflict on trust, autonomy, and task interdependence in self-managing teams”. In: *Academy of Management Journal* 50.4 (2007), pp. 885–900. DOI: 10.5465/amj.2007.26279196.

[57] Lindsay Larson and Leslie DeChurch. “Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams”. In: *The Leadership Quarterly* 31.1 (2020), pp. 1–18. DOI: 10.1016/j.leaqua.2019.101377.

[58] John Lee and Katrina See. “Trust in Automation: Designing for Appropriate Reliance”. In: *Human factors* 46 (Feb. 2004), pp. 50–80. DOI: 10.1518/hfes.46.1.50.30392.

[59] Kwan Lee et al. “Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction”. In: *Journal of Communication* 56 (Dec. 2006), pp. 754–772. DOI: 10.1111/j.1460-2466.2006.00318.x.

[60] Iolanda Leite, Carlos Martinho, and Ana Paiva. “Social robots for long-term interaction: A survey”. In: *International Journal of Social Robotics* 5.2 (2013), pp. 291–308. DOI: 10.1007/s12369-013-0178-y.

[61] Roy J. Lewicki and Chad Brinsfield. “Trust Repair”. In: *Annual Review of Organizational Psychology and Organizational Behavior* 4 (2017), pp. 287–313. DOI: 10.1146/annurev-orgpsych-032516-113147.

[62] Joseph B. Lyons, Izz Aldin Hamdan, and Thy Q. Vo. “Explanations and trust: What happens to trust when a robot partner does something unexpected?” In: *Computers in Human Behavior* 138 (2023), p. 107473. DOI: 10.1016/j.chb.2022.107473.

[63] Poornima Madhavan and Douglas Wiegmann. “Similarities and differences between human-human and human-automation trust: An integrative review”. In: *Theoretical Issues in Ergonomics Science* 8 (July 2007), pp. 277–301. DOI: 10.1080/14639220500337708.

[64] Roger C. Mayer, James H. Davis, and F. David Schoorman. “An Integrative Model of Organizational Trust”. In: *The Academy of Management Review* 20.3 (1995), pp. 709–734. DOI: 10.2307/258792. URL: https://doi.org/10.2307/258792.

[65] Daniel McAllister. “Affect- and Cognition-Based Trust Formations for Interpersonal Cooperation in Organizations”. In: *Academy of Management Journal* 38 (Feb. 1995), pp. 24–59. DOI: 10.2307/256727.

[66] D. Harrison McKnight and Norman L. Chervany. “An Extended Trust Building Model: Comparing Experiential and Non-Experiential Factors”. In: *Information Systems Research: Relevant Theory and Informed Practice.* Ed. by Izak Benbasat. Boston, MA: Springer, 2006, pp. 1–28. DOI: 10.1007/0-387-35489-7_3.

[67] D. Harrison McKnight, Larry L. Cummings, and Norman L. Chervany. “Initial trust formation in new organizational relationships”. In: *Academy of Management Review* 23.3 (1998), pp. 473–490. DOI: 10.2307/259290.

[68] Stephanie M. Merritt et al. “Measuring individual differences in the perfect automation schema”. In: *Human Factors* 57.5 (2015), pp. 740–753. DOI: 10.1177/0018720815581247.

[69] Clifford Nass and Youngme Moon. “Machines and Mindlessness: Social Responses to Computers”. In: *Journal of Social Issues* 56 (Mar. 2000), pp. 81–103. DOI: 10.1111/0022-4537.00153.

[70] Clifford Nass, Jonathan Steuer, and Ellen Tauber. “Computers are social actors”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Boston, MA: ACM, 1994, pp. 73–78.

[71] Doug Newcomb. *Making robot cars more human.* Dec. 2014. URL: http://www.pcmag.com/article2/0,2817,2473149,00.asp.

[72] Traci A. O’Neill et al. “Human–autonomy teaming: A review and analysis of the empirical literature”. In: *Human Factors* 64.5 (2022), pp. 904–938. DOI: 10.1177/0018720820960865.

[73] Abby Ohlheiser. *Ask Siri “what is zero divided by zero?” and she will send you to the burn ward.* June 2015. URL: http://www.washingtonpost.com/news/morning-mix/wp/2015/06/30/ask-siri-what-is-zero-divided-by-zero-and-she-will-send-you-to-theburn-ward/.

[74] Linda Onnasch and Clara Laudine Hildebrandt. “Impact of Anthropomorphic Robot Design on Trust and Attention in Industrial Human-Robot Interaction”. In: *ACM Transactions on Human-Robot Interaction* 11.1 (2022). DOI: 10.1145/3472224.

[75] Maike Paetzel, Giulia Perugia, and Ginevra Castellano. “The Persistence of First Impressions: The Effect of Repeated Interactions on the Perception of a Social Robot”. In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI).* ACM/IEEE, 2020, pp. 73–82. DOI: 10.1145/3319502.3374786. URL: https://doi.org/10.1145/3319502.3374786.

[76] Raul Benites Paradeda et al. “How facial expressions and small talk may influence trust in a robot”. In: *Proceedings of the 8th International Conference on Social Robotics.* Vol. 9979. Lecture Notes in Artificial Intelligence. Kansas City, MO, USA: Springer, 2016, pp. 169–178. DOI: 10.1007/978-3-319-47437-3_17.

[77] Madan M. Pillutla, Deepak Malhotra, and J. Keith Murnighan. “Attributions of trust and the calculus of reciprocity”. In: *Journal of Experimental Social Psychology* 39 (2003), pp. 448–455. DOI: 10.1016/S0022-1031(03)00015-5.

[78] Daniel B. Quinn. “Exploring the Efficacy of Social Trust Repair in Human-Automation Interactions”. MA thesis. Clemson University, May 2018.

[79] David A. Robb et al. “Seeing eye to eye: trustworthy embodiment for task-based conversational agents”. In: *Frontiers in Robotics and AI* 10 (2023). DOI: 10.3389/frobt.2023.1234767. URL: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2023.1234767/full.

[80] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. “Timing is key for robot trust repair”. In: *Social Robotics.* Ed. by Adriana Tapus et al. Berlin: Springer, 2015, pp. 574–583. DOI: 10.1007/978-3-319-25554-5_58.

[81] Paul Robinette, Alan R. Wagner, and Ayanna M. Howard. *The effect of robot performance on human-robot trust in time-critical situations.* Technical Report GT-IRIM-HumAns2015-001. Georgia Institute of Technology, Institute for Robotics and Intelligent Machines, Jan. 2015.

[82] Denise M. Rousseau et al. “Not so different after all: A cross-discipline view of trust”. In: *Academy of Management Review* 23.3 (1998), pp. 393–404. DOI: 10.2307/259285.

[83] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach.* 3rd, Global. Pearson, 2016.

[84] Eduardo Salas, Dana Sims, and Shawn Burke. “Is there a “Big Five” in Teamwork?” In: *Small Group Research* 36 (Oct. 2005), pp. 555–599. DOI: 10.1177/1046496405277134.

[85] Maha Salem et al. “Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust”. In: *Proceedings of the International Conference on Human-Robot Interaction.* 2015, pp. 141–148. DOI: 10.1145/2696454.2696497.

[86] T. Sanders et al. “A Model of Human-Robot Trust: Theoretical Model Development”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* Vol. 55. 1. 2011, pp. 1432–1436. DOI: 10.1177/1071181311551298.

[87] Todd Sanders et al. “The relationship between trust and use choice in human-robot interaction”. In: *Human Factors* 61.4 (2019), pp. 614–626. DOI: 10.1177/0018720818816838.

[88] Rens van der Schoot et al. “A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research”. In: *Child Development* 85.3 (2013), pp. 842–860. DOI: 10.1111/cdev.12169.

[89] Maurice E. Schweitzer, John C. Hershey, and Eric T. Bradlow. “Promises and Lies: Restoring Violated Trust”. In: *Organizational Behavior and Human Decision Processes* 101.1 (2006), pp. 1–19. DOI: 10.1016/j.obhdp.2006.05.005.

[90] Sarah S. Sebo, Prashanth Krishnamurthi, and Brian Scassellati. ““I Don’t Believe You”: Investigating the Effects of Robot Trust Violation and Repair”. In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 2019, pp. 57–65. DOI: 10.1109/HRI.2019.8673304.

[91] Adriana Tapus, Maja J. Mataric, and Brian Scassellati. “Socially assistive robotics: The grand challenges in helping humans through social interaction”. In: *IEEE Robotics & Automation Magazine* 14.1 (2007), pp. 35–42. DOI: 10.1109/MRA.2007.339605.

[92] Edward C. Tomlinson and Roger C. Mayer. “The role of causal attribution dimensions in trust repair”. In: *Academy of Management Review* 34.1 (2009), pp. 85–104. DOI: 10.5465/amr.2009.35713291.

[93] John C. Turner et al. *Rediscovering the Social Group: A Self-Categorization Theory.* Oxford, UK: Basil Blackwell, 1987.

[94] {Anna Sophie} Ulfert and Eleni Georganta. “A model of team trust in human-agent teams”. English. In: *ICMI’ 20 Companion.* United States: Association for Computing Machinery, Inc., Oct. 2020, pp. 171–176. DOI: 10.1145/3395035.3425959.

[95] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. “Meaningful human control and variable autonomy in human-robot teams for firefighting”. In: *Frontiers in Robotics and AI* 11 (2024). DOI: 10.3389/frobt.2024.1323980. URL: https://www.frontiersin.org/articles/10.3389/frobt.2024.1323980.

[96] Ruben S. Verhagen et al. “The Influence of Interdependence on Trust Calibration in Human-Machine Teams”. In: *HHAI 2024: Hybrid AI Systems for the Social Good.* Ed. by Fabian Lorig et al. Vol. 386. Frontiers in Artificial Intelligence and Applications. IOS Press, 2024, pp. 300–314. DOI: 10.3233/FAIA240203. URL: https://doi.org/10.3233/FAIA240203.

[97] E. J. de Visser, R. Pak, and T. H. Shaw. “From “automation” to “autonomy”: The importance of trust repair in human–machine interaction”. In: *Ergonomics* 61.10 (2018), pp. 1409–1427. DOI: 10.1080/00140139.2018.1457725.

[98] E. J. D. de Visser et al. “Towards a theory of longitudinal trust calibration in human–robot teams”. In: *International Journal of Social Robotics* (2019). DOI: 10.1007/s12369-019-00596-x.

[99] Ewart de Visser et al. “Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents”. In: *Journal of Experimental Psychology: Applied* 22 (Aug. 2016). DOI: 10.1037/xap0000092.

[100] Sasha Wald, Kavya Puthuveetil, and Zackory Erickson. *Do Mistakes Matter? Comparing Trust Responses of Different Age Groups to Errors Made by Physically Assistive Robots.* Submitted to IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) 2024. 2024. DOI: 10.48550/arXiv.2408.13153. arXiv: 2408.13153 [cs.RO]. URL: https://arxiv.org/abs/2408.13153.

[101] Vincent R. Waldron. "Apologies". In: *Encyclopedia of Human Relationships.* Ed. by Harry T. Reis and Susan Sprecher. Vol. 3. Thousand Oaks, CA: Sage Publications, 2009, pp. 98–100.

[102] Sumio Watanabe. "A Widely Applicable Bayesian Information Criterion". In: *Journal of Machine Learning Research* 14 (2013), pp. 867–897. URL: http://www.jmlr.org/papers/volume14/watanabe13a/watanabe13a.pdf.

[103] Sumio Watanabe. "WAIC and WBIC for mixture models". In: *Japanese Journal of Statistics and Data Science* (2021). DOI: 10.1007/s41237-021-00133-z.

[104] Adam Waytz, Joy Heafner, and Nicholas Epley. "The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle". In: *Journal of Experimental Social Psychology* 52 (May 2014). DOI: 10.1016/j.jesp.2014.01.005.

[105] Joseph Weber, Deepak Malhotra, and James K Murnighan. "An attributional impetus model of trust development". In: *Advances in the2001 Academy of Management Best Papers Proceedings* (2001). DOI: 10.5465/apbpp.2001.6132979.

[106] Jessica L. Wildman et al. "Trust development in swift starting action teams: A multilevel framework". In: *Group & Organization Management* 37.2 (2012), pp. 137–170. DOI: 10.1177/1059601111434202.

[107] Jeanne M. Wilson, Susan G. Straus, and Bill McEvily. "All in due time: The development of trust in computer-mediated and face-to-face teams". In: *Organizational Behavior and Human Decision Processes* 99.1 (2006), pp. 16–33. DOI: 10.1016/j.obhdp.2005.08.001.

[108] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. "Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 2023. DOI: 10.1145/3544548.3581197.

[109] Kristine T. Wynne and Joseph B. Lyons. "An integrative model of autonomous agent teammate-likeness". In: *Theoretical Issues in Ergonomics Science* 19.3 (2018), pp. 353–374. DOI: 10.1080/1463922X.2016.1260181.

[110] Qihang Zhang, Xuandong J. Yang, and Lionel P. Robert. "What and when to explain? A survey of the impact of explanation on attitudes toward adopting automated vehicles". In: *IEEE Access* 9 (2021), pp. 159533–159540. DOI: 10.1109/ACCESS.2021.3130489.

[111] Xinyi Zhang et al. ""Sorry, it was my fault": Repairing trust in human-robot interactions". In: *Computers in Human Behavior* 139 (2023), p. 107531. DOI: 10.1016/j.chb.2022.107531.

A

Scripts

Round 1 — Teammate's Instructions

Teammate: Hello! I am your teammate for today. My name is Navel, and I'm really looking forward to working with you! I'll be giving you building instructions step by step. While I'm giving instructions, please just listen and follow along—I'll invite you to ask questions once I'm done with each set of steps.
Teammate: Are we ready to start?
[Breakpoint]
Teammate: First, let's get the pieces ready. Please collect these blocks for me:

- 2 long blue rectangular blocks
- 1 long yellow rectangular block
- 1 medium orange rectangular block
- 1 blue cube with a hole in the middle
- 1 green rectangular block with a half-arch cutout
- 1 yellow rectangular block with a half-arch cutout
- 1 big pink triangle
- 1 big orange triangle

Teammate: Do you have all of these pieces with you?
[Breakpoint]
Teammate: Perfect! Let's start building.

- Place the long yellow rectangular block horizontally on the surface.
- Place the medium orange rectangular block directly to the right of the yellow one, so they form a longer line together.
- Take the two long blue rectangular blocks and stand one directly next to the left side of the yellow block and the other one directly next to the right side of the orange block, like two tall pillars.
- Place the blue cube with the hole in the middle, vertically on top of the yellow and orange blocks, between the two blue pillars. Let the hole face you. Align it to the middle between the two pillars.

Teammate: Do you have any questions so far?
[Breakpoint]
Teammate: Great! Let's continue.

- Place the green half-arch block on top of the left blue pillar and the blue cube. Make sure the arch faces downward, like a little bridge. Align it fully with the blue cube such that no parts float in the air.
- Place the yellow half-arch block on top of the right blue pillar and the blue cube, also facing downward.
- Please make sure the two arches meet neatly in the middle, side by side.
- Finally, put the pink triangle on top of the green arch, and the orange triangle on top of the yellow arch, so it looks like two little rooftops!

Teammate: That's the last step. Did you miss any part, or would you like me to guide you through something again?
[Breakpoint]
Teammate: Wonderful! Thank you so much for building with me in this first round, you did really well! Let's look at the results together!

Round 2 — Teammate's Instructions

Teammate: Hello again! Let's get ready for our next build. Please collect the following blocks for me:

- 2 long blue rectangular blocks
- 2 medium pink rectangular blocks
- 1 medium orange rectangular block
- 1 green cube with a hole
- 1 green half cylinder (it looks like a half moon)
- 1 short green cylinder
- 1 short yellow cylinder

Teammate: Do you have all of these pieces with you?

[**Breakpoint**]

Teammate: Perfect! Let's start building.

- Place the short yellow cylinder laid down on the surface, with the circular face towards you.
- Place one long blue rectangular block horizontally, directly to the right of the yellow cylinder.
- Place the short green cylinder laid down directly to the right of the blue block, with the circular face towards you.
- Make sure everything forms a long connected line and that all pieces are roughly the same height.

Teammate: Do you have any questions so far?

[**Breakpoint**]

Teammate: Great, let's continue!

- Place the second long blue rectangular block horizontally on top of the yellow cylinder. Make sure it covers both the yellow cylinder and the long blue rectangular block beneath. The left edge of this blue block should fully cover the yellow cylinder, but not extend further to the left.
- Place one medium pink block horizontally on the right side, next to the blue block you just placed. Make sure the pink block fully covers the green cylinder beneath, but does not extend further to the right. The pink block should also overlap slightly with the blue block below.
- Together, the blue and pink blocks on this layer should form a connected line.

[**Breakpoint**]

Teammate: Now for the top layer.

- Place the green cube with the hole standing on top of the blue block of the second layer. The hole should face you. The right edge of the cube should align exactly with the right edge of the blue block below.
- Place the medium orange block directly on top of the pink block, aligned perfectly.
- Place the second medium pink block on top of the orange block, also aligned perfectly.
- Finally, place the green half cylinder horizontally on top of the blue block of the second layer.

Teammate: That's the final step! Did you miss any part, or would you like me to repeat a section?

[**Breakpoint**]

Teammate: Wonderful! You did a great job completing this round. Let's take a look at the final figure together!

Round 3 — Teammate's Instructions

Teammate: Hello again! Let's get ready for our final build. Please collect the following blocks for me:

- 2 medium pink rectangular blocks
- 2 long blue rectangular blocks
- 2 long yellow rectangular blocks
- 2 short yellow cylinders
- 1 green half cylinder (it looks like a half moon)

Teammate: Do you have all of these pieces with you?

[**Breakpoint**]

Teammate: Perfect! Let's start building.

- Place one medium pink rectangular block horizontally on the surface.
- Place one long blue rectangular block horizontally, directly to the right of the pink block.
- Place the second medium pink rectangular block horizontally, directly to the right of the blue block.
- Make sure all three blocks form one long connected line and that they are the same height.

Teammate: Do you have any questions so far?

[**Breakpoint**]

Teammate: Great, let's continue!

- Place one long yellow rectangular block horizontally on top of the first layer.
- Place the second long yellow rectangular block horizontally, directly to the right of the first yellow block.
- Make sure the two yellow blocks form one long connected line of equal height and are centered over the first layer.

[**Breakpoint**]

Teammate: Now let's finish the structure.

- Place the second long blue rectangular block horizontally in the middle of the yellow line.
- Place the two short yellow cylinders standing upright on top of the blue block. The cylinders should stand side by side, directly touching each other, and positioned in the center of the blue block.
- Finally, place the green half cylinder horizontally on top of the two yellow cylinders so that it covers both cylinders—it should look like a helmet.

Teammate: That's the final step! Did you miss any part, or would you like me to repeat a section?

[**Breakpoint**]

Teammate: Wonderful! You did a great job completing this round. Thank you for being my teammate throughout all the builds. Let's take a look at the final figure together!

B

Participant Instructions and Debriefing

Opening Speech

Hello, and welcome to my experiment! This study looks at how trust develops when people work with either a human or a robot teammate. Today, you will be paired with a [robot/human] teammate to complete a cooperative task.

Your task will be to build figures based on your teammate's instructions. To make it more interesting, you will not see the reference photo yourself, so you will have to rely only on their instructions. At the same time, your teammate will not be able to see your progress. This means you will need to cooperate closely to complete the task together.

For the experiment, I ask you to follow three important rules:

1. Only ask questions when your teammate has indicated they are ready to receive them.
2. Please ask your question in one continuous sentence without pauses or breaks. Think carefully before speaking, as any gaps or delays may cause your question not to be recognized or answered by your teammate.
3. Please avoid sharing personal information about yourself. Small talk is fine, but try not to mention details that could identify you.

If you would like your teammate to move faster, you may let them know when you are ready for the next step. Otherwise, they will continue at a fixed pace.

You will complete three rounds in total, each with a different figure to build. After each round, we will measure your trust in your teammate using short questionnaires. The conversations will also be recorded and transcribed for analysis.

If you are happy to take part, please read and complete the informed consent form before we begin. Feel free to ask me any questions you may have.

Debriefing after the Experiment

I did not mention this during the experiment, but you may have noticed that your teammate gave you an incorrect instruction in the second round. This was intentional and part of the study design. The purpose was to examine not only trust formation, but also how trust is affected by a violation and whether it can be repaired afterwards.

I understand this may have caused some frustration or made you feel uncertain about your performance. Please know that the mistake was planned and not a reflection of your ability to follow instructions. It was an important element of the study to help us measure trust more accurately.

C Surveys

C.1. Informed Consent Form

Comparing Trust Development in Human and Robot Collaboration

You are being invited to participate in a research study titled *Comparing Trust Development in Human and Robot Collaboration.* This study is being conducted by Ching Tsu Guo from TU Delft.

The purpose of this research is to investigate differences in how trust develops between human and robot teammates. Participation will take approximately 30 minutes. The data will be used for the completion and publication of a Master's thesis. During the study, you will cooperate with either a human or a robot teammate (randomly assigned) to complete a task in which you build figures based on instructions provided by your teammate.

As with any online activity, there is always a minimal risk of a data breach. To the best of our ability, your responses in this study will remain confidential. Audio recordings of the task will be directly transcribed, and any personal identifiers (such as names, gender, or other details that could reveal your identity) will be removed from the transcriptions. Data (only the transcriptions and questionnaires) will first be stored temporarily on Microsoft Forms and then moved to TU Delft's Project Data Storage, where it will be kept securely for one year after the conclusion of the study.

Your participation in this study is entirely voluntary. You may withdraw at any time without consequence, and you are free to omit any questions. However, once data has been anonymized, it will no longer be possible to remove individual responses.

If you have further questions, you can contact me.

By reading this statement and agreeing with all the questions, you indicate your consent to participate in the experiment, including the use of anonymized transcriptions of your conversations and your responses to questionnaires during the study.

General Agreement

1. I have read and understood the study information, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.
3. I understand that taking part in the study involves having my audio conversations with the teammate transcribed and completing questionnaires after each round of the experiment.
4. I understand that taking part in the study involves the risk of potential identification if I disclose personal information during the task conversation. I also understand that this risk will be mitigated by the responsible researcher, Ching Guo, who will clean the transcripts and remove

any personal details, as well as by my own effort to focus on the task and avoid sharing personal information.

5. I understand that taking part in the study also involves collecting limited demographic information (age and, if I wish, gender), which may be considered personally identifiable research data (PIRD). These data will only be used to describe the overall demographics of the study. All responses will be anonymized using participant codes (e.g., Participant 001), which minimizes the risk of my identity being revealed.
6. I understand that after the research study the de-identified information I provide will be used for the analysis and reporting of a Master's thesis, and may also be included in academic presentations or publications based on this research.
7. I give permission for the de-identified transcriptions and questionnaire data that I provide to be archived temporarily in Microsoft Teams before being moved to TU Delft Project Data Storage, where it will be securely stored and may be used for future research and learning.

By checking "Yes" on all the above statements, I consent to participate in the study.

C.2. Pre-Study Survey (Demographics)

To better understand the composition of our participant group, we asked a few short demographic questions. These included gender and age range. Responses were anonymous and only used for group-level analysis.

- **Gender:**
  - Woman
  - Man
  - Non-binary
  - Prefer not to say
- **Age Range:**
  - 18–22
  - 23–27
  - 28–34
  - 35–44
  - 45+
  - Prefer not to say

C.3. Interpersonal Trust Questionnaire

Participants completed this questionnaire after each round of the experiment to assess their interpersonal trust toward the teammate. Responses were measured on a 7-point Likert scale ranging from *Strongly disagree* to *Strongly agree*.

Trusting Belief – Benevolence

1. When it comes to my well-being, my teammate really cares.
2. If I required help, my teammate would care enough to help me.
3. I believe that my teammate cares enough to act in my personal interest.
4. When you get right down to it, my teammate cares about what happens to me.

Trusting Belief – Competence

1. My teammate is skillful and effective in giving building instructions.
2. My teammate explains the block-building steps very well.
3. Overall, I have a capable and proficient teammate.

4. Overall, my teammate is competent at his task.

Trusting Intention

1. When an issue that is critical to our building task arises, I feel I can depend on my teammate.
2. I can always rely on my teammate in a task-related issue.
3. My teammate is a person on whom I feel I can rely when the issue is important to my task.
4. I feel I can depend on my teammate on a task-sensitive issue.

C.4. Open-Ended Questions

After each round, participants could provide feedback or reflections about their collaboration experience.

- Do you have any comments or feedback regarding this round and the collaboration?