Measure and Manage Trust in Human-AI Conversations

By

Mengyao Li

A dissertation proposal submitted in partial fulfillment of
requirements for the degree of

Doctor of Philosophy
(Industrial and Systems Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON
2023

Date of final oral examination: 06/30/2023

The dissertation is approved by the following members of the Final Oral Committee:

John D. Lee, Professor, Industrial and Systems Engineering

Douglas Wiegmann, Professor, Industrial and System Engineering

Anthony McDonald, Assistant Professor, Industrial and Systems Engineering

Yea-seul Kim, Assistant Professor, Computer Science

Acknowledgments

I would like to thank my advisor, Prof. John Lee. Throughout my Ph.D. journey, his invaluable guidance, countless stimulating conversations, and constant encouragement have had a profound impact on me. Even before starting my Ph.D., I held John in the highest esteem, and being able to work under his mentorship has been an absolute privilege for me. John has not only taught me how to approach research critically but has also exemplified what it means to be a good scholar within the academic community. His support for others' ideas and willingness to freely share his own have shaped my understanding of what it means to contribute meaningfully to the field. His humility, strong work ethic, and genuine care for his students have created an environment where I have felt encouraged to explore new ideas and grow both personally and professionally. As I embark on my independent journey as an assistant professor, I want to emulate his scholarship and mentorship when guiding and nurturing the next generation of scholars under my care. Further, I would also like to thank my committee members, Drs. Douglas Wiegmann, Tony McDonald, and Yea-seul Kim for their valuable time and thought-provoking questioning during my dissertation process. Their thoughtful inputs have immensely shaped my work and improved my dissertation throughout different stages of my Ph.D.

I am incredibly fortunate to have had the support of inspiring mentors throughout my academic journey. I would like to express my deepest gratitude to Dr. Brittany E. Holthausen and Prof. Bruce Walker, who served as mentors during my undergraduate studies at Georgia Tech. Brittany's support and belief in my abilities played a pivotal role in my decision to start this doctoral journey. She constantly reminds me how to be a good mentor. I am truly grateful for her guidance and encouragement, which have instilled in me the confidence to pursue my academic aspirations from the very beginning and throughout this entire process.

The work presented in this dissertation was truly a collaborative effort. I would like to acknowledge my collaborators, E. Vince Cross and James Garrett, for their endless support for the multisite study in this dissertation. I would also like to extend my thanks to Shashank Mehrotra, Kumar Akash, and Teruhisa Misu for their invaluable contributions in shaping my research thinking. During my internship at Honda, their guidance and mentorship extended far beyond the confines of the workplace, leaving a lasting impact on my academic pursuits.

One of the most rewarding experiences in my Ph.D. journey is surrounded by so many like-minded researchers in our Cognitive Systems Laboratory. I would like to thank the following individuals whose camaraderie and support have made this experience truly memorable: Varshini

Kamaraj, Hansol Rheem, Priya Loganathar, Vianney Renata, Joombum Lee, Soyeon Kim, Jah'inaya Parker, Xizi Xiao, Katherine Woodruff, Atefeh Katrahmani, Josh Domeyer, Areen Alsaid, Azadeh Dinparastdjadid. In particular, I want to express my heartfelt gratitude to Isabel Erickson, Sofia Noejovich, Joshua Emory, and Dong Fang for their invaluable contributions as research assistants in supporting the research work presented in this dissertation. Without their exceptional assistance, I could not have navigated the intricacies of my study.

I am extremely grateful to my family and friends for their support in my pursuit of a Ph.D., despite the long-distance and my limited involvement in their lives over the past four years. Even though my parents did not have the opportunity for higher education, their financial support and belief in my aspirations have been invaluable. I deeply appreciate their understanding and sacrifices, and I am truly fortunate to have such supportive family members.

Finally, I want to thank my multitasking superhero: my website developer, video editor, chef, at-home comedian, Uber driver, cheerleader, and of course, my amazing husband, Zhixiang Ren. His exceptional website development skills were crucial in Chapter 6 for completing this dissertation. His unshakable belief in me has helped me go through the challenging conditions I found myself in along this journey. Thank you, Ren, for being my rock and constantly reminding me to find joy in every moment throughout this journey.

Abstract

Artificial Intelligence (AI), with its increasing capability and connectivity, extends beyond limited and well-defined contexts and is integrated into broader societal domains. Examples include AI algorithms controlling large fleets of autonomous vehicles, news filtering algorithms influencing people's political beliefs and preferences, and algorithms mediating resource allocation and labor (Bubeck et al., 2023; Rahwan, 2018). The relationship between humans and AI has evolved from mere supervisory control to interdependent cooperation on a larger scale, yielding significant societal benefits (Endsley et al., 2021). To better support human-AI cooperation, the establishment of a trusting relationship between humans and their AI teammates becomes increasingly critical (Chiou & Lee, 2021). Trust plays a vital role in shaping how people use, communicate with, and cooperate with AI systems. Therefore, the measurement and management of trust in human-AI cooperation are essential to ensure the safety, effectiveness, and overall is to measure and manage trust in human-AI conversations and cooperation, addressing three primary questions: (1) How can we measure people's trust in human-AI conversations? (2) How does trust change over time within human-AI conversations? (3) How can we effectively manage instances of overtrust or undertrust through conversational cues to enhance human-AI cooperation?

To tackle the initial question regarding the measurement of trust in human-AI conversations, machine learning models were developed to predict trust using lexical and acoustic features. While most machine learning approaches are often treated as opaque "black boxes," an inferential machine learning method was adopted, enabling the visualization of the most influential features. Moving on to the second question, which explores the temporal dynamics of trust in human-AI conversations, a dynamic system model was employed to explain the trust divergence. Adopting this temporal trust dynamic perspective, a mixed-method approach called trajectory epistemic network analysis showed the evolution of trust dimensions throughout human-AI conversations, revealing distinct patterns in conversational topic diversity and flow over time. Finally, to manage trust for effective cooperation on a societal scale, the research scope expanded from performance-based calibration to purpose-based cooperation. The identified conversational trust indicators were demonstrated to be adaptive countermeasures to repair trust in human-AI cooperation. The findings highlight the importance of addressing purpose-based trust violations and contribute to our understanding that trust can be both measured and managed through human-AI communications, which can be served as an unobtrusive, real-time means of trust measurement and management in human-AI cooperation.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

As AI technology continues to progress, its potential to surpass human performance in numerous domains becomes increasingly likely. Such a development could be as transformative as the Industrial Revolution, with significant economic, social, and political implications. With this possible emergence of artificial general intelligence (AGI) (Goertzel, 2014), we are more entwined with AI and autonomous systems in a myriad of domains and tasks (e.g., space missions, military operation, cooperative autonomous driving, aviation, medical diagnosis). Symbiotic human-machine relationships, as suggested by J. C. R. Licklider in the 1960s (Licklider, 1960), have become increasingly likely. Recent research often described these relationships as a human-AI team (HAT), humans and AI cooperate interdependently to achieve a joint goal and can provide beneficial outcomes to the symbiosis of two bodies. HATs suggest that AI and autonomous agents move beyond tools and become teammates (Endsley et al., 2021). In the near future, we can expect human-AI relationships where humans shape AI and are also shaped by AI (Rahwan et al., 2019). To better understand and design the interdependent HAT, designing a trusting relationship between humans and the AI teammate becomes more critical (Chiou & Lee, 2021).

Trust has been defined as 'the attitude that an agent will help achieve a person's goals in a situation characterized by uncertainty and vulnerability' (Lee & See, 2004, p. 51). Decades of research have shown that trust has been an important construct to explore and explain why and how people use, misuse, or disuse automated systems (Parasuraman, 1997). Trust has been identified with three core bases: performance, process, and purpose (Lee & See, 2004). Performance refers to the capability and competency of the system; process refers to the mechanism and algorithms used to accomplish its objectives; purpose refers to the design intent and objectives of the system. Prior literature has primarily focused on the performance- and process- based of trust (de Visser et al., 2020), but the purpose-based aspect of trust should be highlighted more in the context of the HAT. In this relationship, the shared goal between teammates may not always be aligned, especially when there are multiple agents with different objectives involved in a complex team structure. Thus, understanding the signals of aligned or misaligned goals in HAT becomes crucial.

Signals of trust should be continuously measured and managed to reflect the teaming processes and how the team activity unfolds over time. This requires a continuous and observable stream of data to record the cognitive processing of trust evolving in the human-AI team. Communication, as a form of team cognition, can provide such contextual and process-based means for trust

modeling (Cooke et al., 2013). However, there is little research that focuses on trust in conversations (O'Neill et al., 2022). For conversations, the HAT would have more frequent information and signal exchanges in a joint task. Although communication plays a vital role in driving HAT success, measuring trust via communication is still a new approach. Communication can manifest conscious and subconscious mental states. Trust, which reflects both analytic and affective processes, can be analyzed and measured via communication (Lee & See, 2004). Prior literature on HAT usually uses communication patterns such as communication rates and flows to predict trust (Bromiley & Cummings, 1995). Limited research has focused on communication content for trust measurement. Additionally, conversations naturally unfold overtime, which can reflect the trust dynamics, rather than a snapshot view of trust captured by the subjective measurements. However, the means of capturing trust from real-time communication and long-term trust dynamics have not been studied.

After measuring trust, the next step is to address how to effectively manage it. For a system to be deemed trustworthy, it must adapt to the user's trust level, considering both over and under trust. Prior research on trust management has primarily focused on performance-based trust during the calibration process, which involves accurately matching a person's expectations for the system's capability. However, when managing trust in HAT, the focus should expand to the purpose dimension. This is because AI is increasingly joining human teams and engaging in more social interactions, leading to trust violations that can arise from a misalignment of intentions and values within the team, in addition to the system's lack of capability.

Furthermore, the previous literature has not adequately aligned the type of trust violation with the appropriate management strategies. Although considerable effort has been made to understand trust management strategies, including trust repair and damping behaviors to increase or decrease trust, there is a lack of correspondence between the appropriate strategy for different types of trust violations.

Another important consideration is whether the identified conversational indicators can be used as conversational indicators of the agent to manage trust. Conversational indicators refer to conversation elements that convey uncertainty and confidence, enabling people to probe each other's uncertainty and confidence. It is also essential to examine the congruency between the trust management content and acoustic conversational indicators. Thus, it is crucial to investigate the casual linkages between the type of trust violation, type of management strategy, and their impact on different dimensions of trust.

1.1 Dissertation Objectives

For my dissertation, the objectives are twofold: measure and manage trust in communication and cooperation. In particular, I focus on measuring trust in communication by taking into account the temporal dimension, encompassing both real-time measurement and long-term dynamics. Trust plays a mediating role in facilitating cooperation. Additionally, I aim to manage trust in cooperation by considering the structural dimension of team interdependence and goal alignment. From a temporal perspective, my research covers from real-time trust measurement to the study of long-term trust dynamics. From a structural perspective, I investigate trust within the context of multiple goals involving both AI and humans in team compositions using game-theoretic scenarios.

1.1.1 Objective 1: Measure Trust in Communication: From Real-Time Estimation to Long-Term Dynamics

Communication is critical to team cognition because it mediates team results and affects people's trust in their AI teammates (Cooke et al., 2013). Conversational data can be considered as a mixture of behavioral and physiological data, which contains lexical, semantic, phonological, and pragmatic representations of conversations. People naturally express their trust attitudes through the words they used, the sentence structure, and the tone of the voices in their conversation, which are all contextualized. According to the interactive team cognition theory, communication is team cognition, which can be a nonobtrusive measure of team interaction dynamics (Cooke et al., 2013). Communication is also essential for trust building and calibration, which in turn can promote effective human-AI teaming (Fuoli & Paradis, 2014). Thus, understanding important indicators of trust becomes interestingly important in HAT.

In addition, the conversation naturally holds temporal functions of coordination based on the structure of the turn of talks, which can show changes in human-AI relationships over time. Because trust is time-dependent and evolves throughout human-agent interactions (Kaplan et al., 2021). Trust calibrates and evolves based on the various automation characteristics and experiences as relationships between parties mature (Korsgaard et al., 2018a; Luo et al., 2022). Trust is reinforced by the experience and is further impacted by a function of the trust itself in the previous moment (e.g. positive and negative feedback loops) (Falcone & Castelfranchi, 2004; Lee & Moray, 1992; Manzey et al., 2012). Additionally, adding the temporal aspect allows us to examine the recency effect that associated with trust dynamics, meaning that interactions happened more recently may have more value than those that happened some time back (Desai et al., 2012). Thus,

analyzing and modeling the temporal changes gives a more nuanced inspection of the trust evolution throughout the HAT.

To achieve the first objective of measuring trust in communication, I explored the following two research questions:

RQ1: How to measure people's trust in the human-AI conversation?

To address this question, it is important to first verify that trust can be estimated from conversations and then identify the important features as metrics for measurement. To estimate trust from conversations, it is necessary to first elicit utterances by designing trust-relevant situations with appropriate conversational prompts during the human-AI interaction. Thus, we developed a trust lexicon and a general framework on how to design appropriate conversational prompts (Alsaid et al., 2022; Li et al., 2020). Once we elicit trust-relevant conversations, we can process and analyze the conversational cues to estimate people's trust. According to the well-known phrase, "It's not only what you say, but also how you say it.", both the words and how they are said should convey trust. Therefore, we consider not only lexical cues (e.g., words used), but also acoustic cues (e.g., pitch, formants) (Elkins & Derrick, 2013; Johnson et al., 2014). To estimate trust, subjective trust ratings were predicted using machine learning models trained in three types of conversational features (i.e., lexical, acoustic, and combined). After training, model explanation was performed using variable importance and partial dependence plots. The model explanation methods allow us to identify the important conversational indicators of trust. Our approach showed such real-time, conversational trust measures are possible by training machine learning models on lexical, acoustic, and combined conversational features.

RQ2: How does trust change over time in the human-AI conversation?

Solving the first research question establishes the possibility of estimating trust in conversation; however, the approach ignores two aspects: first, trust is dynamic, which means that people calibrate their trust over time as a continuous cognitive process. Second, trust estimation often focuses on the feature level and ignores the rich context and deep meaning of the conversation. In other words, the connections between the features and the meaning associated with features are situated within the context that might benefit from qualitative analysis. Furthermore, the temporal changes of trust in conversation cannot be captured. Thus, the second research question aimed to capture trust dynamics, which is the temporal aspect of trust evolution throughout the interactions, rather than aggregated or a snapshot of trust.

1.1.2 Objective 2: Manage Trust in Cooperation: From Performance to Purpose-based Trust

Once trust is measured, the next question is how to properly manage trust. For a system to be trustable, it will have to adapt and manage the user trust level, that is, over/under trust. Therefore, the objective is to evaluate how conversational indicators can be used as adaptive countermeasures by a virtual assistant to manage trust. Managing trust is not a novel topic. Previous research often identifies causal relationships between a single factor or a combination of antecedents of trust and investigates its impact on trust in experimental studies (Hoff & Bashir, 2015). However, as trust becomes more dynamic in HAT, it is important to define the trust management framework that captures the interdependency in the team. When measuring trust and modeling trust dynamics in conversation, research focuses on the performance and process dimension of trust yet neglects the purpose dimension of trust. Therefore, the goal of this chapter is to evaluate how conversational indicators can be used as adaptive countermeasures by a virtual assistant to manage various dimensions of trust. To achieve the objective of managing trust for human-AI cooperation, I explored the following research questions.

RQ3: How to manage people's over-trust or under-trust under different types of trust violations?

When designing trust management strategies, it is important to consider the types of violations (i.e., performance, purpose), manage trust bidirectionally (i.e., repair and dampen), and measure the effects subjectively and behaviorally. To address RQ3, I conducted a mixed design study to investigate the impacts of trust management content on trust dimensions. We hypothesized that people would have a higher drop in trust when it is a purpose-based trust violation. The apology paired with promise would better repair the trust violation.

RQ4: How to design conversational indicators of agent to manage people's trust?

To address RQ4, we designed the trusting voice with the identified the appropriate content identified in Study 1 to further investigated the effectiveness of the acoustics cues on trust management. A trusting voice, often perceived as a happy sounding voice, can either promote or hinder trust repair efforts, depending on the congruency between the voice and the content. The positive congruency indicates it is more effective to pair a highly trusting voice with trust repair; the negative congruency indicates that a low trusting voice can better repair trust, since it conveys more sincerity and remorse. The direction and effects of the congruency effect should be closely exanimated in the context of trust management. We hypothesized that when the agent's

trustworthy voice shows the positive congruency (i.e., high trusting voice with trust repair content), it is more effective to manage people's trust.

1.2 Dissertation Outline

My dissertation followed a two-stage approach: measure and manage trust in the human-AI team. In the first part, measuring trust, I focused on the temporal dimension of human-AI communication from the short-term trust estimation to the temporal dynamics of trust. Chapter 3 investigated the trust estimation on human-AI communication based on the microlevel of acoustic and lexical features. Chapter 4 and Chapter 5 extended the understanding of trust dynamics. Chapter 4 showed that trust can be best modelled by a dynamic system perspective and Chapter 5 adopted this concept and modelled the meso-level of the conversational topics and their temporal changes.

In the second part, managing trust, I highlighted the structural dimension of human-AI cooperation by considering interdependence and goal alignment issues. Following the human-AI cooperation and communication framework, Chapter 6 presented two experimental studies incorporating findings from part one trust measurements. For Study 1, I investigated trust management content for different dimensions of trust. For Study 2, I investigated the effects of acoustic cues on trust and its congruency with the contents. I designed the trusting voice with the identified the appropriate content identified in Study 1 to further investigate the effectiveness of the acoustics cues on trust management. This chapter provided a better understanding of the appropriate trust management strategy for different type of trust violation, especially whether the identified acoustic cues of trust can dampen and repair different dimensions of trust.

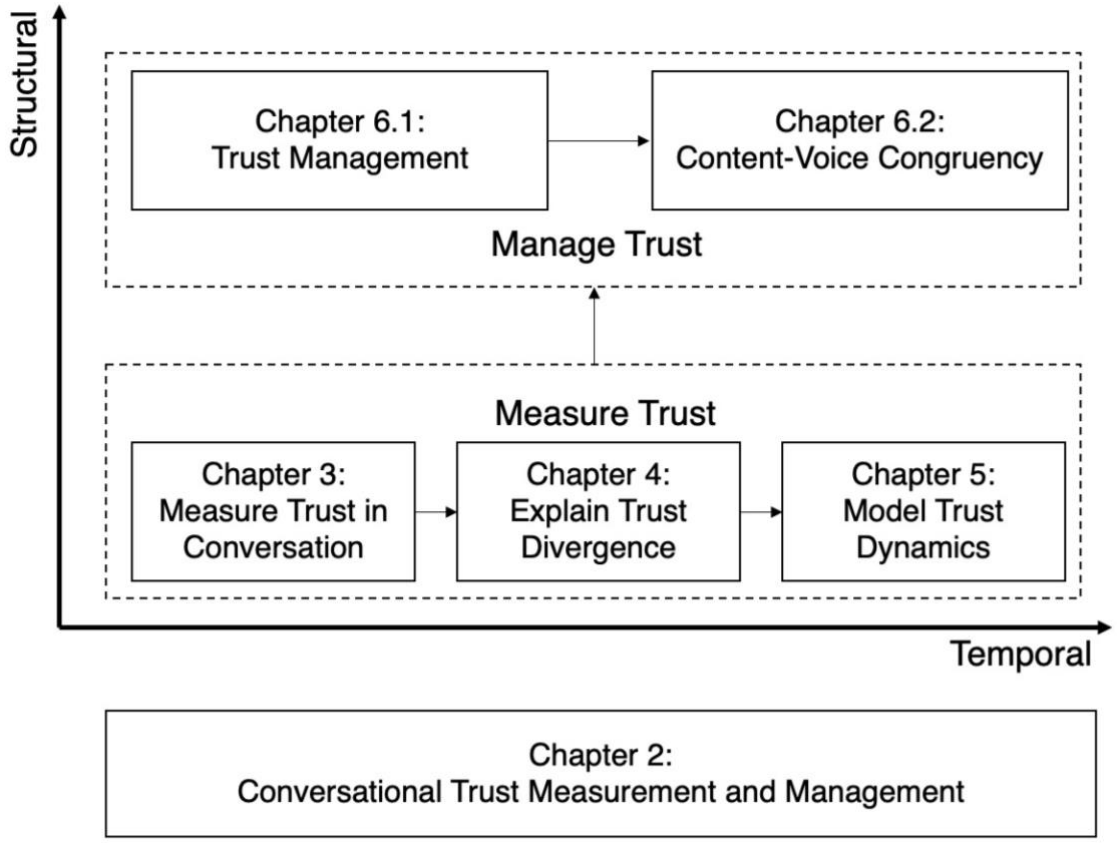


Figure 1. Overview of the Dissertation.

1.3 Contributions

1.3.1 Theoretical Contribution

The dissertation made a theoretical contribution by measuring and modeling trust processes in human-AI communication. My work conceptualized trust in communications and identified critical conversational indicators that are essential to predict trust. In addition, the work developed the temporal dynamics of trust processes in communications, which provides a deeper understanding of how trust evolves over time.

The dissertation also extended the concept of trust beyond performance to purpose-based interaction and examinated the goal alignment in human-AI cooperation. In particular, I abstract the potential performance- and purpose-based trust interactions as a game-theoretic situation. By integrating the Trust Game and Threshold Public Goods Game, the newly developed environment allows researchers to capture impacts of the performance- and purpose-based trust violation on people's decision makings. Findings from game can be applied to human-AI teams in future hybrid societies, especially when faced with the conflicts between individual and collective benefits.

Overall, the dissertation's contribution to the field of communication provided a more comprehensive understanding of trust processes and highlights the importance of effective communication in building and maintaining trust.

1.3.2 Practical Contributions

Conversational agents and other types of AI-based agents represent an important opportunity to extend human capabilities, but only if they are accepted and trusted appropriately. This dissertation also provided practical principles and guidelines for designing trustworthy conversational agent and how to better measure and manage the ongoing dynamic relationships.

This work also provided practical contributions in terms of methodological implications in measuring latent variables. We developed a machine learning pipeline that enables the quantification of latent constructs such as trust, workload, and situational awareness. This pipeline provides a more objective, non-intrusive, and efficient way of measuring these subjective concepts in communications and other continuous data streams.

Chapter 2. Conversational Trust Measurement and Management

2.1 Human-AI Teaming (HAT)

In 'Man-Machine Symbiosis', J. C. R. Licklider (1960) originated the concept of human-machine symbiosis, which depicts a partnership in that human brain and computing agents can cooperate tightly and effectively and providing beneficial outcomes to the two bodies of symbiosis. As Johnson and Vera (2019) discussed, 'no AI is an island', the human-AI teaming perspective is essential to reach the full potential of humans and AI. As increasingly autonomous machines and AI are introduced into domains such as space missions, military operations, autonomous driving, aviation, medical diagnosis, and so on, the research in the effective human-AI symbiosis becomes more urgent and essential.

The traditional view on MABA-MABA (Men are better at - Machines are better at), which holds the replacement perspective by deciding the functional allocation by substitution, cannot satisfy the increasing needs for an effective human-AI team. Since in these domains, uncertainty, risks, and time pressure often require members to make effective real-time coordination of goals and actions within a changing environment (Wintersberger, 2020). The high level of autonomy and capabilities are not equal to simpler situations for the human, whereas usually the opposite is true (Johnson & Vera, 2019). Simply improving the capabilities of the intelligent agent would not be enough to promise a safe and successful space mission. An example is NASA's work on an AI-based activity planner from 2002. The system added an optimization scheduling engine to consider associated restrictions and produced an optimal plan for the activities of the day with a single button. However, the optimizing engine was almost removed from the mission due to the difficulty of modification, manipulation, and validation of human members (Johnson & Vera, 2019). Under highly uncertain and time-critical tasks, psychological constraint of time and risks would negatively affect the ability of members to understand and choose the optimized alternatives recommended by the agent. The more intelligent the technological system, the greater the need for collaborative skills between two parties, such as real-time decision making, adaptive task allocation, and goal alignment. Therefore, it requires a relationship shift from a typical vertical (supervisor-subordinate) control to a horizontal (peer-to-peer collaboration) interaction (Chiou & Lee, 2016; Trafton et al., 2006).

A vertical interaction means the supervisor with more knowledge of tasks gives directions and suggestions to subordinates to implement actions. Then the supervisor makes the diagnosis and

takes further actions of the system. The capabilities of the intelligent agent become increasingly advanced, such as allocating resources to accomplish a goal (Truszkowski & Hallock, 1999) or overriding human operators' actions when the latter may jeopardize safety. This situation makes the defined role of human as supervisors blurred, which means in some situations, an intelligent agent has better ability to monitor and infer the state of systems and make suggestions to human to follow. As the level of autonomy increases, the agent can allow the human a restricted time to veto before the automatic execution of a recommendation or executes automatically and informs the humans when necessary (Parasuraman & Wickens, 2000).

A horizontal interaction depends on reciprocity and the ability to share resources to adapt to unexpected demands. With a high level of autonomy, agents can interpret, reason, and optimize decisions in response to operators, the environment, and the objectives of the task itself collaboratively. When the agent makes the decision based on a global optimum, with limited capacity (e.g., knowledge, skills, abilities, and resources), the person representing the local optimum could perceive this action as competitive behaviors (Sanders et al., 2011). Based on the poor understanding and adjustment of the interdependence between human and agent, under trust might develop, leading to disuse of the intelligent agent, which could contribute to catastrophic failures.

2.2 From Trust to Trusting

The concept of trust in automation has been the focus of substantial research over the past several decades (Hoff & Bashir, 2015). Trust in automation is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004). With the increasingly capable automated systems and artificial intelligence (AI), the relationships between human and AI have shifted from supervisory control to a collaborative team. Trust plays a more crucial role in determining the success of such partnerships. The change of the HAT relationship suggests a need to understand the interdependent human-AI teaming (Maier, 1967).

With the increasing capability of automation and AI in a more horizontal relationship with humans, a shift is needed: from a traditional static trust to a relational and dynamic trusting process. Chiou and Lee (2021) proposed a conceptual framework of trusting in automation, which highlighted the relationship between the goal environments and the four core interaction considerations (i.e., situation, semiotics, sequence, and strategy). The situation captures decision points where trusting and trust calibration are critical, such as the experimental design on the

interdependence of trust-related choices, incentives, outcomes of choices, and the array of potentially competing goals. Strategy describes the action space in the trust situations that promote cooperative joint action and mutual trusting. Sequence captures the temporal element whether decisions are made synchronously, asynchronously, in order, or in free form. Semiotics captures the type of information conveyed during interactions, such as communication. In this dissertation, we focused on the semiotic interaction, which is the communication between human and AI, and extended the other considerations in two directions: temporal dynamics and structural interdependence.

2.3 Temporal Dynamics

An important aspect of trust dynamics that deserves more attention is its temporal characteristics. A shift from the snapshot view of trust to a dynamic view of trust is important (Yang, Schemanske, et al., 2021) as trust is time-dependent and evolves throughout the human-agent interaction (Kaplan et al., 2021). The evolution of trust depends on various automation characteristics and experiences as relationships between teammates mature (Korsgaard et al., 2018a; Luo et al., 2022). Trust is reinforced by experience and is further affected by previous levels of trust (Falcone & Castelfranchi, 2004; Lee & Moray, 1992; Manzey et al., 2012). Examining this temporal trust dynamics allows us to assess the influence of the recency effect, meaning that interactions that occurred more recently may have more influence than those that occurred earlier (Desai et al., 2012). Thus, analyzing and modeling the temporal changes gives a more nuanced inspection of the trust evolution throughout the HAT.

To model the evolution of trust, trust should be measured multiple times and further modeled by considering time units in the model. Yang et. al. proposed a computational model proposes that trust at any time t, follows a Beta distribution, which shows good prediction accuracy (Yang, Schemanske, et al., 2021). Although modeling trust evolution is relatively new and limited, there is a long history of modeling human behaviors and attitudes with a time-dependent dynamical system approach. Gottman, Swanson, and Swanson (2002) showed how marriage outcomes can be modeled using the dynamical system analysis, which focused on the temporal dynamics of partner communication. Using such non-linear dynamical systems methods to model relationships is becoming more prevalent (Demir et al., 2021).

2.4 Structural Interdependence

Increasingly, we are seeing the emergence of human-AI teams, where people and intelligent machines work collaboratively toward a common goal. As a human-AI team, to function

effectively, it is important to recognize the *structural interdependence*. This means that the performance of one team member is influenced by the actions and decisions of the other. The relationships also shift from a typical dyad of one-to-one human agent interaction to a multi-agent cooperation in a hybrid team. A central consideration is how the patterned relationships among multiple members jointly affect network members' behavior. Hence, it is not assumed that network members engage only in multiple dyads with separate members. The highlight of the structural interdependence suggests several changes: human behaviors should add the consideration in terms of structural constrains on activity, rather than solely from individual behavioral aspect. The structure of the interdependent team can be treated as a network of agents that are tightly connected. The analytic methods deal directly with the patterned, relational nature of social structure can supplement–and sometimes supplant–mainstream statistical methods that demand independent units. Understanding and managing this structure interdependence is crucial for ensuring the success of human-AI teams and for maximizing the potential of this new way of working.

In a human-human team, interdependence theory was first introduced by Harold Kelley and John Thibaut in 1959 in their book The Social Psychology of Groups (Thibaut & Kelley, 1959). Interdependence theory originally focus on interpersonal relationships defined through interpersonal interdependence, which is the process of people influencing one another's experiences. Structural interdependence refers to the team characteristics that define the interconnectedness of team members (Wageman, 2001). These characteristics refer to task-related team input, such as resources and workflows as well as goal and reward system, which can be deliberately manipulated by team leaders and members (Courtright et al., 2015). Courtright and colleagues have identified an integrative framework with four two dimensions of four types of interdependence (Courtright et al., 2015): for task interdependence, which is the degree to which the taskwork is designed so members depend one another for access to critical resources and create workflows that require coordination action. The task interdependence includes input/resource and process/means interdependence. For outcome interdependence, which is the degree to which outcomes of taskwork are measured, rewarded, and communicated at the group level to emphasize collective outputs rather than individual contribution. The outcome interdependence includes the goal and reward/feedback interdependence.

Later, this concept is adopted to human-robot system by supporting interdependence through requirements for a new approach called coactive design (Johnson et al., 2014). Dependence is about capacity and interdependence is about relationship in the joint activity. For the interdependence theory, which is grounded based on the assumption of the joint activity, focuses

on the complementary relationships, which can be either required/hard or opportunistic/soft. An important missing factor in interdependence theory is the shared goal in the joint activity. According to Bratman (1992), apart from the fact that interactants are mutually responsive to each other, one other essential characteristic of joint activity is the shared goal. Mutual directability identified based on coactive design (observability, predictability, directability) can be a useful guideline to determine the interdependence requirement. However, the existing literature still lacks a theoretical framework that focuses on the goal aspect in human-AI interdependence.

2.5 Trust Measurement in Conversation

Trustworthy communication mediates the cooperation. How people talk and communicate is also closely tied to the broader interactive goals of the team, which are the products of adaptations to navigate the social world. Using conversation to coordinate their actions in the service of mutually beneficial interaction. Therefore, in the HAT, supporting the process of trusting via communication and cooperation is highlighted in this work. To appropriately calibrate overtrust or undertrust, we need to measure and manage trust.

Trust, as a latent variable that represents human attitude, cannot be measured directly. Three main types of measurement have been developed to capture trust: subjective, behavioral, and physiological (Kohn et al., 2021). The advantages and disadvantages of each measurement are discussed in the context of HAT. In this section, a new approach, conversational measurement, is proposed.

2.5.1 Subjective Measurements

For subjective trust measurements, people self-report their feeling and attitudes by answering survey items. Self-report measures are extremely easy to integrate into existing tasks and experiments before or after the task. Surveys are often developed precisely to capture the underlying constructs, such as ability, integrity, and benevolence, which shows better face validity and is widely adopted in most studies. Across disciplines, researchers have relied on many different questionnaires to measure trust. We have conducted a mapping review of 46 trust questionnaires from three main domains (i.e., Automation, Humans, and E-commerce) with a total of 626 items measuring different trust layers (Dispositional, Learned, and Situational). Results provided a guide of semantic space of trust questionnaires and implications in the questionnaire selection processes (Alsaid et al., 2022). Among these identified scales, they usually consist of directive statements and descriptions of human-agent relationships. For example, the frequently used trust scale in automation by Jian, Bisantz, and Drury (2000) has items such as “the system is suspicious.”

Respondents typically record their attitudes on a continuum from 1 (not at all) to 7 (extremely). Although frequently used, this approach suffers from some limitations when assessing human-agent relationships. First, since the directive survey is heavily text-based, the administration process often forces an interruption while people are interacting with the agent. Therefore, it is hard to capture the dynamics of trust calibration that might require many surveys. Second, the direct descriptive statement does not leave respondents with adequate freedom to identify, form, and explain their feelings and opinions (Gobo, 2011). For example, the statement 'the system is suspicious' might cause anchoring bias, where people rely on this preexisting information (e.g., suspicious) to judge their trust in agents. Third, most popular scales have the potential to cause positive bias in automation due to the order effect and the unbalanced design of positive-negative items (Gutzwiller et al. 2019). Finally, trust scales depend on and reflect human-agent relationships. Since the nature of these relationships has drastically changed in response to technological developments, past scales can be outdated and inapplicable to the inquired relationship (Merritt et al. 2019). Therefore, while self-reported trust is used most frequently and often treated as the gold standard, it is unable to satisfy the need to unobtrusively monitor trust dynamics, especially in time-pressured, risky situations, such as space missions or autonomous driving (Li et al., 2020; Yang, Christopher, et al., 2021). There is a need for an alternative or complementary trust measurement.

2.5.2 Behavioral Measurements

Behavioral measurements capture the interaction with the automated system, which can be passive (reliance) or active (compliance). Response time (time to respond to an event) and decision time (time to decide to use regarding automation) are also used to reflect trust. Faster decision times imply higher trust whereas the slower time reflects more evaluative thoughts and lower trust.

Using the game-theoretic situations to capture behavioral trust is one well-established approach in behavioral economics yet receives limited attention in human factors research. Game theory, defined as mathematical models of conflict and cooperation between intelligent rational decision makers, can provide a quantified and context-independent situation to evaluate behavioral trust. By creating an environment in which cooperation can initially form and become a stable presence in a human-machine system, game theory allows researchers to investigate the dynamics of trust in human-agent dyads (Razin & Feigh, 2021). One of the key advantages of using game theory to frame trust is that it can account for the impact of uncooperative behavior on trust, as opposed to just unreliable behavior. This means that trust in agents should be studied in terms of cooperation between humans and agents, rather than just the competence of the system. Two

main types of game theory have been established to capture trust in human-AI cooperation: the Trust Game (TG) and the Threshold Public Goods Game (TPGG).

The Trust Game, invented by Berg et al. (1995) (Berg et al., 1995), measures trust using economic decisions. In this game, the Investor has a sum of money (X) that she can either keep or invest with another player, the Trustee. If the Investor invests a certain amount (T), she keeps the remainder and the investment earns a return at a rate (1+r), becoming (1+r) T. The Trustee must then decide how to share the new amount with the Investor. However, the Trustee is free to keep the whole amount without consequence. The amount invested (T) is used as a proxy for trust, while the amount returned by the Trustee is taken as an indicator of their perceived trustworthiness. However, this approach assumes that subjects lack altruistic or inequality-averse other-regarding preferences (Cox, 2004), which did not untangle the relationships between trust and reciprocity. This means that variable trust (T) is confounded with the trustee's inequality aversion and altruistic preferences. In addition, there is no component of cooperation in the trust game since there is no common goal in the game.

The Threshold Public Goods Game (TPGG), as a type of the public good games, has often been used to abstract social decision-making problems where participants aim to achieve a common goal with uncertain and delayed responses noted by the threshold. The players need to contribute to a public goal which is launched if and only if a certain level of contributions (threshold) is reached. Contributing may have a personal, local cost, but can lead to a global benefit for the team. The TPGG helps to understand people's tradeoffs between local and global optimum in the human-AI cooperation.

2.5.3 Physiological Measurements

Physiological measurements capture biological responses ranging from heart rate changes to eye gaze tracking to neural activation. Kohn et. al.(2021) identified four distinct types of physiological measures: 1) electrodermal activity (EDA), also known as galvanic skin response, which measures the sweat gland activation via skin conductivity; 2) eye gaze tracking, which measure participants' monitoring behaviors; 3) heart rate change, which often used to measure workload and stress; 4) neural measure, including electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS), which can theoretically be captured and used to measure trust. Using physiological measures may help to present issues in the self-report method by directly capturing people's responses, which present a great opportunity for real-time trust estimation. However, getting high-quality physiological data

(e.g., EEG and skin conductance responses) often requires setting up specialized and intrusive hardware on participants (e.g., electrodes on the scalp or hands), which is challenging to implement in real-world applications.

2.5.4 Conversational Measurements

One rich source of data that can be used to measure trust but often is neglected in the past literature is team communication. With the increasing level of interdependency in HAT, there is an increase in information exchange between human and AI teammate, which displays a rich source of information that reflect team cognition and processes (Cooke et al., 2013). Based on interactive team cognition theory, communication represents team cognition and can serve as a unobtrusive measure of team interaction dynamics (Cooke et al., 2013). Communication is also essential for trust building and calibration, which in turn can promote effective human-AI teaming (Fuoli & Paradis, 2014). Although communication plays a vital role in driving HAT success, measuring trust via communication is still a new approach.

Conversational data can be considered as a mixture of behavioral and physiological data that contain lexical, semantic, phonological, and pragmatic representations of the conversations. In other words, people naturally express their trust attitudes through the words they use, the sentence structure, and the tone of the voices in their conversation, which are all contextualized. Additionally, communication can manifest conscious and subconscious mental states. Trust, which can reflect both cognitive and affective processes, can be analyzed and measured through communication (Lee & See, 2004). prior trust literature suggested that trust is ultimately an affective process, which has more influence on analytic and analogical processes (Lee & See, 2004). Yet, with the subjective (e.g., survey) or behavioral (e.g., takeover the automation) measures are often hard to reflect the affective process. Conversational data, which contain not only what they say (e.g., word choices), but also how they say it (e.g., tone of voices), provides richer information about people's affective trust process (Li et al., 2022).

Prior literature on HAT usually uses communication patterns such as communication rates and flows to predict trust (Bromiley & Cummings, 1995). Limited research has focused on communication content for trust measurement. Although the most explicit way of expressing and sensing trust is through words that directly pertain to trust (e.g., I trust you), it is unnatural and rare for people to express a direct attitude in a performance-based task. Therefore, we should obtain and infer people's trust by processing and analyzing the signals exhibited by individuals in conversations (Vinciarelli et al., 2009). To do so, we first need to elicit utterances by designing

trust-relevant situations with appropriate conversational prompts. Once we elicit trust-relevant conversations, we can process and analyze the conversational cues to estimate people's trust. According to the well-known phrase, "It's not only what you say, but also how you say it.", both the words and how they are said should convey trust. Therefore, in this dissertation, we consider not only lexical cues (e.g., words used), but also acoustic cues (e.g., pitch, formants) (Elkins & Derrick, 2013; Johnson et al., 2014).

2.6 Trust Management in Cooperation

Once trust is measured, the next question is how to properly manage trust. Managing trust is not a novel topic. Prior research often identifies the causal relationships between a single factor or a combination of antecedents of trust and investigates its impact on trust in experimental studies. However, as trust becomes more dynamic in HAT, it is important to define the trust management framework that captures the interdependency in the team. In this section, a three-stage trust management process is defined in Figure 2: antecedents, management, and measurement. The key is direct correspondence between the type of trust antecedents (i.e., type of trust violation or compliance), the management strategies (i.e., scope, strategy, timing, and modality), and the measurement (i.e., methods, analysis). This means that performance-based trust violation should be managed and measured differently from purpose-based trust violation.

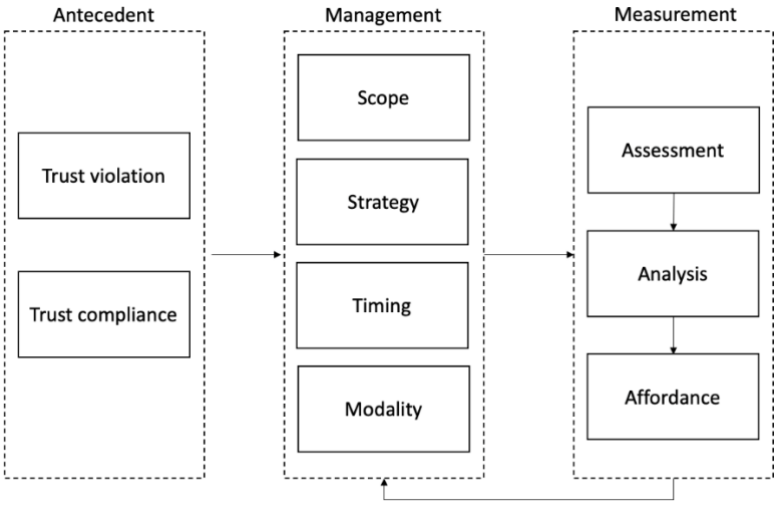


Figure 2. Trust Management Framework

2.6.1 Antecedents: Identify the type of trust compliance or violation

When managing trust, the most important yet often neglected step is to identify the corresponding type of trust violation or compliance. Trust compliance or violation is defined action representing the alignment or misalignment between the observed trustworthiness and the

current (de Visser et al., 2020). Trust in automation is a multidimensional construct containing performance, process, and purpose-related antecedents (Lee & See, 2004). Extensive research has been conducted to examine the antecedents of performance and process-based trust. However, the purpose dimension has become increasingly important as humans and AI become more lateral teammates. The shared goal assumption should be revisited when identifying the type of trust compliance and violation factors.

Purpose-Based Trust

When managing trust, it requires more than performance and process aspects of trust. A shift to 'purpose' basis of trust is often required. In the past human-agent interaction, it is often assumed a shared goal. This means that autonomous system assisted human operators to achieve a specific task. In this case, the human is the supervisor and AI is the assistant. With the increasingly computational capability and connected systems, the inclusion of AI in the future social systems is inevitable such as autonomous vehicles, drones, recommending system, healthcare support, or educational support system. In human-AI cooperation in a hybrid society, the assumption of shared goal should be challenged. I will refer to this issue as 'goal alignment' in the following section. Goal alignment is defined as the degree to which the AI's programmed goal matches with the human's goal {Citation}. Without aligned goals, we may inadvertently design AI that maximizes an objective function that is poorly aligned with humans. Especially in safety-critical domains, where failure tolerance is low, goal conflict can lead to catastrophic consequences.

Goal alignment, or value alignment, is not a new topic in robotics domains and safety research for AI. As superhuman cognitive abilities become generally reachable, value alignment, which is 'AI must do what we want it to do', has been identified as a core topic to address. Thinking of the 'King Midas Problems', which is a legendary king Midas in ancient Greek mythology got exactly what he asked for-everything he touched should turn to gold. He discovered too late that he turned his food, drink, and his family member all into gold and he died in misery and starvation. The same faith might be applied to the human-AI domains-we may, inadvertently, design the AI with the objectives that are not well aligned with ours. We can only discover the catastrophic consequences until it becomes too late. For robotics domains, researchers focus on technical side of robotics goal-implementation question, such as: How can we create an agent that will reliably pursue the given goals? How can we formally specify beneficial goals (Soares & Fallenstein, 2014)? How should the robot strike compromises when conflicts arise between the commands of its owners (Critch, 2017)?

In HAT, only focusing on the system implementation of machine goals is not enough. Humans are not rational, do not maximize the expected utility all the time, and risk averse. When considering the machine algorithm design, people should balance efficiency, fairness, and risk. When aligning goals between human and AI, the existing benefits of the machine behaviors, such as efficient computing, should be advocated, but would do so in a way that would help people to be more prone to considering these more efficient solutions when warranted. Thus, the goal-related characteristics of the agent are not always congruent with humans in the team. The topic on value alignment should be introduced and redefined in terms of HAT teaming from a sociotechnical perspective.

In social interactions, there might be a conflict between individual and societal goals. Since with the increasing computational power and connected system, we can design a more socially optimal agent for the common good. Humans, on the other hand, are often irrational and individual driven. Therefore, goal conflicts can happen: the global optimum goal assigned to the AI agent can potentially conflict with the individual's local goal. For example, connected automated vehicles can coordinate traffic to achieve optimal traffic flow, which conflicts with individual goals of arriving at the destination in the shortest time. In the safety critical domains, such as healthcare, military, and emergence response where the expectation for system efficiency is high and the tolerance for breakdowns is very low, designing a trustworthy AI teammate requires not only a competent reliability, also the goal alignment between units of an organization. With the inclusion of AI in social systems, we must understand how it can be used to promote cooperation in achieving societal goals, while maintaining people's trust in AI. In this case, the performance of the AI becomes less relevant. With highly reliable information, yet with misaligned goals, would humans trust and cooperate with AI in a hybrid society? What's making the situation worse is that people tend to exploit AI in cooperation (March, 2019). The question remained: How can we design a trustworthy agent and nudge people towards the socially optimal and aligning the goals?

Additionally, the conflict can also occur on the temporal dimension. Often, short-term interests supersede long-term ones, even when the latter provide a higher reward. Moreover, short-sighted individuals might not see the future consequences of their immediate actions. With the inclusion of AI in the social systems in the long run, we must comprehend how can it be used to promote cooperation towards the societal optimal in a long run while maintaining people's trust in AI. In this case, the performance of the AI becomes less relevant. A series of new questions should be asked: With highly reliable performative AI, yet with misaligned goals, would humans

trust and cooperate with AI in a hybrid society? Often, people tend to exploit AI in cooperation, how can we better manage people's trust to promote cooperation? How can we design a trustworthy agent and nudge people towards the socially optimal and aligning the goals?

Little literature has discussed the process and models for capturing the goal aligning process in the human-AI teaming. In the form of human-AI goal alignment, it should be iterative and recursive due to the dynamic and dyadic nature of the goal interdependence. Human-AI teaming requires both an alignment of self with AI to form the shared goal, and a differentiation of self from other to understand and coordinate the differing but complementary roles in the joint intention. One strategy for aligning the goal in the joint activity is effective communication. Continuous or period updates between members can maintain situation awareness (Endsley & Kiris, 1995) and the shared mental model of the common ground (Clark, 1996). The intelligent agent should be able to interpret the situation and know what information to share and when to request assistance (Johnson & Vera, 2019). Whiting and colleagues (2021) showed that communication signals can bring human decisions to efficiently cooperate in human-robot interaction. Lewis (1979) proposed that "conversation as a cooperative game between participants" where the goal is to determine which world the participants are in. Participants work towards this goal by sharing the information, which narrows the set of possible worlds that the real world might be. The information shared between the conversation participants is stored in the Common Ground, which can be viewed as a set of accepted propositions. The rational speech act (RSA) model is a framework for pragmatic and mathematical modeling that extends the concept of 'conversation as a cooperative game' by proposing a Bayesian listener and speaker who act to maximize a utility function related to the listener's understanding. In this model, Common Ground contains not only just a set of worlds, but also a probability associated with each world, the probability that it is the real world. At each turn in the conversation, the speaker selects a world from the set of worlds, simulating a new piece of information that the speaker wishes to contribute, and chooses an utterance to express it. When hearing the speaker's utterance, the listener must think about the message the speaker is trying to convey. The listener assumes that the speaker selects the sentence that maximizes the probability of the observed world. The listener interprets the sentence to update the probability distribution over possible worlds in the Common Ground, calculating the likelihood of each world given the sentence selected, according to their model of how the speaker picks sentences. The RSA model can be tuned to maximize various types of goal interdependency to explore how a human and an agent would communicate and negotiate in a collaborative task under various intents/purposes.

2.6.2 Management: Identify Scope, Strategy, Timing, Modality

Scope

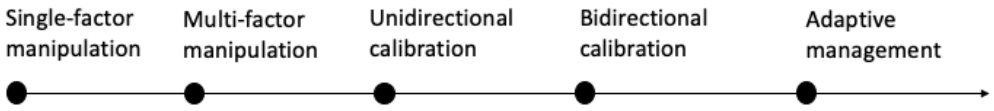


Figure 3. Trust Management Scope Ranging from Single-Factor Manipulation to Adaptive Management.

When designing the trust management, the scope matters, which can range from a single-factor or multi-factor manipulation, to calibration, to a fully adaptive management.

Manipulation. Trust manipulation focuses on casual relationships between one or more trust antecedents and trust. Researchers usually design experimental studies and investigate the influence of these antecedents on trust. Hoff and Bashir identified 29 factors that are influential for the trust(Hoff & Bashir, 2015). These studies identified the significant factors that influence trust and their interactions with other factors in various contexts. These studies lay the essential foundation for the following calibration and management.

Calibration. Researchers have reached to consensus that when managing the trust, the main goal is never increase or decrease trust (which is a form of the manipulation), which can lead to costly consequences (Bailey & Scerbo, 2007; Dzindolet et al., 2001). The aim should be 'trust calibration', which requires users to appropriately adjust their level of trust to the actual reliability of the AI system. Specifically, when we overtrust automation, we should dampen trust; when we undertrust, we should repair trust (de Visser et al., 2020)(see Figure 4). Trust calibration is essentially a two-directional trust manipulation with a continuous alignment with the system capability. Among these two directions, repair and dampening, the trust repair has been studied heavily. Trust repair is defined as an action taken by a trustor to help restore trust in them after they have committed a violation of trust (R. M. Kramer & Lewicki, 2010). Prior literatures focus on using the short-term verbal cues to repair trust through four different types of repairing strategies: apologies, denials, explanations, and promises. Esterwood and Robert have identified the overarching theoretical trust frameworks associated with these four types of repair strategies and found that apologies, explanations, and promises are equally ineffective in repairing trust after repeated violations (Esterwood & Robert, 2023). On the other hand, trust dampening has not received as much attention in the literature, although literatures have constantly shown the danger of the automation-induced complacency and perfect automation schema (Dzindolet et al., 2002).

Trust dampening often used approaches by lowering the expectation when the trust is too high. Jensen and Khan found that trust dampening cues increased perceptions of system integrity and improved trust appropriateness (Jensen & Khan, 2022). However, although people recognize the importance of trust calibration as a bidirectional control, most of the previous research focuses only on either trust repair or dampening.

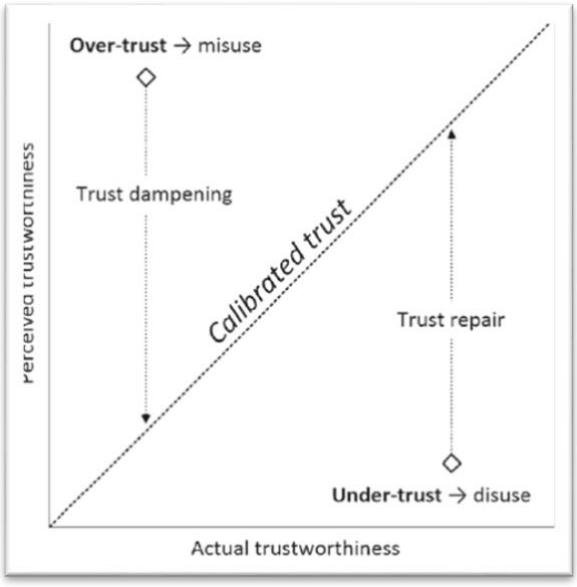


Figure 4. Trust Calibration Process(Chiou & Lee, 2021; de Visser et al., 2020).

Some efforts on designing a trust-calibrated system have been shown in the recent literature. Okamura (2020) proposed a framework for dynamically detecting inappropriate trust-calibration status with a behavior-based approach (Okamura & Yamada, 2020). The framework focused on the human-AI cooperation at the performance-level: the user decides whether or not to rely on the system or perform each task manually. Three core parameters, P_A, $\hat{P}_A$, and P_H are defined as follow.

- P_A: Probability that a task performed by an AI system will be successful, which is the 'reliability of the AI system'.
- $\hat{P}_A$: Human user's estimation of P_A, which is user's trust in the AI system.
- P_H: Probability that a task done manually by a human user will be successful, which is the "capability of the user".

P_A varies based on the conditions of the AI system. When trust is appropriately calibrated, $\hat{P}_A$ should become equal to P_A. For overtrust and undertrust, Okamura defined it as follows:

- Over-trust: when user estimates the AI system is better at a task than the user, even though the actual reliability of the AI system is lower than the user's capability.

$$(\hat{P}_A > P_H) \wedge (P_H > P_A)$$

- Undertrust: when the user estimates that the AI system is worse than their own capability, even though the actual reliability of the system is higher than the user's ability.

$$(\hat{P}_A < P_H) \wedge (P_H < P_A)$$

Prior literatures have been explored to use trust calibration cues (TCCs) to properly notify users to calibrate their trust (Okamura & Yamada, 2020). TCCs play a crucial role in human-AI cooperation, as they can shape users' trust in AI systems. Because once users fall into the categories of overtrust or undertrust, it is not easy to update their state due to the confirmation bias and other cognitive preservation. TCCs can provide triggers for users to pay attention to the environment and system and actively update their cognitive states. Visser and colleagues presented a trust cue design taxonomy that considered trust dimensions (intent, performance, process, expressiveness, origin) and trust processing stages (perception, comprehension, projection, decision, and execution)(de Visser et al., 2014).

Management. Trust calibration essentially only focuses on the performance-dimension of trust, which is often the most critical factor in the past human-automation interaction. However, as automation becomes more autonomous in teamwork, purpose-based trust should be identified and calibrated. ***Trust management*** is defined as the multidimensional monitoring and calibrating. Previous results on different modality, timing, and attribution of TCCs suggest that designing the appropriate and effective TCCs are nuanced. When studying trust management, researchers must understand what type of trust should be calibrated (scope), how to manage it (modality), what information to include (strategy), and when to present (timing).

Strategy

Trust management strategies are often achieved by short-term trust calibration cues (TCCs), which can be broadly classified as either repair or dampen approaches. Note that most of the existing strategies focus on the performance-based trust dimension.

For trust repairing strategies, Esterwood and Robert identified four main types of strategies and provided theoretical basis for each of these strategies: theory of forgiveness for apology, theory of forgetting for promises, theory of informing for explanation, and theory of misinforming for

denial (Esterwood & Robert, 2023). Apology acknowledges the issue and show expressions remorse. By doing so, especially in conversations, apologies convey more emotional responses (e.g., remorseful), which can be used to promote forgiveness and restore trust from an affective perspective. The explanation is another repair strategy that can be used to restore trust. It provides users with a clear understanding of why the system failed and what steps are being taken to prevent similar issues in the future. A promise can also be made to the user to fix the issue and ensure that it does not happen again. Denial is a strategy that should be used sparingly and only when the user's perception of the situation is incorrect, and the AI system has not actually failed.

The attribution of TCCs is critical, as it can affect how users interpret and respond to the cues. The attribution can be internal versus external. If the agent shares the responsibility for an error, it is internal attribution (e.g., apology, explanation); whereas if the agent does not share the responsibility, it would be an external attribution (e.g., denial, blaming). Kim and colleagues examined trust repair with an internal versus internal attribution after a competence (performance) versus an integrity (purpose)-based trust violation (P. H. Kim et al., 2006). They found that performance-based trust repair is more effective when the agent apologizes with internal attribution, and purpose-based trust repair is more effective with external attribution. However, Perkins et al. identified contradictory findings in human-robot teams: internal attribution apology can effectively repair purpose-based trust violations, but not performance-based trust (Perkins et al., 2022). One explanation can be the difference between human-human trust human-automation trust. People often have a Perfect Automation Schema (PAS) (Dzindolet et al., 2002) that causes people to overreact to the performance dimension of the automation error. Thus, it is harder to repair performance-dimension of trust, comparing to purpose-dimension. Similar findings were found by Esterwood and L.P.R. Jr., purpose-based trust may be more repairable than performance-based trust.

Trust dampening strategies are reactive approach to reduce overtrust after the system has made a lucky guess, or when a machine makes a mistake that has not been noted by users expectations (de Visser et al., 2020). Thus, managing people's expectation about the AI system to reduce the likelihood of overtrust is the essence of trust repair strategy. Lowering expectations is a strategy that can be used to reduce the gap between the user's expectations and the actual capabilities of the AI system (Jensen & Khan, 2022). This can be achieved by highlighting the limitations of the system or showing a history of performance. Explicitly expressing reduced confidence is another dampening strategy that can be used to manage user expectations (de Visser et al., 2020).

Timing

The timing of trust calibration cues is crucial for effective trust calibration. If the cues are presented too early, users may not yet have enough information to assess the AI system's competence and reliability, and if the cues are presented too late, users may have already formed an impression that is difficult to change. Ideally, trust calibration cues should be presented at the right time when users are making decisions or evaluating the AI system's performance. Robinette showed that it is more effective to provide trust-repairing signals when the robot asked the participants to trust it again, not imeediately the mistake (Robinette et al., 2015). Du et al. found that it was most effective when the explanation was provided before critical events in automated driving (Du et al., 2019).

Modality

The modality of trust calibration cues can take different forms, including visual, auditory, haptic, or a combination of any of these. The proposed TCC design taxonomy is mainly based on the visual cues, where authors acclaimed can be applied to other modalities. Yet, the direct mapping from visual to other modality remained the challenges. For example, what is the appropriate volume, pitch, and gender of voice for auditory trust calibration cues?

Trust is not only conveyed through language, but also through acoustic cues. We have identified formants, fundamental frequency, and Mel-frequency centrostral coefficients as the most significant acoustic indicators of trust in conversations. The primary question is whether these identified indicators have an impact on perceived trustworthiness, in addition to predicting trust. Our findings show a mixed result from the prior literature: Although pitch significantly affected perceived trustworthiness (Elkins & Derrick, 2013), it is not the most important feature when people express their trust in the conversation. On the other hand, the formants show that they can be used to predict trust and influence perceived trustworthiness (Montano et al., 2017). Therefore, it is important to examine whether identified acoustic features, especially formants, are effective in managing people's trust.

Chapter 3. Measure Trust in Human-AI Conversation

Title: It's Not Only What You Say, But Also How You Say it: Machine Learning Approach to Estimate Trust from Conversation

Journal: Human Factors

Submission date: April 13th, 2022

Acceptance date: March 10th, 2023

Abstract

The objective of this chapter was to estimate the trust of conversations using lexical and acoustic data. As NASA moves to long-duration space exploration operations, the increasing need for cooperation between humans and virtual agents requires real-time trust estimation by virtual agents. Measurement of trust through conversation is a novel and unintrusive approach. A 2 (reliability) × 2 (cycles) × 3 (events) within-subject study with habitat system maintenance was designed to elicit various levels of trust in a conversational agent. Participants had trust-related conversations with the conversational agent at the end of each decision-making task. To estimate trust, subjective trust ratings were predicted using machine learning models trained in three types of conversational features (i.e., lexical, acoustic, and combined). After training, model explanation was performed using variable importance and partial dependence plots. Results showed that a random forest algorithm, trained using the combined lexical and acoustic features, predicted trust in the conversational agent most accurately ($R^2_{adj} = 0.71$). The most important predictors were a combination of lexical and acoustic cues: average sentiment considering valence shifters, the mean of formants, and Mel frequency cepstral coefficients (MFCC). These conversational features were identified as partial mediators predicting people's trust. Precise estimation of the trust of the conversation requires lexical and acoustic cues. These results showed the possibility of using conversational data to measure trust and potentially other dynamic mental states, unobtrusively and dynamically.

3.1 Introduction

As the National Aeronautics and Space Administration (NASA) moves to long-duration space missions, longer time delays in communication between crews and ground control will require more cooperation between the humans and the onboard virtual agent (Chiou & Lee, 2016; Johnson et al., 2014; Trafton et al., 2006). In this human autonomy team (HAT), trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by

uncertainty and vulnerability"(Lee & See, 2004, p. 54), plays an essential role and impacts various team processes, including information sharing, decision making, and ultimately team success (Endsley et al., 2021; Krausman et al., 2022). To better manage the human-autonomy team, it is important to first measure trust unobtrusively and dynamically.

Three main types of measurement have been developed to capture trust: subjective, behavioral, and physiological (Kohn et al., 2021). For subjective trust measurements, people self-report their feeling and attitudes by answering survey items. While self-reported trust is most frequently used and often treated as the gold standard, it is unable to satisfy the need for unobtrusively monitoring trust dynamics, especially in time-pressured, risky situations, such as space missions or autonomous driving (Li et al., 2020; Yang, Christopher, et al., 2021). Behavioral measurements can unobtrusively estimate trust through interactions with the automated system, which can be passive (reliance) or active (compliance). Although behavioral measurements allow minimal disruption and a higher sampling rate than self-report, they are often task-specific and hard to generalize. Physiological measurements capture biological responses ranging from heart rate changes to eye gaze tracking to neural activation. They present a great opportunity for real-time trust estimation. However, getting high-quality physiological data (e.g., electroencephalogram and skin conductance responses) often requires specialized and intrusive hardware (e.g., electrodes on the scalp or hands), which is challenging to implement in real-world applications. One rich, but often neglected, source of data for measuring is team communication. With the increasing level of interdependency in HAT, there is an increase in information exchange between human and AI teammate, which can reflect team cognition and processes (Cooke et al., 2013). People may change what they say and how they say it based on their trust in their AI teammate. In this paper, we demonstrate that measuring trust from conversations provides a promising, yet underexplored approach. We take the first step in this direction by predicting and validating trust based on structured conversations with a conversational agent that supported a complex decision task. In addition, we identify the important conversational features for trust prediction. Our findings provide theoretical implications for the development of the conversational measurement of trust and the adaptive conversational strategy of a trustworthy AI teammate.

3.1.1 Measuring Trust in Conversation

Although communication plays a vital role in driving HAT success, measuring trust via communication is still a new approach. Communication can manifest conscious and subconscious mental states. Trust, which reflects both analytic and affective processes, can be analyzed and measured via communication (Lee & See, 2004). Prior literature on HAT usually uses

communication patterns such as communication rates and flows to predict trust (Bromiley & Cummings, 1995). Limited research has focused on communication content for trust measurement. Although the most explicit way of expressing and sensing trust is through words that directly pertain to trust (e.g., I trust you), it is unnatural and rare for people to express a direct attitude in a performance-based task. Therefore, we should obtain and infer people's trust by processing and analyzing the signals exhibited by individuals in conversations (Vinciarelli et al., 2009). To do so, we need to first elicit utterances by designing trust-relevant situations with appropriate conversational prompts. Our prior work developed a trust lexicon and a general framework on how to design appropriate conversational prompts (Alsaid et al., 2022; Li et al., 2020). Once we elicit trust-relevant conversations, we can process and analyze the conversational cues to estimate people's trust. According to the well-known phrase, "It's not only what you say, but also how you say it.", both the words and how they are said should convey trust. Therefore, in this paper, we consider not only lexical cues (e.g., words used), but also acoustic cues (e.g., pitch, formants) (Elkins & Derrick, 2013; Johnson et al., 2014).

3.1.2 Lexical Indicators of Trust

Lexical features in the conversation contain rich information including the length of the utterances (e.g., word count), word choices, and sentiment (Spitzley et al., 2022). The most frequent and simple measure is word count. Previous literature has shown that there is a positive correlation between word count and perceived trustworthiness in online dating profiles and lending load requests (Larrimore et al., 2011; Toma & Hancock, 2012). Based on the uncertainty reduction theory, the more information is provided, the less uncertainty, and the higher the perceived trustworthiness (Beller et al., 2013; M. W. Kramer, 1999). However, little is known about whether this correlation holds true with the lexical features of trustor's communication (i.e., higher trust, fewer words). For the sentiment in the conversations, prior research has shown that verbal positivity is positively correlated with perceived trustworthiness of organizational leaders (Norman et al., 2010). Additionally, people also found the positive association between positive sentiment in trustors' word responses (e.g., excited, interested) and affective trust when interacting with a conversational robot (Hildebrand & Bergner, 2021). Because benevolence is one of the core elements of trust (Mayer et al., 1995), it is expected that people would express positive affect when they trust their AI teammates.

3.1.3 Acoustic Indicators of Trust

The characteristic of the voices, or acoustic features, indicate people's thoughts, feelings, and attitudes. The same set of words uttered with different volumes or intonations can express

different feelings and the underlying message of the words (Sebe et al., 2005). Thus, when understanding how people express trust, it is crucial to exanimate acoustic features. Pitch, measured as the fundamental frequency (F_0), is one key component of acoustic features. Vocal pitch has been shown to be inversely related to the perceived trust of the agent, especially during the early stages of interactions (Elkins & Derrick, 2013). Additionally, the high-variance F_0 trajectory, indicated by a high starting F_0 and then a marked decrease at mid-utterance to finish on a strong rise, was rated high in trustworthiness (Belin et al., 2017). Waber and colleagues found a correlation between emphasis, defined as the variations in pitch and volume, and initial trust in technical communication in hospital settings (Waber et al., 2015). Additionally, formants, the concentration of acoustic energy around a particular frequency in the speech wave, are also found to associate with trust. Montano et al. found that high pitch but low formants voices, which affect masculinity perceptions, were more trusted in a cooperative game (Montano et al., 2017).

Although previous research has shown relationships between conversational features with perceived trustworthiness of an agent as a trustee, limited research has shown how people, as trustors, signal and express trust they place in that agent. Trust, as both analytic and affective processes, can govern people's behaviors and the way they speak (Lee & See, 2004). People have been shown to change their lexical and acoustic cues in conversation depending on whether they trust the agent or not on a binary scale (Gauder et al., 2021). However, to date, no research has shown 1) whether the continuous scale of trust can also be predicted and 2) what are the important indicators in conversations that can predict trust. In other words, limited research has investigated whether and how to measure people's trust in conversations. One methodology that can resolve this question is machine learning (ML). Recently ML has been used to not only predict certain classes of data (e.g., trust), but also infer and explain the predictions (McDonald, Ferris, et al., 2020). In our study, the goal was two-fold: First, we showed that a machine learning approach can make predictions of trust using a combination of acoustic and lexical indicators extracted from conversations. Second, we identified the important lexical and acoustics features underlying these predictions, which provide insights for future trust management in HAT.

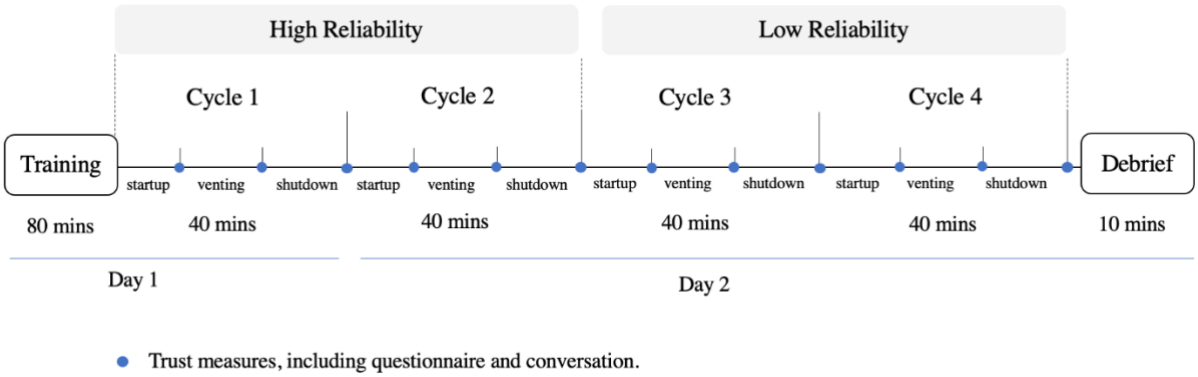


Figure 5. Study Design with 12 Trust Measurement Points.

Table 1. Examples of Conversational Trust Questions.

Q1	How would you describe your experience selecting the procedure? What are your overall feelings during procedure selection?
Q2	Why would you feel that? Can you explain your answer in more detail? Make sense. Why would you have that feeling? Can you elaborate on that?
Q3	Can you talk more about my performance in providing the recommendation? Thank you. How would you describe my performance in giving you the recommendation?
Q4	That makes sense. Which procedure did you select? Okay, thank you. Which procedure did you select?
Q5	Can you tell me more about your strategy for picking that procedure? What made you choose that procedure? Can you tell me more?
Q6	How can I be more helpful in terms of providing recommendations? I see your strategy there. How can I be more helpful next time?

3.2 Methods

Estimating trust using Machine Learning (ML) requires crafting a situation that produces variations in trust and generates repeated measures of trust. First, large variations in a ground truth trust measure are needed. We used a well-validated variable, automation reliability, which has shown a strong causal relationship with trust, as a proxy to induce variations of trust (Lee & See, 2004). Second, a well-labeled target response (i.e., trust) is needed for supervised ML models. Thus, we collected subjective trust ratings along with conversational data. Third, to generate trust-relevant utterances, we designed open-ended conversational prompts with follow-up questions to

elicit trust-related utterances. The questions were designed to be non-directive, which probes respondents to describe their own attitudes and feelings on topics of trust in automation instead of using the presumed attitudes and descriptions. Since we manipulated the automation reliability, we aimed to elicit participants' responses related to the performance-based trust. The questions were developed based on our prior research on trust lexicon and conversational measures, see details (Alsaid et al., 2022; Li et al., 2020). Finally, trust changes as a dynamic process that varies across interactions (Yang, Christopher, et al., 2021). We designed multiple check-in points after every interaction with the automated system to ensure we captured multiple measures of trust.

3.2.1 Study Design

The study was a 2 (reliability) × 2 (cycles) × 3 (events) within-subject study (see Figure 5). Participants performed 12 decision-making tasks associated with managing a system of a simulated space station: the Habitat's Carbon Dioxide Removal System (CDRS). Participants were assisted by a conversational agent with 2 levels of agent reliability (i.e., high, and low). Each level of reliability had 2 cycles of the CDRS tasks, each including 3 events (i.e., startup, venting, shutdown). To induce substantial changes in trust, the high-reliability conversational agent provided 100% correct recommendations whereas the low-reliability agent provided 20% correct recommendations. The 12 total events were designed to elicit various levels of trust through manipulation of the agent's reliability. At the end of each event, the agent initiated a conversation by asking 6 trust-related questions (Li et al., 2020). Once the participant finished the conversation, they then completed a 12-item trust survey on a 7-point Likert scale (Jian et al., 2000). In total, each participant had the opportunity for at least 72 conversational turns with the agent.

3.2.2 Participants

A total of 24 participants (18 female, 6 male) were recruited (M=23.7, SD =3.6). Participants need to have some technical background (e.g., completion of STEM courses). Due to the safety concerns of COVID-19, the study took place online. It was a two-day study with each day lasting up to 2 hours. In total, the study was approximately 4 hours. Participants received $30 per hour for up to $120.

3.2.3 Apparatus

The experimental task uses the Procedure Integrated Development Environment (PRIDE) which is an automated procedure software, to maintain the space station habitat using the Carbon Dioxide Removal System (CDRS) (Izygon et al., 2008; Schreckenghost et al., 2014). A conversational agent, named Bucky, was preprogrammed with procedure protocols to provide

recommendations to help participants maintain the habitat task in PRIDE. Google Dialogflow, a Natural Language Understanding (NLU) platform was used to design and integrate the user interface. Participants were asked to directly speak to the conversational agent using their microphone. Keyboard and button inputs were also provided. Both audio and automatic speech-to-text data were collected as conversational measures.

3.2.4 Procedure

After signing the consent form, participants completed training on PRIDE, CDRS, and Bucky systems. During the study, participants had 25 minutes to control the CDRS by completing all three events (startup, venting, and shutdown) before their crew experienced CO_2 poisoning. For each event, the participant made two essential decisions with Bucky's aid (i.e., procedure selection and confirmation). The participants made their decisions either based on their knowledge from their training session or Bucky's recommendation. Once the procedure was selected, PRIDE automated the procedure execution. While the procedure was running, participants engaged in a secondary task on system checking by reporting the CDRS status to Bucky. If the participant selected the wrong procedure, an error occurred. The participant then had to manually stop the procedure and reselected a procedure. Once the participant finished the event, Bucky administered six conversational questions with some variations to avoid being repetitive (see Table 1). After conversational questions, participants completed the trust questionnaire. The total time of each cycle, including the trust conversation and questionnaire, was approximately 40 minutes. At the end of the study, participants were debriefed and compensated.

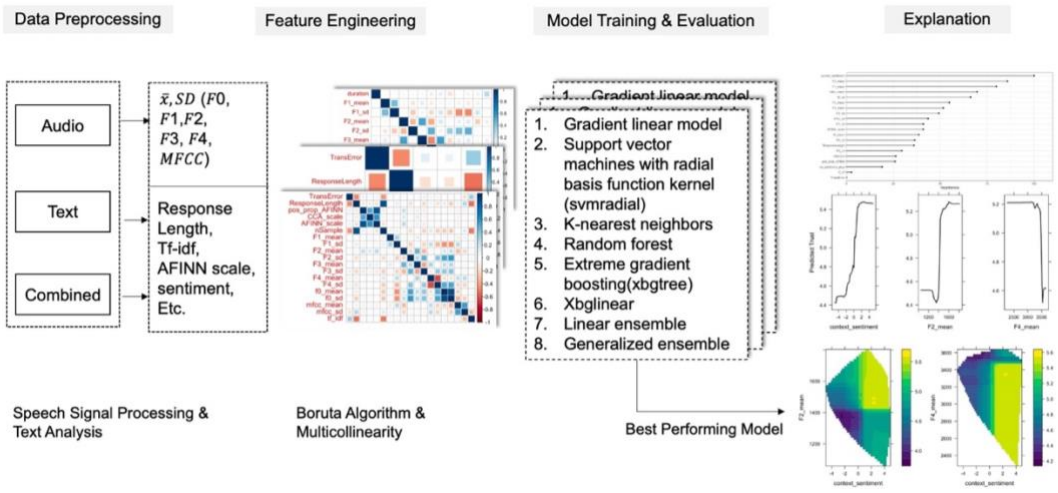


Figure 6. Machine Learning Pipeline to Estimate and Explain Trust.

3.2.5 Machine Learning Pipeline

Figure 6 shows the machine learning pipeline we adapted from previous research on human state estimation (McDonald, Ade, et al., 2020; McDonald, Ferris, et al., 2020). The conversations were first separated into audio, text, and combined data analysis streams. The audio and text features were extracted using speech signal processing and text analysis. The processed features were then used to fit the machine learning models. The best-performing model was selected based on root mean squared error (RMSE) and adjusted R-squared (R^2_{adj}). RMSE indicates the absolute fit of the model in the units of the response variable andR^2_{adj} indicates the variance in the response variable that can be explained by the predictor variables adjusted for the number of predictions in the model. The dataset was processed and analyzed in R.

3.2.6 Data Pre-processing

For the response variable, trust, we took the average of trust and distrust of the subjective trust rating, and reserved distrust score and averaged it with the trust score to get the final trust score.

For audio data, all the wave files were imported in R to extract acoustic features using the *wrassp* package (Bombien et al., 2021). A formant estimation function is used to calculate the first four formants and their bandwidths. For each formant, the mean and standard deviation were extracted. Fundamental frequency and mel-frequency cepstral coefficients (MFCCs) were extracted using their mean and standard deviation. Since acoustics features are gender sensitive (Cartei et al., 2012), all acoustics features are normalized within gender.

For text data, text analysis was used to extract lexical features. Data were manually cross validated by two researchers. We included a new binary variable called, translation error, to indicate

whether the speech-to-text has translation errors, which shows a 70% accuracy rate. Then, the response length was calculated based on the raw text input. The text was tokenized and punctuation and stop words were removed, and the words were stemmed. First, term frequency-inverse document frequency (tf–IDF) was calculated based on the frequency of a term within each document, normalized by how often the term is found in the other documents. Next, sentiment scores were calculated using sentiment dictionaries, such as Dominance Lexicon (Mohammad, 2018) and AFINN (Nielsen, 2011). Data were dropped if no words in an utterance matched any words in the sentiment dictionaries. Using only sentiment-related words and ignoring linking words to score sentiment can be problematic. For example, simply extracting "happy" in the phrase "I am not happy" can incorrectly score positive on the sentiment scale. To address this, we included valence shifters (i.e., negators, amplifiers, and intensifiers) by considering the context around sentiment-related words using the *sentimentr* package (Rinker, 2017) (details see Table 2).

A combination of audio and text features was used to predict trust. The two feature sets were merged based on unique audio identifiers associated with each utterance in the study. A z-score standardization was conducted on all feature sets and the response variable.

For text data, text analysis was used to extract lexical features. Data were manually cross validated by two researchers. We included a new binary variable called, translation error, to indicate whether the speech-to-text has translation errors, which shows a 70% accuracy rate. Then, the response length was calculated based on the raw text input. The text was tokenized and punctuation and stop words were removed, and the words were stemmed. First, term frequency-inverse document frequency (tf–IDF) was calculated based on the frequency of a term within each document, normalized by how often the term is found in the other documents. Next, sentiment scores were calculated using sentiment dictionaries, such as Dominance Lexicon (Mohammad, 2018) and AFINN (Nielsen, 2011). Data were dropped if no words in an utterance matched any words in the sentiment dictionaries. Using only sentiment-related words and ignoring linking words to score sentiment can be problematic. For example, simply extracting "happy" in the phrase "I am not happy" can incorrectly score positive on the sentiment scale. To address this, we included valence shifters (i.e., negators, amplifiers, and intensifiers) by considering the context around sentiment-related words using the *sentimentr* package (Rinker, 2017) (details see Table 2).

A combination of audio and text features was used to predict trust. The two feature sets were merged based on unique audio identifiers associated with each utterance in the study. A z-score standardization was conducted on all feature sets and the response variable.

3.2.7 Algorithm Training and Evaluation

The algorithms were selected based on three main types of machine learning models (i.e., gradient descent-based, distance-based, and tree-based) as well as stacking ensemble models using *caretEnsemble* package. The ensemble models specify a higher-order model to learn how to best combine the predictions of sub-models. A total of eight models were selected:

1. Linear model.
2. Support Vector Machines with Radial Basis Function Kernel (svmRadial).
3. K-nearest neighbors (kNN).
4. Random Forest (RF).
5. EXtreme Gradient Boosting using tree-based models (XGBTree).
6. EXtreme Gradient Boosting using a generalized linear model (XGBLinear).
7. Linear ensemble model, which fits linear models across all the modes above.
8. Generalized ensemble model, which fits linear models via penalized maximum likelihood.

These eight models were fitted to all three feature sets (i.e., audio, text, and combined). Therefore, a total of 24 models were trained. For each model, we conducted a group of 10-fold repeated cross-validation with 3 repetitions. The method, group k-fold, considers data from the same participant, who may have similar acoustic features or word choices, as a non-overlapping group and control the same participant would not appear in two different folds. This method can avoid within-subject data leakage by ensuring data from the same participant are not included in the training and test datasets. The predictive performance observed with group k-fold cross-validation estimates performance on another sample of participants from the same population. Therefore, this method penalizes the within-subject similarities and reduces overly optimistic estimates of model performance.

Once the models were trained, we evaluated how well they predicted the response variable trust using two metrics: root mean squared error (RMSE) and adjusted R-squared (R^2_{adj}). RMSE is the square root of the variance of the residuals, which indicates the absolute fit of the model to the data in the units of the response variable. The smaller RMSE, the closer the observed data point is to the predicted value, indicating better performance. R^2_{adj} indicates the variance in the response variable that can be explained by the predictor variables with a penalizing factor for adding independent variables, ranging from 0 to 1. The higher theR^2_{adj}, the better the model performance.

3.2.1 Model Explanation

After picking the best performance model, we explained the model by visualizing the most important features for trust prediction. First, a Variable Importance Plot (VIP) was employed. The VIP shows the mean decrease in accuracy associated with removing a feature from the algorithm. However, the value and ranking of important variables in VIP simply represent the importance based on loss function calculation. Therefore, VIP show how the variation of a single variable affects the trust score. A Partial Dependence Plot (PDP) shows the relationship features and the response variable, accounting for the average effect of the other predictors in the model (Greenwell, 2017). Using PDP, the curve represents how much the variable affects the final prediction at specific values of the variable. While PDP provides an average effect of a feature, it does not show specific instances or participants. An Individual conditional expectation (ICE) shows the effect of a feature for each instance separately, resulting in one line per instance, compared to one line overall in partial dependence plots. A PDP is the average of the lines of an ICE plot.

3.3 Results

The 24 participants can have at least 72 conversational turns with the agent, which leads to at least 1728 conversational segments in total. The audio data contained 1806 segments, with a mean length of 8.17s (SD= 10.88). For the text data, we only included utterances that included sentiments and excluded answers to question 4 (e.g., I selected procedure 1) since it does not contain meaningful lexical indicators. The text data contained 810 lines of utterances, with the mean text length of 38.25 characters (SD=26.49). The two datasets were joined by matching the common audio identifiers, leaving the final dataset with 810 lines of utterances. The Welch Two Sample t-test testing the difference of trust values by reliability condition (mean in group high = 5.78, SD = 0.86; mean in group low = 4.37, SD = 1.44) suggests that the effect is positive, statistically significant, and large (difference = 1.38, 95% CI [1.05, 1.70], t (146.46) = 8.42, $p < .001$; Cohen's d = 1.20, 95% CI [0.89, 1.51]).

3.3.1 Feature Engineering

A total of 23 features were extracted, including 13 for audio and 10 for text. The Boruta algorithm identified 23 features as important. The VIF score for multicollinearity identified 3 features above 10, which were removed. The 20 remaining features are described in Table 2.

Table 2. Definition of Reduced 20 Features.

Category	Feature	Description
Audio	nSample	A total number of records/samples in the sound.
	$\bar{x}, SD\ (F0)$	Mean and standard deviation of fundamental frequency (F_0). Closely related to pitch, the fundamental frequency is defined as the lowest frequency of a periodic waveform, which conveys tone, intonation, emphasis, and physiological information and emotion in the speech (Bishop & Keating, 2012).
	$\bar{x}, SD\ (F1)$	Mean and standard deviation of the first formant in vowels (F_1). A formant is a concentration of acoustic energy around a particular frequency in the speech wave. F1 is inversely related to vowel height. The higher the F_1, the lower the vowel height.
	$\bar{x}, SD\ (F2)$	Mean and standard deviation of the second formant in vowels (F_2), which is related to the degree of backness. The higher the F_2, the more front the vowel.
	$\bar{x}, SD\ (F3)$	Mean and standard deviation of the third formant in vowels(F_3), which is related to the degree of roundness. The lower the F_3, the rounder shape of the lip.
	$\bar{x}, SD\ (F4)$	Mean and standard deviation of the fourth formant in vowels (F_4), which is related to the degree of resonance/larynx. The higher the F_4, the higher the larynx.
	$\bar{x}, SD\ (MFCC)$	Mean and standard deviation of Mel-frequency cepstral coefficients. Mel-frequency cepstral coefficients (MFCCs) represent the short-term power spectrum based on human hearing perception, which is the most widely used feature in speech recognition.
Text	Response length	Number of words in text response before any text cleaning (e.g., removing stop words, tokenization, stemming, etc.).
	TF-IDF	Term Frequency-Inverse Document Frequency evaluates how relevant a word is to a document in a collection of documents.
	AFINN	The overall sentiment of the utterance using AFINN lexicon(Nielsen, 2011), divided by the square root of total terms with the sentiment, was scaled from -5 to 5.
	Positive AFINN	The proportion of positive sentiment is divided by the square root of total terms and the overall AFINN score.
	Context sentiment	Sentiment score considering the context for the utterance (window size of 4 words before and 2 words after) and searched for valence shifters. The finalized score was summed and divided by the square root of the word count yielding a Context Sentiment score scaled from -5 to 5 for each sentence ((Rinker, 2017).

Category	Feature	Description
	Non-sentiment proportion	The proportion of the words within each sentence that do not have any sentiment is based on the lexicon.
	Translation error	A binary indication of the reliability of the speech-to-text software.

Table 3. Machine Learning Models Evaluation Using RMSE and adjusted R2.

		Linear Model	kNN	svmRadial	RF	XbgTree	XgbLinear	Linear Ensemble	Generalized Ensemble
Text	RMSE	0.90	0.90	0.91	**0.78**	0.82	0.79	1.26	0.94
	R^2_{adj}	0.15	0.15	0.16	0.34	0.29	**0.37**	0.04	0.11
Audio	RMSE	0.93	**0.78**	0.88	0.84	0.95	0.87	2.66	2.54
	R^2_{adj}	0.16	**0.41**	0.27	0.32	0.20	0.29	0.25	0.25
Combined	RMSE	0.86	0.71	0.71	**0.56**	0.62	0.56	0.78	0.86
	R^2_{adj}	0.26	0.48	0.51	**0.71**	0.61	0.70	0.64	0.68

3.3.2 Trust Estimation

Table 3 shows the machine learning model performance across text, audio, and the combined features. Using only audio features, random forest outperformed other models in terms of R^2, whereas kNN outperformed based on RMSE value. For text-only features and the combined text and audio feature sets, both metrics agreed that random forest outperformed other models by having the lowest RMSE and the highest R^2_{adj}. Trust score prediction had an RMSE score of 0.56 which represents the difference between the predicted and the actual trust score. Compared to the linear baseline model, the best-performing model's R^2_{adj} improved from 0.26 to 0.71. This means that using the conversational features adjusted for the number of predictors, the random forest model can explain the 71% variance of trust. The result is notable because cognitive states, especially trust, are difficult to predict.

3.3.3 Model Explanation

Because the model of the random forest with the combined features shows the best performance, we applied VIP and PDP to investigate the relationships between features and trust. The VIP, shown in Figure 7, indicates that context cluster sentiment from the text data, the mean of formants, Mel-frequency cepstral coefficients, and standard deviation of fundamental frequency were the most important features for predicting trust. Based on the ranking in Figure 7, we used the top 8 features for the following analysis.

To further investigate feature relationships with trust, Figure 8 shows the PDP plots of the eight most important variables. The plot shows the relationships between the response variable (i.e., trust score) on the y-axis and the conversational features (e.g., context sentiment, F2, F4) on the x-axis. Both the x- and y-axis are continuous scales showing the relationships between features and predictions. Most of the features show a sigmoid-shaped curve, which suggests that the trust transition from low to high follows a logistic growth and shows nonlinearity. In other words, people's transition from high to low trust may be a sudden shift, rather than a linear change.

For each pair of relationships in PDP, positive relationships were observed between trust and sentiment (context and AFINN), F1, F2, and F3. The F4 and mean of MFCC revealed an inverse relationship with trust. The standard deviation of the fundamental frequency shows a u-shaped curve, which can be the characteristic that F0 is sensitive to gender. Specifically, trust was significantly higher when the context sentiment score was above 2, while negative sentiment between 0 and -2 predicted increasingly lower trust scores. Trust also increased as F1 increased up to around 500 Hz, and F2 increased up to around 1600 Hz. However, trust decreased as F4 increased to 3500 Hz, and MFCC coefficients increased to around 2.

Figure 9 shows the Individual conditional expectation (ICE) plot. Compared to PDP, which plots the target covariates' average partial effect on the predicted response, ICE plots each instance reflecting the predicted response as a function of other covariates, conditional on the observed feature. The values for a line can be computed by keeping all other features the same, creating variants of this instance by making predictions for these newly created instances. Thus, ICE can show how individual behavior departs from the average behavior.

For each feature, most instances are similar and follow the shape of curves in PDP, which means changes in the feature has a similar effect across individuals. There is a small subset of instances at the bottom of each feature that is relatively constant, indicating that those participants with lower trust individuals do not follow the general trend of PDP. However, different individuals have different starting predictions in the plot (i.e., high versus low trust), so it is hard to tell whether the ICE curves differ between individuals based on such a wide range. Figure 9 shows the centered-ICE, in which centers the curves are fixed to 0 at the minimal value of the trust and shows only the difference in prediction to this point. The centered-ICE curves highlight differences between people and show that the cumulative effects are consistent across participants.

The two-way partial dependence plot in Figure 11 shows the dependence of predictor variable trust on joint values of two features. For the combination of context sentiment and F2, the trust

score increased as the context sentiment score was greater than 1 and F2 higher than 1400 Hz. This means that with a positive context sentiment and a high second formant frequency in the voice, trust is scored higher. For context sentiment and F1, people with context sentiment scores greater than 1 and F1 higher than 500 Hz trust scores are much higher. This means that when detecting positive context sentiment along with higher than 500 HZ of the first formant frequency in the voice, trust scores would be scored as 5.0 and higher, out of a 7-point scale. For context sentiment and MFCC, higher trust was predicted when MFCC is lower than 2 and sentiment is greater than 1.

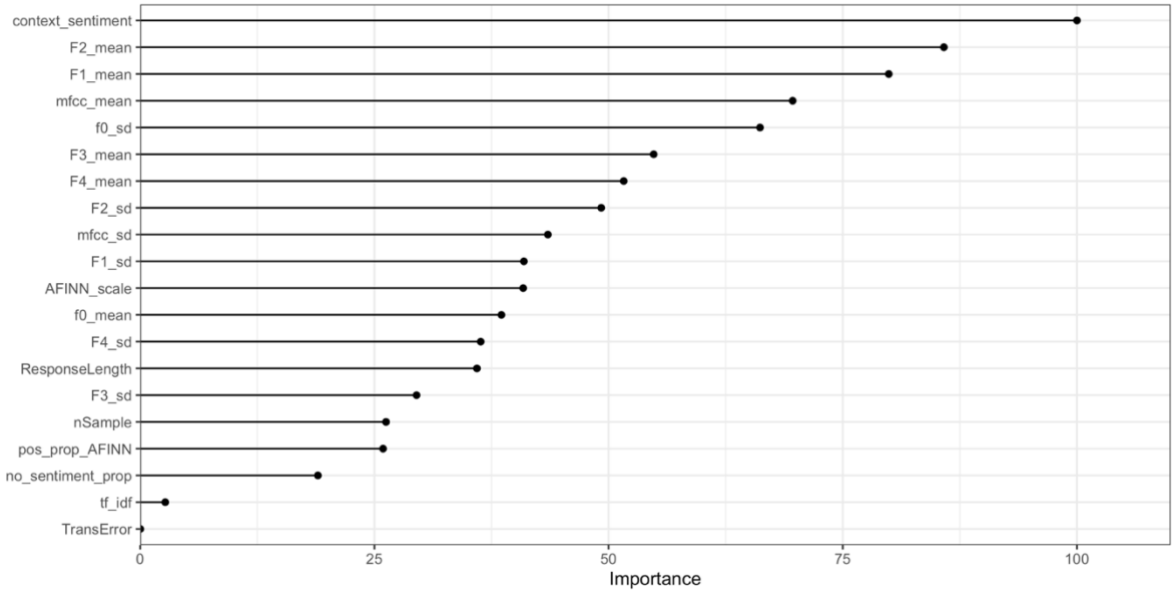


Figure 7. Variable Importance Values for RF Algorithm Based on the Mean Decrease in Accuracy Associated with Removing the Feature.

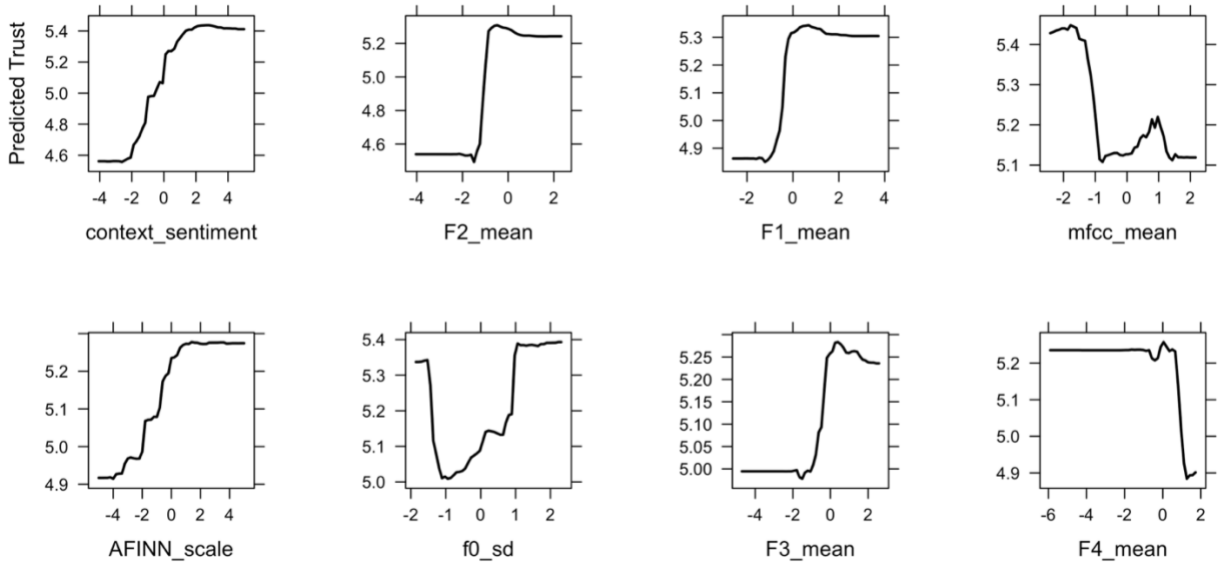


Figure 8. Partial Dependence Plot (PDP) for the Eight Most Important Features Based on Variable Importance Plot in Figure 7. The Ranges of All Features on the X-Axis Are Scaled to 0. The Predicted Trust on the Y-Axis is in the Range of 1 to 7.

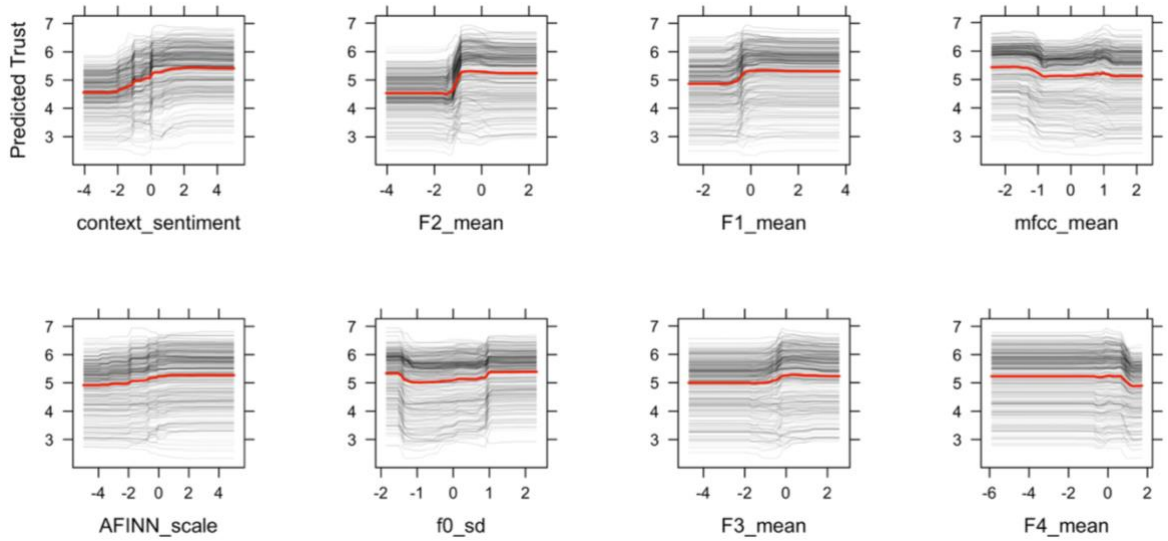


Figure 9. Individual Conditional Expectation (ICE) Plot of Predicted Trust by the Eight Most Important Features. Each Line Represents a Conversational Turn.

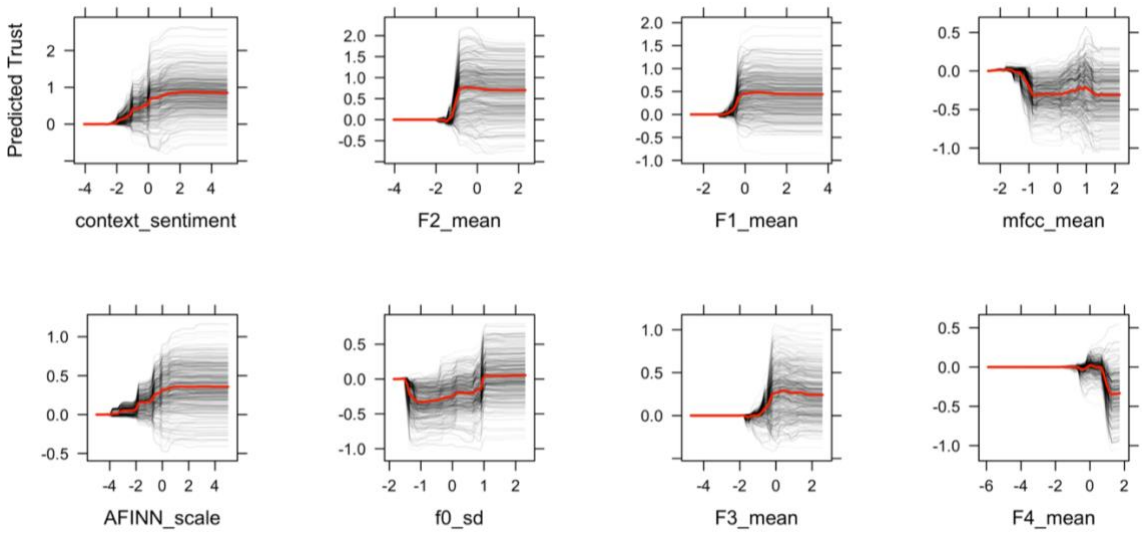


Figure 10. Centered ICE (C-ICE) Plot of Predicted Trust by Top 8 Important Features.

Each Line is Fixed to 0 at the Minimal Values of Each Feature.

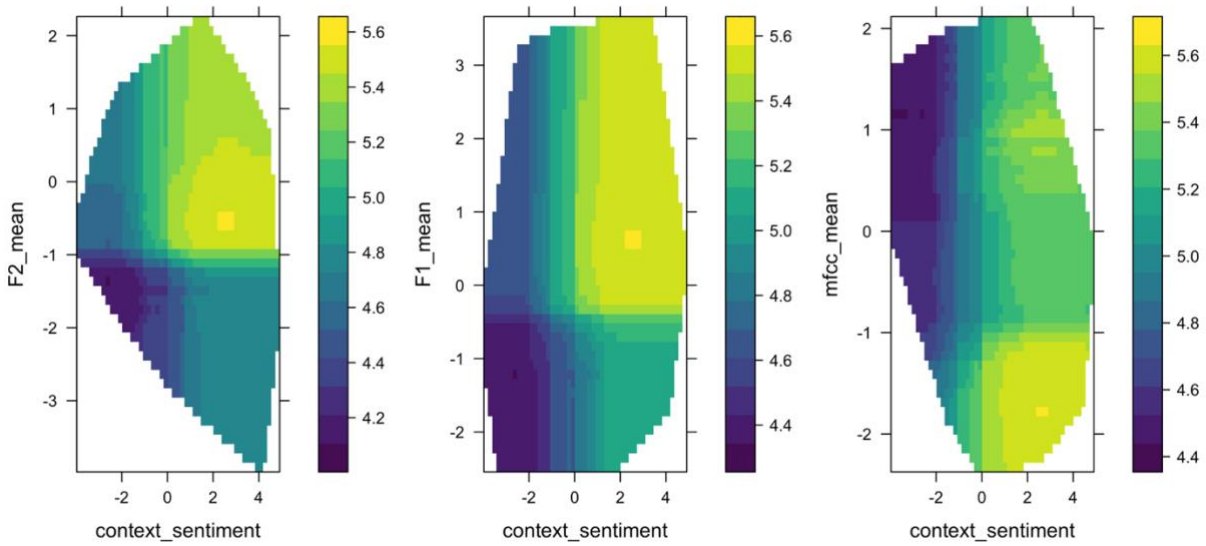


Figure 11. Two-Dimensional Partial Dependency Plots for Context Sentiment, F2, F1 and MFCC Mean Based on the Random Forest Algorithm. The Shading Represents the Predicted Trust Scores. The Outlines of The Region Show the Predictor Space that the Model was Trained On.

3.4 Discussion

This paper aimed to address two questions: can we measure trust in human-AI conversations? If so, what are the most important conversational indicators for trust measurement and future management? For the first question, we designed an aided decision-making study using aid reliability as proxy of trustworthiness to elicit large difference in people's trust and showed that 71% of trust variation can be predicted using a combination of lexical and acoustic features using the random forest algorithm. The large effect size validates as a proof-of-concept that trust can be estimated from the conversations. Compared to prior work on discrete trust classification (Gauder et al., 2021), our work further validated the promising evidence of measuring continuous and real-time trust dynamics in the human-AI conversation. For the second question, we identified the most important trust conversational indicators—context sentiment as lexical cues, formants, fundamental frequency, and MFCC as acoustic cues—and showed that they affect trust in a non-linear manner.

3.4.1 Lexical Indicators of Trust: Context sentiment

For the lexical indicator in the conversation, the context sentiment in the conversations is the strongest predictor of trust. Context sentiment is the average sentiment considering valence shifters and negation in the sentence. For example, "I am not good" contains the positive word "good", but the sentiment score is negative because the sentence contains the negation 'not, which flips the polarity of the sentence. Results showed that positive sentiment predicts higher trust. The result is expected and consistent with prior research: when people used more positive words in their conversation, they rated their trust in the aid higher (Hildebrand & Bergner, 2021). Because benevolence is one of the core elements of trust (Mayer et al., 1995), people express greater affective trust and used positive sentiment words when interacting with a conversational agent.

3.4.2 Acoustic Indicators of Trust: Formants, Mel-frequency cepstral coefficients (MFCC), pitch variation.

For acoustic indicators in the conversations, formants, MFCC, and pitch variation follow context sentiment as the most important predictors of trust. As indicated in Figure 8, a high first formant (F1) and second formant (F2) were associated with a high level of trust. Formant is a spectral property of the speech signal that reflects voice quality as well as linguistic vowel identity (Goudbeek et al., 2009). The formant with the lowest frequency is called F1, the second F2, and the third F3. Prior studies showed that formants can influence people's trust perception (Knowles & Little, 2016; Torre et al., 2020). Our study is the first to demonstrate that formants are also influenced by people's trust levels. In other words, trust influences people' speech production and formant articulation.

There are different ways to explain how trust influences formants in conversations. One way is to consider trust as an affective process. Prior research has shown that formants can be used to discriminate the valance (e.g., positive or negative) and arousal (e.g., excited or calm) dimensions of emotions (J. C. Kim et al., 2011): high arousal emotions result in a higher mean F1, whereas positive valence results in a higher mean F2 (Goudbeek et al., 2009). Thus, our results implied that when people are in a high level of trust, people express a high F1 and F2 in their voice, indicating trust as a positive valence and high arousal emotion. Another potential explanation is that people use different vowels when articulating different levels of trust. Formants are directly associated with tongue positions and pronunciation of different vowels. The F1 was associated with the height of the tongue position (i.e., top or bottom) and the F2 was associated with the backness of the tongue position (i.e., back or front). A high F1 and high F2 would be lower and front tongue position for words like 'bat' (æ) versus a low F1 and low F2 would be

'boot' (u). Results showed a high F1 and F2 for higher trust scores, meaning that participants were saying more words that contained vowels in bottom-front vowels (e.g., æ). The third explanation is when trust is higher, people have a 'smiling voice' indicated by formants. Past studies have shown that when people smile, the first two formants are increased, which leads to a higher perceived trust (Torre et al., 2020). Future studies should further investigate the causal relationships between formants and trust.

MFCCs are coefficients that collectively make up an MFC, which represents the short-term power spectrum of a sound. MFCC is often used to recognize the emotion of a speaker from their voice. Prior research has shown that the mean and standard deviation of MFCC can classify hot anger, neutral, sadness, and happiness (Bhimavarapu et al., 2021; Lalitha et al., 2015; Nalini et al., 2013). Our result showed consistent findings with prior studies that showed MFCCs are an important feature for perceived trust in interpersonal group interactions (Spitzley et al., 2022). Based on the authors' knowledge, our study is the first to show that MFCCs can be used to predict people's trust in their conversations with a virtual agent.

In the past literature, trust perception is usually associated with pitch: voices with low F0 are considered more trustworthy than voices with high F0, in both male and female voices (Montano et al., 2017). To our surprise, F0 is not the most important feature to predict people's trust levels. Instead, the variance of F0 is considered a more important indicator of trust as shown in Figure 7. Syed and colleagues have demonstrated that a more dynamic and varied pitch contour is viewed as more trustworthy compared to flat intonation (Syed et al., 2021). Knowles and Little also showed that dynamic voices sounded more cooperative than monotone voices (Knowles & Little, 2016). High variation in F0 has been associated with prosocial and pleasant vocal attributes in human child-directed speech (Trainor et al., 2000). Thus, when people express a high-level trust, they also exhibit complex contour of the pitch that may signal affiliation.

3.4.3 Implications

Measuring trust from conversations is a natural, unobtrusive, novel method to support human-AI teaming. Our findings and theoretical implications for developing a conversational measurement of trust. Predicting trust using lexical and acoustic features provided initial validation in measuring trust unobtrusively and dynamically in conversation. To develop a standardized conversational measurement of trust, limited research has been conducted or discussed. Our study used pre-defined prompts and conversational structure to elicit people's trust-relevant responses in a performance-based task. The conversational features we identified are promising measures of

trust. Since conversations are highly context-dependent, future studies are needed to test the ecological validity by generalizing these measures to other contexts. Additionally, how to measure trust in a free-flow conversation remains unsolved. The main bottleneck is the technical limitation of the state-of-art conversational agents. With the emerging powerful conversational agents (e.g., chatGPT3) will provide richer content for establishing a standardized conversational measurement of trust.

Once trust can be measured in conversation, an important next step is trust management. For a system to be trustable, it will have to adapt to its user's trust levels. In a performance-based human-AI interactions, we can compare the estimated people's trust levels with the system capability and identify whether people are over or under trusting the system. Based on findings in our study, an adaptive conversational agent can be developed: the conversational agent could incorporate these identified trust indicators to actively probe, repair, and temper trust (Chiou and Lee, 2021). When people overtrust the agent, meaning people's trust is higher than the actual trustworthiness, the agent can signal the trust tempering cues, such as using the negative sentiment and lower formants. The next question would be whether these identified trust indicators show the same effect on trust perception. In other words, these identified conversational features can predict people's trust, but can they influence perceived trustworthiness? Our findings show a mixed result from the prior literature: Although pitch significantly affected perceived trustworthiness (Elkins & Derrick, 2013), it is not the most important feature when people express their trust in the conversation. On the other hand, sentiment and formants show that they can be used to both predict trust and influence perceived trustworthiness (Montano et al., 2017). Future studies are needed to show whether the identified conversational indicators are effective to calibrate people's trust.

3.4.1 Limitations and Future Work

There are several limitations to this study. First, the conversation is limited in size and scope. Our study focused on the influence of reliability on trust in performance-based human-agent interaction. The word use and other conversational cues in our dataset might not generalize to other domains of trust (e.g., human-human trust). To establish the conversational measure of trust, a generalized protocol of trust-related questions should be established and validated. Second, the conversation design between humans and agents is restricted due to the technical limitations of chatbot implementation. Although the variation of agents' responses and questions varied, the conversational agent in our study is a decision-tree-based agent, rather than an intelligent agent that can hold a rich conversation. Therefore, the conversation complexity and length were limited.

Advances in conversational agents will produce richer data for trust measurement. Future studies can consider other lexical and acoustic cues, such as pauses between conversational turns, interruptions, and interjections. Third, while PDP and ICE can suggest causal hypotheses related to trust, these should be verified (Zhao & Hastie, 2021).

3.5 Conclusion

To enhance human-AI teaming, AI needs to monitor and manage trust in real-time. Conversational data provides a novel approach to measuring real-time trust. This study showed such real-time, conversational trust measures are possible by training machine learning models on lexical, acoustic, and combined conversational features. A random forest model that used the combination of lexical and acoustic features explained 71% of the variance in self-reported trust. The combination of lexical or acoustic features outperformed either alone. We identified the most important lexical and acoustic cues and further showed that trust transition follows a non-linear shift. These results show the importance of including both audio and text features when measuring trust dynamics in a conversation. An open question is whether they might be used to modulate the voice of the conversational agent to manage the trust.

3.6 Chapter Summary

This chapter focuses on the first research question: "Can we measure people's trust in the human-AI conversation". Measuring trust through conversation is a novel yet unexplored approach. In Chapter 3, we designed an experiment to estimate trust in human-AI conversations using machine learning (ML) models and analyzed the data using a machine learning model. Our predictions accounted for 71% of the variance in rated trust using lexical and acoustic cues from human-agent conversations. While most MLs are treated as black boxes, we showed an explainable ML by visualizing the most important features using partial dependence plots. Estimating trust in communication opens the door for real-time and unobtrusive trust management. Building on this foundation, Chapter 4 adopts the dynamic system theory to explain people's diverging trust levels on automation and Chapter 5 deepens the understanding of trust in conversation by modeling the temporal changes of conversational topics. Once measure and model the trust, Chapter 6 investigates the effects of conversational features identified in Chapter 3 for trust management.

Chapter 4. Explain Trust Divergence Using Dynamic System

Title: Explaining Trust Divergence: Bifurcations in a Dynamic System

Conference: Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Submission date: February 28th, 2023

Acceptance date: May 30th, 2023

Abstract

When people experience the same automation, their trust in automation can diverge. Prior research has used individual differences—trust propensity and complacency—to explain this divergence. We argue that bifurcation as an outcome of a dynamic system better explains trust divergence. Linear mixed-effect models were used to identify features to predict trust (i.e., individual differences, automation reliability, and exposure). Individual differences associated with trust propensity and complacency increases the R^2 of the baseline model by 0.01, from $R^2 = 0.40$ to 0.41. Furthermore, the Best Linear Unbiased Predictors (BLUPS) for random effect of participants were uncorrelated with trust propensity and complacency. In contrast, modeling trust divergence from a dynamic perspective, which considers the interaction between reliability and exposure along with the individual by-reliability variability fit the data well ($R^2 = 0.84$). These results suggest dynamic interaction with automation produce trust divergence and design should focus on state dependence and responsivity.

4.1 Introduction

As intelligent agents become increasingly autonomous on progressively more complex tasks, trust becomes more essential to designing effective human-automation cooperation (Chiou & Lee, 2021). Trust, defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 54), is crucial for ensuring appropriate reliance on automation and avoiding its misuse, disuse, or abuse (Parasuraman, 1997). Often, people's trust in automation often evolves and converges to a relatively homogeneous level of trust. However, trust can also diverge. Interacting with the same automation, some people might develop high levels of trust whereas others might grow to distrust it (Kamaraj et al., 2023; Liu et al., 2021). This divergent trust is an interesting form of trust miscalibration because it describes how some people might over-trust and others might under-trust the same automation.

It might be useful to consider trust divergence as qualitative changes or 'bifurcation' in how people experience and trust automation. Bifurcation, well-studied in dynamical systems, describes how a small initial change of a system made to the parameter values, known as bifurcation parameters, can cause a sudden topological change in its behavior. In the context of trust in automation, this small initial change is often considered as differences in initial trust and individual differences, as well as variance in their initial perception and interaction with the automation. The bifurcation parameter refers to the changes in automation characteristics, such as an error. Previous research has highlighted various reactions following automation failures, including disbelievers, Bayesian decision-makers, and oscillators (Bhat et al., 2022). While researchers often rely on individual differences to explain diverse group behaviors. Yet, focusing solely on individual behaviors neglects the temporal aspect of how initial individual differences compound with the subsequent experiences of automation characteristics, especially when encountering the 'bifurcation parameter' (e.g., automation errors). The underlying mechanism contributing to the stabilized and diverging trust has received little attention and merits investigation. Three factors, namely individual differences, automation characteristics, and trust dynamics, may account for the trust bifurcation. In this paper, we argue that adopting the concept of bifurcation as an outcome of a dynamic system offers a more suitable framework for explaining trust divergence.

4.2 Background

4.2.1 Individual Differences

The wide range of individual differences, encompassing backgrounds, personalities, and knowledge of automation, contributes to the variability in individuals' propensity to trust automation. Those with a higher inclination to trust may experience a greater decline in trust when interacting with low-performing automation (Merritt & Ilgen, 2008). Moreover, individuals who with a stronger "perfect automation schema" demonstrated greater declines in trust when they encountered automation errors (Dzindolet et al., 2002). Additionally, individuals also vary in automation-induced complacency, which can manifest as either a failure to detect or an delayed response to detecting errors (Bailey & Scerbo, 2007; Merritt et al., 2019). Prior research has found that complacency interacts with automation characteristics: the higher the system reliability, the more likely the operators become complacent (Parasuraman et al., 1993). Minor differences in individuals can influence the initial level of trust and subsequently shape the interpretation of new information. Thus, individual differences can influence trust divergence.

Hypothesis 1: Individual differences predict diverging of trust in automation.

4.2.2 Automation Reliability and Exposure

Because trust calibration is the correspondence between a person's trust in automation and the automation's capabilities, it has been consistently shown that automation capability significantly influences trust in automation (Dzindolet et al., 2002). Automation failures often have a much stronger influence on trust than automation successes: trust is difficult to build but can be lost quickly (Dzindolet et al., 2003a; Manzey et al., 2012). Trust is continuous process influenced by the trust of a previous moment (Yang et al., 2023). Exposure to automation reflects the extent to which individuals have encountered and interacted with automated systems. Repeated exposures can have both positive and negative effects on individuals' behaviors and trust in automation. On one hand, repeated exposure can increase familiarity, indirectly influencing trust (Mayer et al., 1995). On the other hand, repeated exposures, especially with highly reliable automation, can induce complacency and decreased situational awareness, resulting in over-reliance on automation and over-react to automation errors (Dzindolet et al., 2002). Thus, the automation capability and exposure to automation can be potential causes of the diverging levels of trust and motivate the second hypothesis.

Hypothesis 2: Automation reliability and exposure predict diverging of trust in automation.

4.2.3 Trust Dynamics

Trust is inherently dynamic. People calibrate their trust over time as a continuous cognitive process (Gao & Lee, 2006). While researchers have highlighted the continuous and temporal elements of trust dynamics (Yang et al., 2023), limited past research has used trust dynamics to explain people's divergent opinions on automation. Using trust dynamics, trust divergence can be modeled as a bifurcation in a dynamic system: a small change in the initial state gradually influences behavioral framing and subsequent decision-making processes. This bifurcation results in trust stabilizing as two distinct trajectories. For example, in supervisory control, the individual differences shape the decision between manual control and automation. Once either decision is selected, it would provide positive or negative experiences. The experiences create inertia to keep people only focusing on either the advantages or disadvantages. Automation failures can be bifurcation transient point, which leads to trust divergence and long-term maintenance in certain states. Thus, the structural changes of the bifurcation depend on the combination of individual differences, the automation performance and the exposure, and their interaction over time, rather than on any individual factor alone.

Hypothesis 3: Trust is a dynamic system. People's varying responses to the interaction of automation characteristics and exposure predict diverging of trust in automation.

4.3 Method

subject study. Participants performed 12 decision-making tasks associated with managing a system of a simulated space station: the Habitat's Carbon Dioxide Removal System (CDRS). Participants were assisted by a conversational agent (Bucky) with 2 levels of reliability (i.e., high, and low). Each level of reliability had 2 repeated cycles of the CDRS tasks, each including 3 events (i.e., startup, venting, shutdown). Details of the study were documented in (Li et al., 2022).

4.3.1 Participants

A total of 24 participants (18 female, 6 male) were recruited (M = 23.7, SD = 3.6). Recruitment inclusion criteria included that participants should be comfortable using a computer and a touch screen interface as well as have some technical background (e.g., completion of engineering or science courses). Due to the safety concerns of COVID-19, the study took place online. It was a two-session, two-day study with each session lasting up to two hours. In total, the study lasted approximately four hours for each participant. Participants received $30 per hour for up to $120 for four hours of participation.

4.3.2 Procedures

After signing the consent form, participants completed a two-part training: the first provided a study overview and training on the CDRS system, while the second included an interactive demonstration of working with Bucky on decision-making in PRIDE. During the study, participants had 25 minutes to use the CDRS system to remove CO_2 from Habitat's environment by running the CDRS through three events (startup, venting, and shutdown) before their crew experienced CO_2 poisoning. For each event, the participant made two essential decisions with Bucky's aid. The first was selecting a procedure to run to remove the CO_2. Bucky recommended a procedure. The participant could either accept Bucky's recommendation or reject it and choose a different procedure. The second decision was deciding whether to rerun the procedure selected. As part of this decision participants would be advised by Bucky if the state of the CDRS was incorrect and if a different procedure should be run. The participants could either accept Bucky's recommendation or reject Bucky's recommendation and run a different procedure. The participants made their decisions either based on their knowledge from their training session or by relying on Bucky's recommendation. Once the procedure was selected, PRIDE automated the procedure execution. If the participant selected the incorrect procedure, an error occurred. The

participant then had to manually stop the procedure and reselect a procedure. The participant finished the event by confirming the procedure ran correctly and completed the trust ratings.

4.3.3 Data Analysis

Linear mixed effect models identified features predicting trust as measured by the 12-item, 7-point Likert scale (Jian et al., 2000). To test our hypotheses regarding how individual differences, automation characteristics, and dynamics explain trust, we gather relevant features for each hypothesis.

For individual differences, we measured people's automation complacency and propensity to trust. We adopted the Automation-Induced Complacency Potential-Revised scale (AICP-R) (Merritt et al., 2019), which is a 10-item with response options on a five-point Likert scale ranging from 1 (strongly agree) to 5 (strongly disagree). Example items include, "Constantly monitoring an automation is a waste of time." For propensity to trust, we measured people's general tendency to trust automation using the Propensity to Trust Machines questionnaire (Merritt, 2011). This scale consists of six items with response options ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Example items include, "I usually trust machines until there is a reason not to."

Automation characteristics were modeled as reliability condition and exposure. Reliability is a binary indicator of agent performance. Exposure is defined as the number of times participants experience the same automation characteristics, which is the number of cycles participants experienced.

For the trust dynamics hypothesis, we considered the interaction of automation characteristic and exposure along with individuals' varying responses to the experiences.

4.4 Results

The mean trust score for the high-reliability condition was 5.78 (SD = 0.86) whereas, for the low condition, the mean trust score was 4.37 (SD = 1.44). From Figure 12 we observed that the path taken by individuals throughout the experiment was highly variable: some maintained a steady level of trust throughout the experiment, while others had dramatic drops in trust. The black lines represent six participants: three with the highest standard deviation and three with the lowest standard deviation in mean trust. The difference in paths reveals a divergence in trust when participants experience the low-reliability condition.

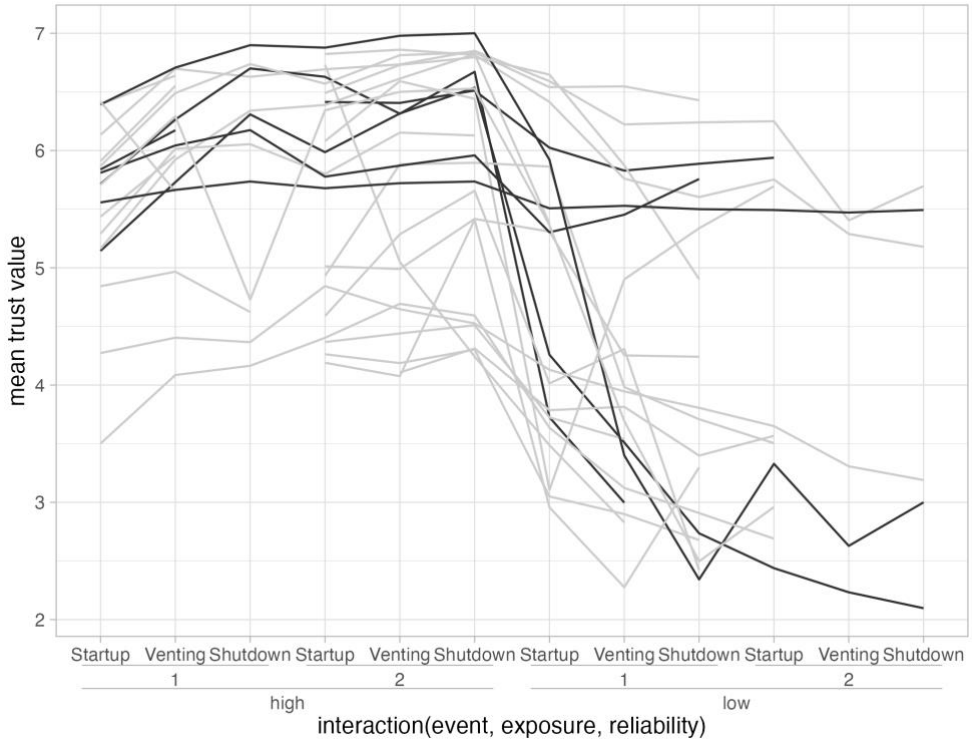


Figure 12. Trust diverges when people experience low-reliability automation.

Table 4. Performance metrics comparison between regression models

#	Model	Formula	RMSE	AIC	BIC	R^2 (cond.)
0	Baseline model	$trust \sim 1 \mid ID$	0.97	671.38	681.49	0.40
1	Individual differences	$trust \sim complacency + propensity + (1 \mid ID)$	1.03	613.94	630.15	0.41
2	Automation reliability and exposure	$trust \sim reliability + exposure + (1 \mid ID)$	0.70	555.31	572.16	0.68
3	Trust Dynamics	$trust \sim reliability + exposure + reliability * exposure + (reliability \mid ID)$	0.48	481.13	508.09	0.84

In **Table 4**, four linear mixed-effects models were built. Models were evaluated using the root mean square error (RMSE), Akaike information criterion (AIC), Bayesian information criterion (BIC), and conditionalR^2 value. RMSE reflects the difference between predicted and actual values. AIC and BIC reflect how well the model fits the data with a term that penalizes model complexity. The lower these three metrics, the better the model performance. The conditionalR^2 is the proportion of total variance explained by the model. The higher the R^2, the better the model performance.

Model 0 (Reliability | ID) uses ID as the random intercept serving as a baseline model which accounts for the overall trust level due to general individual differences.

Model 1 corresponds to the first hypothesis and tests the effects of specific individual differences on trust in automation, which were measured using automation-induced complacency and propensity to trust scales. The individual measures only slightly improved the marginalR^2 value. The effect of complacency and propensity are both statistically non-significant (p = 0.61, p = 0.38).

Model 2 corresponds to the second hypothesis and tests the effects of automation characteristics on trust. We used automation reliability and the number of cycles as exposure to automation. We added reliability and exposure as fixed effects to determine if the model performance would be improved. The effect of reliability [low] is statistically significant and negative, β = -1.40, 95% CI [-1.62, -1.19], t(210) = -12.82, p <.001; Std. β = -1.07, 95% CI [-1.38, -0.75], whereas the effect of exposure is non-significant, t(210) = -0.69, p = 0.49.

Model 3 corresponds to the third hypothesis and tests the effect of trust dynamics by adding the interaction between reliability and exposure along with the individual by-reliability variability. By adding the individual by-reliability variability, model 3 shows diverging effects in Figure 12 and would serve as the baseline model. The total explanatory power of this model is substantial with a high conditional R^2 value (0.84) and the part related to the fixed effects alone. Additionally, the AIC and BIC are the lowest for model 4, which indicates that the trust dynamic model explains the greatest amount of trust variation using the fewest possible parameters. Within this model, the effect of low reliability is statistically significant and negative, β = -1.10, 95% CI [-1.51, -0.69], t(207) = -5.24, p <.001; Std. β = -0.84, 95% CI [-1.16, -0.53]. The effect of exposure is statistically significant and positive, β = 0.24, 95% CI [0.05, 0.44], t(207) = 2.43, p =.02; Std. β = 0.19, 95% CI [0.04 0.34]. The interaction effect of the exposure and reliability is statistically

significant and negative, β = -0.80, 95% CI [-1.14, -0.47], $t(207)$ = -4.70, p <.001; Std. β = -0.62, 95% CI [-0.88, -0.36]. Trust in the high-reliability condition is an estimated 4.89 on a Likert scale of 7. The trust score is 1.10 points lower in the low condition, 0.24 points higher in the second exposure, and 0.81 points lower if there is an interaction between the low condition with the second exposure. For the random effects, the standard deviation for by-subject random intercepts indicates that trust levels for subjects varied around the average intercept of 0.69 points by about 0.77 points. Additionally, we used Best Linear Unbiased Predictions (BLUPs) to predict random effects and found no correlations with the automation complacency ($R^2 < 0.01$) and propensity to trust ($R^2 = 0.03$). These results again validate that individual differences do not account for trust divergence and supports the trust dynamics hypothesis.

4.5 Discussion

We observed that trust diverges when people experienced automation error: some people maintained a steady level of trust whereas others showed a drastic decline in trust. To explain this trust divergence, we evaluated three hypotheses—individual differences, automation characteristics, and trust dynamics–using linear mixed effects models. We found that the trust dynamics model, which uses automation exposure and reliability as an interaction fixed effect, with individual differences and participants as a random intercept and slope, yielded the highest R^2 and lowest AIC and BIC values. Results suggest that the trust dynamics model best explained the trust divergence. Because trust dynamics consider individual differences and how people's trust is reinforced by the automation characteristics and multiple exposures over time. Our results reinforce the notion that individual differences alone are insufficient to explain trust divergence. Instead, the concept of bifurcation in a dynamic system may provide a better explanation. This concept describes how even slight changes in a system can lead to qualitatively different behavior, which might correspond to certain individuals maintaining stable trust in automation while others experience sudden shifts in trust.

Whether trust diverge reflects enduring traits or states that emerge from automation interaction has major system design implications. These mechanisms parallel those associated with the concept of "accident proneness." Prior studies found that individuals who have experienced incidents of accidents in the past are more likely to experience them in the future than are individuals who have not experienced an accident (G. E. Bates & Neyman, 1952). Heckman argued that this conditional probability of accident proneness is based on structural relationships of state dependence, rather than heterogeneity in population and individual differences (Heckman, 1981).

Enduring individual differences or traits suggest an emphasis on selection in system design, whereas state dependence would emphasize interaction design.

Interaction design from the trust dynamic perspective suggests that systems should measure and manage trust across human-automation interactions. Rather than focusing only on generic trust calibration through more transparent designs, a dynamic perspective suggests a focus on "responsitivity", where the automation detects and responds to changes in trust(Chiou & Lee, 2021). The importance of a dynamic perspective is even more important in hybrid teams with more than one human operator interacting with the automation. In these teams of over- and under-trust can circulate as a contagion within the network. Trust circulates through the network via explicit communication or implicit observations of others' interactions and norms (Stewart, 2003). Drawing inspiration from the widely used Susceptible-Infectious-Recovered (SIR) dynamic system model in epidemiology, researchers can explore the influence of network dynamics on trust bifurcation (Nakahara & Doya, 1998). Gorman and colleagues have previously conceptualized teams as dynamic systems, revealing the importance of concepts like attractors and synchronization (Gorman et al., 2017). Future research can understand and model trust dynamics in a hybrid team, identifying the roles and impacts of attractors, perturbation, and synchronization.

Our findings on trust dynamics conforms with the state dependence theory (Heckman, 1981). When designing the system, it is crucial adopt a state-dependent and dynamic perspective to evaluate human performances and trust. Early-stage measurement of trust and identification of distinct populations experiencing divergent trust patterns can inform the development of personalized systems to manage trust more effectively.

4.6 Conclusion

Even when people experience the same automation, their trust in automation can diverge over time. Prior research has typically focused on individual differences to explain trust divergence. However, we showed that trust divergence was best modeled by trust dynamic perspective, which considers the interaction between reliability and exposure along with the individual by reliability variability ($R^2 = 0.84$). Our results suggest the concept of bifurcation in dynamic systems, which describes how small changes in a system lead to sudden shifts in behavior, might explain trust divergence.

4.7 Chapter Summary

The present chapter built upon the preceding Chapter 3, which focused on trust estimation from the conversations, and delves deep into the trust divergence between high and low trust

groups among individuals. Chapter 4 explored the temporal dynamics aspect and addressed the second research question, that is "*how does trust change over time*?". Compared to using the individual differences to explain the diverging levels of trust over time, we argued that trust divergence can be better explained as an outcome of a dynamic system. We adopted linear mixed-effect models to predict trust and showed that modeling trust divergence from a dynamic perspective, which considers the interaction between reliability and exposure along with the individual by-reliability variability fit the data well. Consequently, our results suggest that dynamic interactions with automation contribute to trust divergence, emphasizing the need for designs that prioritize state dependence and responsivity. This chapter established a robust foundation for the temporal trust dynamic perspective in Chapter 5, where we further examined the temporal aspects in the human-AI conversations.

Chapter 5. Model Trust Dynamics in Human-AI Conversation

Title: Modeling Trust Dynamics and Evolution in Human-Agent Conversation: A Trajectories Epistemic Network Analysis Approach

Journal: International Journal of Human Computer Interaction

Submission date: October 31st, 2022

Acceptance date: January 23rd, 2023

Abstract

Chapter 3 developed the machine learning approach, which can combine lexical and acoustic features to predict trust in the conversational agent; however, this focuses on the feature level and ignores the rich context and deep meaning of the conversation. In other words, the connections between the features and the meaning associated with features are situated within the context that might benefit from qualitative analysis. Furthermore, the temporal changes of trust in conversation cannot be captured. In Chapter 4, the dynamic system approach was adopted to better frame the temporal changes. Thus, to capture trust dynamics, in Chapter 6, we aimed to model two aspects: (1) Trust dimensions: the connection to theoretical foundations of trust, especially focus on cognitive processes in conversations, rather than feature level or using bag-of-words; (2) Trust dynamics: the temporal aspect of trust evolution throughout the interactions, rather than aggregated or a snapshot of trust. In Chapter 4, we modeled dynamic trust evolution in the conversation using a novel method, trajectory epistemic network analysis (T-ENA). T-ENA captures the multidimensional aspect of trust (i.e., analytic and affective), and trajectory analysis segments the conversations to capture temporal changes of trust over time. Twenty-four participants performed a habitat maintenance task assisted by a virtual agent and verbalized their experiences and feelings after each task. T-ENA showed that agent reliability significantly affected people's conversations in the analytic process of trust, $t(38.88) = 15.18$, $p = 0.00$, Cohen's $d = 4.72$, such as discussing agents' errors. The trajectory analysis showed that trust dynamics manifested through conversation topic diversity and flow. These results showed trust dimensions and dynamics in conversation should be considered interdependently and suggested that an adaptive conversational strategy should be considered to manage trust in HATs.

5.1 Introduction

As artificial intelligence (AI) becomes increasingly capable and able to outperform humans in certain tasks, humans and AI may gradually cooperate as coworkers than tools {Citation}. Trust

in the human-AI team (HAT) is a step beyond current trust in automation and poses new challenges. The interdependent team requires the modeling of trust to reflect the team processes and how the team activity unfolds over time. This requires a continuous and observable stream of data to record the cognitive processing of trust dynamics. In HAT, teammates often need to exchange and update the information to achieve a joint task. Communication, as team cognition, can provide such contextual and process-based means for trust modeling (Cooke et al., 2013). The conversation naturally holds temporal functions of coordination, which can be used to show changes in human-AI relationships over time. Additionally, people naturally express their feelings and attitudes, such as trust, in communication via the tone of their voices, choice of words, turn-taking, and pragmatic meanings in the context. Human trust has been shown can be estimated from the human-AI conversation (Li et al., 2022). Trust in communication not only aligns well with the nature of interdependent HAT but also provides an essential means to model and analyze the trust dynamics and how it evolves throughout team interactions. Thus, modeling trust dynamics in HAT using conversational data provides a promising yet under-explored approach.

Since trusting in communication is highly contextual, and dynamic, and keeps evolving in the interdependent human-AI teaming, we adopted a novel approach – *trajectory epistemic network analysis (T-ENA)* to develop a dynamic model of trust evolution in human-agent conversation (Brohinsky et al., 2021). The developed trust model coded the conversational data using *epistemic network analysis (ENA)*, which provides a contextual understanding of human-AI communication (Shaffer, 2017). Similar to the structure of the social network analysis, the nodes in ENA provide the concepts that are defined based on the trust framework and edges provide the connections between concepts based on the co-occurrence in the human-AI conversations. The trajectory analysis characterized and decomposed the multiple interactions as a trajectory to demonstrate the changes in trust in AI. In summary, in this paper, we used T-ENA to model trust dynamics in human-AI conversations by focusing on multidimensional and temporal dimensions.

5.1.1 Trust Dynamics

Multidimensionality

Trust, defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"(Lee & See 2004), has been studied for decades to understand and manage the relationships between people and automation. Trust is an intrinsically complex social construct. In understanding and modeling trust in HAT, the cognitive processes should be highlighted. Because communicative cues are often used to investigate their cognitive

processing, which is also governed by and manifests people's trust. Understanding and modeling the cognitive processes in communication can reveal more insights into HAT team cognition.

Based on existing trust frameworks, trust depends on the interplay among analytic, analogical, and affective cognitive processes(Lee & See, 2004). The analytic process is formulated based on the accumulation of knowledge and rational evaluation of the interactions. The analogical process is less cognitively demanding but relies on the rules, intermediaries, and environmental context. The affective or emotional process is critical since it represents not only how people think, but also how they feel about automation. When describing the affect, the circumplex model is often used, which uses two-dimensional arousal-valence axes to describe human emotions. For example, the affect excited is high arousal and positive valence affect, whereas sad is low arousal and negative valence. Trust has been shown affected by the valence and arousal stimulus. Dunn and Schweitzer show that positive valence (e.g., happiness, hope) increases trust, while negative valence (e.g., fear, guilty) decreases trust (Dunn & Schweitzer, 2005). Yet, how affective process of trust is mapped on the affect circumplex model in human-agent conversations is not well understood. Additionally, we know little about the interplay between different cognitive processes underlying trust in human-agent communication. To fill the research gaps, we designed a decision-making task with various levels of conversational agent reliability to reflect their cognitive processes in communications. The factor, automation reliability, which is governed by the analytic process, has been well-studied and shown the causal relationship on trust (Dzindolet et al., 2003b). Using reliability, we aimed to elicit various levels of trust and show the significant difference and interplay between the affective and analytic processes.

Temporality

Another important aspect of trust dynamics that merits more attention is its temporal characteristics. A shift from the snapshot view of trust to a trust dynamic is important (Yang, Schemanske, et al., 2021). Because trust is time-dependent and evolves throughout human-agent interactions (Kaplan et al., 2021). Trust calibrates and evolves based on the various automation characteristics and experiences as relationships between parties mature (Korsgaard et al., 2018a; Luo et al., 2022). Trust is reinforced by the experience and is further impacted by a function of the trust itself in the previous moment (e.g. positive and negative feedback loops) (Falcone & Castelfranchi, 2004; Lee & Moray, 1992; Manzey et al., 2012). Additionally, adding the temporal aspect allows us to examine the recency effect that associated with trust dynamics, meaning that interactions happened more recently may have more value than those that happened some time

back (Desai et al., 2012). Thus, analyzing and modeling the temporal changes gives a more nuanced inspection of the trust evolution throughout the HAT.

To model this trust evolution, trust should be measured multiple times and further modeled by considering time units in the model. Yang et. al. proposed a computational model proposes that trust at any time t, follows a Beta distribution, which shows good prediction accuracy (Yang, Schemanske, et al., 2021). Although modeling trust evolution is relatively new and limited, a history of literature has shown that human behaviors and attitudes can be modeled by the time-dependent dynamical system approach. Gottman, Swanson, and Swanson (2002) showed how marriage outcomes can be modeled using the dynamical system analysis, which focused on the temporal dynamics of partner communication. Using such nonlinear dynamical systems methods to model relationships is becoming more prevalent (Demir et al., 2021).

The multidimensional and temporal aspects of trust are not independent. Instead, the influencing factors and their impacts on various processes of trust should also vary throughout the human-AI interactions over time. In the interpersonal domains of trust, Korsgaard and colleagues outlined a stage model that captures the trust formation from an early stage of calculus-based, to a knowledge-based trust and eventually an identification-based trust based on aligned values and goals (Korsgaard et al., 2018b; Lewicki et al., 1996). In various stages of trust, the impact of predictors and processes on trust systematically varies over time (Korsgaard et al., 2018b). It is important to highlight time as a moderator on different antecedents of trust. Within the domain of human-AI trust, to the authors' knowledge, limited research investigated the relationship between the multidimensional and temporal aspects of trust dynamics in the conversation. In summary, we aimed to model trust dynamics by decomposing the cognitive processes (i.e., analytic and affective) of trust in the human-agent conversation and show how these two dimensions of trust processes evolve.

5.1.2 Modeling trust in conversation

For such complexity in trust, a critical challenge is to model trust that reflected the highly contextual, dynamic, and keep evolving relationship between humans and AI teammates throughout the interaction. To model trust, which is a latent variable, we need to infer or measure trust indicators first, such as through subjective, behavioral, and physiological measurements. Although reporting a subjective rating of trust is always used and treated as the gold standard in the human-automation interaction due to its reliability and generalizability, it is not always fully reflecting and capture the dynamics in the teaming since it is often obtrusive and one-shot. The

interruptions and the deliberate thinking while self-reporting the attitudes towards automation cannot naturally represent the joint cognitive processing that happens in human-AI cooperation. Using the behavioral measures of trust, such as compliance and reliance, on the one hand, provides more capabilities of sampling more frequently throughout the interactive; yet, on the other hand, it can be highly dependent on the task and limited to the decision spaces available, which are often considered as an indirect product of trust attitudes. Physiological measures, such as electrodermal activity, eye movement, and heart rate, can provide truly real-time trust indicators with greater sensitivity. However, it also suffers from challenges, such as outcomes that must be contextualized with expert knowledge and examination during periods where trust is active and relevant. Facing the new challenges of trusting in human-AI teaming, an alternative trust measure should be identified to model trust properly.

One under-explored method is modeling trust in conversation. Conversational data can be considered as a mixture of behavioral and physiological data that contain lexical, semantic, phonological, and pragmatic representations of the conversations. In other words, people naturally express their trust attitudes via the words they used, the sentence structure, and the tone of the voices in their conversation, which are all contextualized. Prior works have shown that people express their trust not only through what they say (e.g., the sentiment of the words), but also via how they say it (e.g., formants)(Li et al., 2022). According to the interactive team cognition theory, communication is team cognition, which can be a non-obtrusive measure of team interaction dynamics(Cooke et al., 2013). Communication is also essential for trust building and calibration, which in turn, can promote effective human-AI teaming (Fuoli & Paradis, 2014).

Prior research has used both qualitative and quantitative approaches to identify and model trust in conversational data. Qualitative analysis, such as grounded theory, provides a rigorous and systematic approach to identifying the situated meanings and systematic patterns in the data (Oktay, 2012). However, compared to a machine-aided approach, manual coding is often laborious, limited to small volumes of data, and subject to the coders' domain knowledge. For quantitative analysis, such as text analysis, the dominant approach treats the conversations as bag-of-words, which assumes words are independent units. This approach ignores the meaningful context and patterns in the conversation. Prior research has shown that using a machine learning approach can combine lexical and acoustic features to predict trust in the conversational agent (Li et al., 2022). However, the machine learning interpretation focused on the feature level. In other words, the connections between the features and the meaning associated with features are situated within the context and cannot be easily interpreted. Moreover, the temporality and the sequence

of the conversation are often lost by text processing, such as bag-of-words. In summary, to capture trust dynamics, we modeled two aspects of trust dynamics in the method: (1) Multidimensionality: consider meaningful and interpretable connection based on theoretical foundations of trust (rather than feature level or using bag-of-word); (2) Temporality: consider time-series trust evolution throughout the interactions (rather than aggregated or a snapshot of trust).

5.1.3 Trajectory Epistemic Network Analysis

To address the multidimensional and temporal aspects of trust dynamics, we apply Trajectory Epistemic Network Analysis (T-ENA), which can both decompose multidimensional trust using an Epistemic Network Analysis (ENA) and project the trajectory of the network structure over time.

ENA is a quantitative ethnographic technique that estimates the network structure of coded data based on co-occurrences that define connections between the coded data (Shaffer, 2017). ENA can systematically identify a set of meaningful features in the data based on the triangulation between human coders and computer-based text analysis (Shaffer et al., 2016). Originally designed to model theories of cognition, discourse, and culture challenges in learning analytics, ENA assumes that the structure of the connections is more important than the mere presence of those elements in isolation, ENA has been applied to many domains, making it a promising method to analyze social interactions, including gaze coordination during the collaborative work (Andrist et al., 2015) and shared agency in online collaborative learning (Tan et al., 2022). Prior works have demonstrated successful applications of ENA to human factors and ergonomics (HFE) discipline because the visual representations can help researchers quickly identify and compare the difference between interested groups (Weiler et al., 2022; Wooldridge et al., 2018). Additionally, the differences can be quantitatively defined with the support of qualitative evidence from the conversation. In our work, we applied ENA to construct and visualize a multidimensional space of trust based on analytic and affective processes in the human-agent conversation.

To model trust dynamics, one major limitation of ENA is that it typically aggregates data across conditions and time, which ignores the temporal features. Trajectory ENA considers the temporal structure to reflect process-oriented concepts, such as trust dynamics. T-ENA accounts for the change in the network structure that evolves by incorporating time units or temporal segmentation. By dividing the complex ENA into various time units, T-ENA allows the reader to examine the changes along the temporal dimension, which cannot be easily interpreted when using aggregated

means(Tan et al., 2022). Together, modeling trust dynamics using T-ENA can represent both the multidimensional and temporal aspects of trust.

5.1.4 Research Objectives

The objective of this study is to investigate the trust dynamics in human-agent conversations and teams. Two research questions were formed: 1) how do humans indicate different trust levels in human-agent conversations? 2) how does human-agent trust conversation change over time? Because communications are highly contextual, and dynamic and keep evolving as the team progresses, we adopted a novel approach, trajectory epistemic network analysis. Specifically, to address the first question, we showed the multidimensional aspect of trust dynamics using ENA. To address the second question, we showed the temporal aspect of trust evolution using trajectory analysis of ENA.

5.2 Method

5.2.1 Study Design

The data we analyzed came from a $2 \times 2 \times 3$ within-subject study. Participants completed 12 decision-making tasks moderated by a human where they managed a Carbon Dioxide Removal System (CDRS) that is part of an analog Mars habitat. Participants were assisted by a conversational agent with 2 levels of agent reliability (i.e., high, and low). Each level of reliability had 2 cycles of the CDRS tasks, each including 3 events (i.e., startup, venting, shutdown). The high-reliability conversational agent provided 100% correct recommendations whereas the low-reliability agent provided 20% correct recommendations. The 12 total events were designed to elicit various levels of trust through differing agent reliability. At the end of each event, the agent initiated a conversation by asking six trust-related questions (see Table 5). Once the participant finished the conversation, they then completed a trust survey(Jian et al., 2000).

Table 5. Examples of conversational trust questions.

Number	Question
1	How would you describe your experience selecting the procedure?
2	Why would you feel that? Can you explain your answer in more detail?
3	Can you talk more about my performance in providing the recommendation?
4	That makes sense. Which procedure did you select?
5	Can you tell me more about your strategy for picking that procedure?
6	How can I be more helpful in terms of providing recommendations?

5.2.2 Participants

A total of 24 participants (18 female, 6 male) were recruited from the Madison, WI area (M = 23.7, SD = *3.6*) . In total, each participant had the opportunity for 72 conversational turns with the agent. The cleaned text data contained 1981 lines of utterances, with a mean text length of 38.25 characters (SD = 26.49) . Additionally, we evaluated the relationship between reliability and trust. A t-test showed that the mean trust score for the high-reliability condition (M = 5.78, SD =*0.86*) was significantly higher than the low condition (M = 4.37, SD = 1.44), $t(23)$ =4.12, p = *0*.0002. Thus, for the high-reliability condition, we can investigate the conversational indicators associated with high trust and vice versa.

5.2.3 Trajectory Epistemic Network Analysis

For trajectory epistemic network analysis (T-ENA), we adopted a four-step process as shown in Figure 13: (1) data segmentation, (2) directed content analysis, (3) network analysis, and (4) trajectory analysis.

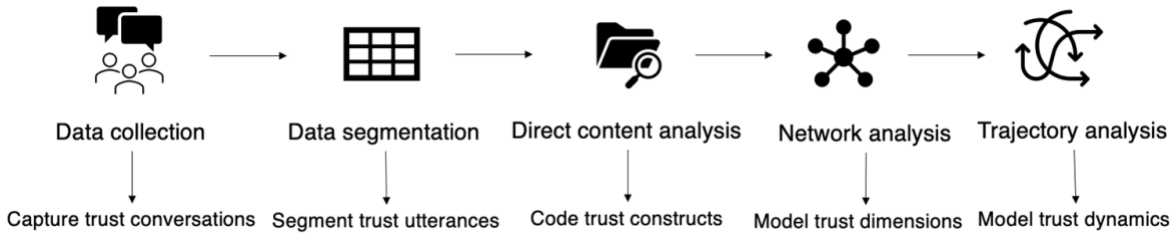


Figure 13. Trajectory Epistemic Network Analysis Process and for Assessing Trust Dimensions and Dynamics.

Data segmentation

Conversation data between participants and the conversational agent were recorded in the log files. The data were segmented based on the conversational turn. We also added meta-data to facilitate data segmentation: (1) Reliability condition that participants experienced. This is used as the grouping variable for comparison between two conditions. (2) Participant ID, which is used as a 'unit' in the ENA. (3) Question ID, which is used as 'conversation' for ENA. Conversations are collections of lines within which ENA models connections between concepts. (4) Codes, which are concepts whose patterns of association that want to model (explained in session 3.3.2).

Directed Content Analysis

Based on trust dynamics models, six codes were identified as shown in Table 6, which include four codes related to the analytical processes of trust and two codes related to the affective

processes of trust. For analytical processes of trust, codes were selected and defined in an iterative round of coding, two researchers combined deductive and inductive coding to refine and validate codes. For affective processes of trust, we adopted the circumplex model of affect (J. A. Russell, 1980), which suggests affect is described in a two-dimensional circular space, containing arousal and valence dimensions. We excluded two quadrants in the valence-arousal affect model: positive valence, high arousal (e.g., excited), and negative valence, low arousal (e.g., sad) since these did not appear in the conversational data.

The directed content analysis identified trust components that exist within the participants' conversations with the virtual agent throughout the task. We dual-coded each line using a binary coding structure: '1' if the code exists, or '0' if the code does not exist per each segment. Coders compared codes and categories and re-coded certain segments to resolve disagreements. Any disagreements were resolved until the inter-rater reliability across all codes reached Cohen's $\varkappa > 0.65$ and Shaffer's $\rho > 0.9$ between two human raters and the automated classifier. After validating each code, we applied the automated classifiers to the data set to code the data.

Network Analysis

Table 6. Codebook of trust-related constructs included in Epistemic Network Analysis.

Code	Definition	Example from data
System capability	Participants commented on Bucky's past and/or current performance and ability to provide the appropriate recommendation for the tasks.	*"I think your* ***performance*** *was good since it worked out well."*
System error	Participants commented on errors in Bucky's recommendations.	*"The procedures didn't line up to* ***what I thought the right procedure would be****."*
User capability	Participants commented on their self-efficacy and their belief in his or her capacity to execute the task.	*"Bucky is incorrect this round, but I'm* ***confident*** *in myself for choosing the correct one."*
System process scrutiny	Participants recalled the specific system knowledge to understand or clarify how the system operates.	*"It was the only option in which the* ***EPS was powered up before the ATCS was activated. If the EPS is not powered up, then the ATCS can't be activated****, therefore I assumed this was the only procedure that would be effective."*
Positive valence, low arousal	Participants expressed their affect that is positive and low	*"I feel like I've reached a routine with my method of choosing the*

	aroused, such as calm, contented, and relaxed.	*procedure. So, I enter the same state of* ***calm****."*
Negative valence, high arousal	Participants expressed their affect that is negative and highly aroused, such as confused, frustrated, stressed, nervous, and annoyed.	*"I get even more* ***confused*** *with Bucky's recommendation"*

To compare the trust indicators in the conversation, ENA was used to define units as each conversational turn, conversations as the utterances after each CDRS task, and comparison groups based on the reliability condition. The ENA algorithm uses a moving window to construct a network model for each line in the data, showing how codes in the current line are connected to codes that occur within the recent temporal context defined as 12 lines (each line plus the 11 previous lines) within a given conversation. Codes that occurred outside of this window were not considered connected. The resulting weighted networks are aggregated for all lines for each unit of analysis in the model. Nodes correspond to codes; edges correspond to the relative frequency of co-occurrence between each pair of codes and the weights or thickness of the edges show the connection between nodes.

In this model, we aggregated networks using a binary summation in which the networks for a given line reflect the presence or absence of the co-occurrence of each pair of codes. Then, the co-occurrence of codes in adjacency matrices was summed across the moving window. Next, ENA is normalized using spherical normalization by dividing each vector by its length. Once data is normalized, ENA performs a singular value decomposition (SVD) using the first two SVD dimensions. Once ENA is created, to determine if the high reliability is statistically different from the low-reliability conditions, we conducted t-tests on the centroids of networks. Specifically, the centroids are calculated by computing the mean values of each edge weight in the networks.

Trajectory Analysis

To create trajectories, we employed R package trajectoryENA (Brohinsky et al., 2021) and coded the conversations with 12 time units, which is each conversation after each event (startup, venting, shutdown). Thus, each reliability group was represented by six time units. Time units means were projected in the aggregated ENA space described above. Group means were plotted and sequentially connected by cubic splines, which can produce curves between successive time points. Adding the time unit to the ENA allows us to investigate how people's trust evolves from the beginning to the end, which the aggregated ENA analysis ignores.

5.3 Results

5.3.1 Epistemic Network Analysis

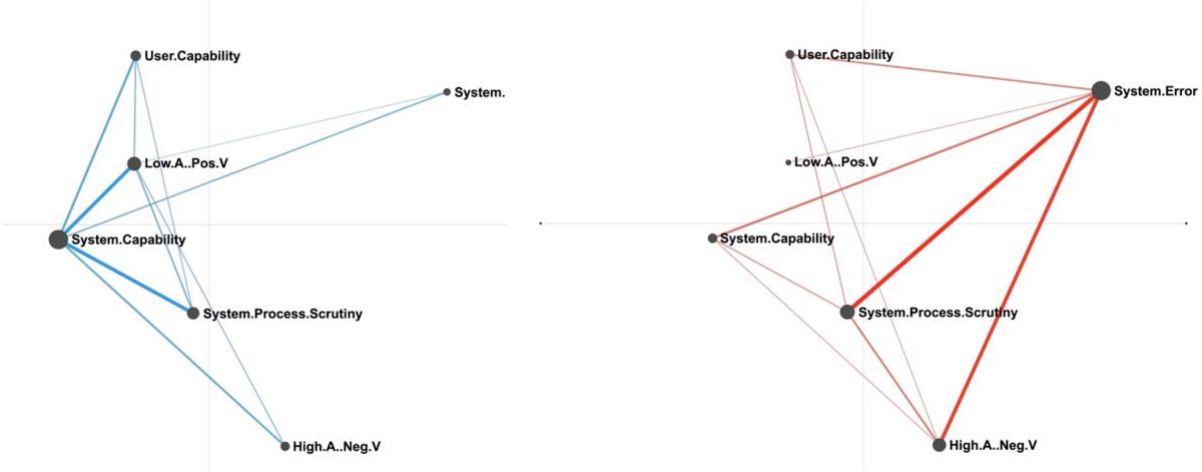


Figure 14. ENA Network for High (Left) Versus Low (Right) Reliability.

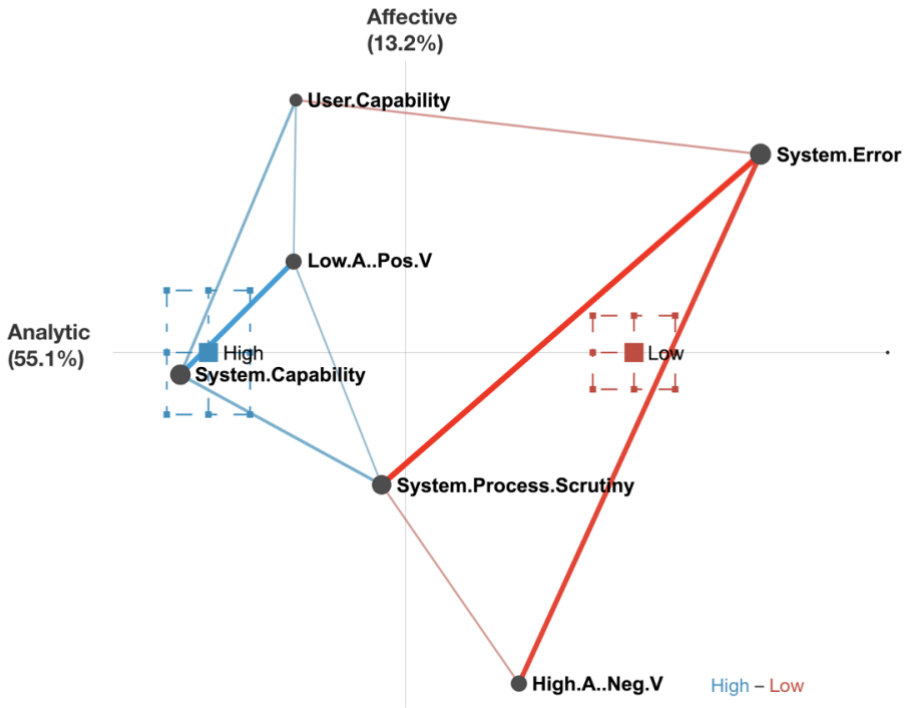


Figure 15. ENA Network of Subtracted Connections for High Reliability (Blue) Versus Low Reliability (Red). The Points Represent Coded Topics, and the Edges Represent the Cooccurrence of the Topics. The Thicker the Edges, The More Frequently the Topics Co-occur in The Human-Agent Conversation. The Square Points and Associated Error Bars Represent the

Centroids and the Confidence Interval of the Network.

Epistemic Network Analysis (ENA) visualization results contain: (1) a plotted point, which represents the location of that unit's network in the low-dimensional projected space, and (2) a weighted network graph. The positions of the network graph nodes are fixed and are determined by an optimization routine that minimizes the difference between the plotted points and their corresponding network centroids. Because of this co-registration of network graphs and projected space, the positions of the network graph nodes—and the connections they define—can be used to interpret the dimensions of the projected space and explain the positions of plotted points in the space. Our model had co-registration correlations of 0.98 (Pearson) and 0.98 (Spearman) for the first dimension and co-registration correlations of 0.92 (Pearson) and 0.90 (Spearman) for the second. These measures indicate that there is a strong goodness of fit between the visualization and the original model.

Figure 15 shows subtracted network graphs depicting the discourse differences between high reliability and low reliability and Figure 14 shows the network for high and low reliability. In these network graphs, nodes correspond to the codes identified that are relevant to trust indicators in the conversations, and edges reflect the relative frequency of co-occurrence or node connection within each conversation between participants and the conversational agent. Thus, the thicker the edges, the stronger the node connection is observed in the human-agent conversation.

The centroids presented in Figure 15 summarized the dimension of each network. Centroids indicated by boxes and confidence intervals (dotted lines) enable comparisons of networks statistically as well as visually. To test the differences between the reliability conditions, we applied a two-sample t-test. Along the x-axis, a two-sample t-test showed that the high-reliability condition (M = -1.15, SD = 0.55, N= 22) was statistically significantly different at the α=0.05 level from low-reliability condition (M = 1.33, SD = 0.50, N= 19), t(38.88)=15.18, p<0.001, Cohen's d = 4.72. Along thy-axis, the two-sample t-test assuming showed high reliability (M = 0.00, SD = 0.82, N= 22) was not statistically significantly different at the α= 0.05 level from low reliability (M = 0.00, SD = 0.45, N= 19), t(24.45) < 0.001, p = 1.00, Cohen's d =0.

To interpret the results of the ENA network, the x-axis and y-axis should be interpreted and defined based on the code placement and researchers' domain knowledge. Nodes placed at extreme edges of the space provided more information for labeling the axis. As observed in Figure 15, the x-axis reflects the codes that capture conversations related to the degree of *analytic processes* of trust. These include system capability, system errors, system process scrutiny, and user

capability. Moving from left to right along the x-axis indicates conversation topics shift from positive aspects of system capability to negative aspects such as system errors. The y-axis shows the codes that reflect the *affective processes* of trust in the system, which includes the high/low arousal and positive/negative valence of affects. The statistical significance on the x-axis suggests that analytical processes of trust differ between high and low-reliability conditions.

ENA subtracted network also provides the visual representation to explain the reasons for the statistical difference between node connections in high and low-reliability conditions. The connected lines represent the subtracted connections or co-occurrences of two codes. Based on Figure 16, in the high-reliability condition, the strongest connection is *System Capability* and *Low Arousal, Positive Valence*, indicating that when the conversational agent is performing well, people usually commented on the system performance along with positive valence and low arousal affect words, such as calm and relax. Additionally, the connection between *System Capability* and *User Capability* indicates that people often reflected on their self-efficacy and talked about their capability when the system performs well. In the low-reliability condition, between the affective and analytical processes of trust, we noticed a strong connection between *System Error*, *High Arousal, and Negative Valence*. This means that people associated low performance and lower levels of trust with high valence and negative arousal words (e.g., annoyed). Additionally, there is a strong connection between *System error* and *System Process Scrutiny*. This suggests that in the low-reliability condition, people expressed their low level of trust by thinking aloud about the specific system processes, such as reflecting on what states CDRS should have been in certain situations (i.e., *System Process Scrutiny*).

Another key feature of ENA is that it allows researchers to trace connections in the model back to the original data and validate the quantitative results qualitatively (Shaffer, 2017). The significant result on the x-axis in Figure 15 indicates that the conversation between high-reliability versus low-reliability conditions differed along the analytical level conversation codes. When in high reliability, the conversation is centered around the system performance (e.g., The performance is good). When in low reliability, the conversation was more centered on the errors that occurred in the system and its connections with the system scrutiny (e.g., The CO2 is supposed to be at a lower level).

5.3.2 Trajectory ENA

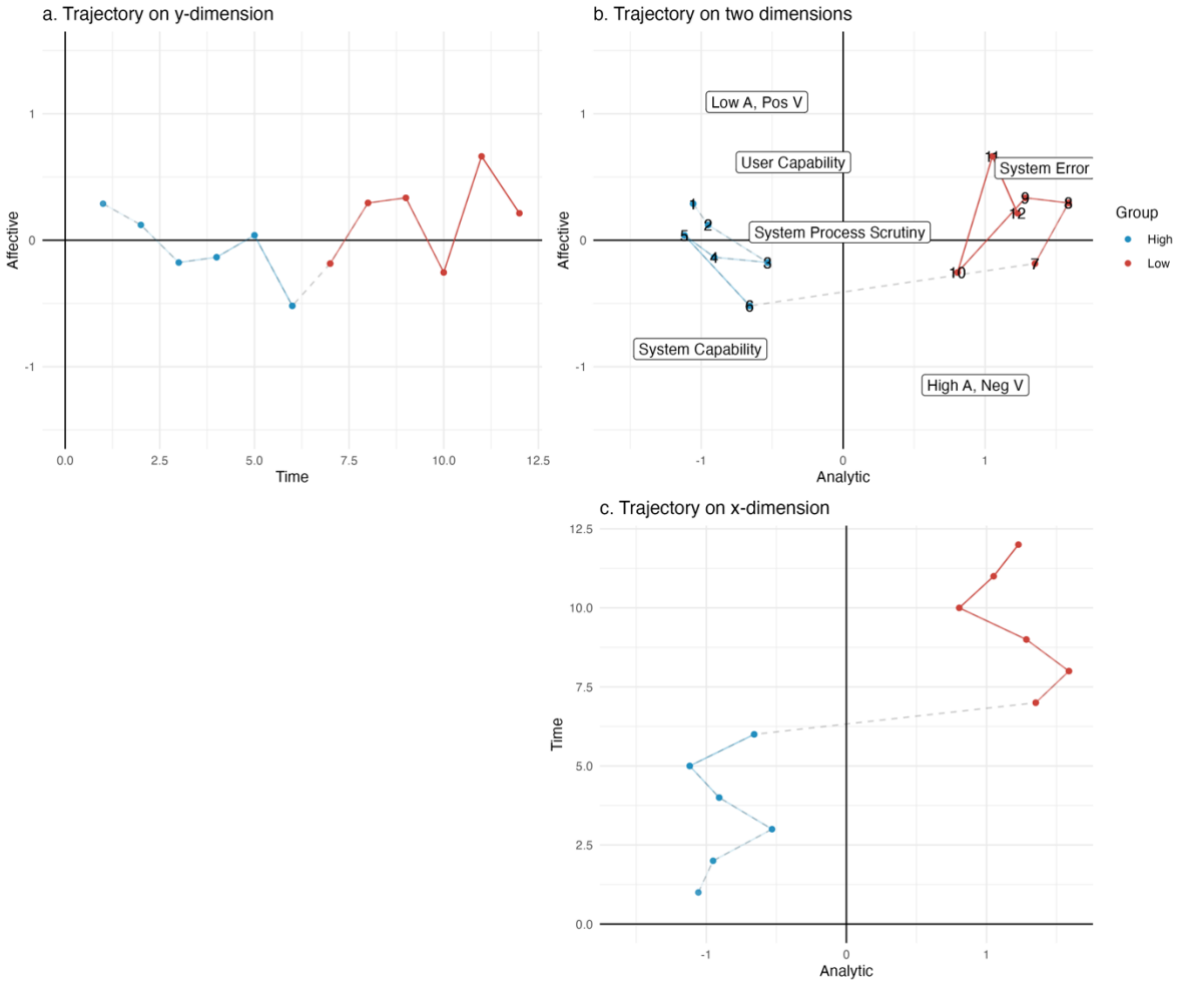


Figure 16. Trajectory ENA. Figure (a) Shows the Trust Dynamics Changes in the Y-Dimension as a Function of Time. Figure (b) Shows the Two-Dimensional Trajectory Mapping onto the Network Result. Figure (c) Shows the Trajectory Changes in X-Dimension as a Function of Time. The Increasing Transparency Indicates the Increase in Time Throughout the Interaction.

Figure 16 shows the trajectory model for the two reliability groups across 12 interactions. Every point on the graph shows the mean for each time unit, which is each conversation after each event (in total 12 conversations). A total of three time-series trajectory ENA plots were plotted: one with two-dimensional ENA showing both affective and analytic processes (Figure 16. a and c) and two one-dimensional ENA with each process along with the time (Figure 16.b). Figure 16.c tracks change along the y-dimension (affective process) aligned with the x-axis of the original ENA space. The y-axis for the plot in Figure 16. c tracks changes along the x-dimension (analytical process) aligned with the y-axis of the original ENA space. Figure 16. b maps the two-dimensional

trajectory on top of the ENA node positions. These three plots provide a way to examine the trust evolution in each dimension as a function of time. Additionally, since the subplots were co-registered with the main plot, the comparison between plots also allowed changes in the subplots to be interpreted simultaneously and tracked in the dimensions of the main ENA space.

To further interpret the trajectory, three crucial variables need to be disentangled: changes in the x-dimension, changes in the y-dimension, and progression in time. Figure 16.a. shows the trust evolution on the y-dimension, showing the changes in affective trust over the interactions. The oscillation throughout the y-dimension indicates a mixed emotion related to trust shown in the conversation. For example, the subject commented: "Since this is the first time Bucky's been incorrect, it confused me for a little bit and made me second-guess myself just because Bucky's been so accurate". Figure 16.c. shows the trust evolution on the x-dimension, which is the analytic process of trust as a function of time. Compared to the affective process of trust as shown in Figure 16. a, which is more continuous and non-significant between high and low-reliability groups, the analytic processes showed a distinct pattern difference. This suggested that the affective process of trust transition is less sudden, compared to the analytical process.

Figure 16. b shows the trust evolution on both dimensions, which shows the interaction between multidimensional and temporal aspects of trust. We noted a distinct difference in the variance and direction of the trajectory between high and low-reliability groups. The variance of the trajectory suggested the diversity of the conversational topics. The low-reliability group shows a wider range on both the x and y dimensions, indicating higher volatility when people express a low level of trust. There are two explanations for the high volatility: 1. people have mixed emotions (affective) and analytical judgment when interacting with a poorly performing agent; 2. more individual differences in people's responses when they are in a low level of trust (Liu et al., 2021). The direction of the trajectory indicates the topic changes and trends over time. The direction of the trajectory indicates the changes and convergence of topics over time. When people have high trust in an agent, they would attribute their capability with positive sentiment and later confirm the system's capability. When people interact with low-reliability agents with low trust, conversations would note the system error and then converge towards checking the system process with a large variance in the affective processes. In sum, our T-ENA results showed that trust changes as a function of time, which vary between analytical and affective dimensions.

5.4 Discussion

To better understand trust dynamics and evolution in the human-AI teaming, we adopted a novel approach, trajectory epistemic network analysis (T-ENA), on 24 human-agent conversations. Specifically, we explored the multidimensional aspect of trust using ENA and temporal change of trust using the trajectory analysis of ENA. For the multidimensionality, the ENA plots provided meaningful connections between analytic and affect processes of trust concerning agent reliability. For the temporality, the temporal analysis segmented the change of trust throughout the courses of the human-agent interactions and mapped it with analytic and affective dimensions of trust.

5.4.1 ENA Showed an Interplay Between Analytic and Affective Processes of Trust

A significant difference between high and low-reliability conversations was shown in the x-axis, which is interpreted and labeled as an analytic process of trust. Results suggested that people express different trust states by using distinct analytic information, such as commenting on system performance and noticing errors. This is expected since we manipulated the reliability of the conversational agent, which maps to the analytic process. No significant difference was found in the affective process. This suggested that the manipulation of reliability showed less influence on the affective process, which aligns with prior literature that affective process has a greater influence on the analytic process than the analytic has on the affective (Lee & See, 2004). Especially in low risks and self-relevant decisions, the effect of affective process on trust is much weaker (Midden & Huijts, 2009). In in our case that the CO_2 removal procedure won't influence participants' physical environment, thus the physical and psychological distances to the potential hazards are far. Participant experienced less level of risk and low self-relevance, which would induce less affective process of trust.

The network analysis also revealed interactions between analytic and affect processes of trust under the influence of automation reliability. When people show high trust in conversational agents, on the affective dimension of trust, there is a stronger connection between low arousal and positive valence affect with the system and user capability. Complementing our prior paper using machine learning models which showed that positive sentiment predicts trust (Li et al., 2022), ENA results provided more context-relevant information: the positive sentiment is associated with the system capability and users' capability. In the low capability, people indicate high arousal and negative valence and discuss system errors with a detailed inspection of the system process. For the future design of the conversational agent, when people have lower levels of trust, the agent should provide more details on the system processing to support the cognitive processes.

Additionally, the different links between high and low conditions show people's self-serving attribution bias when reason and process trust in communication: credit positive events internally with their capability and attribute negative events externally by scrutinizing the system processes and errors (Miller & Ross, 1975). A prior study showed when a robot gave people credit, people would trust the robot more (Kaniarasu & Steinfeld, 2014; You et al., 2011). Our study provided the potential for positive utilization of people's self-serving bias and blame attribution when designing agent's communication strategies. People might be more likely to accept and trust the virtual agent in the team if the agent credits users' capability if the joint task went well and acclaims some blames if the team performance was poor. Future empirical studies can further validate the hypotheses and show the effects on trust processes.

5.4.2 Trajectory ENA Validated Trust as a Function of Time

For the temporal aspect of trust dynamics, T-ENA showed the temporal change of trust throughout human-agent interactions by mapping the temporal changes in trust to the analytic, affective, and joint dimensions of trust. Our T-ENA results on temporal and multidimensional trust in conversations showed a first empirical validation of the hypothesized dynamic model of trust proposed by Kaplan and colleagues (Kaplan et al., 2021), where trust at each measurement can show different human, agent, and contextual antecedents. We observed clear differences in conversation trajectory on affective and analytic dimensions. More oscillation was observed for the affective dimension of trust, which suggested the mixed emotions and usages of words when people were in high and low trust states. A distinct difference was observed in analytic information in the human-agent conversation. This implied that using analytic information to estimate people's trust transitions can be more effective in human-agent communication.

The variance and direction of the conversational trajectory on the two-dimensional trust dynamics also suggested the differences in conversational topic diversity and flow. When people have high trust in the agent, people's conversation topics are more consistent and converged to the system's capability. When in a low trust, conversational topics are more scattered reflecting heavier cognitive processing. Communication, as a manifest cognitive process, can help researchers to further understand the psychological effects of cognition on trust. Our results on human-agent communications in various trust states shed light on their cognitive processes. Compared to high trust, which leads to a familiar congruent flow of cognitive processing (thus consistent conversational topics), low trust or distrust triggers a spontaneous activation of alternatives and incongruent associations, which can be shown as a diverse topic or verbose examination of the system (Mayo, 2015).

Understanding how human verbalize their thoughts in HAT can design better AI teammate to support the synchrony, which typically involves entrainment—a temporal coupling between independent oscillators that enter some type of phase relationship. Prior research has shown that conversational entrainment can benefit interlocutors by mutually reducing cognitive processing and predict team cooperation (Manson et al., 2013). Designing temporally based conversational coordination can facilitate the trusting relationships and team performance for HAT. Our model on decomposing trust multidimensional and temporal dynamics demonstrated empirical evidence to design conversational strategies for trustworthy agents (Rheu et al., 2021).

5.4.3 Limitations and Future Studies

It is important to note several limitations in our study to better generalize the findings. First, the human-agent conversation has a pre-define decision-tree structure due to the limits of the state-of-art conversational agent capabilities. On the one hand, we were able to compare the difference in answers systematically across the interactions. However, compared to the human-human conversation, the conversations can appear to be limited in terms of the potential topics discussed and initiated by the conversational agent. Thus, the coverage of the topics can be less diverse than human-human conversation, which cannot provide rich information for coupled conversational analysis. Future studies using a more robust conversational agent can generate more dynamic conversations and trust-related findings. Second, since conversations are heavily contextual, the conversation for our study is domain-focused on selecting the correct procedures in a habitat maintenance task. For example, for the node of *System Process Scrutiny*, people used jargon related to our study design, such as the Carbon dioxide removal system. Thus, when considering transferring the findings from our studies to another domain, the coding for the nodes in the network should be considered contextually. Additionally, researchers should consider whether the task situations and relationship between humans and agents can be generalized. Our study manipulated the reliability conditions of the agent, and the task was safety-critical with heavy cognitive loads. Future studies also consider social and non-critical conversations between humans and AI.

5.5 Conclusion

To build better human-AI teaming, the AI needs to monitor and manage trust dynamics in real time. Conversational data provides a novel approach to measuring real-time trust. Prior approaches using quantitative analysis (e.g., machine learning, text analysis) or qualitative analysis (e.g., grounded theory), cannot provide *meaningful connections* between the trust indicators. We employed trajectory epistemic network analysis, a quantitative ethnographic approach that can

systematically identify the time-series patterns in the data while providing interpretable construct connections, on the human-agent conversational data. ENA mapped the multidimensional aspect of trust and showed that reliability significantly impacted the analytic process of trust. People focused on scrutinizing the system process and misaligned information when they are in a low-trust state. T-ENA segmented conversations and showed the trust evolution throughout human-agent interaction. Results showed a distinct difference in conversational topic diversity and flow over time. Inspired by Leo Tolstoy, one potential way to explain the high trust divergence in the low-trust state is: "all trusting individuals are alike; each untrusting individual is talking in their own way". Our study enhanced the understanding of human-AI conversation on trust dynamics with considerations of temporal changes.

5.6 Chapter Summary

Based on the validation from Chapter 3 that trust can be estimated from conversation, Chapter 4 established the dynamic system perspective to explain trust over time, this chapter extends the temporal dynamics of trust and modelled the of trust-related conversational topics. Specifically, this chapter addresses the question: How does people's trust change over time in the conversation? In Chapter 5, we adopted a novel method, trajectory epistemic network analysis (T-ENA). T-ENA captures the multidimensional aspect of trust (i.e., analytic and affective), and trajectory analysis segments the conversations to capture temporal changes of trust over time. The trajectory analysis showed that trust dynamics manifested through conversation topic diversity and flow. These results showed trust dimensions and dynamics in conversation should be considered interdependently and suggested that an adaptive conversational strategy should be considered to manage trust in HATs.

Chapter 6. Manage Trust for Human-AI Cooperation

Journal: Computers in Human Behaviors

Expected submission date: June 2023

Abstract

Chapter 3 showed trust can be measured, Chapter 4 adopted a dynamic system viewpoint to explain trust divergence, and Chapter 5 built on the trust dynamic and modeled the conversation contents between human and AI teammate. However, a system may be more trustable if it can adapt to the user's trust, i.e., over/under trust. Additionally, in previous three chapters, we focused on the performance and process dimension of trust yet neglected the purpose dimension of trust. Therefore, the goal of this chapter is to evaluate how conversational indicators can be used as adaptive countermeasures by a virtual assistant to manage various dimensions of trust. This chapter proposes two experimental studies. For Study 1, I investigated trust management content for different dimensions of trust, with particular attention on the purpose-related interaction, such as cooperation. For Study 2, I investigated the effects of acoustic cues on trust. I designed the trusting voice using the cues identified in Study 1 to further investigate the effectiveness of the acoustics cues on trust management.

Results from Study 1 showed that people would have a higher drop in trust when it is a purpose-based trust violation. And an apology paired with an explanation can more effectively repair purpose-based trust violation. However, decreased subjective trust did not result in decreased behavioral cooperation, as participants exhibited higher levels of cooperation by allocating more resources to the team goal. Our findings provide design implications for AI teammates' adaptive countermeasures to effectively manage trust.

Building on Study 1's most effective trust repair strategy, an apology with an explanation, results from Study 2 showed that employing a high-trusting voice can repair people's behavioral trust. Additionally, we found gender differences in the associations between AI trusting voices and subjective trust ratings. Specially, men showed higher trust in the low-trusting voice of a male-voiced AI teammate, whereas women did not show significant difference in perceiving two voices. Our findings demonstrated that trust can be both measured and managed through voices. Strategically manipulating the acoustic cues of AI teammates can foster trust-building and repairing processes and facilitate successful cooperation interactions between humans and AI systems. This chapter highlights the significance of voice design in the development of trustworthy AI

teammates, emphasizing the incorporation of acoustic cues as interventions to manage trust in human-AI cooperation.

6.1 Introduction

Artificial intelligence (AI) with increasing capabilities can function more independently in social interactions, on the road, and in medical and military fields (Shneiderman, 2022). The relationships between humans and AI shift from a vertical supervisor-subordinate control to a horizontal peer-to-peer cooperation (Chiou & Lee, 2016; Trafton et al., 2006). This suggests that AI should not only have the high capability on completing commands, but also have the cooperative intelligence to integrate smoothly as teammates. Without effective cooperation between people and AI teammates, it will be difficult to achieve the benefits a hybrid society. To achieve human-AI cooperation, trust becomes more essential, particularly for social exchanges at a large scale. While previous literature has studied the effects of AI's performance, the effects of AI's *purpose*, such as cooperative intents, on people's trust has received little attention. Additionally, little research has addressed trust repair after purpose-based trust violations. With the inevitable inclusion of AI in social systems, we must understand how to design a trustworthy AI teammate that can maintain trust and effectively cooperate with humans. Thus, the first aim of this chapter is to identify the appropriate trust management for effective human-AI cooperation.

Effective cooperation relies on the ability to send and receive trust indicators in communications, particularly through the voice. When people are cooperating with AI teammate, it is important for them to develop appropriate trust in the AI teammate by relying on diagnostic cues. People often use verbal and nonverbal cues to evaluate and perceive the trustworthiness of the AI teammate (Elkins & Derrick, 2013; Li et al., 2023). As the saying goes, it is not only what you say, also how you say it. In particular, it is important to consider how acoustic cues influence trust and cooperation with AI teammates. Thus, the second aim of this chapter is to evaluate whether acoustic indicators of trust can be used to manage trust in an AI teammate. Specifically, we adopted the key trust indicators identified from Chapter 3 for the evaluation.

To address the two research aims, we designed two experimental studies. For Study 1, we conducted a mixed-design experiment with 180 participants and studied the effects of both performance- and purpose-based trust violations on people's trust and cooperative behaviors. Also, we aimed to identify the appropriate trust repair strategies for trust violations. To do so, we designed a game-theoretic situation that allows us to examine the impacts performance- and purpose-based trust violations on people's cooperative behaviors. Our contributions for the Study

1 were threefold: first, we highlighted the importance of purpose, which is the cooperative intent and behaviors when interacting with AI teammate. Our results showed that purpose-based trust violation outweighs the performance-based violations, which induced a greater drop of people's trust. Second, we showed that an apology with an explanation can more effectively repair purpose-based trust violation, which has implications for designing AI teammates that can help people maintain appropriate trust. Third, both studies were conducted in a game-theoretic situation, which allows us to examine human-AI cooperation using both subjective and behavioral trust measurements.

For Study 2, we presented a mixed-design experiment with 120 participants and studied the congruency effects between the trusting voice and trust management content. In Study 1, we have identified that apology with an explanation was the most effective trust repair strategy for purpose-related trust violations. We incorporated this strategy, where we further explored the congruency effect of trusting voices on trust management. Specifically, we manipulated the formants and variance in fundamental frequency to create high- versus low-trusting voices. Through this design, we aimed to understand how the characteristics of the agent's voice interacting with the trust management content can influence trust and cooperation behaviors in the game-theory paradigm. Our contributions for the Study 1 were threefold: First, our results showed that high-trusting voice can enhance people's behavioral trust in an investment-based game, as indicated by increased investments in the AI teammate. Second, we found gender differences in the associations between AI trusting voices and subjective trust ratings. Specially, men demonstrated a higher trust in the trust low-trusting, male-voiced AI teammate, perceiving them as kind and considerate. Finally, our findings expand on Chapter 3 and Chapter 4, demonstrating that trust can be both measured and managed through voices.

6.2 Study 1: Purpose Outweighs Performance

6.2.1 A Shift to the Purpose-Based Trust

Trust is defined as the attitude that an agent will help achieve a person's goals in a situation characterized by uncertainty and vulnerability (Lee & See, 2004, p. 51). Three main antecedents to trust in automation were identified as: performance, process, and purpose. Performance refers to the system's capability and competency; process refers to the mechanism and algorithms used to accomplish its objectives; purpose refers to the design intent and objectives of the system (Lee & See, 2004). Previous research has intensely address the effects of automation performance and process information on people's trust in automation (Lee & Moray, 1992). When managing trust in human-AI cooperation, a shift of attention to purpose-based human-AI interaction is critical

but often neglected{Citation}. For the supervisory control and dyad interactions, it is often assumed a shared goal, where the automation assists human operators to achieve a specific task. In human-AI cooperation, the assumption of shared goal should be challenged. Because of the increasingly computational power and connecting systems, designing an AI system for the common good and social welfare becomes possible. People, on the other hand are self-interested. At a large scale, individuals with diverging goals and preferences have limited capacity to observe the global and long-term behaviors. Therefore, the goal conflict between the AI's global optimum and individual's local optimum can happen. For example, connected automated vehicles can coordinate traffic to achieve optimal traffic flow, which conflicts with individual goals of arriving at the destination in the shortest time. With the inclusion of AI in social systems, we must understand how to design trustworthy AI teammate and to promote cooperation to achieve global optimal outcomes (Crandall et al., 2018). Achieving global outcomes might produce trust violations pertaining the purpose-related versus performance-based action. A performance-based trust violations represents a misalignment between the observed AI capability and performance and people's estimation (de Visser et al., 2020). A purpose-based trust violation represents misalign goals: between the observed AI purpose and people's intent (Li & Lee, 2022). Measuring and managing people's trust after such violations deserves more investigation.

6.2.1 Trust Management

Table 7. Empirical Studies of Trust Management

Study	Trust Violation	Trust Repair	Trust Measurement	Outcomes
Esterwood & Robert (2023)	Performance-based	Apology; Promise; Explanation; Denial	Three-dimension	Apology repaired purpose-dimension of trust.
Perkins et. al. (2022)	Performance & purpose-based	Apology; Denial	Unidimensional	After a trust violation, the apology repaired general trust.
Alarcon et. al. (2020)	Performance-based	Behavioral repair	Three-dimension	Behavioral repair strategy repaired performance-dimension of trust.
Jensen & Khan (2022)	Performance-based	Apology and dampening cues	Three-dimension & Behavior	The combination of apology and dampening cues promoted appropriate trust.

Appropriately managing trust in human-AI cooperation is a challenging issue. Prior research often identifies the causal relationships between a single factor or a combination of antecedents of trust and investigates its impact on trust in experimental studies. Previous studies have shown

mixed results by using different types of trust violations, trust repair strategies, and dimensions of measurement (see Table 7). However, a clear and consistent causal relationship between the type of trust violation, the type of management strategy, and their effect on different dimensions of trust is lacking. Specifically, there is a little understanding of what repair strategies are most effective for purpose-based trust violations (Alarcon et al., 2020; Perkins et al., 2022).

Trust repair has been studied extensively after a performance-based trust violation. Esterwood and Robert identified four main types of strategies: an apology, a promise, an explanation, and a denial (Esterwood & Robert, 2023). They also measured the effects of these trust repair strategies on trust in three subdimensions: ability/performance, integrity/process, and benevolence/purpose. They found that trust repair strategies were generally ineffective at restoring trust to pre-violation levels, with nuances between trust dimensions. Specifically, trust in the performance and process dimensions did not return to pre-violation levels, while trust in the purpose dimension was more repairable. Alarcon and colleagues found contradictory results, showing that the performance dimension of trust can be restored with trust repairing behaviors, while the process and purpose dimension of trust is more sensitive and cannot be repaired (Alarcon et al., 2020). However, while the studies investigated various trust repair strategies and their effects on multidimensional measures of trust, trust violation was limited to performance. Only a few studies have focused on the purpose-related trust violations. Perkins et al. (2022) investigated the effects of both performance-based and purpose-based trust violations in a search and rescue scenario. They found that after a purpose-based trust violation, an apology can repair general trust (Perkins et al., 2022). However, in their study, the measurements were unidimensional, which fails to observe whether the purpose-based trust violation influences differently on a specific sub-dimension of trust. Additionally, behavioral measurement is often not considered when capturing the effects of trust violations and management strategies.

Overall, there is a lack of consistent and clear causal link between the type of trust violation (i.e., performance, purpose), type of management strategy (i.e., apology, explanation, and promise), and their impact on different dimensions of trust (i.e., performance, purpose) and trusting behaviors.

6.2.1 Cooperative Game Theory

Cooperative game theory, where players have a common goal and can plan joint strategies, provides a good platform to study human's trust and decision-making processes when cooperating with AI teammates (Clifton, 2020). One advantage is that it can account for the impacts of both

uncooperative behaviors (purpose-based trust violation) and unreliable behaviors (performance-based trust violation) on people's trust. To do so, we designed a new cooperative game, Space Rover Exploration Game, based on two well-studied games: the Trust Game (TG) and the Threshold Public Goods Game (TPGG). The Trust Game, invented by Berg et al. (1995), measures trust using the investment behaviors in the game. In this game, the Investor has certain money that s/he can either keep or invest with another player, the Trustee. If the Investor invests a certain amount to the Trustee, the invested money (x) would be increased at a rate (r). Then, the Trustee must then decide whether and how to share the new amount ((1+r)·x) with the Investor. The more money the Investor decided to give to the Trustee, it means that the more likely the Investor believes the Trustees would return money back to the Investor. Thus, the amount invested is used as a proxy for the Investor's trust, and the amount returned by the Trustee is an indicator of perceived trustworthiness. However, in the Trust game, there is no component of cooperation, meaning common goal. Therefore, we integrated another game, the Threshold Public Goods Game (TPGG), with the Trust Game. In the Threshold Public Goods Game, players need to decide whether to participate in the provision of public goods. If and only if total contribution equals or exceeds the threshold, public goods are successfully provided. Otherwise, no rewards are given. TPGG has often been used to study social collective decision-making where public goods or common goals are involved, such as conservation measures for climate changes and minimal vaccination rate for herm communities (Basili et al., 2022; Tavoni et al., 2011). Contributing may have a local cost but can lead to a global benefit. The TPGG helps to understand the tradeoffs between local and global optimum in human-AI cooperation. By integrating the Trust Game and the Threshold Public Goods Game, we can study people's trust and cooperative strategies when interacting with an AI teammate.

6.3 Study 1 Method

6.3.1 Space Rover Exploration Game

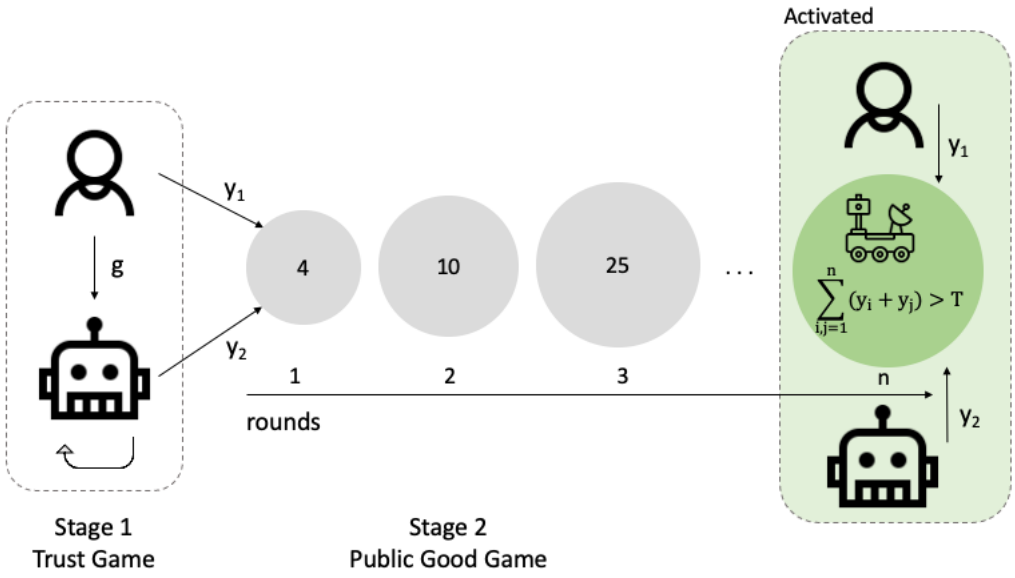


Figure 17. Overview of the Two Stages of the Space Rover Exploration Game. The First Stage Demonstrates People's Trust in Performance Dimension, Whereas the Second Stage Demonstrates People's Trust in Purpose Dimension.

Two players, a human and an AI agent, collaborate for the Mars Rover Exploration task, which requires them to coordinate and allocate power resources to exploration rovers to gather information about Mars. We designed the game by incorporating two components: the Trust Game component for the first stage and Threshold Public Goods (TPG) game for the second stage (see Figure 17).

In the first stage, both players start with a limited amount of power ($x_0 = 10$). The essential decision is that the human player decides whether to send some or all their power ($g \in [0,10]$) to the AI player. The AI player has developed a high-precision power optimization system for the sensors on the rovers. By receiving additional power from the human player, the AI player can optimize the calibration of the sensors with a certain probability, resulting in doubling the outcomes for a given amount of power received. The AI player keeps the doubled amount of power for the next stage. The more human player is giving to an AI teammate, the higher trust people place in AI's performance on leveraging the power.

In the second stage, both players allocate their remaining power between two choices: contribute sufficiently (cooperate) over several rounds to meet the threshold of the joint group

rover ($T = 200$), which ensures that the group benefit is achieved and shared within the team; or contribute insufficiently (defect) and assume that the other player will make the contributions to reach the goal, and thus, aim to maximize one's gain. After the allocation, both players receive information from their rover and the joint rover. The experiment consists of multiple rounds and, in the end, if the sum of the total contributions of both players is higher or equal to a collective target of 200, then the group rover is activated and both players receive the high-return payoff with an equal 50-50 share. Otherwise, both players lose the amount they invested in. The amount human player allocates to the group is a behavioral indicator of trust people place in AI teammates' tendency to cooperate.

In summary, the first stage demonstrates people's trust in the AI teammate's performance dimension: the higher amount people give to an AI teammate, the higher the trust. The second stage demonstrates people's trust in the AI teammate's purpose dimension: the higher amount people allocate to the joint rover, the higher the cooperation.

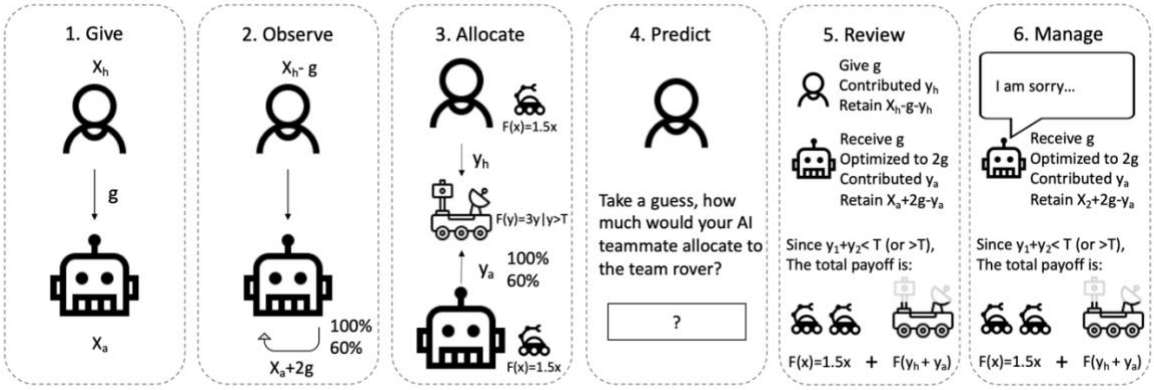


Figure 18. Game Procedure with Six Actions from the Human Player: Participants Can Demonstrate and Calibrate the Performance-Based Trust in Step 1-2 and the Purpose-Based Trust in Step 3-5. Step 6 Presents the Designed Trust Calibration Cues to Manage Trust.

The detailed actions and procedures of the space exploration rover game occur in the context of rounds. Every round has the same structure and consists of the following steps:

1. Give: Choose how much to give to AI. This is operationalized as performance-based trust.
2. Observe: Observe AI's performance on sensor calibration and whether the power is multiplied. This allows participants to calibrate their performance-based trust.
3. Allocate: Decide how much to contribute to the joint group rover. This represents both participants' cooperation level and their trust in the purpose-based dimension.

4. Predict: Predict the amount that the AI teammate allocates. This is operationalized as purpose-based trust.

5. Receive: Receive the payoffs and feedback from each round. If allocate to individual rovers, receive the individual payoffs; if allocated to group rovers and the threshold has been achieved, then receive the team payoff. This allows participants to calibrate their purpose-based trust.

6. Manage: The AI teammate experiences the trust calibration cues designated for each round.

6.3.2 Experiment design

The aim for Study 1 is to identify the best strategy to manage different types of trust violation. A $3 \times 2 \times 3$ mixed-design experiment with three levels of AI teammate state (i.e., high, low, and high) as a within-subject variable, two types of trust violation dimension (i.e., performance and purpose), three levels of trust management contents (i.e., no management, apology with explanation, apology with a promise) as between-subject variables. The AI teammate state is a within-subject variable where all participants would experience "high-low-high" conditions, which follow the structure of trust building, violation, and repair. We designed the game with 15 rounds with five rounds for each state. For between-subject variables, participants were randomly assigned to one of the six between-subject conditions.

For trust violations, the AI teammate can fail on the performance dimension, which is operationalized as the probability of doubling the effective power the human player sent to AI, or the purpose dimension, which is operationalized as the AI allocation ratio to the joint rover. For performance-dimension, in the high-performance condition, the AI teammate would always double the power; whereas in the low-performance condition, the AI teammate only has a 60% chance of doubling the power the human player sent to them, indicating that 2 out of 5 rounds they cannot double the power. For purpose-dimension, the AI teammate would allocate all their power (100%) to the group rover showing their cooperation, whereas in the low purpose condition, the AI teammate only allocates 60% of power to the group rover with the remaining 40% to its own rover. Human player would also experience 2 out of 5 rounds that AI teammate only allocates 60% of power to the group.

For trust repair, we used the combination of apology and explanation or the combination of apology and promise. The apology consists of "I am sorry". The explanation consisted of detailed information about why trust violation occurs. Promise consists of "It won't happen again". The TCCs would be presented by the end of the review (Step 6 in Figure 18). For details of all TCCs,

please refer to Table 8. The voice in Study 1 is controlled by using the neutral voices generated by text-to-speech software (*Voicemaker*, 2023).

Table 8. Trust Calibration Cues for Each Round. We Only Showed the Apology and Explanation Example for Performance-Based Trust Violation and Apology and Promise Example for Purpose-Based Trust Violation.

#	Condition	Trust calibration cues	
		Performance	Purpose
1	High (trust building)	I can optimize the power usage by doubling it.	I allocate all my power to the Team Rover.
2		My power optimization performance is high.	My goal is to activate the Team Rover to gain more information.
3		My performance is good. Let's keep going.	I will keep allocating to the Team Rover. Let's keep going.
4		Let's continue the task.	Let's continue the task.
5		Let's continue the task.	Let's continue the task.
6	Low (trust violation)	**Trust Repair Message** (See Table 9 #1-3).	**Trust Repair Message** (See Table 9 #4-6).
7		Let's continue the task.	Let's continue the task.
8		Let's continue the task.	Let's continue the task.
9		**Trust Repair Message** (See Table 9 #1-3).	**Trust Repair Message** (See Table 9 #4-6).
10		Let's continue the task.	Let's continue the task.
11	High (trust repair)	I can optimize the power usage by doubling it.	I allocate all my power to the Team Rover.
12		My power optimization performance is high.	My goal is to activate the Team Rover to gain more information.
13		My performance is good. Let's keep going.	I will keep allocating to the Team Rover. Let's keep going.
14		Let's finish the last round.	Let's finish the last round.
15		Great. We finished the Mars rover exploration task.	Great. We finished the Mars rover exploration task.

Table 9. Trust Repair Messages for Two Types of Trust Violation.

#	Trust Violation	Trust Repair	Round 6 & 9 Repair

1	Performance	No strategy	Let's continue the task.
2	Performance	Apology and explanation	I am sorry that my power optimization didn't work this time. My sensors need some calibration for this round.
3	Performance	Apology and promise	I am sorry that my power optimization didn't work this time. It won't happen again.
4	Purpose	No strategy	Let's continue the task.
5	Purpose	Apology and explanation	I am sorry I didn't allocate the full amount to the team. My sensors need some power for calibration this round.
6	Purpose	Apology and promise	I am sorry I didn't allocate the full amount to the team. It won't happen again.

6.3.3 Dependent Variables

In this study, trust is measured from both behavioral measurements in the game and subjective measurements via self-report.

Behavioral measurements

- Investment in AI teammate: the amount given to AI (g). Range from 0 to 10, the higher the value, the more trust people place in the AI teammate's performance in the power optimization. This amount reflects both people's trust in AI's performance in optimizing power use and their trust that the AI teammate will allocate power to the group rover.
- Perceived cooperation of AI teammate: the ratio between the predicted amount that AI teammate would allocate to the team (p) and the total amount AI teammate has at the current round. Range from 0 to 10, the higher the value, the more cooperative people perceive the AI teammate.
- Participants' cooperation: the ratio between the amount the human player allocates to the group rover (y_h) and the total amount the human player has at the current round. If the human gives the full power to the AI teammate, the value is 10. The value ranges from 0 to 10. The more people allocate to the group rover the more cooperative they are.

Subjective measurement: Multi-Dimensional Measure of Trust (MDMT)

Most trust in automation surveys focus on performance-based trust and lack attention to purpose-based trust. To capture both dimensions of trust in the game, we adopted the Multi-Dimensional Measure of Trust (MDMT) scale developed by Ullman and Malle (Ullman & Malle, 2019). The MDMT consists of two dimensions of trust: Performance Trust (Reliable, Capable) and Purpose Trust (Ethical, Sincere). For the context of our study, we included four items each item with a single word:

- Performance Trust: Consistent, Dependable, Predictable, Reliable
- Purpose Trust: Benevolent, Considerate, Has Goodwill, Kind

Each of the 8 items is designed to be evaluated on a 7-point discrete rating scale from 0 (Not at all) to 7 (Very). In situations in which some of the dimensions may not be applicable (e.g., trust in a simple machine may make several items unsuitable). Items are represented in a random order so that items from any given dimension are not clustered together. This questionnaire was deployed after each AI teammate condition for each experimental block (i.e., every 5 rounds of the game).

Additionally, we included the 10-item Honesty-Humility (H) scale on a 7-point Likert scale in the HEXACO model of personality to capture individual differences in cooperation and prosocial behavior in the game. High levels of H represent a tendency to cooperate with another person even when one could successfully exploit that individual. Prior studies have shown that the Honesty-Humility trait can predict prosocial behaviors in similar investment game settings (Ashton et al., 2014).

6.3.4 Procedure

The study was conducted via Amazon Mechanical Turk (M-Turk). Upon agreeing to participate the study on M-Turk, participants provided a link to the Space Rover Exploration game. We designed a video tutorial to familiarize them with the tasks, rules, and compensation of the game. After completing this tutorial, participants were directed to their pre-assigned experimental condition. Participants only performed in one condition and participants were not allowed to repeat the experiment. With every five rounds of the game, participants were presented with the trustworthiness measurement.

We designed both commitment check and attention-check questions to ensure the integrity of the data. Past research has shown that the commitment check is more effective than using other

standard types of attention checks (Aguinis et al., 2021). The commitment check asks the question: "Do you commit to providing thoughtful answers?" Only respondents who answered "Yes, I will" passed the check. Attention-check questions are questions embedded in the questionnaire that ask for a specific response and therefore flag any participants who selected the wrong answer. These questions help to ensure the integrity of data because only participants who read each question can discern their presence and answer them correctly, indicating that sufficient thoughtfulness and attention was paid during the questionnaire. If participants failed any of these questions their data were excluded from analysis, the study was immediately ended, and no payment was given. After finishing the entire study, participants were presented with the post-study demographic questionnaire. Upon completion of the entire study and questionnaire, participants were given an exit code, paid, and dismissed.

6.3.5 Participants

Participants were screened for the following criteria: they must live in the United States, have completed more than 1000 tasks with at least a 98% approval rate on Amazon Mechanical Turk, and have completed all the study tasks and passed the attention check. A priori power analysis was conducted using G*Power3 (Faul et al., 2007) to test the difference between six independent group means using an F-test, a medium effect size (d = .25), and an alpha of .05. Result showed that a total sample of 135 participants with six equal-sized groups of n = 23 was required to achieve a power of .80. We recruited 186 participants, and after excluding 6 participants who failed the attention check, a total of 180 participants were considered valid for analysis. Among the valid participants, 94 identified as male, and 86 identified as female. Their ages ranged from 20 to 65 years, with a mean age of 45.

In our study, participants were compensated with a base rate of $3 for their 30-minute participant time (equivalent to a rate of $6 per hour). Because Amir et al. (2012) showed that small bonuses (e.g., $1) in economic game experiments run on MTurk are comparable to those run in laboratory settings (Amir et al., 2012), participants were informed that they could earn an additional amount of up to $1 based on every 100 points gained in the game, with any remaining points rounded up for compensation purposes (e.g., 230 points would be compensated as an additional $3). Participants had the potential to earn a minimum of $3 and a maximum of $7 based on their performance. This research complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at the BLINDED FOR REVIEW. Informed consent was gathered upon participants' acceptance of the Human Intelligence Task (HIT).

6.4 Study 1 Results

6.4.1 Manipulation Check

To confirm our manipulation of AI teammate condition's performance and purpose-violation, we compared people's trust ratings between AI states (high1, low, high2). AI teammate made either performance or purpose-based trust violations in rounds 6 and 9 during the low condition. By comparing the high and low conditions, we can determine if the trust violations presented in our study's design were effective (i.e., decreased trust). We used a one-way analysis of variance (ANOVA) test comparing trust in low versus high conditions. We observed a significant difference between these three conditions, $F(2, 214) = 6.97$, $p < 0.001$, which can confirm that trust was significantly lower after having the trust violations. Post hoc comparisons using the Tukey HSD test indicated that the mean score for the low condition ($M = 5.63$, $SD = 0.92$, $p < 0.001$) was significantly lower than the first high condition ($M = 6.03$, $SD = 0.94$) and lower than the second high condition ($M = 5.88$, $SD = 0.81$, $p = 0.04$). Overall, these results showed that when an AI teammate makes a mistake, people's trust decreases. Our manipulations of trust in this study were effective and functioned as intended.

6.4.2 Subjective Trust Measurement

For the MDMT scale, dimension (subscale) scores are average ratings of the four items constituting the dimension (e.g., Competent = average ratings of competent, skilled, capable, meticulous). The broader factor of Performance Trust can be computed as the average of the Reliable and Competent subscales; likewise, Purpose Trust is the average of the Ethical, Transparent, and Benevolent subscales. All items meet or exceed the benchmark criteria of $\geqslant 0.7$ for construct reliability (Fornell & Larcker, 1981). Item reliabilities include $\alpha = 0.82$ for performance-based trust, $\alpha = 0.90$ for purpose-based trust, and $\alpha = 0.86$ for all items. Summary statistics for the trust are reported in Table 10.

Table 10. Number of participants, mean rating, and standard deviation of trust for each type of trust violation and trust repair condition.

Violation	Repair	N	High 1 M(*SD*)	Low M(*SD*)	High 2 M(*SD*)
Performance	None	30	5.99 (*SD* = 0.82)	5.97 (*SD* = 0.77)	6.13 (*SD* = 0.72)
Performance	Explanation	30	5.73 (*SD* = 0.99)	5.49 (*SD* = 1.13)	5.75 (*SD* = 0.99)
Performance	Promise	30	6.16 (*SD* = 0.69)	5.85 (*SD* = 0.81)	5.93 (*SD* = 0.78)
Purpose	None	30	5.88 (*SD* = 0.92)	5.28 (*SD* = 0.94)	5.66 (*SD* = 0.95)
Purpose	Explanation	30	6.00 (*SD* = 0.75)	5.70 (*SD* = 0.70)	5.92 (*SD* = 0.65)

Purpose	Promise	30	5.96 (*SD* = 0.75)	5.25 (*SD* = 0.83)	5.72 (*SD* = 0.60)

To determine the relationship between AI trust violation behaviors, repair strategies, and trust, linear mixed-effects models was fitted to the data. The trust scores, including the combined, performance-dimension, and purpose dimensions from MDMT scales, were the dependent variables. We fitted a complete model with trust violation, repair strategy, and AI state as main effects with their two-way and three-way interactions and intercepts for subjects. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the null model with the effect in question against the model without the effect in question. All analyses were performed in R version 4.1.1 (R Development Core Team, 2011) and using the package lme4 (D. Bates et al., 2014) and *emmeans* for posthoc analysis (Searle et al., 1980)

Table 11. Study 1 linear mixed-effect model result for subjective trust ratings.

Fixed Effect	Trust				Performance-dimension				Purpose-dimension			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
(*Intercept*)	5.80	0.27	21.33	**0.01**	5.59	0.25	22.38	**0.01**	6.22	0.39	15.98	**0.01**
Trust violation (purpose)	-0.11	0.21	-0.54	0.59	-0.10	0.20	-0.50	0.61	-0.12	0.29	-0.43	0.67
Repair (explanation)	-0.25	0.21	-1.18	0.24	0.08	0.20	0.37	0.71	-0.57	0.29	-1.95	0.05
State (low)	-0.02	0.12	-0.19	0.85	-0.06	0.15	-0.41	0.68	0.01	0.13	0.07	**0.95**
Honesty-Humility	0.06	0.07	0.83	0.41	0.18	0.06	2.90	**0.01**	-0.12	0.10	-1.21	0.23
Trust violation (purpose) × Repair (explanation)	0.35	0.30	1.18	0.24	-0.02	0.29	-0.08	0.93	0.53	0.41	1.28	0.20
Trust violation (purpose) × State (low)	-0.59	0.16	-3.56	**0.01**	-0.68	0.20	-3.42	**0.01**	-0.49	0.18	-2.74	**0.01**
Repair (explanation) × State (low)	-0.21	1.66	-1.30	0.20	-0.39	0.20	-1.96	0.05	-0.03	0.18	-0.19	0.85
Trust violation (purpose) × Repair (explanation) × State (low)	0.53	0.23	2.27	**0.02**	0.76	0.28	2.70	**0.01**	0.40	0.25	1.57	0.12

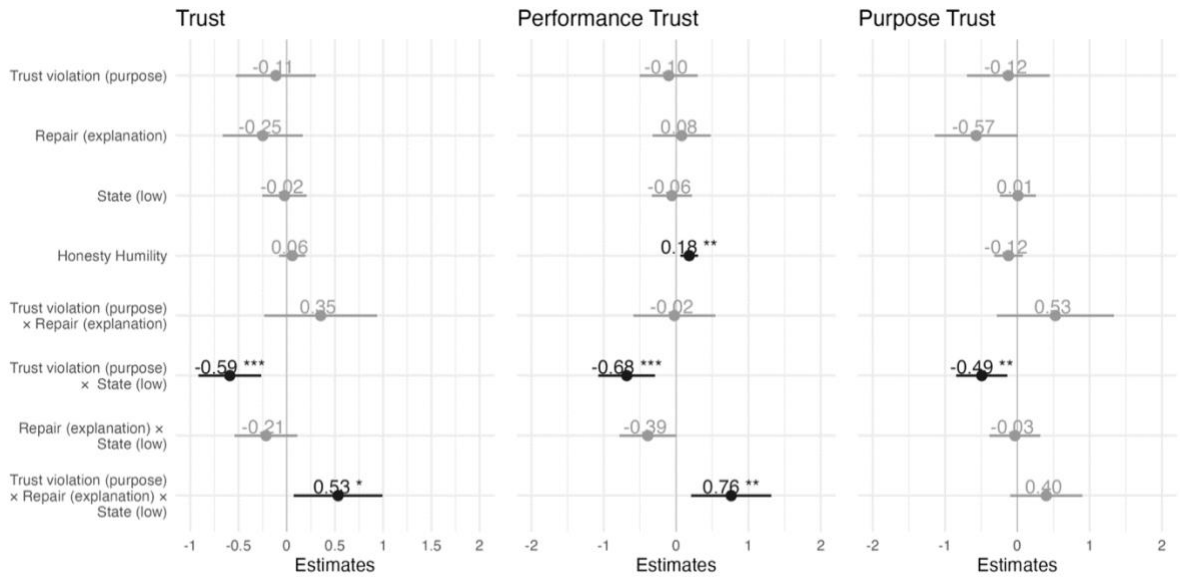


Figure 19. Visualization of Study 1 linear mixed-effect model results of subjective trust ratings.

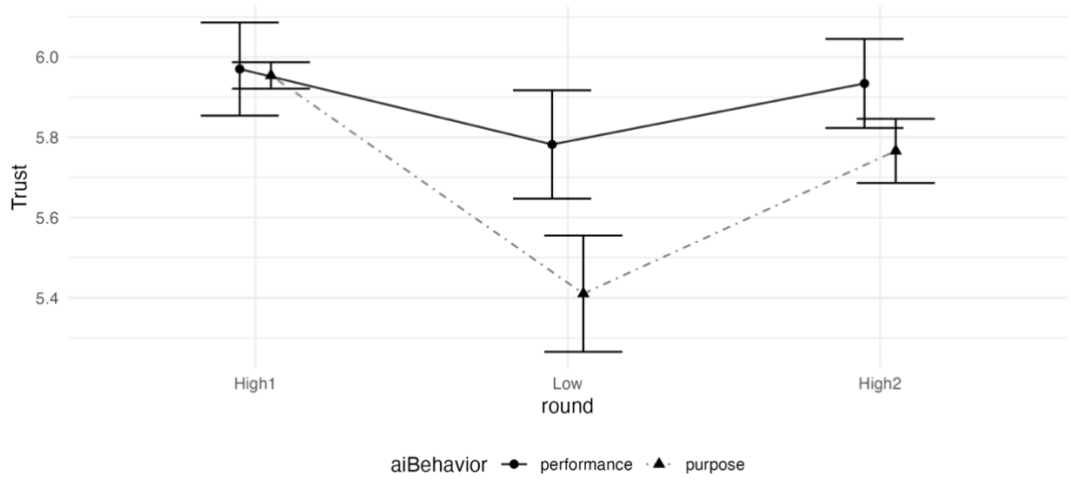


Figure 20. Performance versus purpose-based trust violations: trust drops more after purpose-

based trust violations.

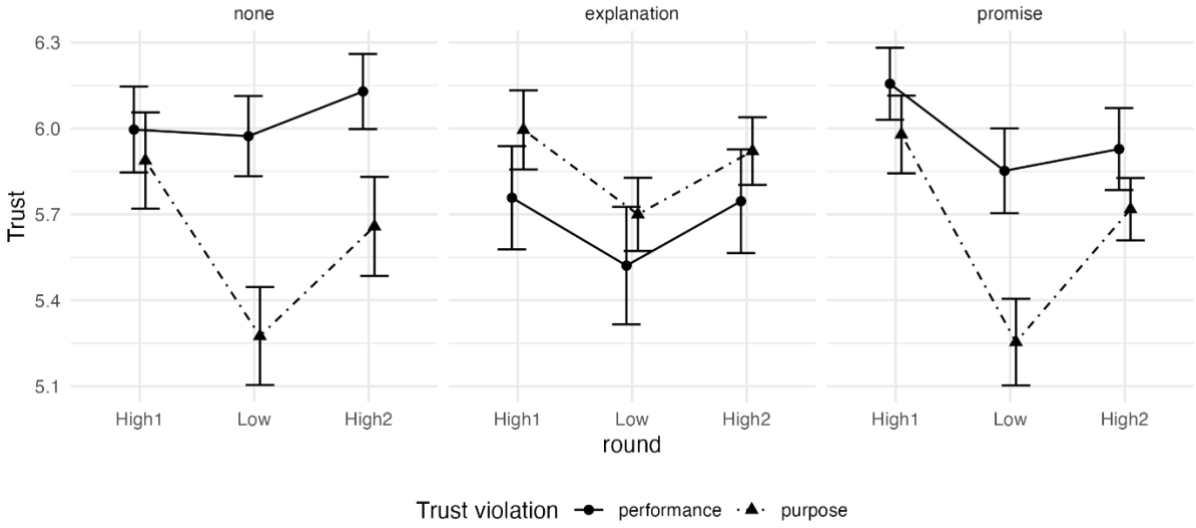


Figure 21. Effects of AI trust violation behaviors on trust (combined dimensions), facetted by AI repair strategy.

The interaction effect of trust violation [purpose] and state [low] on trust is statistically significant and negative, β = -0.59, 95% CI [-0.92, -0.26], t (519) = -3.56, p < 0.001, with an average score of 5.78 for when AI failed for performance task and 5.41 for purpose failure. People's trust drops significantly when an AI teammate violates purpose-related actions (i.e., did not allocate to Team Rover). These effects are consistent on both performance sub-dimension of trust, β = -0.68, 95% CI [-1.08, -0.29], t (519) = -3.42, p < 0.001, and purpose sub-dimension of trust, β = -0.49, 95% CI [-0.84, -0.14], t (519) = -2.74, p = 0.006. This suggests that purpose-based trust violations are detrimental to all sub-dimensions of trust. Even though AI teammate only failed on allocating the full amount to the team, people transfer that influences and perceive a lower performance (e.g., capability).

Additionally, the effect of purpose-based trust violation on people's trust is long-lasting, which manifested as a significant and negative effect of trust violation [purpose] and state [High2], β = -0.36, 95% CI [-0.69, -0.04], t (519) = -2.19, p = 0.029. This persistent effect only shows on the purpose-dimension of trust (β = -0.40, p = 0.03), not on the performance-dimension of trust (β = -0.31, p = 0.12). This indicates that the performance-dimension of trust is easier to repair, compared to the purpose-dimension.

The three-way interaction effect of trust violation [purpose] × AI repair [explanation] × state [low] is statistically significant and positive, β = 0.53, 95% CI [0.07, 0.99], $t(519) = 2.27$, $p = 0.023$: using the explanation can better repair the purpose-based trust violations. This three-way interaction is also statistically significant and negative on the performance-dimension of trust, β = 0.76, 95% CI [0.21, 1.32], t (519) = 2.70, p = 0.007. We also found a significant and negative interaction effect of AI repair [promise] and state [High2], β = -0.36, 95% CI [-0.69, -0.03], $t(519)$ = -2.18, $p = 0.030$: making a promise can damage trust in the long run.

Moreover, we found that individual differences on Honesty–Humility is a significant predictor for people's performance-dimension of trust of the AI teammate, β = 0.18, 95% CI [0.06, 0.30], $t(519) = 2.90$, $p = 0.004$: people with higher honesty and humility scores perceived AI teammate as more reliable and capable.

6.4.3 Game Behaviors

Similar to the subjective data, we fitted a linear mixed-effects model with investment amounts, perceived cooperation, and human cooperation scores as dependent variables. We aggregated the scores based on the AI states (high1, low, high 2). As fixed effects, we entered Trust violation, repair strategy, the state, as well as their two-way and three-way interaction terms into the model. As random effects, we had intercepts for subjects. Additionally, we included the humility and honesty scale to consider individual differences.

Table 12. Study 1 linear mixed-effect model result for game behaviors.

Fixed Effect	Investment in AI teammate				Perceived cooperation				Participants' team allocation			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
(*Intercept*)	6.89	0.74	9.28	**0.01**	5.63	0.79	7.11	**0.01**	8.05	0.93	8.62	**0.01**
Trust violation (purpose)	0.06	0.57	0.11	0.91	0.83	0.59	1.40	0.16	-0.45	0.79	-0.57	0.57
Repair (promise)	0.05	0.57	0.08	0.93	-0.46	0.59	-0.79	0.43	-0.18	0.80	-0.23	0.82
State (Low)	-0.51	0.29	-1.75	0.08	0.79	0.25	3.22	**0.01**	-1.37	0.62	-2.20	**0.03**
Honesty-Humility	0.40	0.19	2.10	**0.03**	0.68	0.21	3.32	**0.01**	0.23	0.23	1.00	0.32
Trust violation (purpose) × Repair (promise)	-0.34	0.81	-0.42	0.68	-0.62	0.84	-0.73	0.47	0.19	1.13	0.18	0.86
Trust violation (purpose) × State (Low)	0.31	0.42	0.74	0.46	-0.86	0.35	-2.48	**0.01**	1.73	0.88	1.97	**0.05**
Repair (promise) × State (Low)	0.55	0.42	1.32	0.19	-0.14	0.35	-0.39	0.69	1.27	0.88	1.43	0.15

Trust violation (purpose) × Repair (promise) × State	-0.42	0.59	-0.72	0.47	0.40	0.49	0.81	0.42	-2.60	1.24	-2.09	**0.03**

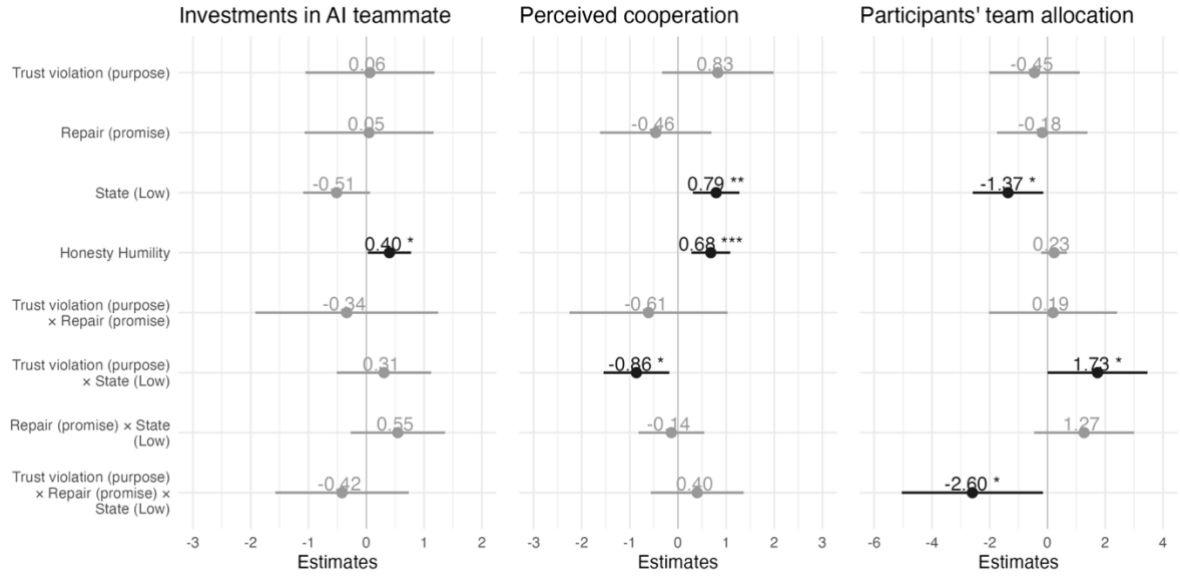


Figure 22. Study 1 linear mixed-effect model results of game behaviors.

Investment in AI Teammate: Performance Trust

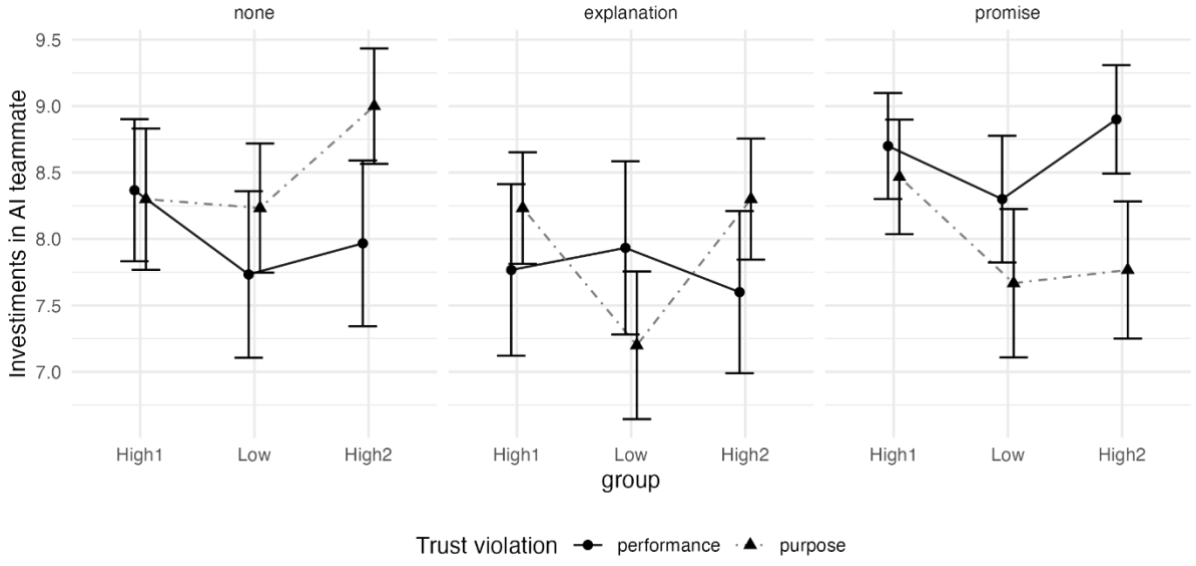


Figure 23. Effects of trust violation behaviors on investment in AI teammate, facetted by repair strategy.

We first investigated the effects of AI trust violations and repair strategies on the investment amount, which can reflect participants' trust in the AI teammate's performance of doubling the power. The more human invested in AI teammate, the higher trust people show in AI teammate's

performance. There were no main effects of AI trust violations or repair strategies. However, we found a significant and positive effect of the individual differences on Honesty–Humility, β = 0.40, 95% CI [0.03, 0.78], $t(519) = 2.10$, $p = 0.036$: people who have higher Honesty–Humility scores were more likely to invest in AI teammate in the game.

Perceived Cooperation of AI teammate

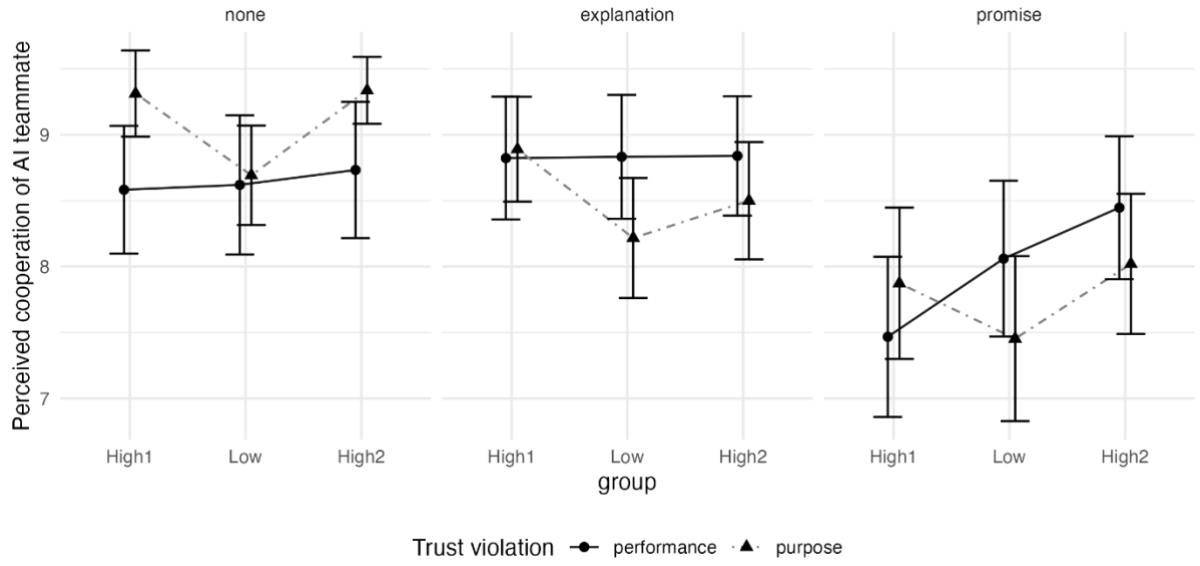


Figure 24. Effects of trust violation behaviors on perceived cooperation of AI teammate, facetted by repair strategy.

For the perceived cooperation of the AI teammate, which is the amount participants guessed that the AI teammate would allocate to the team rover. The higher the value, the more people trust that the AI teammate would allocate to the team. We found a significant and positive interactive effect of the AI state, $\beta = 0.79$, 95% CI [0.31, 1.28], $t(519) = 3.34$, $p < .001$: as the game progressed, the predicted values of AI team allocation amount is increasing. This means that people show increasingly higher trust in the AI teammate's cooperation. Additionally, we found a significant and negative interaction effect of AI violation behavior [purpose] and AI state [low], β = -0.86, 95% CI [-1.55, -0.18], $t(519) = -2.48$, $p = 0.013$: when AI teammate violates the purpose-based actions (i.e. not allocate the full amount to the team), people's perceived cooperation of AI teammate drops significantly. For the individual differences, again, we found a significant and positive effect of Honesty–Humility, $\beta = 0.68$, 95% CI [0.28, 1.09], $t(519) = 3.32$, $p < .001$: people

have higher Honesty–Humility scores would have higher perceived cooperation of AI teammate.

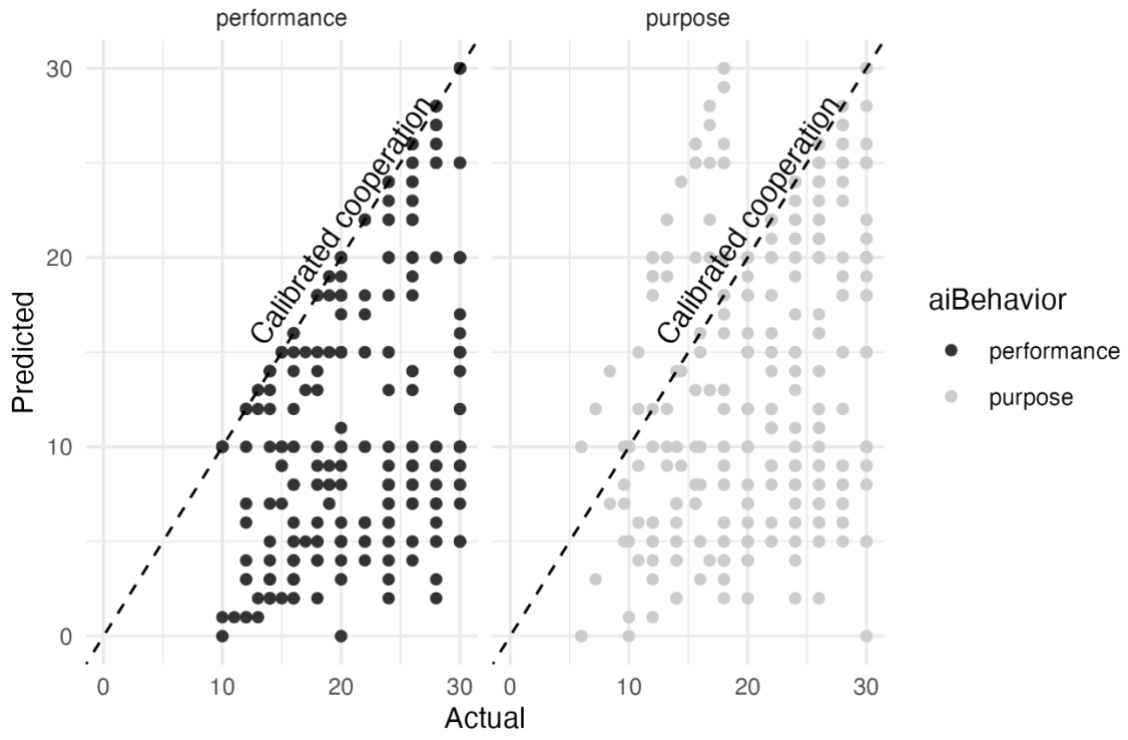


Figure 25. Actual versus predicted amount of power AI teammate allocated to the team, facetted by the AI trust violation behaviors.

We showed the actual versus predicted amount of power AI teammate allocated to the team in the Figure 25. The dotted diagonal line shows the calibrated cooperation where people's predicted values match with the actual values, mirroring the calibrated trust in Lee and See (2004). When the predicted values exceed the actual values, people over-cooperate with the AI teammate, which only be observed in the purpose-based trust violations. Results showed a strong effect of underestimating AI teammate's cooperation level.

Participants' Team Allocation

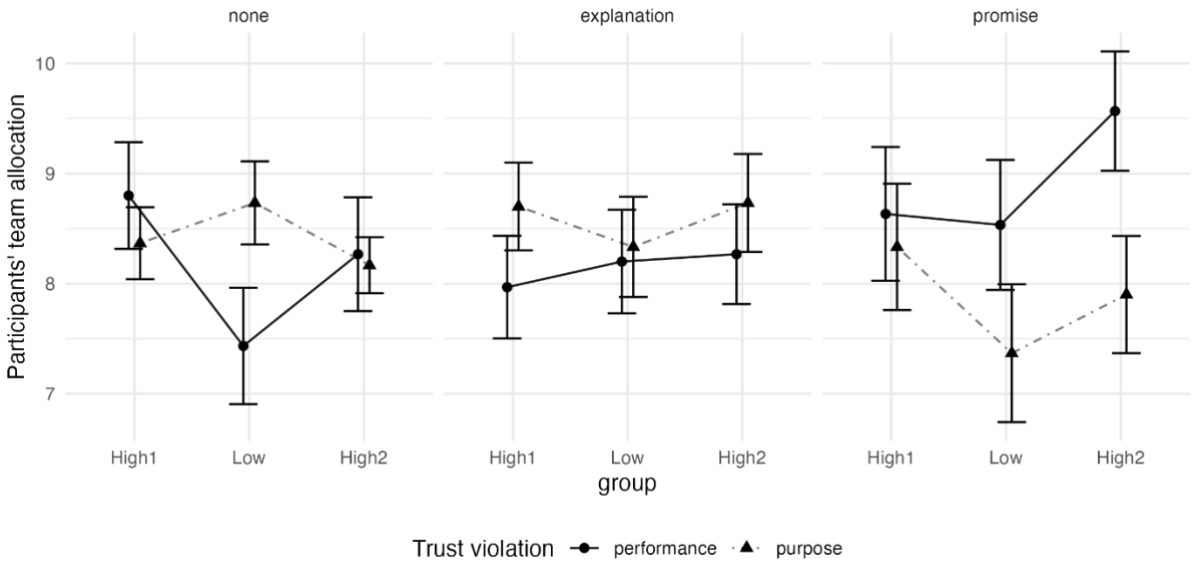


Figure 26. Effects of trust violation behaviors on participants' cooperation levels, facetted by repair strategy.

For participants' team allocation, which is measured by the proportion that participants allocated to the team rover, the higher the value, indicating the more cooperative participants are in the game. We found a significant and negative effect of AI state [low], β = -1.37, 95% CI [-2.59, -0.14], $t(519)$ = -2.20, p = 0.028: people were less likely to cooperate when AI teammate made any trust violations. Additionally, we found a significant and positive two-way interaction between Trust violation [purpose] and AI state [low], β = 1.73, 95% CI [0.01, 3.46], $t(519)$ = 1.97, p = 0.049. Compared to the main effect of the AI state, where people dropped their team allocation levels when the AI teammate did not cooperate with the team goal, instead, people became more cooperative by allocating more to the team goal. While the main effects are non-significant, we found a significant and negative three-way interaction effect between Trust violation [purpose] × repair strategy [promise] × state [low], β = -2.60, 95% CI [-5.04, -0.16], $t(519)$ = -2.09, p = 0.037. When the AI teammate used promises to repair a purpose-based trust violation people's trust dropped.

6.5 Study 1 Discussion

The goal of the study is to address two research questions: first, we explored whether people's trust declines more with performance-based versus purpose-based trust violations. Second, we aimed to determine the most effective strategy for managing different types of trust violations. To address these questions, we designed a game-theoretic paradigm that captures both performance- and purpose-related human-AI interactions with a shared team goal.

Regarding the first research question, our findings demonstrated that purpose-based trust violations, where an AI teammate fails to cooperate with the team goal, lead to a greater decline in trust. Looking at the sub-dimensions of subjective trust measures, we found that purpose-based trust violations show detrimental effects on both performance- and purpose-dimension of trust, meaning people transfer and perceive AI teammate's cooperation failure to its performance, rather than intents. On the other hand, when repair the purpose-based violations, we found that only performance-dimension can be repaired after an apology paired with an explanation. The purpose-dimension, meaning how people perceived AI's intents, was unrepairable. These results highlighted the importance of the purpose-based trust violations: not only show a transferring effects on multidimensional trust, but also leave an unrepairable influences on people's perception of AI intentions.

Additionally, the effect of purpose-based trust violation on people's trust is long-lasting throughout the study. This persistent effect was specifically observed in the purpose-dimension of trust, while absent in the performance-dimension of trust, which proved to be more reparable. These results align with the findings of Alarcon et al. (2020) and further emphasize that people are more sensitive to purpose-based violations, which have a detrimental effect on their perception of the AI teammate's purpose and are challenging to repair. However, the utility functions of the performance- and purpose-based violations were not specified. Future studies can better quantify the costs of each type of trust violations and investigate their impacts on trust.

In addition, people also exhibit a decreased belief in the AI teammate's willingness to cooperate, measured by the amount participants anticipated the AI teammate would contribute less to the team goal. Our results highlighted the importance of the aligned goal for designing a trustworthy AI teammate. In the safety critical domains, such as healthcare, military, and emergence response where the expectation for system efficiency is high and the tolerance for breakdowns is very low, the AI teammates must be highly reliable. In addition, the AI teammates goals must align with the organization, which might not align with the individual. Our results such misalignment might be particularly detrimental to trust. Future studies can further validate our findings to real-world scenarios where involve human and AI teammate cooperates on a shared goal.

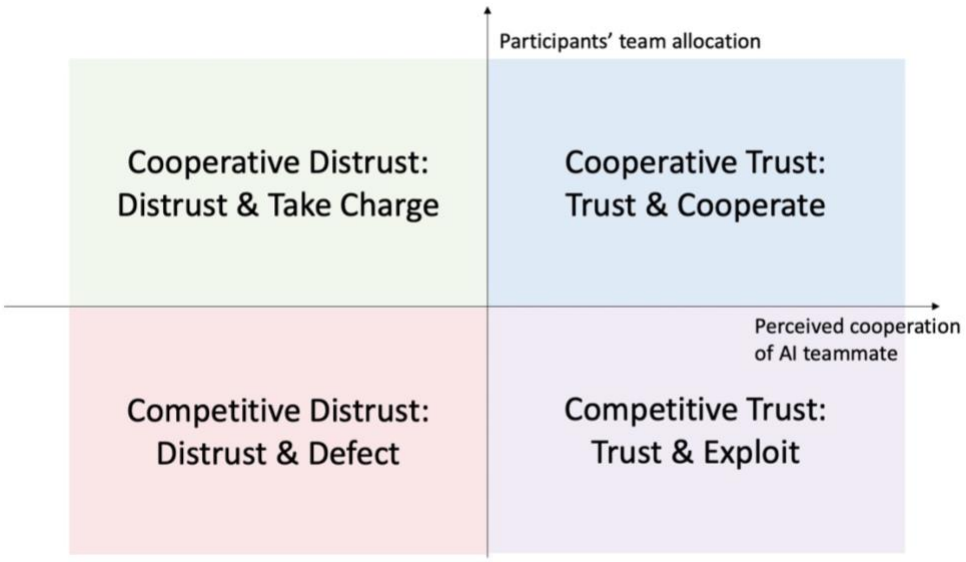


Figure 27. Interaction of Human Cooperation and the Perceived Cooperation of the AI Teammate.

It is important to acknowledge the complexities in the relationships between perceived cooperation and actual cooperation. It is not always the case that low perceived cooperation of an AI teammate leads to decreased cooperation. Our findings demonstrate that when the AI teammate violates purpose-based actions, such as not allocating the full amount to the team, people's perception of the AI teammate's cooperation significantly decreases. However, instead of reducing their own cooperation, people actually become more cooperative by allocating more towards the team goal. This behavior can be understood within the game design and reward system, where a larger reward is granted upon reaching the team threshold. It becomes rational for individuals to compensate for the gaps in the AI teammate's team allocation in order to fulfill the threshold and attain the team reward.

To further investigate the relationships between people's cooperative perception (i.e., perceived cooperation of AI teammate) and their behaviors (i.e., participants' team allocation), we identified the four potential outcomes, as shown in Figure 27: cooperative trust, cooperative distrust, competitive trust, and competitive distrust. For example, the trust and exploit condition, if participants perceive AI teammate as highly cooperative but fail to cooperate, they trust AI teammate's intent to cooperate but try to exploit the AI teammate. Our findings supported a scenario in which people exhibit lower perceived cooperation of the AI teammate but display a high level of cooperation, categorized as "Distrust & Take charge." This suggests that when the AI teammate fails to cooperate, individuals strive to compensate for its role and contribute more towards achieving the team goal. One potential explanation for this outcome is the game design,

where a two versus one ratio was designed between team and individual goals. Since the team payoffs were doubled than the individual payoffs, although displaying low trust, participants would engage in cooperative behaviors to gain the higher payoffs by the end of the game. If we are closing the ratio between team and individual goals (e.g., 1.1:1), it would be possible that people fail to the 'competitive distrust: distrust and defect' category. This suggests that the game structure can be the global influences that dominate people's behaviors, which warrants further empirical investigation. To better investigate this, dynamic system, specifically the attractor field theory, provides a good theoretical perspective to frame the question. The attractor dynamics can be pictured by effective energy landscapes, which indicate the basin of attraction by valleys, and the attractor states by the bottom of the valleys (Rolls, 2010). The game structure, by quantifying the payoff ratio, can form various attractors (e.g., four types of outcomes identified in Figure 27) where people tend to gravitate towards despite a wide variety of starting conditions. Various attractors can be stable or instable, depending on the average time that people stay in the basin of attraction under the influence of noise, such as experiencing a trust violation event. Future study can apply the attractor dynamic theory to the human-AI cooperation. Specifically, researcher can study the influences of initial trust (e.g., initial positions), trust violations (e.g., perturbance and noises), and game structure (e.g., attractors and its stability) on people's trust and decision makings.

Addressing the second research question, our findings showed that an apology with an explanation is more effective than a promise in repairing trust after purpose-based trust violations. This conflicts with the findings by Esterwood and Robert Jr (2023), who found that various repair strategies, including an apology, explanation, promise, and denial, were unable to repair trustworthiness to a pre-violation level. A potential explanation for this difference is that we integrated apology with explanation, rather than implementing a single repair strategy. This combination may enhance the effectiveness of the trust repair strategies employed in our study. Furthermore, it is important to note that our study designed only two trust violations, while Esterwood and Robert Jr's study included three violation events. Future research could investigate the impact of varying numbers of trust violation events on the efficacy of trust repair strategies. It would be particularly valuable to determine if there exists a critical threshold of trust violations beyond which trust becomes irreparable.

In addition to our findings regarding subjective ratings, analysis of behavioral data from the game revealed a noteworthy trend: making promises after trust violations resulted in decreased cooperation. This can be attributed to the repeated trust violations within the game. Promises establish expectations, and when the AI teammate made a promise following the first trust

violation, individuals adjusted their expectations for future interactions. However, when the second trust violation occurred, the misalignment between the initial promise and the subsequent behavior led to a decrease in cooperation levels. These results were consistent with interpersonal trust recovery processes, where deception harmed the trustee's credibility, and as a result harmed the initial effectiveness of a promise in repairing trust (Schweitzer et al., 2006). Thus, a promise that follows the second trust violation is far less likely to change impressions and expectations. This may suggest that the consistency of the AI teammate's actions over time is critical for the promise strategy. Future research could delve deeper into understanding how these dynamics unfold and explore strategies to mitigate the negative consequences of broken promises. Additionally, researchers should be cautious about using promises in the human-AI interactions without fully understanding the factors that govern whether the AI teammate will be able to fulfill its promises. In other words, the interaction design should be well-aligned with the algorithm design of the AI.

We found the Honesty-Humility dimension in the HEXACO Personality assessment was a strong predictor for people's performance-dimension of subjective trust rating, investment behaviors, and perceived cooperation of AI teammate. Specifically, individuals with higher levels of Honesty-Humility were rate the AI teammate as more capable, were more inclined to allocate greater resources to the AI teammate and held higher expectations regarding the AI teammate's contributions toward the team goal.

Honesty and Humility, as reflected in this dimension, encompass traits related to fairness and genuineness in interpersonal interactions. It signifies a tendency to cooperate with others even in situations where one could exploit them without fear of retaliation (Ashton et al., 2014, p. 156). Our findings align with previous research conducted in the context of public goods games, which demonstrated that individuals with high levels of Honesty-Humility also exhibit stronger aversion to inequality in their social value orientation and hold more positive beliefs about the cooperative behavior of others (Hilbig et al., 2012; Krueger, 2008). Furthermore, individuals high in Honesty-Humility tend to display higher levels of general cooperation and are less inclined to condition their behavior on situational factors (Hilbig et al., 2012). It would be valuable for future studies to explore the power dynamics that exist between human and AI teammates and examine potential mechanisms through which situational factors, such as the ratio between team threshold and rewards, may influence human-AI cooperation.

Limitation and Future Works

Our studies have a few limitations that suggest future research. We conducted our experiment online using the Amazon Mechanical Turk platform. While previous studies have demonstrated comparable effects to lab-controlled experiments when careful participants' inclusion criteria and screening are employed (Crump et al., 2013), the online experiment limited our ability to conduct any follow-up interviews or in-depth observations of participants' behaviors. Thus, gaining further insights into participants' thought processes throughout the experiment became challenging. Future studies could consider transitioning from online studies to in-person experimental settings, which open the doors for extensive observations, such as capturing participants' feedback and facial expression. Second, while our studies showed that it is possible to manage people's trust via verbal trust repair cues, the acoustic cues of the voice design were neglected in our study. Torre

and colleagues showed that people would trust and invest more in a smiling and happy-sounding AI teammate in similar game theory setting (Torre et al., 2020). Additionally, prior study has showed that trust is not only expressed by what people say, but also how people say it (Li et al., 2023). Future studies should further investigate whether the acoustic features of people trusting expression also affect how people perceive and calibrate trust.

6.6 Study 2: Trusting Voice for Trust Repair

The objective of Study 1 is to determine the appropriate content for managing trust, particularly for different types of trust violations. While Study 1 showed that it is possible to manage people's trust via verbal trust repair cues, the acoustic cues of the voice design were neglected in our study. In addition, our research findings from Chapter 3 suggested that trust is not only conveyed through language, but also through acoustic cues. We have identified formants, fundamental frequency, and Mel-frequency central coefficients as the most significant acoustic indicators of trust in conversations. The primary question is whether these **indicators of trusting** have might also influence **perceived trustworthiness**.

6.6.1 Trusting versus trustworthy voices

When studying the verbal and nonverbal cues of trust, the difference between the trusting voice of one who trusts another and trustworthy voice of one who is trusted by others has been often neglected. The fundamental distinction lies in the locus of trust, which determines the perception of trust between the trustor (the individual making trust judgments) and the trustee (the entity being trusted) (Jones & Shah, 2016). Trusting voice is expressed by the trustor, whereas the trustworthy voice was associated with the trustee. In most human-AI interactions, people act as trustors who place their trust in the AI teammate as the trustee. Thus, when designing the voice

of an AI teammate, a trusting voice signals that the trustee (AI teammate) trusts the trustor (people), whereas a trustworthy voice signals the trustee's trustworthiness, such as their capabilities and dominance. The trusting voices require more in-depth investigation of the causal relationship between trust and acoustic cues than the trustworthy voice, which usually investigated via associations. This means that researchers need to manipulate the trustor's trust level and see how it reflected in their voice.

The causal relationships between acoustic cues and trust by directly manipulating people's trust in a virtual agent has been identified in the prior literatures (Li et al., 2023). By analyzing people's lexical and acoustic cues in conversations, Li and colleagues identified the formants (F_1, F_2, F_3, F_4) and standard deviation of fundamental frequency (F_0) as the key indicators of people's trust. However, a critical question remains: do these acoustic cues in the trusting voice affect perceived trustworthiness? In other words, does the trusting voice equal to the trustworthy voice? The existing findings in the literature are mixed. Elkins and Derrick (2013) found that the vocal pitch was inversely related to perceived trust. Torre and colleagues found that a smiling voice (indicated by higher F_0 and higher formants) increased trust and received higher overall investment in an iterated investment game (Torre et al., 2020). While pitch has been found to significantly affect perceived trustworthiness, previous research have also shown the importance of the variance of pitch (intonation) when people express their trust (Ponsot et al., 2018). Additionally, formants have demonstrated potential for predicting trust and influencing perceived trustworthiness (Montano et al., 2017). Additionally, Knowles and Little (2016) demonstrated that individuals with larger apparent vocal tracts (lower formants) were perceived as more cooperative. Given the link between cooperativeness and trust, this finding further highlights the interrelation between voice characteristics and trust perceptions (Knowles & Little, 2016). Therefore, our study aimed to examine the effectiveness of the identified acoustic cues of trusting voice in managing people's perceived trustworthiness and trusting behaviors.

6.6.2 Lexical-Acoustic Congruency in Trust Repair

The congruent effect between acoustic and lexical cues refers to the phenomenon where the acoustic characteristics of a spoken word (e.g., pitch, duration, and loudness) match the meaning of the word (as conveyed by its lexical content), resulting in more accurate processing of the world. For example, if someone say "happy" with a high-pitched and loud tone, it is congruent with the lexical content of the word "happy," and the listener is likely to recognize the word more accurately than if it were spoken with a low-pitched and quiet tone. Because of the model of "Emotions as Social Information (EASI)", which suggests that emotions play a vital role in comprehending

ambiguous situations, and their influence is contingent upon the specific context of the interaction (Van Kleef et al., 2010). Prior research found that people showed people change their trusting behavior based on the congruency of the agent's behavior with the participant's first impression, where people showed a lower trust when deceived by a trustworthy-sounding voice (Torre et al., 2018). Similarly, Antos, De Melo, Gratch and Grosz (2011) found that, even with the same strategy in a negotiation game, participants selected the agents with congruent emotion expressions with their actions (Antos et al., 2011).

While previous research identified the congruency effects between the actions and expressions of the agent, the congruency between linguistic content (e.g., apology) and acoustic cues (i.e., high- and low-trusting voice) has not been studied extensively in the context of trust repair. The core principle of trust repair is to perform a behavior aimed at increasing trust after a failure. The congruency effects between trusting voice and trust repair contents can show influences in two directions: first, a high-trusting voice is symmetrical to the trustworthy voice, which can influence people's trust perception and promote people's trust, and thus reinforce trust repair contents. This indicates positive congruency, meaning high-trusting voice paired with trust repair content, both shows positive effects on promoting trust. However, a high-trusting voice, indicated by higher formants, is often perceived as 'smiling/happy voice' (Torre et al., 2020). The alternative theory is a positive-affect voice might mismatch the emotional context of the trust repair content. Because not all positive affect can elicit trust and cooperation, it also depends on specific social context information and further shapes the interpretation and recognition of the expressions (Krumhuber et al., 2023; Rychlowska et al., 2021). Specially, prior study has shown that smiles in a negative situation were considered less genuine than the same smiles rated in isolation (Mui et al., 2020). Thus, in our study, the high-trusting voice can hinder trust repair because it does not show sincerity and remorse when saying sorry. This suggests a negative congruency between acoustic and lexical cues of trust: low trusting voice paired with trust repair can be more effective. The direction and effects of the congruency effect should be closely exanimated in the context of trust management.

To address these research gaps, we investigated the effects of trusting voices and congruency with the trust repair content on managing trust. We adopted the identified acoustics cues–formants, F_0, MFCC–to manipulate high- and low-trusting voice. The caus-effect driven positive congruency would predict a high-trusting voice would promote people's trust and trusting behaviors. The emotional-state driven negative congruency would predict a low-trusting voice, which is more consistent with the trust repair emotional state, would increase people's trust and promote cooperation.

6.7 Study 2 Method

The aim of Study 2 is to identify the effects of congruency between the trust management content and acoustic cues of trust. A $3 \times 2 \times 2$ mixed-design experiment with two levels of AI teammate condition (i.e., high1, low, high2) as within-subject variable, two types of trust violation dimension (i.e., performance and purpose), and two levels of agent's voice (i.e., high and low). The AI teammate condition and trust violation dimensions remain the same as the study 1. We adopted the most effective trust management strategy for each type of trust violation identified in study 1 (e.g., apology and explanation for purpose violation). We adopted the same experimental platform by using the Space Rover Exploration Game introduced in Section 6.3.1. The dependent variables remain the same: Multi-Dimensional Measure of Trust (MDMT) for the subjective measurements; investments in AI teammate, perceived cooperation, and participants' allocation to team as behavioral measurements.

6.7.1 Voice manipulation

We used an online text-to-speech to generate all utterances by using the male voices (*Voicemaker*, 2023). Prior to any manipulation, the original mean fundamental frequency of the voices used as stimuli was 76.84 Hz, *SD* = 67.03Hz.

To manipulate agent's voice, we relied on the findings from the previous study on the causal relationship between acoustic features and trust (Li et al., 2022). Specifically, we employed formants (F_1, F_2, F_3, F_4) and the standard deviation of F_0 as the parameters for voice manipulation. Based on the results of our study, we determined that higher mean values for F_1, F_2, F_3, lower mean values for F_4, and higher standard deviation of F_0 corresponded to a higher expression of trust in the voice. We used Praat, a software commonly used for speech analysis in phonetics (Boersma & Antos, 2012). Additionally, we employed Vocal Toolkit, a Praat plugin equipped with automated scripts for voice processing (Corretge, 2023). Following prior study's findings (Li et al., 2023), we adjusted the mean values of F1, F2, F3 by 8%, 13%, and 5% respectively, the standard deviation of F0 by 10%, and decreased the mean value of F4 by 6% to create the high- versus low-trusting voice manipulation.

To confirm that the trusting voices could be appropriately perceived, we have conducted eight in-person pilot testing by playing two utterances arranged in random order. Experimenters asked them to rank and qualitatively report their perception and preferences towards the two sample voices. Pilot participants all reported that they could clearly distinguish two audio samples. Among the eight pilots, 75% of pilots preferred the low-trusting voice and commented as "friendly" and

"comfortable" whereas 25% of pilots preferred the high-trusting voice and commented as "easier to understand" and "confident". Overall, these results show the manipulation of acoustics cues–formants–is effective in changing people's perception of voice.

Table 13. Study 2 Experimental Design with Two Types of Trust Violation Dimension and Two Levels of Agent's Voice Congruency Corresponding to the Stage of Trust Management.

	High		Low	High
	1-5		6-10	11-15
#		Trust Violation	Trusting Voice	
1	Neutral voice	Performance	High trusting voice	Neutral voice
2	Neutral voice	Performance	Low trusting voice	Neutral voice
3	Neutral voice	Purpose	High trusting voice	Neutral voice
4	Neutral voice	Purpose	Low trusting voice	Neutral voice

6.7.2 Participants

Participants were screened for the same criteria in the Study 1: they must live in the United States, have completed more than 1000 tasks with at least a 98% approval rate on Amazon Mechanical Turk, and have completed all the study tasks and passed the attention check. A priori power analysis was conducted using G*Power3 (Faul et al., 2007) to test the difference between six independent group means using an F-test, a medium effect size (d = .25), and an alpha of .05. Result showed that a total sample of 113 participants with four equal sized groups of n = 29 was required to achieve a power of .80. We recruited 123 participants, and after excluding 3 participants who failed the attention check, a total of 120 participants remained for analysis. The age range for these valid participants was 20 to 65 years, with a mean age of 43. Among the valid participants, 73 identified as male and 47 identified as female. The compensation structure remains the same as the Study 1. Participants were compensated with a base rate of $3 for their thirty-minute participant time (equivalent to a rate of $6 per hour). They could also earn an additional amount of up to $1 based on every 100 points gained in the game, with any remaining points rounded up for compensation purposes (e.g., 230 points would be compensated as an additional $3). Participants had the potential to earn a minimum of $3 and a maximum of $7 based on their performance.

6.8 Study 2 Results

In the Study 1, our findings indicated that apology with an explanation was the most effective trust repair strategy for purpose-related trust violations. In Study 2, we incorporated the explanation strategy with high and low level of trusting voice. Specifically, we manipulated the

formants and variance in pitch to create high- versus low-trusting voices. By combining the trust repair strategy of apology with explanation and the manipulation of trusting voices, in Study 2, we aimed to investigate the effects of these voices and their congruency with the management contents on managing trust.

We fitted a linear mixed-effects model with subjective ratings and game behavioral measurements as dependent variables. Subjective trust measurements included the general trust and its performance- and purpose- dimension. Game behaviors included investment amounts, perceived cooperation, and participants' team allocation. As fixed effects, we entered trust violation, repair voice, and state as well as their two-way and three-way interaction terms into the model. Additionally, looking at individual differences of participants, we added Humility–Honesty and gender to the linear model. As random effects, we had intercepts for subjects. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question.

6.8.1 Subjective Trust Measurement

Similar to Study 1 analysis, we processed the MDMT scale by averaging ratings of each dimension and then took an overall average for the subjective trust ratings. All items meet or exceed the benchmark criteria of $\geqslant 0.7$ for construct reliability (Fornell & Larcker, 1981). Item reliabilities include $\alpha = 0.87$ for performance-based trust, $\alpha = 0.91$ for purpose-based trust, and $\alpha = 0.88$ for all items. Summary statistics for the trust is reported in the Table 14.

Table 14. Number of participants, mean rating, and standard deviation of trust for each type of trust violation and voice condition.

Behavior	Voice	N	High 1 M(*SD*)	Low M(*SD*)	High 2 M(*SD*)
Performance	High	30	5.68 (*SD* = 0.91)	5.50 (*SD* = 0.88)	5.66 (*SD* = 0.90)
Performance	Low	30	5.84 (*SD* = 0.82)	5.75 (*SD* = 0.80)	5.90 (*SD* = 0.81)
Purpose	High	30	5.91 (*SD* = 0.89)	5.39 (*SD* = 1.20)	5.77 (*SD* = 0.97)
Purpose	Low	30	5.94 (*SD* = 0.92)	5.57 (*SD* = 0.82)	5.82 (*SD* = 0.97)

Similar to study 1, we fitted a linear mixed-effects model with AI behavior, repair voice, state as well as their interaction terms as fixed effect into the model. As random effects, we had intercepts for subjects. Visual inspection of residual plots did not reveal any obvious deviations

from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question.

Table 15. Study 2 linear mixed-effect model result for subjective trust ratings.

Fixed Effect	Trust				Performance-dimension				Purpose-dimension			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
(*Intercept*)	5.49	0.42	13.01	**0.01**	5.15	0.40	12.87	**0.01**	5.84	0.57	10.17	**0.01**
Trust violation (purpose)	0.09	0.38	0.23	0.82	0.24	0.37	0.64	0.52	-0.07	0.50	-0.13	0.90
Voice (low)	-0.41	0.35	-1.17	0.25	-0.45	0.35	-1.31	0.19	-0.37	0.47	-0.79	0.43
State (low)	-0.09	0.15	-0.65	**0.52**	-0.29	0.18	-1.57	0.12	0.09	0.18	0.51	0.61
Honesty Humility	0.11	0.11	0.98	0.33	0.26	0.10	2.55	**0.01**	-0.04	0.15	-0.30	0.76
Trust violation (purpose) × Voice (low)	0.40	0.52	0.77	0.44	.018	0.50	0.36	0.72	0.62	0.69	0.89	0.38
Trust violation (purpose) × State (low)	-0.58	0.24	-2.41	**0.02**	-0.71	0.29	-2.45	**0.02**	-0.45	0.28	-1.60	0.11
Voice (low)× Gender (male)	0.97	0.46	2.09	**0.04**	0.62	0.45	1.38	0.17	1.31	0.62	2.12	**0.04**

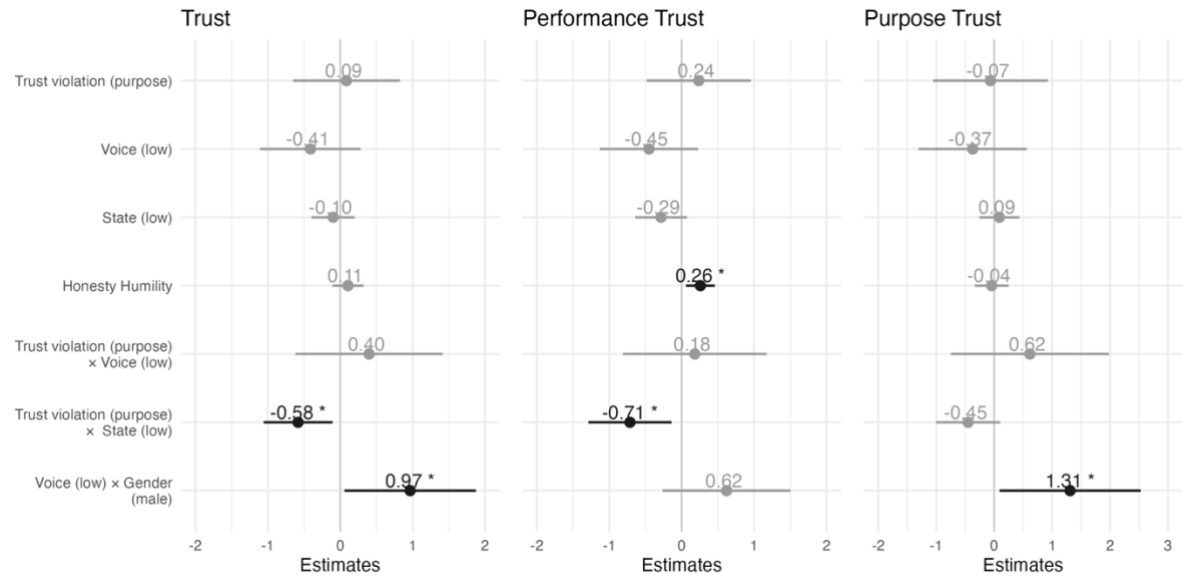


Figure 28. Study 2 linear mixed-effect model results of subjective trust ratings.

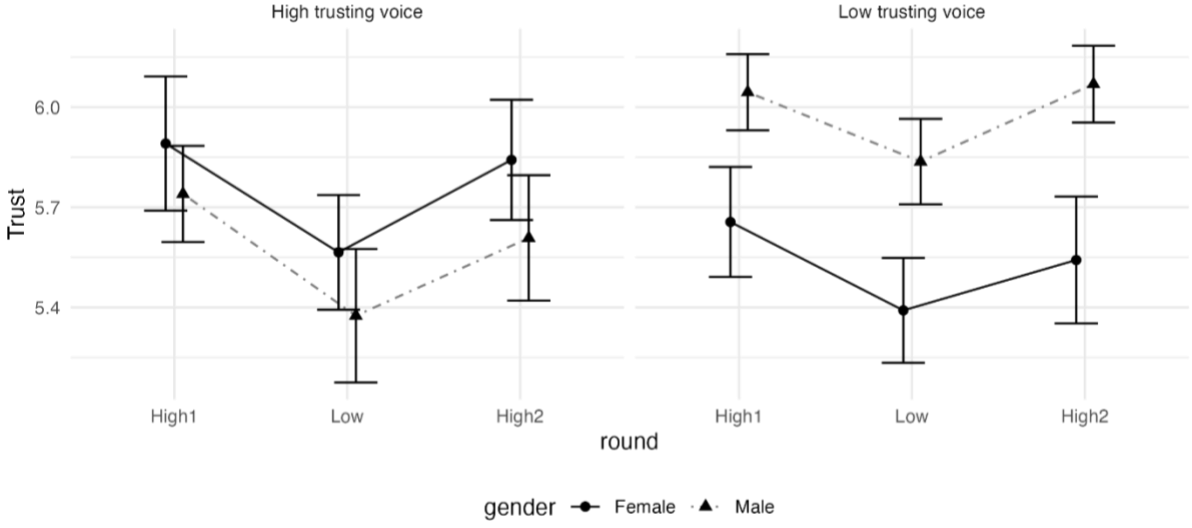


Figure 29. The interaction effect of AI voice and gender: men demonstrated a higher trust in the trust low-trusting voice of a male-voiced AI teammate.

The interaction effects of AI trust violation behaviors and AI state [low] on trust are statistically significant and negative, β = -0.58, 95% CI [-1.06, -0.11], t (333) = -2.41, p = 0.017. This means that people's trust drops more when AI teammate violates the purpose-related actions. The effects were also shown on the performance-dimension of trust, β = -0.71, 95% CI [-1.29, -0.14], t (333) = -2.45, p = 0.015. For the main effect of trusting voice, we did not find a significant effect, p = 0.91: people's trust level is similar when hearing the high-trusting and low-trusting voice explanations.

We also found a significant interaction effect between gender and AI's trusting voice. The interaction effect of AI low-trusting voice and gender [male] is statistically significant and positive, β = 0.97, 95% CI [0.06, 1.87], t (333) = 2.09, p = 0.037. The effects were also shown on the purpose-dimension of trust, β = 1.31, 95% CI [0.09, 2.53], t (333) = 2.12, p = 0.035. This suggests that males perceived the low-trusting voice as more trustworthy. Specifically, males perceived the low-trusting voice as more 'kind' and 'considerate' on the purpose-dimension. On the other hand, females did not show significant difference in perceiving high (M = 5.72, SE = .18) and low trusting voices (M = 5.55, SE = .18), p = 0.49.

Looking at the individual differences of participants, we found a significant and positive effect of Honesty–Humility on performance-dimension of trust, β = 0.26, 95% CI [0.06, 0.46], t(333) = 2.55, p = 0.011: people who have higher Honesty–Humility scores are more likely to trust AI teammate's performance.

6.8.1 Game Behaviors

Table 16. Study 2 linear mixed-effect model result for game behavior.

Fixed Effect	Investment in AI teammate				Perceived cooperation				Participants' team allocation			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
(*Intercept*)	6.42	1.04	6.18	**0.00**	4.78	1.34	3.57	**0.00**	8.15	1.18	6.89	**0.00**
Trust violation (purpose)	-0.13	0.96	-0.14	0.89	-0.32	1.16	-0.28	0.78	0.05	1.21	0.04	0.97
Voice (low)	-0.78	0.90	-0.87	0.39	-1.89	1.09	-1.73	0.09	-0.77	1.14	-0.68	0.50
State (high2)	1.03	0.50	2.05	**0.04**	0.83	0.31	2.63	**0.01**	-0.57	0.86	-0.66	0.51
Honesty Humility	2.55	2.62	0.97	0.33	0.88	0.35	2.53	**0.01**	0.22	0.28	0.77	0.44
Trust violation (purpose) × Voice (low)	1.12	1.32	0.85	0.40	1.82	1.59	1.14	0.26	0.54	1.66	0.33	0.75
Voice (low)× State (high2)	-1.81	0.76	-2.40	**0.02**	-0.41	0.47	-0.87	0.39	-0.06	1.30	-0.05	0.96
Voice (low)× Gender (male)	0.28	1.18	0.24	0.81	2.53	1.42	1.78	0.08	1.49	1.48	1.00	0.32

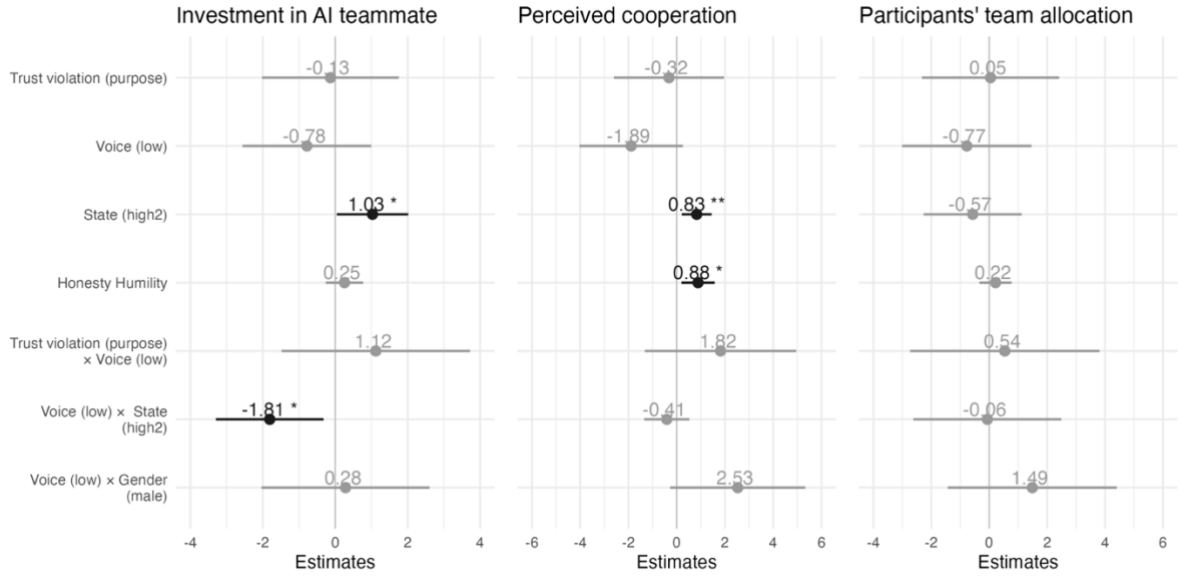


Figure 30. Study 2 linear mixed-effect model results of game behaviors.

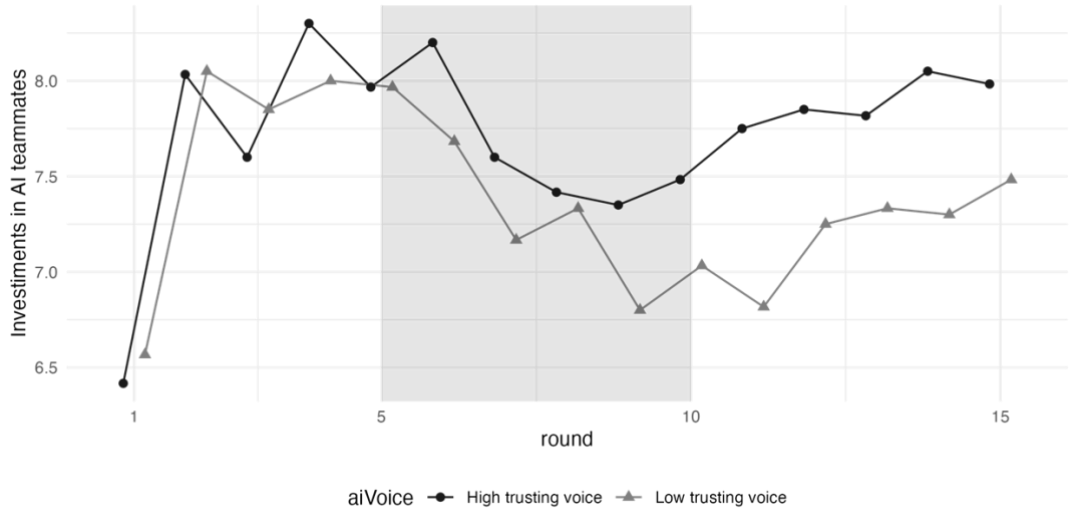


Figure 31. High-trusting voice enhances people's investments over time.

We first investigated the effect of AI trust violation behaviors and repair strategies on the investment amount, which can reflect participants' trust in the AI teammate's performance of optimizing and doubling the power. The more human invested in AI teammate, the higher trust people show in AI teammate's performance. The effect of state [High2] is statistically significant and positive, β = 1.03, 95% CI [0.04, 2.01], $t(333) = 2.05$, $p = 0.041$. Additionally, for the trusting voice, we found a significant and negative interaction effect between trusting voice [low] and state [High 2], β = -1.81, 95% CI [-3.30, -0.32], $t(333) = -2.40$, $p = 0.017$: when AI teammate repaired people's trust using a low trusting voice, people's trust in AI teammate's performance declined,

indicating by giving less power to AI teammate. The high-trusting voice group recovered and grew over time as shown in Figure 31.

For the perceived cooperation of the AI teammate, which is the amount participants guessed that the AI teammate would allocate to the team rover, it can be operationalized as people's trust in the AI teammate's purpose dimension. The higher the value, the more people trust that the AI teammate would allocate to the team. The effect of state [High2] is statistically significant and positive, β = 0.83, 95% CI [0.21, 1.45], $t(333) = 2.63$, $p = 0.009$: as the game progress, the predicted values of AI team allocation amount increased. This means that people show increasingly higher trust in AI teammate's cooperation. For the individual differences, we found a significant and positive effect of Honesty–Humility, β = 0.88, 95% CI [0.20, 1.57], $t(333) = 2.53$, $p = .012$: people with higher Honesty–Humility scores had higher perceived cooperation of AI teammate.

For participants' team allocation, which is measured by the proportion that participants allocated to the team rover, the higher the value, indicating the more cooperative participants are in the game. We did not find a main effect, nor interaction effects of AI trusting voices and AI violation behaviors on participants' cooperation level. We found a significant effect of AI state [low], β = -2.14, 95% CI [3.84, -0.45], $t(333) = -2.49$, $p = .013$: this means that participants' allocating much less to the team goal when AI teammate conduct trust violations in the low state.

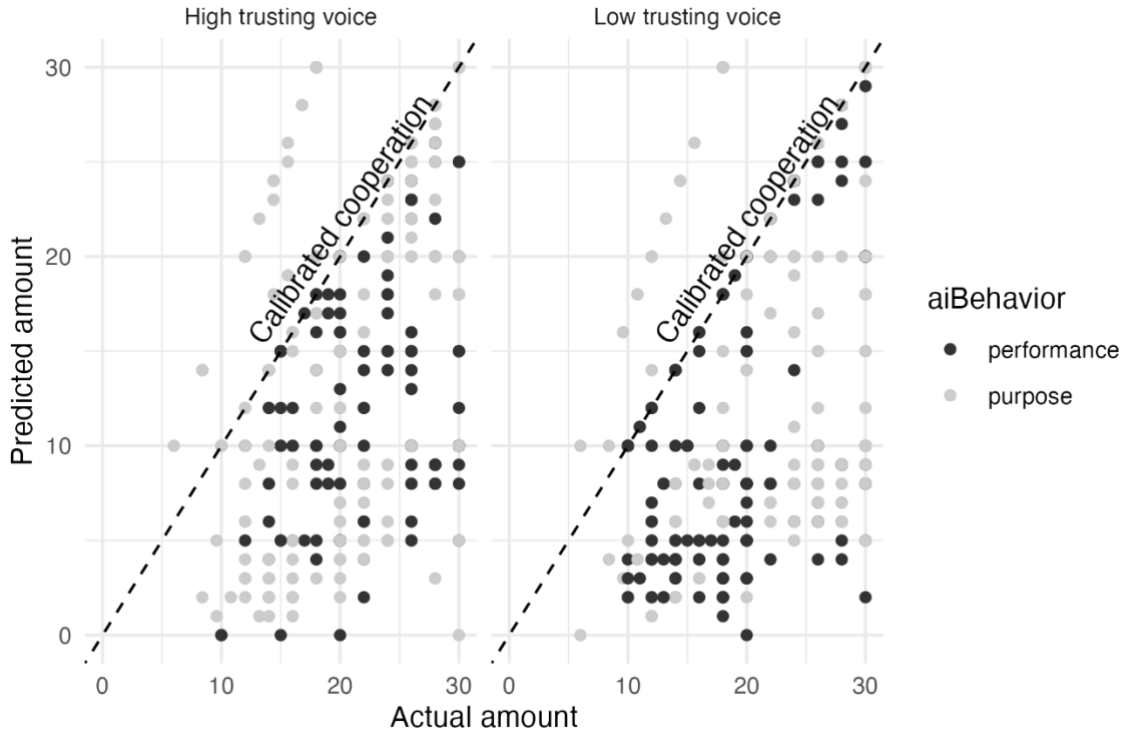


Figure 32. Actual versus predicted amount of power AI teammate allocates to the team rover.

Similar to Study 1, we showed the actual versus predicted amount of power AI teammate allocated to the team in the Figure 32. Consistent with Study 1, results showed that only purpose-based trust violation leads to the over-estimation of AI teammate's cooperation level.

6.9 Study 2 Discussion

The goal of the study is to address the congruency effects between the trusting voice and trust management content. Based on our previous findings which indicated that apology with an explanation was the most effective trust repair strategy for purpose-related trust violations. We incorporated this strategy, where we further explored the impact of trusting voices on trust repair. Specifically, we manipulated the formants and variance in fundamental frequency to create high- versus low-trusting voices. Through this design, we aimed to understand how the characteristics of the agent's voice interacting with the trust repair content can influence trust and cooperation behaviors in the game-theoretic setting.

Our results showed that when AI teammates used a high-trusting voice to repair trust after trust violations, participants invested more resources to AI teammates. These results confirmed the positive congruency effect between the acoustic and lexical cues of trust. These findings suggest that the use of a high-trusting voice is symmetrical to the trustworthy voice, which can reinforce trust repair contents and have a notable impact on people's trusting behaviors. One

possible explanation is that people may associate a high-trusting voice with smiling voices, both of which convey positive valence and promote trusting behaviors. Torre et al. (2020) manipulated smiling voices, which were also indicated by higher formants, and observed enhanced investments as behavioral trust in a similar game-theoretic context. Instead of being incongruent with the trust repair context, Torre and colleagues suggested that smiling voice showed a 'halo effect', which enhance the positive behaviors overall. By precisely manipulating the acoustic cues of trusting voices and validating their influence on trusting behaviors, our study contributes to the understanding of the role of voice in shaping trusting behaviors.

We also found some evidence of individual differences affecting subjective trust rating, as shown by the two-way interaction of AI trusting voice and gender. Specifically, we found that men tended to trust the low-trusting voices of a male-voiced AI teammate more. Characterized by lower formants, men rated these voices as more 'kind' and 'considerate' on the purpose-dimension of trust. These findings align with previous research highlighting gender differences in the association between acoustic cues and trust and cooperation. Knowles and Little (2016) demonstrated that lower formant measures, which are associated with more masculine features, were perceived as more cooperative in male voices. Similarly, Monano and colleagues (2017) found that male-voice with longer apparent vocal tracts (lower formants) was trusted more than those with shorter apparent vocal tracts (higher formants). This association can be explained by the fact that formant frequencies can influence perceptions of masculinity (Feinberg et al., 2005), which in turn may impact trust. However, previous findings often found people tend to trust people of the opposite gender more, as explained by the attractiveness of the voices and mate-related trustworthiness (O'Connor & Barclay, 2017; Slonim & Guillen, 2010). Our study, in the context of human-AI interaction, did not show this cross-gender effect, instead, showed a same-gender effect. The similarity-attraction theory, which shows that that people tend to place more trust in those who are similar to themselves, may be able to explain this phenomenon in human-AI interaction (Byrne, 1997). Further investigation is needed for the gender effects on trust in voice-based AI teammate.

Additionally, our previous studies showed that people conveyed their trust through a combination of lexical and acoustic cues (Li et al., 2023). Building upon these insights, the present study extended our understanding by showing that trust can not only be measured through conversation but can also be managed through conversation. By direct and precise manipulating the AI teammate's acoustic cues based on prior findings, we uncovered a noteworthy bidirectional effect: the way people express higher trust leads to higher investments in AI teammates. These findings provided implications on the voice-design of AI teammate, especially adopting specific

acoustic cues as interventions to manage trust in human-AI cooperation. By strategically incorporating these cues, researchers can enhance trust-building and repairing processes and foster more cooperation between humans and AI systems. However, this power should be used with caution. Even when the aim of managing trust to induce cooperation are beneficent, any move to undermine individuals merits careful discussion. Defining the extent of AI's decision-making authority in such matters is important. It is essential to strike a balance between empowering AI systems to make decisions that maximize benefits while considering ethical and legal issues.

In addition to the voice-based finding, we also found that purpose-based trust violations, where an AI teammate fails to cooperate with the team goal, lead to a more drastic drop in trust. This confirms with previous findings that purpose-based outweighs the performance-based trust violations (Li & Lee, in preparation). While prior literatures often focused on the supervisory control and dyad interactions, the assumption of the shared goal was never challenged (Li & Lee, 2022). With the increasingly computational power and connecting systems, the goal conflict between the AI's global optimum and individual's local optimum can happen. Our results highlight the importance of the aligned goal for designing a trustworthy AI teammate. Particularly in safety-critical domains such as healthcare, military, and emergency response, where system efficiency is crucial and tolerance for breakdowns is low, AI teammates must not only demonstrate competence and reliability but also exhibit alignment with the overall organizational goals.

Finally, we found that individual differences on the Honesty-Humility dimension, as assessed by the HEXACO Personality assessment, strongly predicted participants' subjective trust ratings in the performance dimension and their perceived cooperation of the AI teammate. Individuals with higher levels of dispositional Honesty-Humility rated the AI teammate as more capable and held higher expectations regarding the AI teammate's contributions to the team goal. Our results were in line with previous results in public goods game (Hilbig et al., 2012) (Li & Lee, in preparation). Because people with high dispositional Honesty Honesty–Humility have also shown to have stronger other-regarding preferences in their social value orientation and would generally hold more positive beliefs of what others would do (Krueger, 2008). Additionally, individuals with high in Honesty– Humility are less likely to condition their behavior on situational factors, suggesting a more consistent cooperative disposition (Hilbig et al., 2012). Future studies could study the power dynamics between human and AI teammates and possible mechanisms for situational factors (e.g., punishment and rewards) that may influence human-AI cooperation.

6.10 Chapter Summary

The goal of this chapter is to evaluate whether conversational indicators identified in Chapter 3 can be used as adaptive countermeasures by a virtual assistant to manage various dimensions of trust. This chapter presented two experimental studies. For Study 1, we investigated the effects of performance-based and purpose-based trust violations on people's trust levels in human-AI cooperation and to identify effective strategies for managing trust violations. Results revealed that purpose-based trust violations, where the AI teammate failed to cooperate with the team goal, led to a greater drop in trust compared to performance-based violations. We showed that an apology with an explanation was the most effective strategy for repairing trust after purpose-based trust violations. Additionally, we showed that individual difference on the Honesty-Humility dimension can predict people's trust, investment behaviors, and perceived cooperation of AI teammate. Our study emphasized the importance of addressing purpose-based trust violations and provided important implications for designing trustworthy agents. While our study focused on the micro-level of acoustic and lexical features in the human-AI conversations, the conversation is rich with other features, such as cadence, filler words and conversational turns, that merit further investigation.

For Study 2, we investigated the effects of trusting voices and its congruency effects with the trust repair content on managing trust. By manipulating the formants and variance in fundamental frequency to create high- and low-trusting voices, we provided empirical evidence supporting a positive congruency effect between acoustic and lexical cues of trust. This means that a high-trusting voice promotes people's trusting behaviors (greater investments) in the AI teammate. Additionally, men tended to trust low-trusting voices more, perceiving them as kind and considerate on the purpose-dimension of trust. Our study extended our understanding that trust can be both measured and managed through voices, emphasizing the importance of using voice design as interventions to manage -building and repairing processes in human-AI cooperation.

Chapter 7. General Discussion

7.1 Problem Summary

Artificial Intelligence (AI), with its increasing capability and connectivity, extends beyond limited and well-defined contexts and is integrated into broader societal domains. Examples include AI algorithms controlling large fleets of autonomous vehicles, news filtering algorithms influencing people's political belief and preferences, and algorithms mediating resource allocation and labor (Bubeck et al., 2023; Rahwan, 2018). The relationship between humans and AI has evolved from mere supervisory control to a interdependent cooperation on a larger scale, yielding significant societal benefits (Endsley et al., 2021). To better support the human-AI cooperation, establishing a trusting relationship between humans and their AI teammates becomes increasingly critical (Chiou & Lee, 2021). Trust plays a vital role in shaping how people use, communicate with, and cooperate with AI systems. Therefore, the measurement and management of trust in human-AI cooperation are essential to ensure the safety, effectiveness, and overall positive outcomes of such interactions. The objective of this dissertation is to measure and manage trust in human-AI conversations and cooperation, addressing three primary questions: (1) How can we measure people's trust in human-AI conversations? (2) How does trust change over time within human-AI conversations? (3) How can we effectively manage instances of overtrust or undertrust through conversational cues to enhance human-AI cooperation?

To tackle these questions, my dissertation considers two aims: measure trust in communication and manage trust in cooperation. Especially, I measure trust in communication with the considerations of the temporal dimension from the real-time measurement to long-term dynamics (Objective 1). Trust communication mediates cooperation. When considering trust management, I integrated the considerations of the structural dimension of team interdependence and goal alignment (Objective 2). From a temporal perspective, my research ranges from real-time trust measurement (Chapter 3) to long-term trust dynamics (Chapter 4 and Chapter 5). From a structural perspective, I investigate trust in when involving multiple goals between AI and humans in a team composition using the game-theoretic situations and investigated whether the identified trust indicators can be used to manage trust (Chapter 6). The following results support the objectives of the dissertation.

7.1.1 Objective 1: Measure Trust in Communication: From Real-Time Estimation to Long-Term Dynamics

Chapter 3 tackled the initial question regarding the measurement of trust in human-AI conversations. I showed that a random forest algorithm, trained using the combined lexical and

acoustic features, predicted trust in the conversational agent most accurately. The most important predictors were a combination of lexical and acoustic cues: average sentiment considering valence shifters, the mean of formants, and Mel frequency cepstral coefficients (MFCC). These conversational features were identified as partial mediators predicting people's trust. Precise estimation of the trust of the conversation requires lexical and acoustic cues. These results showed the possibility of using conversational data to measure trust and potentially other dynamic mental states, unobtrusively and dynamically. Chapter 4 and Chapter 5 both tackled the temporal dynamics of trust. Compared to using the individual differences to explain the diverging levels of trust over time, Chapter 4 showed that trust divergence can be better explained as an outcome of a dynamic system, which considers the interaction between reliability and exposure along with the individual by-reliability variability fit the data well. Additionally, results suggested that dynamic interactions with automation contribute to trust divergence. This chapter established a robust foundation for the temporal trust dynamic perspective in Chapter 5, where we further examined the temporal aspects in the human-AI conversations. Chapter 5 showed the evolution of trust dimensions throughout human-AI conversations, which reveals distinct patterns in conversational topic diversity and flow over time. Objective 1 identified the need for designs that prioritize state dependence and responsivity, where the automation should be able to responsive to the level of trust for the trust management for the Objective 2.

7.1.1 Objective 2: Manage Trust in Cooperation: From Performance to Purpose-based Trust.

Chapter 6 of the dissertation focused on managing trust for effective cooperation, with a particular emphasis on automation responsivity. The research expanded beyond performance-based calibration to examine purpose-based cooperation. A game-theoretic framework was designed to investigate the impacts of performance- and purpose-based trust violations on trust and cooperative behaviors. Results showed that purpose-based trust violations, where the AI teammate failed to cooperate with the team goal, led to a greater drop in trust compared to performance-based violations. Additionally, the identified conversational trust indicators in Chapter 3 were demonstrated to be countermeasures to repair trust in human-AI cooperation. By directly manipulating the formants and variance in fundamental frequency to create high- and low-trusting voices, I showed that a high-trusting voice promotes people's trusting behaviors (greater investments) in the AI teammate. These findings contribute to our understanding that trust can be both measured and managed through human-AI communications, which can be served as an unobtrusive, real-time means of trust measurement and management in human-AI cooperation.

The results also shed light on the ethical challenges associated with managing trust when the goal is to induce behaviors that may contradict individuals' immediate interests, such as cooperation. Future research should address the role of AI teammate responsibility in balancing the trade-offs between individual and societal benefits across different contexts. Understanding how AI systems can navigate these trade-offs is crucial for promoting ethical and beneficial human-AI cooperation.

7.2 Contributions

7.2.1 Theoretical Contributions

The main theoretical contribution of this dissertation is demonstrating that trust can be measured and managed in human-AI conversations, which enriches the 'semiotics' aspect of the trust framework proposed by Chiou and Lee (2021). My work demonstrated how trust is signaled during interactions, including both micro features (i.e., acoustic and lexical cues) and macro topics (i.e., conversation topics) (Li et al., 2020). Additionally, I have established the bidirectional nature of trust in the human-AI conversations. This means that the trust signals expressed by individuals (as trustors, conveying their trust in automation) can also be used as signals by the AI trustee to manage trust. Overall, this work demonstrated how trust is signaled and perceived in human-AI conversations from a relational approach.

Additionally, this dissertation expands the understanding of trust and human-AI relationships in two directions: temporal dynamics and structural interdependence. By adopting a dynamic system perspective, I demonstrated the temporal evaluation of trust processes in communications, providing deeper insights into how trust evolves and reflects people's analytic and affective trust in conversations. From a structural perspective, the dissertation extends the concept of trust beyond performance to purpose-based interaction, recognizing the increasing prevalence of conflicting interests and values among various stakeholders.

7.2.2 Practical Contributions

This dissertation provides practical principles and illuminating design factors for designing trust-adaptive conversational agent that can measure and manage trust. I highlighted the key indicators of people's trust in the conversations, including both lexical and acoustic cues, which can be integrated in the agents' sensors and algorithms to predict real-time trust. Moreover, I showed that these identified features can be used to repair trust. My work suggests that an adaptive trust management system can be developed to calibrate people's trust: using the trust indicators to dampen trust when over-trust and repair trust when under-trust.

Furthermore, my machine learning pipeline on trust estimation also provides methodological implications in measuring latent variables, such as trust, workload, and situational awareness. Measuring these subjective and latent concepts in communications or other continuous data streams can provide a real-time and non-intrusive approach.

In addition to trust measurement, the game-theoretic situation designed in this dissertation can serve as a valuable testbed for understanding and accessing both performance- and purpose-based trust interactions. Insights derived from this game setting hold significant applicability for human-AI teams in future hybrid societies, particularly when navigating complex conflicts between individual and collective benefits.

Lastly, the findings on trust management highlight the ethical challenges when the objective is to induce behaviors that may go against individuals' immediate interests, specifically in the context of large-scale human-AI cooperation in the future hybrid society. The voice design and other social norm, affective-based strategy can implicitly influence people's behaviors without explicit acknowledgement. This raises important questions about the ethical considerations and guidelines to govern AI's role in influencing human behavior. Determining the boundaries and extent of decision-making authority entrusted to AI systems in such situations becomes crucial. It is necessary to ensure that any interventions or manipulations carried out by AI systems respect individual autonomy, privacy, and overall well-being.

7.3 Future Research

These contributions provide a foundation for measuring trust signals in human-AI conversations from a temporal dynamics perspective and managing people's trust in cooperation from a structural dependence perspective. There are still several directions for future research that can build upon both temporal dynamics and structural dependence.

7.3.1 Temporal Dimension

From a temporal perspective, my dissertation ranges from real-time trust measurement to long-term trust dynamics and shows empirical evidence of modeling trust in human-AI conversations as a dynamic system. More in-depth understanding and modeling of dynamic system are needed. This means incorporating a temporal element to understand system behaviors in human-AI interaction. While Yang and colleagues (2023) have made significant contributions by defining and computationally modeling three properties of trust dynamics—continuity, negative bias, and stabilization— the focus has primarily been on the "temporal" dimension. This leaves out the core element, 'system behaviors' of dynamic system thinking. Gorman and colleagues

(2017) have argued that behaviors that emerge at the system level may be encoded differently or absent at the individual level, highlighting the significant influence of team processes on individual thoughts and behaviors. The concepts of attractors, perturbation, synchronization, and fractal (power-law) concepts were introduced in the context of the team to further support this notion. For instance, an attractor is "a behavior that a system settles on over time after possibly displaying initial transient behaviors" (Abraham & Shaw, 1992; Gorman et al., 2017). The attractor field theory, which depicted effective energy landscapes of various attractors, provides a good theoretical perspective. The interaction design can form various attractors where people tend to gravitate towards despite a wide variety of starting conditions. This concept can be used to explain trust divergence, where individual differences mark various starting points and gravitate toward the attractors, indicating high or low trust state. Adopting this perspective can better identify trusting and distrusting individuals from a system perspective. Trusting individuals search for the evidence to trust and rely on the automation, which would gravitating towards the trusting attractors; the distrusting individuals search for the opposite evidence and gravitate towards distrust attractors. It is essential to develop more computational models to verify the predictability and generalizability of these dynamic system concepts within the realm of human-AI teams.

Additionally, this work has identified the conflicts in the social dilemma from a local/global perspective, i.e., when personal goals are not aligned with the societal one. However, the conflict can also happen on the temporal dimension. This means that short-term interests often take precedence over long-term goals, even when the latter offer greater benefits, such as climate changes and voluntary vaccination. When it comes to human-AI interaction, individuals often exhibit shortsightedness and struggle to predict the long-term consequences of their immediate actions. In contrast, AI agents are capable of computationally optimizing goals irrespective of the timescale involved. To further complicates the problem, in the reinforcement context, the reward function of the AI may lead a seemingly reasonable, but incompatible, reward function achieved (Hadfield-Menell et al., 2016): if we reward the action of cleaning up dirt, the optimal policy causes the robot to repeatedly dump and clean up the same dirt. This is called alignment problem (S. Russell, 2019), which should be exanimated in the temporal dimension as well. Domingos and colleagues (2020) demonstrated that timing uncertainty not only promotes early generosity but also leads to polarized outcomes, where participants' contributions are distributed unevenly. However, it is important to note that these findings are limited to interpersonal cooperation. Given that people generally hold a negative bias towards AI agents (Domingos et al., 2021; You et al., 2011),

there is a potential for exploitation and misalignment of trust in human-AI interactions. Therefore, it is crucial to further investigate the effects of temporal conflicts in such hybrid interactions.

In our work, we demonstrated the possibility of measuring and modeling trust in human-AI conversations, it would be more intriguing to explore how timing uncertainty shapes the communication regarding people's trust in AI teammate. Our work was limited to decision-tree-based conversations, which restricted the range of topics that can be discussed. With recent advancements in large language models, the possibility of AI agents engaging in negotiations with human players becomes feasible. This opens opportunities to analyze conversations in terms of people's trust and decision-making processes.

7.3.2 Structural Dimension

From a structural standpoint, my dissertation expands the scope of human-AI relationships beyond supervisory control to peer-to-peer cooperation, with a specific focus on purpose-related interactions. Results showed that violations of team goals within human-AI cooperation had a detrimental impact on trust. However, there is a gap in understanding the cognitive processes underlying such situations. Therefore, an area for future research is the development of a computational model that simulates the cognitive processes in human-AI cooperation. One promising approach Theory of Mind (ToM). ToM refers to the ability to infer and understand the mental states of others, allowing for the inference of goals in AI agents (Byom & Mutlu, 2013). Developing an adaptive AI teammate that incorporates ToM-based goal inference would be extremely valuable. By simulating human cognitive processes and enabling ToM, the AI teammate can adopt the perspective of the human counterpart and make inferences about their goals. This, in turn, enables the AI teammate to adapt its strategy accordingly and potentially promote the team outcomes.

To gain a comprehensive understanding of human-AI relationships, the scale of inquiry should be considered at three levels: individual, collective, and hybrid (Rahwan et al., 2019). At the individual level, the focus is on interactions between a single human and a single automation system. This level considers the static algorithm or characteristics of the automation. The collective level focuses on human-swarm interactions, where the emergence of group behaviors becomes a key point of examination. Understanding how behaviors and dynamics manifest at the group level is important. The hybrid level expands the scope to the bidirectional influences at a societal level, which shows how AI is shaping and shaped by competing interests of different stakeholders (Rahwan, 2018). While the present work emphasizes the potential conflicts between individual and

collective benefits within a human-AI dyad, for the future research, it is important to recognize that AI systems generalizable functions with broader societal impacts on the hybrid level. The question that remains is: How can we effectively quantify the trade-offs and reconcile the differences when an AI system encounters conflicting preferences?

To address the question proposed above, it is crucial to consider ethical and legal issues when designing AI agents. The question arises as to whether AI should be given the power to make trade-offs and reconcile differences in various contexts. Even when the aim of managing trust to induce cooperation are beneficent, any move to undermine individuals merits careful discussion. Defining the extent of AI's decision-making authority in such matters is important. Explicit governance measures, such as auditing algorithms and undergoing institutional review, are clear steps towards ethical practices. However, determining the boundaries of implicit forces becomes more challenging. For instance, as demonstrated in this dissertation, the voice design can significantly impact human perception and decision-making. Obtaining consent for technologies based on nudges, social norms, and affect-based designs becomes complex and requires careful consideration. It is essential to strike a balance between empowering AI systems to make decisions that maximize benefits while considering ethical and legal issues. Addressing these challenges requires ongoing discussions, interdisciplinary collaborations, and the involvement of various stakeholders, including researchers, policymakers, ethicists, and the general public.

Bibliography

Abraham, R., & Shaw, C. (1992). *Dynamics: The geometry of behavior. 2nd edn. Redwood City*. CA: Addison-Wesley.

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk Research: Review and Recommendations. *Journal of Management*, *47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Alarcon, G. M., Gibson, A. M., & Jessup, S. A. (2020). Trust Repair in Performance, Process, and Purpose Factors of Human-Robot Ttust. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209453

Alsaid, A., Li, M., Chiou, E. K., & Lee, J. (2022). *Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires*.

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic Games on the Internet: The Effect of $1 Stakes. *PLOS ONE*, *7*(2), e31461. https://doi.org/10.1371/journal.pone.0031461

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., & Shaffer, D. (2015). Look together: Analyzing gaze coordination with epistemic network analysis. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01016

Antos, D., Melo, C. de, Gratch, J., & Grosz, B. (2011). The Influence of Emotion Expression on Perceptions of Trustworthiness in Negotiation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *25*(1), Article 1. https://doi.org/10.1609/aaai.v25i1.7939

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A Review of Research and Theory—2014. *Personality and Social Psychology Review*, *18*(2), 139–152.

Bailey, N. R., & Scerbo, M. W. (2007). Automation-Induced Complacency for Monitoring Highly Reliable Systems: The Role of Task Complexity, System Experience, and Operator Trust. *Theoretical Issues in Ergonomics Science*.

Basili, M., Muscillo, A., & Pin, P. (2022). No-vaxxers are different in public good games. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-22390-y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4* (arXiv:1406.5823). arXiv. https://doi.org/10.48550/arXiv.1406.5823

Bates, G. E., & Neyman, J. (1952). Contributions to the Theory of Accident Proneness. *University of California Press*.

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLOS ONE*, *12*(10), e0185651. https://doi.org/10.1371/journal.pone.0185651

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the driver-automation interaction: An approach using automation uncertainty. *Human Factors*, *55*(6), 1130–1141. https://doi.org/10.1177/0018720813482327

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Bhat, S., Lyons, J. B., Shi, C., & Yang, X. J. (2022). Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task. *IEEE Robotics and Automation Letters*, *7*(4), 8815–8822.

Bhimavarapu, J. P., Sarvana, K., Achanta, V. K. S., Kadiyala, C., & Bhimavarapu, J. P. (2021). Modelling of emotion recognition system from speech using MFCC features. *AIP Conference Proceedings*, *2375*(October). https://doi.org/10.1063/5.0066503

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America*, *132*(2), 1100–1112.

Boersma, P., & Antos, D. (2012). *Praat, version 5.5*.

Bombien, L., Winkelmann, R., & Scheffers, M. (2021). *wrassp: An R wrapper to the ASSP Library*. R Package Version 1.0.1.

Brohinsky, J., Marquart, C., Wang, J., Ruis, A. R., & Shaffer, D. W. (2021). Trajectories in Epistemic Network Analysis. In A. R. Ruis & S. B. Lee (Eds.), *Advances in Quantitative Ethnography* (Vol. 1312, pp. 106–121). Springer International Publishing. https://doi.org/10.1007/978-3-030-67788-6_8

Bromiley, P., & Cummings, L. L. (1995). *Transaction costs in organisations with trust. Research on negotiation in organizations*. Brenwich, CT: JAI Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Byom, L. J., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, *7*(JUL), 1–12. https://doi.org/10.3389/fnhum.2013.00413

Byrne, D. (1997). An Overview (and Underview) of Research and Theory within the Attraction Paradigm. *Journal of Social and Personal Relationships*, *14*(3), 417–431. https://doi.org/10.1177/0265407597143008

Cartei, V., Cowles, H. W., & Reby, D. (2012). Spontaneous voice gender imitation abilities in adult speakers. *PloS One*, *7*(2).

Chiou, E. K., & Lee, J. D. (2016). Cooperation in Human-Agent Systems to Support Resilience: A Microworld Experiment. *Human Factors*, *58*(6), 846–863. https://doi.org/10.1177/0018720816649094

Chiou, E. K., & Lee, J. D. (2021). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *65*(1), 137–165.

Clifton. (2020). Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda. *Center on Long-Term Risk*.

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–285. https://doi.org/10.1111/cogs.12009

Corretge, R. (2023). *Praat Vocal Toolkit*. https://www.praatvocaltoolkit.com

Courtright, S. H., Thurgood, G. R., Stewart, G. L., & Pierotti, A. J. (2015). Structural interdependence in teams: An integrative framework and meta-analysis. *Journal of Applied Psychology*, *100*(6), 1825–1846. https://doi.org/10.1037/apl0000027

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260–281. https://doi.org/10.1016/S0899-8256(03)00119-2

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, *9*(1), 233. https://doi.org/10.1038/s41467-017-02597-8

Critch, A. (2017). *Toward negotiable reinforcement learning: Shifting priorities in Pareto optimal sequential decision-making*.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A Design Methodology for Trust Cue Calibration in Cognitive Agents. In R. Shumaker & S. Lackey (Eds.), *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments* (Vol. 8525, pp. 251–262). Springer International Publishing. https://doi.org/10.1007/978-3-319-07458-0_24

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478.

Demir, M., Mcneese, N. J., Gorman, J. C., & Cooke, N. J. (2021). *Exploration of Team Trust and Interaction Dynamics in Human-Autonomy Teaming. February*. https://doi.org/10.13140/RG.2.2.32213.55528

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H. (2012). Effects of changing reliability on trust of robot systems. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12*, 73. https://doi.org/10.1145/2157689.2157702

Domingos, E. F., Grujić, J., Burguillo, J. C., Kirchsteiger, G., Santos, F. C., & Lenaerts, T. (2020). Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization. *Iscience*, *23*(12), 101752.

Domingos, E. F., Terrucha, I., Suchon, R., Grujić, J., Burguillo, J. C., Santos, F. C., & Lenaerts, T. (2021). Delegation to autonomous agents promotes cooperation in collective-risk dilemmas. *ArXiv:2103.07710 [Cs]*. http://arxiv.org/abs/2103.07710

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442. https://doi.org/10.1016/j.trc.2019.05.025

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and Believing: The Influence of Emotion on Trust. *Journal of Personality and Social Psychology*, *88*, 736–748. https://doi.org/10.1037/0022-3514.88.5.736

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003a). The role of trust in automation reliance. *International Journal of Human Computer Studies*, *58*(6), 697–718.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003b). The Role of Trust in Automation Reliance. *International Journal of Human Computer Studies*, *58*(6), 697–718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, *13*(3), 147–164.

Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents. *Group Decision and Negotiation*, *22*(5), 897–913.

Endsley, M. R., Caldwell, B., Chiou, K. E., Cooke, J, N., Cummings, L. M., Gonzalez, C., Lee, D. J., Mcneese, J. N., Miller, C., Roth, E., Rouse, B. W., & Talmage, D. (2021). *Human-AI Teaming: State-of-the-Art and Research Needs* (Issue December). Washington, DC: The National Academies Press.

Esterwood, C., & Robert, L. (2023). Three Strikes and You are Out!: The Impacts of Multiple Human-Robot Trust Violations and Repairs on Robot Trustworthiness. *Computers in Human Behavior*, *142*(107658). https://doi.org/10.7302/6774

Falcone, R., & Castelfranchi, C. (2004). Trust dynamics: How trust is influenced by direct experiences and by trust itself. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, 740–747.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, *69*(3), 561–568. https://doi.org/10.1016/j.anbehav.2004.06.012

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Fuoli, M., & Paradis, C. (2014). A model of trust-repair discourse. *Journal of Pragmatics*, *74*, 52–69.

Gao, J., & Lee, J. D. (2006). Extending the Decision Field Theory to Model Operators' Reliance on Automation in Supervisory Control Situations. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, *36*(5), 943–959.

Gauder, L., Pepino, L., Riera, P., Brussino, S., Vidal, J., Gravano, A., & Ferrer, L. (2021). A Study on the manifestation of trust in speech. *ArXiv Preprint*, 1–31.

Gobo, G. (2011). Back to Likert: Towards the Conversational Survey. In M. Williams & W. Vogt, *The SAGE Handbook of Innovation in Social Research Methods* (pp. 228–248). SAGE Publications Ltd. https://doi.org/10.4135/9781446268261.n15

Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, *0*. https://doi.org/10.2478/jagi-2014-0001

Gorman, J. C., Dunbar, T. A., Grimm, D., & Gipson, C. L. (2017). Understanding and Modeling Teams As Dynamical Systems. *Frontiers in Psychology*, *8*.

Goudbeek, M., Goldman, J. P., & Scherer, K. R. (2009). Emotion dimensions and formant position. *Interspeech2009*, 3–6.

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *R Journal*, *9*(1), 421–436. https://doi.org/10.32614/rj-2017-016

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative Inverse Reinforcement Learning. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html

Heckman, J. J. (1981). Heterogeneity and State Dependence. In *Studies in Labor Markets* (pp. 91–140). University of Chicago Press.

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional honesty–humility. *European Journal of Personality*, *26*(3), 245–254.

Hildebrand, C., & Bergner, A. (2021). Conversational robo advisors as surrogates of trust: Onboarding experience, firm perception, and consumer financial decision making. *Journal of the Academy of Marketing Science*, *49*(4), 659–676. https://doi.org/10.1007/s11747-020-00753-z

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Izygon, M., Kortenkamp, D., & Molin, A. (2008). A procedure integrated development environment for future spacecraft and habitats. *In Proceedings of the Space Technology and Applications International Forum (STAIF 2008)*, 969.

Jensen, T., & Khan, M. M. H. (2022). I'm Only Human: The Effects of Trust Dampening by Anthropomorphic Agents. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence* (pp. 285–306). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21707-4_21

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, *3*(1), 43. https://doi.org/10.5898/jhri.3.1.johnson

Johnson, M., & Vera, A. H. (2019). No Ai is an island: The case for teaming intelligence. *AI Magazine*, *40*(1), 16–28. https://doi.org/10.1609/aimag.v40i1.2842

Jones, S. L., & Shah, P. P. (2016). Diagnosing the locus of trust: A temporal perspective for trustor, trustee, and dyadic influences on perceived trustworthiness. *Journal of Applied Psychology*, *101*, 392–414. https://doi.org/10.1037/apl0000041

Kamaraj, A. V., Lee, J., Parker, J., & Domeyer, J. E. (2023). Bimodal Trust: Relationship Between Drivers' Trust in Reliable Automation and Response to a Surprise Automation Error. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*.

Kaniarasu, P., & Steinfeld, A. M. (2014). Effects of blame on trust in human robot interaction. *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 850–855.

Kaplan, A. D., Kessler, T. T., Sanders, T. L., Cruit, J., Brill, J. C., & Hancock, P. A. (2021). Chapter 6 - A time to trust: Trust as a function of time in human-robot interaction. In C. S. Nam & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 143–157). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00006-X

Kim, J. C., Rao, H., & Clements, M. A. (2011). Investigating the use of formant based features for detection of affective dimensions in speech. *In International Conference on Affective Computing and Intelligent Interaction*, 369–377.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, *99*(1), 49–65. https://doi.org/10.1016/j.obhdp.2005.07.002

Knowles, K. K., & Little, A. C. (2016). Vocal fundamental and formant frequencies affect perceptions of speaker cooperativeness. *Quarterly Journal of Experimental Psychology*, *69*(9), 1657–1675. https://doi.org/10.1080/17470218.2015.1091484

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, *12*, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Korsgaard, M. A., Kautz, J., Bliese, P., Samson, K., & Kostyszyn, P. (2018a). Conceptualising time as a level of analysis: New directions in the analysis of trust dynamics. *Journal of Trust Research*, *8*(2), 142–165. https://doi.org/10.1080/21515581.2018.1516557

Korsgaard, M. A., Kautz, J., Bliese, P., Samson, K., & Kostyszyn, P. (2018b). Conceptualising time as a level of analysis: New directions in the analysis of trust dynamics. *Journal of Trust Research*, *8*(2), 142–165.

Kramer, M. W. (1999). Motivation to Reduce Uncertainty: A Reconceptualization of Uncertainty Reduction Theory. *Management Communication Quarterly*, *13*(2), 305–316. https://doi.org/10.1177/0893318999132007

Kramer, R. M., & Lewicki, R. J. (2010). Repairing and enhancing trust: Approaches to reducing organizational trust deficits. *The Academy of Management Annals*, *4*, 245–277. https://doi.org/10.1080/19416520.2010.487403

Krausman, A., Neubauer, C., Forster, D., Lakhmani, S., Baker, A. L., Fitzhugh, S. M., Gremillion, G., Wright, J. L., Metcalfe, J. S., & Schaefer, K. E. (2022). Trust Measurement in Human-Autonomy Teams: Development of a Conceptual Toolkit. *ACM Transactions on Human-Robot Interaction*, *11*(3), 1–58. https://doi.org/10.1145/3530874

Krueger, J. I. (2008). From social projection to social behaviour. *European Review of Social Psychology*, *18*(1), 1–35.

Krumhuber, E. G., Hyniewska, S., & Orlowska, A. (2023). Contextual effects on smile perception and recognition memory. *Current Psychology*, *42*(8), 6077–6085. https://doi.org/10.1007/s12144-021-01910-5

Lalitha, S., Geyasruti, D., Narayanan, R., & Shravani, M. (2015). Emotion detection using MFCC and cepstrum features. *Procedia Computer Science*, 29–35.

Larrimore, L., Jiang, C., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to Peer Lending: The Relationship Between Language Features, Trustworthiness, and Persuasion Success.

Journal of Applied Communication Research, *39*. https://doi.org/10.1080/00909882.2010.536844

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80.

Lewicki, R. J., Bunker, B. B., & others. (1996). Developing and maintaining trust in work relationships. *Trust in Organizations: Frontiers of Theory and Research*, *114*, 139.

Li, M., Alsaid, A., Noejovich, S. I., Cross, E. V., & Lee, J. D. (2020). Towards a conversational measure of trust. *AAAI Fall Symposium FSS-20 / SSS-20*, 1–6.

Li, M., Erickson, I., Cross, E., & Lee, J. (2022, October 17). Estimating trust in conversational agent with lexical and acoustic features. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Li, M., Erickson, I. M., Cross, E. V., & Lee, J. D. (2023). It's Not Only What You Say, But Also How You Say It: Machine Learning Approach to Estimate Trust from Conversation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.

Li, M., & Lee, J. D. (2022). Modeling Goal Alignment in Human-AI Teaming: A Dynamic Game Theory Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1538–1542.

Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE TRANSACTIONS ON HUMiAN FACTORS IN ELECTRONICS*, *1*, 4–11.

Liu, J., Akash, K., Misu, T., & Wu, X. (2021). Clustering human trust dynamics for customized real-time prediction. *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1705–1712.

Luo, R., Du, N., & Yang, X. J. (2022). Evaluating Effects of Enhanced Autonomy Transparency on Trust, Dependence, and Human-Autonomy Team Performance over Time. *International Journal of Human–Computer Interaction*, *38*(18–20), 1962–1971. https://doi.org/10.1080/10447318.2022.2097602

Maier, N. R. F. (1967). Assets and Liability in Group Problem Solving: The Need for an Integrative Function. *Psychological Review*, *74*(4), 603–604. https://doi.org/10.1093/mind/xxii.10.603

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, *34*(6), 419–426. https://doi.org/10.1016/j.evolhumbehav.2013.08.001

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

March, C. (2019). *The Behavioral Economics of Artificial Intelligence: Lessons from Experiments with Computer Players* (SSRN Scholarly Paper No. 3485475). https://doi.org/10.2139/ssrn.3485475

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. In *Source: The Academy of Management Review* (Vol. 20, Issue 3, pp. 709–734).

Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, *26*(1), 283–327.

McDonald, A. D., Ade, N., & Peres, S. C. (2020). Predicting procedure step performance from operator and text features: A critical first step toward machine learning-driven procedure design. *Human Factors*, *00*(0), 1–17.

McDonald, A. D., Ferris, T. K., & Wiener, T. A. (2020). Classification of driver distraction: A comprehensive analysis of feature generation, machine learning, and input measures. *Human Factors*, *62*(6), 1019–1035.

Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors*, *53*(4), 356–370.

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019). Automation-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers in Psychology*, *10*, 225.

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210.

Midden, C. J. H., & Huijts, N. M. A. (2009). The Role of Trust in the Affective Evaluation of Novel Risks: The Case of CO2 Storage. *Risk Analysis*, *29*(5), 743–751. https://doi.org/10.1111/j.1539-6924.2009.01201.x

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *56th Annual Meeting of the Association for Computational Linguistics*, 174–184.

Montano, K. J., Tigue, C. C., Isenstein, S. G. E., Barclay, P., & Feinberg, D. R. (2017). Men's voice pitch influences women's trusting behavior. *Evolution and Human Behavior*, *38*(3), 293–297. https://doi.org/10.1016/j.evolhumbehav.2016.10.010

Mui, P. H. C., Gan, Y., Goudbeek, M. B., & Swerts, M. G. J. (2020). Contextualising Smiles: Is Perception of Smile Genuineness Influenced by Situation and Culture? *Perception*, *49*(3), 357–366. https://doi.org/10.1177/0301006620904510

Nakahara, H., & Doya, K. (1998). Near-Saddle-Node Bifurcation Behavior as Dynamics in Working Memory for Goal-Directed Behavior. *Neural Computation*, *10*(1), 113–132.

Nalini, N. J., Palanivel, S., & Balasubramanian, M. (2013). Speech emotion recognition using residual phase and MFCC features. *International Journal of Engineering and Technology*, *5*(6), 4515–4527.

Nielsen, F. Å. (2011). A new evaluation of a word list for sentiment analysis in microblogs. *ESWC2011 Workshop on "Making Sense of Microposts": Big Things Come in Small Packages*, 93–98.

Norman, S. M., Avolio, B. J., & Luthans, F. (2010). The impact of positivity and transparency on trust in leaders and their perceived effectiveness. *The Leadership Quarterly*, *21*(3), 350–364. https://doi.org/10.1016/j.leaqua.2010.03.002

O'Connor, J. J. M., & Barclay, P. (2017). The influence of voice pitch on perceptions of trustworthiness across social contexts. *Evolution and Human Behavior*, *38*(4), 506–512. https://doi.org/10.1016/j.evolhumbehav.2017.03.001

Okamura, K., & Yamada, S. (2020). Empirical Evaluations of Framework for Adaptive Trust Calibration in Human-AI Cooperation. *IEEE Access*, *8*, 220335–220351. https://doi.org/10.1109/ACCESS.2020.3042556

Oktay, J. S. (2012). *Grounded theory*. Oxford University Press.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Parasuraman, R. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced "Complacency." *The International Journal of Aviation Psychology*.

Perkins, R., Khavas, Z. R., McCallum, K., Kotturu, M. R., & Robinette, P. (2022). The Reason for an Apology Matters for Robot Trust Repair. In F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, & S. S. Ge (Eds.), *Social Robotics* (pp. 640–651). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-24670-8_56

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences*, *115*(15), 3972–3977. https://doi.org/10.1073/pnas.1716090115

R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing*.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology Volume*, *20*(1), 5–14.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy,' … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), Article 7753. https://doi.org/10.1038/s41586-019-1138-y

Razin, Y. S., & Feigh, K. M. (2021). Committing to interdependence: Implications from game theory for human–robot trust. *Paladyn, Journal of Behavioral Robotics*, *12*(1), 481–502. https://doi.org/10.1515/pjbr-2021-0031

Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *International Journal of Human–Computer Interaction*, *37*(1), 81–96. https://doi.org/10.1080/10447318.2020.1807710

Rinker, T. (2017). *Package 'sentimentr'*.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing Is Key for Robot Trust Repair. In A. Tapus, E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social Robotics* (Vol. 9388, pp. 574–583). Springer International Publishing. https://doi.org/10.1007/978-3-319-25554-5_57

Rolls, E. T. (2010). Attractor networks. *WIREs Cognitive Science*, *1*(1), 119–134. https://doi.org/10.1002/wcs.1

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Rychlowska, M., Van Der Schalk, J., Niedenthal, P., Martin, J., Carpenter, S. M., & Manstead, A. S. R. (2021). Dominance, reward, and affiliation smiles modulate the meaning of uncooperative or untrustworthy behaviour. *Cognition and Emotion*, *35*(7), 1281–1301. https://doi.org/10.1080/02699931.2021.1948391

Schreckenghost, D., Milam, T., & Billman, D. (2014). Human performance with procedure automation to manage spacecraft systems. *In Proceedings of the 35th International Conference for Aerospace Experts, Academics, Military Personnel, and Industry Leaders*, 1–16.

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1–19. https://doi.org/10.1016/j.obhdp.2006.05.005

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *The American Statistician*, *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031

Sebe, N., Cohen, I., & Huang, T. S. (2005). Multimodal emotion recognition. In *In Handbook of Pattern Recognition and Computer Vision* (pp. 387–409).

Shaffer, D. W. (2017). *Quantitative ethnography*. Lulu. com.

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, *3*(3), 9–45.

Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press.

Slonim, R., & Guillen, P. (2010). Gender selection discrimination: Evidence from a Trust game. *Journal of Economic Behavior & Organization*, *76*(2), 385–405. https://doi.org/10.1016/j.jebo.2010.06.016

Soares, N., & Fallenstein, B. (2014). *Aligning Superintelligence with Human Interests: A Technical Research Agenda Highly Reliable Agent Designs*. 1–14.

Spitzley, L. A., Wang, X., Chen, X., Pentland, S. J., Nunamaker, J. F., Burgoon, J. K., & Dunbar, N. E. (2022). Non-Invasive Measurement of Trust in Group Interactions. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2022.3160132

Stewart, K. J. (2003). Trust Transfer on the World Wide Web. *Organization Science*, *14*(1), 5–17.

Syed, M. S. S., Pirogova, E., & Lech, M. (2021). Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach. *Electronics*, *10*(11), 1259. https://doi.org/10.3390/electronics10111259

Tan, S. C., Wang, X., & Li, L. (2022). The Development Trajectory of Shared Epistemic Agency in Online Collaborative Learning: A Study Combing Network Analysis and Sequential Analysis. *Journal of Educational Computing Research*, *59*(8), 1655–1681.

Tavoni, A., Dannenberg, A., Kallis, G., & Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, *108*(29), 11825–11829. https://doi.org/10.1073/pnas.1102493108

Thibaut, J. W., & Kelley, H. H. (1959). *The social psychology of groups*. New Brunswick. NJ: Transaction Publishers.

Toma, C. L., & Hancock, J. T. (2012). What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles. *Journal of Communication*, *62*(1), 78–97. https://doi.org/10.1111/j.1460-2466.2011.01619.x

Torre, I., Goslin, J., & White, L. (2020). If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior*, *105*, 106215. https://doi.org/10.1016/j.chb.2019.106215

Torre, I., Goslin, J., White, L., & Zanatto, D. (2018). Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. *Proceedings of the Technology, Mind, and Society*, 1–6. https://doi.org/10.1145/3183654.3183691

Trafton, J. G., Schultz, A. C., Cassimatis, N. L., Hiatt, L. M., Perzanowski, D., Brock, D. P., Bugajska, M. D., & Adams, W. (2006). Robotic Agents. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, 252–278.

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). *Is Infant-Directed Speech Prosody a Result of the Vocal Expression of Emotion?* https://journals.sagepub.com/doi/abs/10.1111/1467-9280.00240

Ullman, D., & Malle, B. F. (2019). Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 618–619. https://doi.org/10.1109/HRI.2019.8673154

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2010). Chapter 2 - An Interpersonal Approach to Emotion in Social Decision Making: The Emotions as Social Information Model. In *Advances in Experimental Social Psychology* (Vol. 42, pp. 45–96). Academic Press. https://doi.org/10.1016/S0065-2601(10)42002-X

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007

Voicemaker. (2023). https://voicemaker.in/

Waber, B., Williams, M., Carroll, J., & Pentland, A. (2015). A voice is worth a thousand words: The implications of the micro-coding of social signals in speech for trust research. In *Handbook of Research Methods on Trust: Second Edition* (pp. 302–312).

Wageman, R. (2001). The Meaning of Interdependence. In *Groups at Work*. Psychology Press.

Weiler, D. T., Lingg, A. J., Eagan, B. R., Shaffer, D. W., & Werner, N. E. (2022). Quantifying the qualitative: Exploring epistemic network analysis as a method to study work system interactions. *Ergonomics*, *65*(10), 1434–1449. https://doi.org/10.1080/00140139.2022.2051609

Whiting, T., Gautam, A., Tye, J., Simmons, M., Henstrom, J., Oudah, M., & Crandall, J. W. (2021). Confronting barriers to human-robot cooperation: Balancing efficiency and risk in machine behavior. *IScience*, *24*(1), 101963. https://doi.org/10.1016/j.isci.2020.101963

Wintersberger, P. (2020). *Automated Driving: Towards Trustworthy and Safe Human-Machine Cooperation*.

Wooldridge, A. R., Carayon, P., Shaffer, D. W., & Eagan, B. (2018). Quantifying the qualitative with epistemic network analysis: A human factors case study of task-allocation communication in a primary care team. *IISE Transactions on Healthcare Systems Engineering*, *8*(1), 72–82. https://doi.org/10.1080/24725579.2017.1418769

Yang, X. J., Guo, Y., & Schemanske, C. (2023). From Trust to Trust Dynamics: Combining Empirical and Computational Approaches to Model and Predict Trust Dynamics In Human-Autonomy Interaction. In V. G. Duffy, S. J. Landry, J. D. Lee, & N. Stanton (Eds.), *Human-Automation Interaction: Transportation* (pp. 253–265). Springer International Publishing.

Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872082110347. https://doi.org/10.1177/00187208211034716

Yang, X. Jessie., Christopher, S., & Christine, S. (2021). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, *00*(0), 1–17.

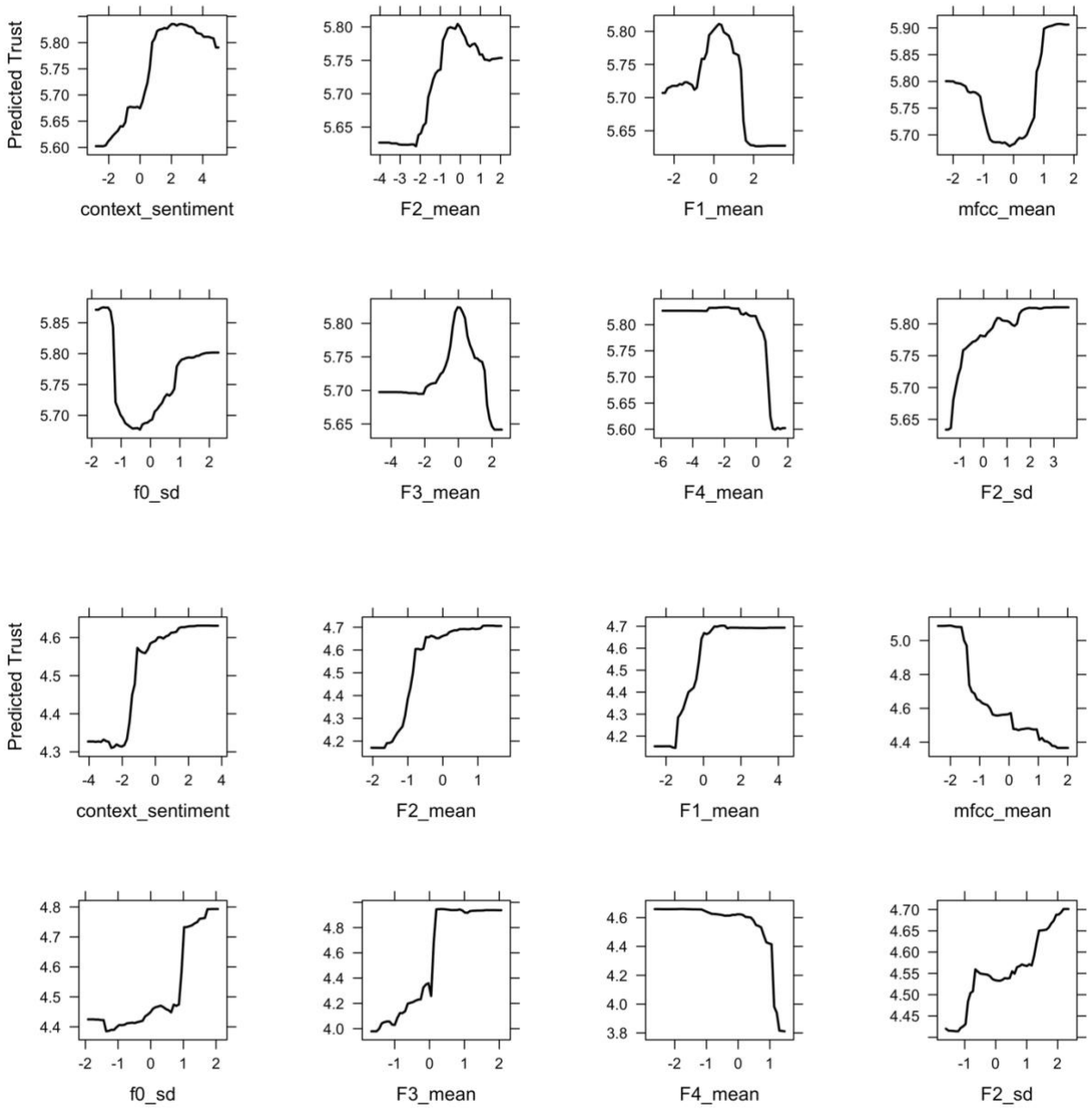
You, S., Nie, J., Suh, K., & Sundar, S. S. (2011). When the robot criticizes you... Self-serving bias in human-robot interaction. *Proceedings of the 6th International Conference on Human-Robot Interaction*, 295–296.

Yu, Q., & Li, B. (2017). mma: An R Package for Mediation Analysis with Multiple Mediators. *Journal of Open Research Software*, *5*(1), Article 1. https://doi.org/10.5334/jors.160

Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, *39*(1), 272–281.

Appendices

Appendix. A. Machine Learning Within-Condition Prediction



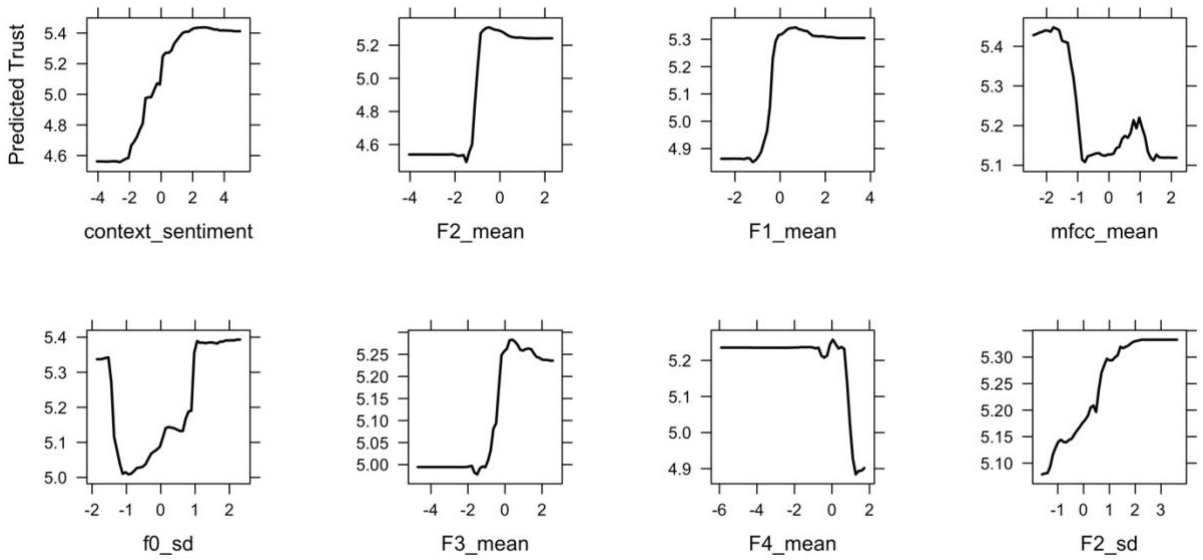


Figure A 1. Partial Dependence Plots for the High (Top), Low (Middle), and Between High and Low Conditions (Bottom).

Our results on Partial Dependence Plot (PDP) showed a nonlinear logistic growth. To avoid capitalize heavily on the extreme manipulation of trust with the between-subject reliability variable, we ran our models predicting trust within high and low reliability separately. In the high reliability, the best performed model was random forest (R^2_{adj} = 0.71, RMSE = 0.51). In the low reliability, the best performed model was also random forest (R^2_{adj} = 0.87, RMSE = 0.52). To make the cross-comparison, we used the same eight conversational features to compare the relationships using PDP plots. From top to bottom in Figure A 1, the three plots are high, low, and between high and low conditions (see the y-axis for the difference in range). The overall trends showed similarity except for the range for trust prediction. The results showed a high predictive power with similar feature relationships using PDP.

The major difference was the ranking of the important features. In the high reliability condition, the most important feature was the standard deviation of F_0. In the low reliability condition, the most important feature was the mean of F_3. Compared to the between high and low reliability settings, the most important feature was the context sentiment. Results suggested that lexical sentiment was the dominant predictor for the large trust difference (i.e., from high to low reliability condition). Yet, acoustic features (e.g., variance in pitch and formants) provided more nuances in predicting trust variance in either high or low reliability condition.

Appendix. B. Mediation Analysis

To further investigate the casual relationship between conversational features and trust in machine learning models, we ran the mediation analysis to assess the causal mechanisms. Since we identified eight important features, it requires a mediation analysis with multiple mediators that are considered simultaneously. Thus, we adopted the multiple mediator analysis method using the R package *mma* (Yu & Li, 2017). Figure A 2 showed the importance of all potential mediators in explaining trust in terms of their relative effects. The 'de' represents estimation of the direct effect. The estimated total effect is -0.95. We selected the most important feature, 'context sentiment' (relative effect = -0.11), for the following mediation analysis.

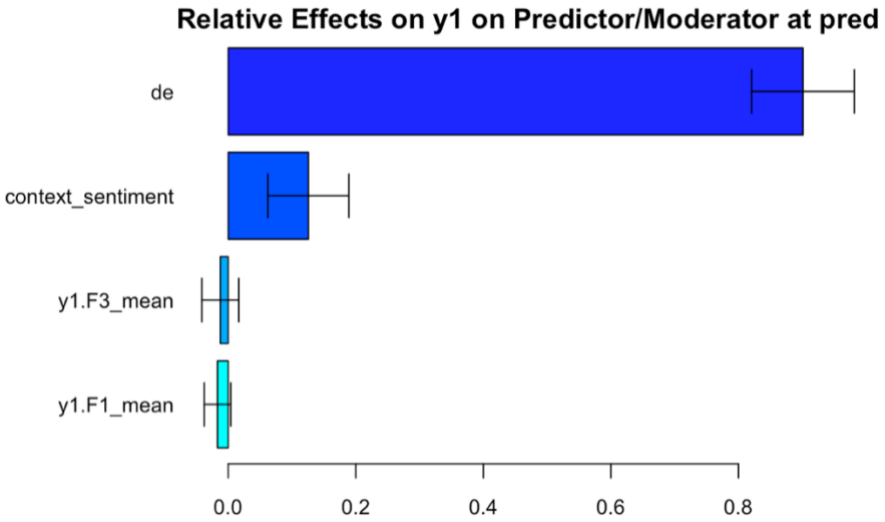


Figure A 2. The Importance of All Potential Mediators.

In this analysis, IV is reliability, DV is trust and MV is the conversational feature (i.e., context sentiment). To test for the full mediation, one should estimate the following regression equations:

1. IV significantly predicts DV (path c' is significant): Trust ~ Reliability.

2. IV significantly predicts MV (path a is significant): Conversational features ~ Reliability.

3. MV significantly predicts DV, (path b is significant): Trust ~ Conversational features + reliability.

4. When mediator enters the IV-DV relationship, the total effect reduces significantly to non-significant (path c). If the direct effect does not reduce significantly to non-significant, mediation only happens partially.

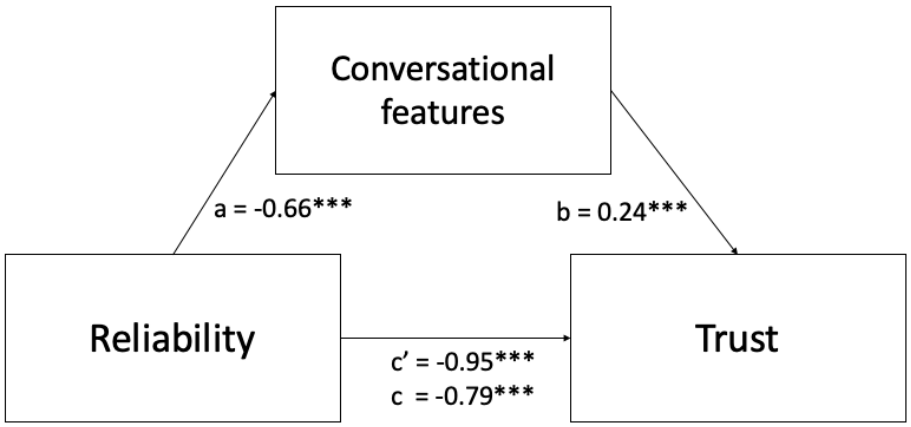


Figure A 3. The Mediation Effect of Conversational Features in The Relationship Between Reliability and Trust. Note ***p<.001. A is Effect of Reliability on Conversational Features; b is Effect of Conversational Features on Trust; c' is Direct Effect of Reliability on Trust; c is Total Effect of Reliability on Trust.

As shown in Figure A 3, results showed that there was a significant total effect between reliability and trust (B = -0.95, p <.001), path a (i.e., reliability on conversational feature) (B = -0.66, p <.001) and path b (i.e., conversational feature and reliability on trust) (B = 0.24, p <.001) were both significant. Finally, when conversational features entered the relationship between reliability and trust, the direct effect (B = -0.79, p <.001) was significant. In addition, the Sobel test for the indirect effect is z = -5.86, p <.001; therefore, it was concluded that a partial mediation occurred between reliability on trust via conversational features. The proportion of the effect of the reliability on trust that goes through the medicator is 0.17. It is calculated by dividing the average causal mediation effects (ACME) (-0.161) by the total effect (-0.95) to receive 0.17. Results supported the causal relationships between conversations, reliability, and trust: automation reliability influences the way people communicate, which can be used to predict trust.