Computational Accounts of Trust in Human AI Interaction

by

Zahra Zahedi

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved November 2023 by the
Graduate Supervisory Committee:

Subbarao Kambhampati, Chair
Erin Chiou
Siddharth Srivastava
Yu Zhang

ARIZONA STATE UNIVERSITY

December 2023

ABSTRACT

The growing presence of AI-driven systems in our daily lives calls for the development of efficient methods to facilitate interactions between humans and AI agents. At the heart of these interactions lies the notion of trust, a key element shaping human behavior and decision-making. It is essential to foster a suitable level of trust to ensure the success of human-AI collaborations, while recognizing that excessive or misplaced trust can lead to unfavorable consequences. Human-AI partnerships face distinct hurdles, particularly potential misunderstandings about AI capabilities. This emphasizes the need for AI agents to better understand and adjust human expectations and trust.

The thesis explores the dynamics of trust in human-robot interactions, acknowledging that the term encompasses human-AI interactions, and emphasizes the importance of understanding trust in these relationships. This thesis first presents a mental model-based framework that contextualizes trust in human-AI interactions, capturing multi-faceted dimensions often overlooked in computational trust studies. Then, we use this framework as a basis for developing decision-making frameworks that incorporate trust in both single and longitudinal human-AI interactions. Finally, this mental model-based framework enables us to infer and estimate trust when direct measures are not feasible.

To the guardians of my heart, my guiding moon and nurturing sun, my beloved Mom and Dad, with boundless love and divine grace.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Subbarao Kambhampati, for his invaluable support and guidance throughout my doctoral journey. Under his excellent mentorship, I have acquired numerous valuable skills and insights that have significantly enhanced my academic journey. He inspired me to push beyond my boundaries and challenged me to think more deeply and critically.

Next, I extend my sincere thanks to Dr. Erin Chiou for her pivotal feedback and guidance on my work. Her unique perspective and view have significantly contributed to the improvement of my research. I also wish to express my appreciation for my other committee members: Dr. Siddharth Srivastava and Dr. Yu Zhang. Their invaluable suggestions and thought-provoking insights have been instrumental in shaping my research. I would also like to take the chance to thank Dr. Dave Smith. He has been a valuable source of insightful discussions and feedback, and his kindness and generosity in sending us New Year treats and making visits are greatly appreciated.

My experience at the Yochan lab has been nothing short of remarkable, thanks to the incredible and talented labmates I've had the privilege to work with. First and foremost, I'd like to acknowledge my mentor, friend, and brother at heart, Sarath. His unwavering support, guidance, teachings, and friendship have played a pivotal role in advancing my Ph.D. journey, and I am truly grateful for his all-encompassing support. Next, I would like to extend my appreciation to Sailik, with whom I had the opportunity to work during the first two years of my Ph.D. He was like a lantern, illuminating complex concepts that were initially foreign to me when I began my Ph.D. journey. Through our collaboration, I gained valuable knowledge and skills.

I am deeply grateful to Sachin, Alberto, and Ram for their unconditional friendship and support. Their constant presence and encouragement have been a significant

source of strength for me. In addition, I would like to express my appreciation for Anagha, Yantian, Tathagatha, Utkarsh and Lydia, whose friendship and presence have been a source of inspiration throughout my time at Yochan. The current Yochanites, including Mudit, Lin, Karthik, Siddhant, Matthew, Anil, and Kaya, have also left an unforgettable mark on my lab experience, have shaped wonderful memories.
I would also like to thank my entire SCAI family especially Monica Dugan, Pamela Dunn, and Christina Sebring, for their continuous support.

I'm thankful for the remarkable opportunity I had as an intern at Honda Research Institute USA, Inc., where I collaborated with brilliant colleagues and mentors. I want to express my gratitude to my mentors, Akash Kumar, Shashank Mehrotra, Terry Misu, and my colleagues, Kevin, Dmitri, David, Hiu, Faizan, and Reza, for contributing to my enjoyable and rewarding experiences in this new environment.

I am genuinely thankful for the incredible friendships outside the lab, including Mansooreh, Azadeh, Mohammad, Faezeh, Rana, Zahra Soltani, Hooman, and Roozbeh. Their love, joy, support, and friendship have added immeasurable richness to my life.

Last but not least, I cannot overstate my deep gratitude to my family. Their constant love and support have not only been the foundation of my success but have also been a constant source of strength and motivation. I consider myself blessed to have them in my life and want to extend a special thanks to my beautiful and lovely sister, Fatemeh. Her divine blessing presence has brought nothing but beauty, joy, and unconditional love to my life. Having her alone is a source of lifelong gratitude. All through her heavenly ether that surrounds me with a sense of cherished bonds, making me feel at home, safe, and supported.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

In recent years, the proliferation of AI-powered systems has significantly increased their presence across many aspects of our daily lives, with their transformative effects being observed in diverse fields such as social media, robotics, finance, autonomous vehicles, and intelligent healthcare applications. As these AI systems become more pervasive, it is critical to establish effective and intuitive interaction mechanisms between humans and AI agents. A fundamental aspect that significantly influences the success of these interactions is the level of trust that humans place in these systems.

Trust is a complex concept, instrumental in shaping human behavior and decision-making processes. While AI systems have to engender trust in human in the loop, the complex task of engendering an appropriate level of trust does involve a delicate balance. Excessive trust or misplaced trust in AI systems can lead to detrimental outcomes such as automation bias and complacency Parasuraman and Manzey (2010). On the other end of the spectrum, insufficient trust may result in ignoring the system's capabilities, consequently hindering performance Lee and See (2004); Chen *et al.* (2018).

Among these challenges, the dynamics within human-AI teams deserve special attention. In contrast to homogeneous human teams where members have a well-developed understanding of each other's roles and capabilities, human-AI teams often grapple with potential misconceptions about the AI's abilities. This highlights a critical requirement for building lasting trust- the capacity of autonomous agents to comprehend and, if necessary, correct human expectations about their capabilities.

In addressing these needs, recent advancements in the realm of human-aware

planning provide a valuable tools Sreedharan *et al.* (2022). Techniques such as explicable planning and model reconciliation offer powerful insights into how AI agents can effectively model and interact with human trust, thereby guiding user expectations accurately Chakraborti *et al.* (2017); Zhang *et al.* (2017).

These insights set the stage for the next critical step - designing trust-aware AI systems or frameworks. Crucial to this endeavor is the creation of generalized and consistent formalizations of trust that can be applied across various decision-making contexts. These formalizations should serve as a foundation for understanding, estimating, and fostering an appropriate level of trust in human end-users.

By integrating these insights and addressing the challenges associated with trust in AI systems, we can pave the way for the development of AI systems that not only understand and adapt to human trust but also promote effective collaboration and trust-building in human-AI teams.

The concepts, theories, and prior works in the area of trust from both Psychological and Human Factors Perspectives as well as computational accounts serve as the foundational background for this research. A deep dive into various dimensions and characteristics of trust, and the way they are treated in computational models, reveals a significant gap in the literature. This gap lies in bridging the human factor view of trust and the computational models of trust, which in many instances might not sufficiently capture the multi-dimensional aspects of trust integral to human-AI interactions.

In this thesis, I will describe my main contribution made thus far within the space of various computational accounts of trust in human-AI interaction as part of my Ph.D. research. I have made significant strides towards developing this formalization through a mental model-based framework. This framework (see Figure 1.1) effectively contextualizes trust in human-AI interactions and captures multiple dimensions of

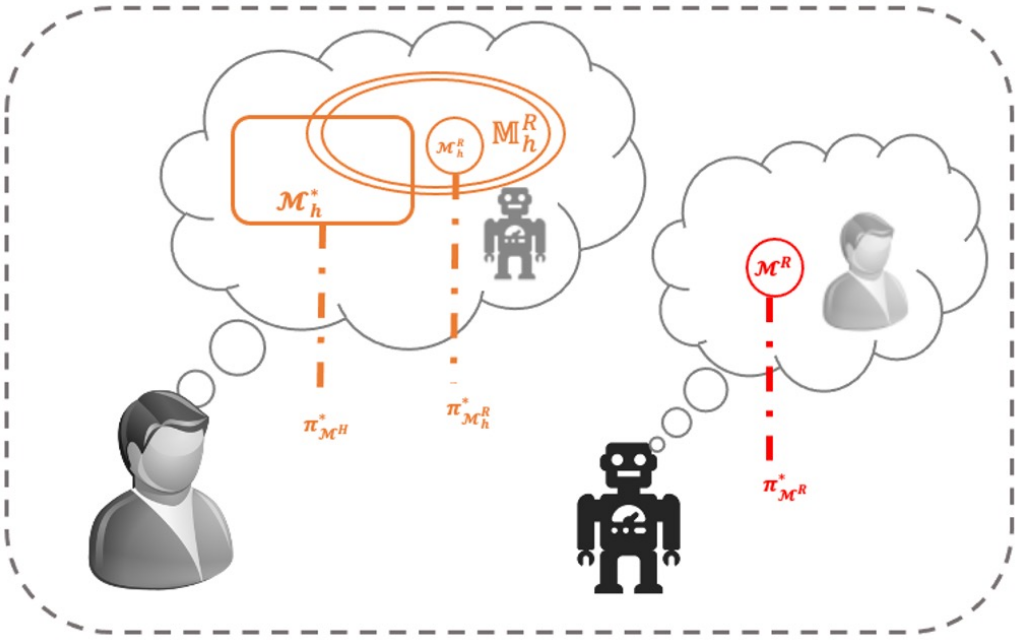


Figure 1.1: A schematic representation of the mental model based framework of trust. $\mathcal{M}^R$ is the task model that the AI agent ascribes to itself; $\mathbb{M}_h^R$ is set of models the human may ascribe to the agent; $\mathcal{M}_h^*$ is the human's task model that captures their expectations about the task.

trust, often overlooked in many computational trust studies. It forms the basis for my works towards developing decision-making frameworks that incorporate trust in both single and longitudinal human-AI interactions. Moreover, it serves as a tool for inferring and estimating trust, especially when direct measures may not be feasible.

As we delve into the dynamics of trust in this research, it's important to note that the terms 'human-AI interaction' and 'human-robot interaction' often carry similar connotations and may be used interchangeably throughout this work. Both encompass the essential elements of human-machine collaboration, focusing on how people understand, interact with, and develop trust in artificial systems. The emphasis on 'human-robot interaction' in the ensuing chapters is primarily due to its embodiment of the sequential decision-making nature inherent in our AI system under consideration. A robot, acting as the physical avatar of AI, provides a tangible interface for human

engagement, reflecting the real-time, dynamic interaction that mirrors human-human interaction in a more recognizable sense. So, even if 'human-robot interaction' becomes the prevalent term in our discussion, it's crucial to remember that we are always, at the core, exploring the intricate layers of trust in human-AI relationships.

1.1 Thesis Outline

The subsequent sections of this thesis document are structured as follows:

- **Chapter 2:** We will go into more depth regarding the psychological and human factors viewpoints of trust, and previous research related to computational accounts of trust in the context of human-AI interaction. We then emphasize the importance of developing a comprehensive framework that unifies considerations of the human factor with computational trust models.

- **Chapter 3:** This chapter lays the foundation by describing the human-AI interaction setting and establishing the mental model-based framework that we will explore throughout the thesis. It also presents a methodology to encapsulate trust according to the proposed mental model and illustrates how this formalization effectively captures various dimensions of trust as delineated by human factor researchers.

- **Chapter 4:** Here, we delve into the dynamics of a single interaction between a human and a robot. We explore scenarios where warranted trust is not present and how that might influence human-robot task approaches. In the absence of warranted trust, humans often tend to monitor the robot excessively. This chapter presents a solution for efficient monitoring, formalizing the problem within a game-theoretic framework. Through various human subject studies, we examine the effectiveness of our proposed monitoring strategy in practice and

discern the natural strategies individuals are likely to adopt.

- **Chapter 5:** We introduce a formal model in this chapter that encapsulates the probabilistic relationship between a human's trust level and their tendency to monitor an AI agent. This model aims to enhance decision-making frameworks in human-robot interactions by integrating the trust-monitoring dynamic. To get the model parameters, we conduct a human subject study, which provides insights into the impact of trust on human monitoring behavior in robot-assisted tasks.

- **Chapter 6:** We will focus on my contribution within longitudinal interaction between human and the robot. With iterative interactions between the human and the robot, the robot can indeed focus on trust-building or performance optimization. We propose a computational model designed to capture and modulate trust in longitudinal human-robot interactions. By integrating the human's trust and expectations into its planning, the robot can effectively build and maintain trust throughout the interaction horizon. As the human's trust in the robot grows, they may opt to reduce monitoring or refrain from intervening to stop the robot. Therefore, once the required level of trust is established, the robot can shift its focus to maximizing the team's goals. The reasoning concerning trust levels is modeled as a meta-process that influences individual planning tasks, ensuring the robot adapts its behavior to maintain trust throughout the entire interaction.

- **Chapter 7:** We operationalize our mental model based theory of trust to model trust evolution, formalize appropriate levels of trust and model the human's reliance on the AI agent. Furthermore, we focus on the comprehensive evaluation of mental model based framework of trust to validate the central aspect of it,

which examines that changes in trust (as measured by the questionnaire) can be achieved by altering the human's belief about the agent, as predicted by our mental model theory. Additionally, by controlling different aspects of the model associated with various perceptions of trust, we can evaluate whether changing a specific part of the model influences a corresponding change in the associated information of trust (i.e. performance, process, and purpose).

- **Chapter 8:** We conclude the thesis with a summary of how the works presented achieve the goals of this thesis, reflect on various aspects of the presented works, and highlight the key takeaways, as well as avenues for future directions.

Chapter 2

BACKGROUND

In this chapter, we will explore the psychological and human factors perspectives of trust, delve into prior work in computational trust accounts in human-AI interaction, and highlight the need for a comprehensive framework that integrates human factor considerations with computational trust models. By bringing these facets together, we can promote AI systems that understand, adapt to, and foster human trust, paving the way for more effective collaboration in human-AI teams.

2.1 Psychological and Human Factors Perspectives of Trust

Trust, a complex and extensively studied concept, spans various research fields, such as psychology, sociology, philosophy, political science, economics, and human factors. Scholars from these disciplines have attempted to understand trust and develop comprehensive ways to define it. In the realm of human factors, trust has been extensively researched, particularly concerning its role in guiding interactions with various technologies Hoff and Bashir (2015).

Various perspectives on trust in connection with automation have been proposed by researchers. Some researchers view trust as an attitude or expectation Rotter (1967); Rempel *et al.* (1985); Barber (1983), while others approach it as an intention or willingness to act Johns (1996); Moorman *et al.* (1993); Mayer *et al.* (1995). Among these viewpoints, the most widely accepted definition emphasizes vulnerability as a critical element, stating that trust involves willingly taking risks by delegating responsibility to another party Mayer *et al.* (1995); Rousseau *et al.* (1998). Additionally, some definitions go further, portraying trust as an outcome of behavior or as a state

of vulnerability or risk Deutsch (1960); Meyer (2001). Despite the significance of trust in cooperative relationships, establishing a singular, all-encompassing definition remains challenging, as definitions differ between trust as a belief, attitude, intention, or behavior Lee and See (2004). Among diverse perspectives, Lee and See's definition stands out as one of the most widely used and accepted, defining trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" Lee and See (2004). According to Ajzen and Fishbein's framework, trust primarily affects reliance as an attitude rather than a belief, intention, or behavior. Trust is grounded in beliefs and perceptions, influencing attitudes and leading to specific intentions that translate into behaviors Ajzen (1980). Therefore, Lee and See consider the basis of trust as information-based, guiding expectations regarding how well the trustee can achieve the goals. This information supports trust and can be described using three levels of attributional abstraction: purpose, process, and performance information, which describe the agent's goal-oriented characteristics Lee and See (2004).

In quantifying and measuring trust, researchers also have proposed foundational works aligned with the three information types of trust. For instance, the Muir scale describes trust in terms of reliability, predictability, faith, and overall trust Muir (1994). Many extensions of this scale have been proposed. Other works, such as Cummings *et al.* (2008) and Ullman and Malle (2018), base trust measurement on perceived reliability, technical competence, understandability, faith, and personal attachment. An empirical scale with twelve factors, developed by Jian *et al.* (2000), measures non-directed feelings of trust in automated systems. Some researchers also consider self-reported trust scales that focus on explicit and implicit predictors, including factors like propensity to trust machines and implicit attitudes toward automation Merritt *et al.* (2013).

Furthermore, researchers also emphasize that trust has two key properties: the vulnerability of the user; and the ability to anticipate the impact of the AI model's decisions Jacovi *et al.* (2021). The aspect of vulnerability holds significant weight, especially in the realm of Human-AI trust. This vulnerability arises from the user's dependence on the AI system's capabilities and decisions, particularly when the outcomes can greatly affect the user's well-being or goals. Another property that is imperative in understanding trust is the user's ability to anticipate the impact of the AI model's decisions. Anticipation is essentially an expectation about future behavior and is closely aligned with the views of researchers who perceive trust as an attitude or expectation. The user's anticipation, and therefore trust, can change as they gain more information about the AI model's reliability, competence, and other attributes.

In conclusion, the study of trust, particularly within the domain of human-AI interaction, is multifaceted, underpinned by both psychological insights and human factor research. The information types that informs trust, encapsulated by performance, process, and purpose, each offer a unique perspective on how trust is formed and maintained. Performance focuses on the outcomes, process considers the methods by which results are achieved, while purpose looks at the objectives that guide actions. These information types, interconnected yet distinct, provide a comprehensive view of trust's complexity.

Further, the two critical properties of trust - vulnerability and anticipation - add depth to our understanding. The vulnerability of a user, especially in the face of decision-making AI systems, and the ability to anticipate an AI's actions and potential impact, embody the essence of trust in this context. These properties underscore the significant role of trust in human-AI interactions, demonstrating how trust in AI is not merely about the technology's capabilities but deeply entwined with the human experience of dependence and expectation.

As research progresses in this area, these information types and properties of trust will persist as guiding principles. Continuing to examine and improve these concepts is essential to better understand trust thereby facilitating the successful integration and user acceptance of AI systems. Acknowledging the significance of these three information types - performance, process, and purpose - along with the two properties of trust - vulnerability and anticipation - moves us towards successfully bridging the gap between humans and AI. By fostering computational frameworks that assimilate these insightful aspects of trust, we are more likely to design AI systems that resonate with human needs and expectations, thereby paving the way for a future where AI and humans can work together in harmony.

2.2 Prior Art in Trust in Human Robot Interaction

A number of works have studied computational accounts of trust in the context of human-robot interaction. Research in this area generally falls into three primary categories: (1) Trust inference, which leverages observed human behavior to predict trust levels Desai (2012); Kok and Soh (2020), (2) Trust utilization, which harnesses estimated trust to guide and optimize robot behavior, and (3) Trust calibration and repair, which deals with adjustment and restoration of trust levels.

Within the realm of trust inference, significant strides have been made. One notable contribution is the Online Probabilistic Trust Inference Model (OPTIMo) Xu and Dudek (2015), along with its various extensions Guo *et al.* (2020); Soh *et al.* (2020). OPTIMo encapsulates trust as a latent variable within a dynamic Bayesian network, thereby capturing the relationships between trust, its influencing factors, and its time-bound evolution. The model employs a technique for real-time trust estimation based on the robot's task performance, human intervention, and trust feedback Xu and Dudek (2015).

An essential extension to OPTIMo is a trust inference model that utilizes Bayesian inference with a Beta-distribution to capture both positive and negative attitudes towards robot performance Guo *et al.* (2020). This model enriches the understanding of how trust evolves based on multifaceted human responses to robot interactions.

Additionally, Bayesian reasoning for trust inference has been explored non-parametrically using Gaussian processes and Recurrent Neural Network (RNN) Soh *et al.* (2020). A hybrid approach was also proposed where trust is viewed as a task-dependent latent function. These methodological advancements present a nuanced view of trust, reflecting its intricate dynamics and its dependence on context and performance.

Furthermore, a 'trust transfer function' was developed by Lee and Moray (1992) to describe the dynamics of trust. Studies by Desai *et al.* (2013, 2012) also provide insight into trust inference based on the negative impacts of robot failure. These contributions provide a comprehensive understanding of the factors influencing trust and how it can be inferred from various aspects of human-robot interaction.

Trust utilization primarily focuses on integrating estimated trust into a mechanism that dynamically modifies robot behavior, in order to enhance team collaboration efficiency. Examples of such works include those that estimate trust and trustworthiness Floyd *et al.* (2015) using various parameters like reputation function Xu and Dudek (2012), or OPTIMo Xu and Dudek (2016). An extension of OPTIMo employing a time series trust model Wang *et al.* (2015) has been utilized in multi-robot scenarios, informing decisions about manual or autonomous robot control Wang *et al.* (2018). Furthermore, researchers have proposed a POMDP planning model that enables robots to form policies by considering human trust as a latent variable Chen *et al.* (2018, 2020). Likewise, Nikolaidis *et al.* (2017) have demonstrated that factoring in human adaptability can enhance their trust in robots. In swarm robotics, trust has been utilized for task reassignment to trusted team members Pierson and Schwager (2016);

Pippin and Christensen (2014) and for minimizing misinformation from less trusted robots Liu *et al.* (2019).

Given the inevitability of robot failures and the associated impact on human-robot interaction, there is a growing interest in devising methods for trust calibration and repair Billings *et al.* (2012); Baker *et al.* (2018); Tolmeijer *et al.* (2020). Traditional methods such as apologies, promises De Visser *et al.* (2018); Robinette *et al.* (2015); Sebo *et al.* (2019), and consistent trustworthy behaviors Schweitzer *et al.* (2006) have been extensively studied. The significance of perceived shared intention in trust recovery has also been explored Dennett (1987). Additionally, transparency enhancements, like detailed explanations, have been shown to calibrate trust and augment performance Wang *et al.* (2016).

In conclusion, research into trust in human-robot interactions is extensive, and encompassing aspects of trust inference, utilization, and calibration. We could see that the vulnerability aspect of trust is inherent in these studies in different capacities, considering that trust in robots implies potential risks to the users, while the element of anticipation, a human's psychological expectancy of a robot's performance, drives the development and adaptation of these models. However, while these works predominantly focus on the performance information of trust, reflecting the reliability of robots in achieving a goal, the process and purpose information have largely been overlooked. The process information, representing how the robots operate to achieve their goal, and the purpose information, signifying the intended goal or the 'why' behind a robot's actions, are equally critical in understanding and developing trust. As we continue to integrate robots into diverse aspects of our lives, recognizing and addressing these information types and properties of trust becomes increasingly important. The dynamic nature of trust demands continuous refinement of these models. Therefore, as we strive for successful and productive human-robot interactions, a more

comprehensive understanding and shaping of this complex trust relationship become paramount.

2.3 Bridging the Gap: Human Factors vs. Computational Models

Despite the substantial progress in human factor research on human-automation trust and computational accounts of trust in human-AI interaction, a significant gap persists between these two perspectives. Each offers a unique viewpoint, and the integration of both is a crucial step towards achieving a comprehensive understanding of trust in AI.

Human factor research explores trust from a psychological and cognitive perspective. This approach has provided a wealth of insights into the intricacies of human behavior, particularly regarding how individuals perceive, develop, and maintain trust in automated systems. By examining human cognition and behavior, this field has delved into the multidimensional nature of trust that can help in shaping AI systems that align more closely with human expectations, leading to increased user acceptance and more effective human-AI collaborations.

Conversely, computational models aim to tackle trust from a technical perspective, utilizing algorithms and data-driven methods to infer and manage trust in AI systems. These models often concentrate on quantifying trust based on observed behaviors and performance metrics, taking a more mathematical and objective approach to a largely subjective concept. However, this focus on quantifiable data tends to neglect other integral information types of trust, notably the process and purpose aspects, posing challenges in fully grasping the intricate and dynamic nature of trust in human-AI interactions.

Therein lies the challenge: the inherently multifaceted and dynamic nature of trust in human-AI interactions requires an integrated approach. Each perspective, human

factors and computational models, captures a portion of the picture, but neither can fully encompass the complexity of trust on its own. Bridging this gap calls for an innovative approach that marries the in-depth psychological and cognitive insights from human factors research with the precision and scalability of computational models.

In my research, I am addressing this need by developing a comprehensive mental model-based framework for understanding trust in AI. This framework aims to effectively encapsulate all information types that informs trust, contextualizing them within both human cognitive processes and computational methods. By doing so, it bridges the gap between human factors and computational models, leading towards more nuanced, accurate, and ultimately more useful models of trust in human-AI interactions.

Chapter 3

MENTAL MODEL BASED THEORY OF TRUST

3.1 Interaction Setting

The basic interaction setting we will consider throughout the thesis consists of two agents types, an AI agent (denoted by R) and the human user trying to derive useful decisions from the agent (denoted by H). We will model each agent that is part of the scenario as a model-based agent: the agent uses a model to not only derive its own decisions but to form expectations about what other agents might or should do. Note that we use the term "model" loosely to mean any formal model that encodes, among other things, an agent's beliefs about task objectives, state of the world and how the world may evolve on its own or in response to an agent action. There exist a wide variety of options for what the models might be (for example they may be MDPs Puterman (2014), POMDP or symbolic models like PDDL Geffner and Bonet (2013)). We only require that the models used by an agent are in a form that they can use to derive the required decision or expected decisions.

We will provide a detailed description of each type of model and detailed description of a possible symbolic representation of a model.

Model of the agent $\mathcal{M}^R$: This is the model the AI agent ascribes to itself. This determines what actions the agent believes they could perform and the objectives and preferences they are trying to satisfy.

Set of the human's models of the agent $\mathbb{M}_h^R$: This set consists of the models the

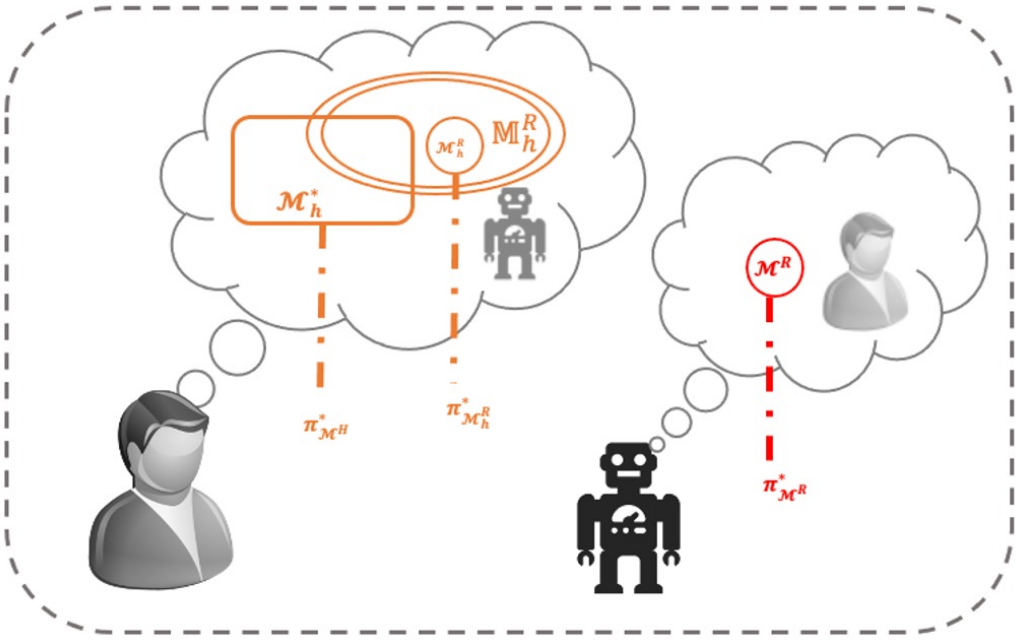


Figure 3.1: A schematic representation of the mental model based theory of trust. $\mathcal{M}^R$ is the task model that the AI agent ascribes to itself; $\mathbb{M}_h^R$ is set of models the human may ascribe to the agent; $\mathcal{M}_h^*$ is the human's task model that captures their expectations about the task.

human may ascribe to the robot, such that each model $\mathcal{M}_h^R \in \mathbb{M}_h^R$ is, as far as the human is concerned, could be the actual model being used by the agent to derive its decisions. Each model is also associated with some likelihood ($P_{\mathbb{M}}$) that corresponds to the human's degree of certainty that a specific model corresponds to the true model used by the agent.

Human's task model $\mathcal{M}_h^*$: This model is meant to capture the expectations the human may have about the task that is independent of what they believe the agent is capable of performing. In our framework of trust, this model will mostly act as a way to capture the human's expectations about the idealized way of completing the task. Such expectation may be formed from their beliefs about the human's own ability to complete the task or may even come from other sources (either from observing

more experienced users, or the expectations may be carried over from institutional expectations). One important point to note here is that this is the human's belief about the ideal way of solving the task and need not reflect the true optimal ways of solving the task for the robot. Additionally, to allow Bayesian reasoning, we will assume that at least some of the models in the set $\mathbb{M}_h^R$ can generate solutions comparable to those that are generated by $\mathcal{M}_h^*$, however, the prior likelihood on those models may be quite low.

Although our representation of model can capture any formal model, a common model representation scheme that we use through out the works is STRIPS style planning problem (Geffner and Bonet, 2013). In such cases, a model may be represented as a tuple of the form $\mathcal{M} = \langle \mathcal{D}, \mathcal{I}, \mathcal{G} \rangle$, consisting of domain $\mathcal{D} = \langle \mathcal{F}, \mathcal{A} \rangle$, where F being a finite set of fluents that define the state of the world $s \subseteq \mathcal{F}$, $\mathcal{A}$ the finite set of actions, $\mathcal{I}$ the initial state, $\mathcal{G}$ the goal, where $\mathcal{I}, \mathcal{G} \subseteq \mathcal{F}$. A plan for such models is a sequence of actions $\pi = \langle a_1, ..., a_k \rangle$, which when executed in the initial state leads to the goal $\rho_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$. The cost of a plan π is given by $C(\pi, \mathcal{M}) = \sum_{a \in \pi} c_a$ if $\rho_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$, otherwise the cost is ∞. The likelihood of a plan given a model will take the form $P_l : \mathcal{M} \times \pi \rightarrow [0, 1]$. While we will try to be agnostic to likelihood functions, a fairly common approach Fisac *et al.* (2018); Baker *et al.* (2009) is a noisy rational model based on the Boltzmann distribution: $P_l(\mathcal{M}, \pi) \propto e^{-\beta \times C(\pi)}$ Sreedharan *et al.* (2021). The cheapest plan with highest probability is called optimal plan $\pi^* = arg\min_\pi C(\pi, \mathcal{M})\ \forall \pi$ such that $\rho_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$.

3.2 Exploring Trust Within Mental Model Framework

With the basic interaction setting in place, we are ready to provide a preliminary definition of trust and a way to model the user's choice to rely on a specific AI agent.

Trust: Trust as a contextual measure is best understood in terms of one's expectation on an agent to satisfy some specific contract as opposed to thinking of it as a general measure associated with the agent Jacovi *et al.* (2021). In our model, the contract corresponds to the agent's behavior meeting some specific performance guarantee generally related to the quality of the solution that can capture different information types of trust; performance, process, and purpose. This contract, denoted as $\mathcal{C}$, will be formed by the human based on their model $\mathcal{M}_h^*$. Now the *trust measure* ($\mathcal{T}(\mathcal{C})$), i.e., the numeric quantity reflecting the degree of trust the human places on the agent in the context, will be defined to be directly proportional to a monotonically increasing function over the likelihood the human places on the agent to satisfy the contract, i.e.,

$$\mathcal{T}(\mathcal{C}) \propto \mathcal{F}(P(\mathcal{C}^H))$$

Where the likelihood the human would believe the robot can satisfy the contract $P(\mathcal{C}^H)$ ($\mathcal{C}^H$ is the random variable corresponding to the human belief that the contract will be satisfied) is, in itself, controlled by the human's belief about the robot's model. Specifically, we will assume that the human reasoning can be captured using the probabilistic graphical model presented in Figure 3.2(A) and the likelihood is given as

$$P(\mathcal{C}^H) = \sum_{\mathcal{M} \in \mathbb{M}_h^R} P(\mathcal{C}^H|\mathcal{M}) \times P_{\mathbb{M}}(\mathcal{M})$$

Where $P(\mathcal{C}^H|\mathcal{M})$ is the likelihood human associates with a model $\mathcal{M}$ coming up with a solution that can satisfy a given contract $\mathcal{C}$.

3.2.1 Three Information types of Trust Within Mental Model Framework

In the domain of human factor research, it is posited that appropriate trust hinges upon the understanding of performance, process, and purpose. Thus, it is paramount

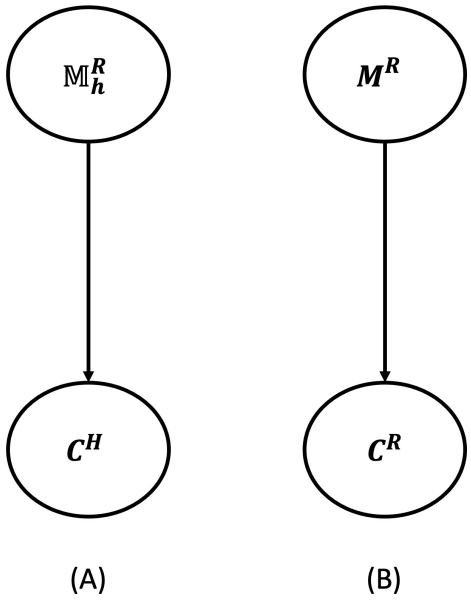


Figure 3.2: A graphical model representing the probabilistic reasoning that is performed in this setting: Subfigure (A) captures the reasoning performed at the human's end with $\mathcal{C}^H$ being the random variable corresponding to the human's belief that a contract $\mathcal{C}$ will be satisfied. Similarly subfigure (B) represents the reasoning performed at the human's end, where $\mathcal{C}^R$ captures whether the robot achieves the contract C.

for a computational trust model to encompass these three information types. Our mental model-based framework is designed to encapsulate these information types of trust. Here, we delve deeper into the alignment of these trust information types with our mental model-based framework. Initially, we will define each information and illustrate how they are encapsulated in our model.

Of paramount importance in our framework is recognizing that the formation of trust in humans is tightly bound with their mental models. The trust a human places in a robot is fundamentally linked to the interplay between their mental models ($\mathcal{M}_h^*$ and $\mathbb{M}_h^R$). To logically reason about trust, we must therefore focus our attention on the section of the framework capturing all models related to the human.

To delve deeper into the mechanics of our model, consider one of the human's model of the robot that belongs to the set of human models of the robot, $\mathcal{M}_h^R \in \mathbb{M}_h^R$.

This model is represented as $\langle \mathcal{D}_h^R, \mathcal{I}_h^R, \mathcal{G}_h^R \rangle$. Here, $\mathcal{D}_h^R$ symbolizes the domain of the model, $\mathcal{I}_h^R$ refers to the initial conditions, and $\mathcal{G}_h^R$ encapsulates the goal state. This goal state, $\mathcal{G}_h^R$, comprises both the internal AI agent's goal and the goal as articulated by the human.

With this understanding, we can now map these components of our model to the three information types of trust:

1. **Performance:** Performance information depicts the automation's functions and speaks to the competency or proficiency of the automation as indicated by its success in fulfilling human objectives Lee and See (2004). This information corresponds to the domain of the model, $\mathcal{D}_h^R$.

2. **Purpose:** Purpose information elucidates why the automation was conceived and assesses whether the agent has a positive orientation towards the human Lee and See (2004). This information is associated with $\mathcal{G}_h^R$ within our model, which consists of both the AI agent's internal goal and the human's specified goal.

3. **Process:** The process information explores the operational procedures of the robot and evaluates the quality of actions based on human expectations Lee and See (2004). It correlates with the extent to which a plan derived from the model matches the human's desired plan. To quantify this alignment, we can use cost as a metric. Thus, the process can be seen as the probability that a model's optimal solution aligns with the human's optimal solution, $\pi^*_{\mathcal{M}_h^*}$. This probability is mathematically defined via the Boltzmann distribution of the cost difference between the two plans, expressed as $P_e(\mathcal{M}_h^R, \pi^*) \propto e^{-\beta(C(\pi^{\mathcal{M}_h^*}) - C(\pi^*)}$.

3.3 Concluding Remarks

In this chapter, we formalized a mental model based theory of trust that contextualizes trust in human-AI interactions and captures multiple information types of trust. This framework can be used as a basis to infer human trust and utilized as a foundation for any future trust-aware decision-making frameworks.

Throughout the remainder of this thesis, we delve into the practical application of our mental model-based framework. We focus on developing decision-making frameworks that can leverage our mental model based framework of trust to guide both single and longitudinal interactions. This offers comprehensive insights into the dynamics of trust under different scales and contexts of human-robot interactions. Through these works, we will shed light on the effects of immediate robot actions on the trust level and discern the strategies individuals are likely to adopt in practice and how to utilize trust evolution that can happen over repeated interactions. We also utilize our mental model-based framework for trust inference where the trust measure is not observable.

Chapter 4

SINGLE INTERACTION

In this chapter, we consider scenarios involving a robot (R), who is making and executing a plan (or policy) in the world, and a human supervisor (H), who monitors the robot's action and is held responsible for R's behavior. When the supervisor trusts the robot, they do not need to always spend their valuable resources such as time and cognitive effort in monitoring or intervening in the robot's plan (or execution of these plans). While it is possible to develop trust in longitudinal settings Chen *et al.* (2018); Xu and Dudek (2015), in one-off interactions, where no warranted trust exists Jacovi *et al.* (2021), conventional wisdom often guides the supervisor to spend all their time monitoring the robot's behavior to ensure that it adheres to their expectations

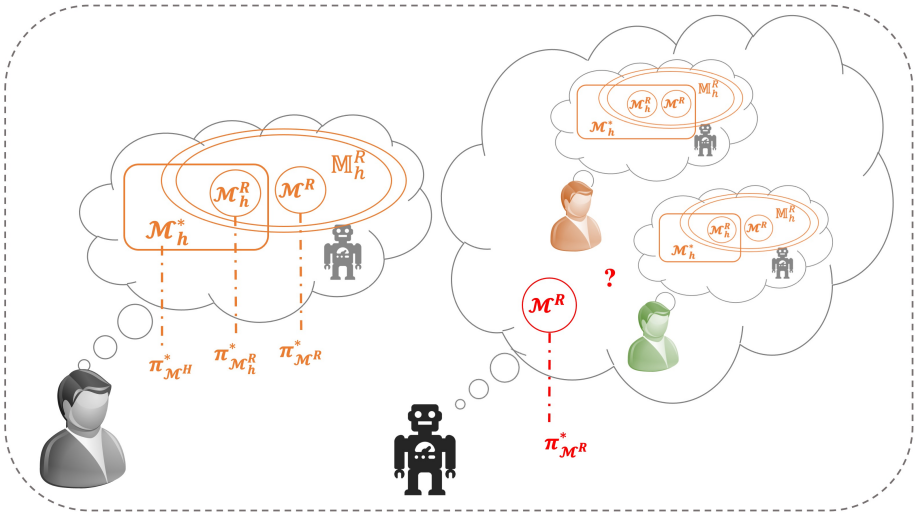


Figure 4.1: A simplified schematic representation of the interaction according to mental model based framework of trust. At least one of the models in $\mathbb{M}_h^R$ is executable in human model $\mathcal{M}_h^*$. The robot on the other hand is uncertain about which of the human supervisor's model of the robot ($\mathcal{M}_h^R$, $\mathcal{M}^R$ or both) is executable in $\mathcal{M}_h^*$.

(making it too resource-intensive for the human). In this chapter, we seek a solution for resource-efficient monitoring strategies for human in the absence of trust. There are cases when a robot's expectation may deviate from its supervisor's expectations. First, a robot may have side-goals that do not align with the supervisor's expectation (the robot internal goal in $\mathcal{G}^R$ might not match human expectations). For example, an autonomous car ride-sharing (or, in general, robot-as-a-service) may have certain expectations from its supervisor (eg. travel on shortest routes) but may need to adhere to passenger's expectation (eg. avoid hilly roads) that are in conflict with one another. Second, the worker robot may not be aware of the human's exact model $\mathcal{M}_h^*$ that describes the safety requirements the supervisor has in mind.

Hence, when the human does not observe the robot's plan or its execution, the robot may choose to execute a less costly plan that is deemed unsafe (by the human). To handle such scenarios, we formally model the inference problem related to the finding a monitoring strategy for the human supervisor that saves their valuable resources (time, cognitive overload) while ensuring that the robot sticks to the expected behavior and achieves the goal.

Specifically, we introduce a notion of trust that a human supervisor H places on a worker robot R when H chooses to *not observe* R's plan or its execution, by modeling the interaction in a game-theoretic framework of trust motivated by Sankaranarayanan *et al.* (2007). In our case, the robot is unaware of the human's exact model $\mathcal{M}_h^*$, but has knowledge about set of human model of the robot $\mathbb{M}_h^R$ that captures all the possible set of safety constraints the human might have, i.e., $\mathcal{M}_h^R \in \mathbb{M}_h^R$ and $\exists \mathcal{M}_h^R$ where $\mathbb{M}_h^R \cap \mathcal{M}_h^*$ (A simplified schematic representation of the interactions and models is shown in Figure 4.1). This uncertainty about the human's model that R has can be reflected in the utilities of the players, making our formulated game a Bayesian one. Without prior interaction (and thus, a lack of trust) if H does not observe R,

R will always deviate to a plan that is less costly for itself. In this chapter, we show that H can devise a probabilistic observation strategy that ensures that (1) R does not deviate from executing the safest plan (i.e., executable in all the models of $\mathbb{M}_H^R$) and also, (2) H saves valuable resources (such as time, effort, etc.) as opposed to continually monitoring R.

In addition to providing a novel type of service that can assist H on when to supervise R to ensure expected behavior, we also explore if such a service is useful in practice by performing human studies and to figure out what are the natural strategies they would follow. First, we show that in such supervision or monitoring scenarios, humans may either be risk-averse (ensuring that the robot does the right thing, no matter the monitoring cost) or risk-taking (in the hope to minimize their cost, will choose to cut down their monitoring time). These results justify the Bayesian modeling of our human player in the game-theoretic framework for the supervision scenario. Second, we show, in contrast to work in existing human-aware planning scenarios where humans are asked to monitor the robot all the time Kulkarni *et al.* (2016); Dragan *et al.* (2013), humans often deviate to more split-time strategies where some of the time, that is originally meant for monitoring, can be used for other tasks and still ensure that the robot adheres to constraints.

Thus, it makes sense to analyse the supervision scenario formally and provide human agents with optimal monitoring strategies that let them maximize their utility while ensuring the supervised agent R does not execute behavior that is either unsafe or fails to achieve the goal. Lastly, we conduct another human study when the optimal strategy is suggested to the human, and demonstrate that suggesting the optimal strategy as computed by our approach will help the human to come up with better monitoring strategy.

4.1 Game Theoretic Formulation

Before describing the game-theoretic formulation–the actions and the utilities of the agents– we first clearly highlight the assumptions made about the two agents.

4.1.1 Assumptions About the Agents

The human H

who is a supervisor in our setting, has the following characteristics:

1. H has a particular model of the robot R, denoted as $\mathcal{M}_h^R \in \mathbb{M}_h^R$, where the solution generated by that model matches the one generated by $\mathcal{M}_h^*$ (i.e. $P_e(\mathcal{M}_h^R, \pi^*) = 1$).

2. Upon observation of the plan that R comes up with or its execution, if H believes the plan is risky (i.e., is inexecutable or unsafe in their model $\mathcal{M}_h^*$), H can stop the execution at any point in time. If H stops the robot R from executing its plan, H incurs some cost of inconvenience for not having achieved the team goal G or because H should stop the robot and make the robot to do the safe plan. This seems pragmatic because H, being the supervisor, will be held responsible for it.

3. H has a positive cost for observing the robot's plan or the plan's execution.

The Robot R

who is the agent being monitored, has the following capabilities and assumptions associated with it:

1. R is uncertain about which human's model of it is executable in $\mathcal{M}_h^*$, but knows that there exist at least one in the set of possible models $\mathbb{M}_h^R$ that is executable.

Table 4.1: Normal-form game matrix for modeling the robot-monitoring scenario. R (H) is the row (column) player.

	$O_{P,\neg E}$	$O_{\neg P,E}$	NO-OB
π_{pr}	$-C_P^H(\pi_{pr}) - I_P^H(\pi_{pr})$, $-C_P^R(\pi_{pr}) - C_{\tilde{E}}^R(\pi_{pr}) - C_{\tilde{G}}^R$	$-C_E^H(\tilde{\pi}_{pr}) - I_E^H(\hat{\pi}_{pr})$, $-C_P^R(\pi_{pr}) - C_E^R(\tilde{\pi}_{pr}) - C_{\tilde{G}}^R$	$-V_I^H(\pi_{pr})$, $-C_P^R(\pi_{pr}) - C_E^R(\pi_{pr})$
π_s	$-C_P^H(\pi_s) \overbrace{-I_P^H(\pi_s)}^{0}$, $-C_P^R(\pi_s) - C_E^R(\pi_s)$	$-C_E^H(\pi_s) \overbrace{-I_E^H(\hat{\pi}_s)}^{0}$, $-C_P^R(\pi_s) - C_E^R(\pi_s)$	$\overbrace{-V_I^H(\pi_s)}^{0}$, $-C_P^R(\pi_s) - C_E^R(\pi_s)$

2. R, given a sequential decision making problem, can come up with two plans– (1) a safe plan (π_s) that is executable in all models $\in \mathbb{M}_h^R$ and (2) a probably risky plan (π_{pr}) that is executable in a subset of $\mathbb{M}_h^R$ but in-executable (or unsafe) in the other models.

3. There are costs for coming up with the plans π_s and π_{pr} and executing them. Also, since R may have to work on other goals or cater to the needs of other supervisors, it would like to execute π_{pr} if it can get away with it.

4. It incurs a cost for not achieving the team's goal G. This happens when the human observes the plan or execution and stops it midway (due to safety concerns).

5. The robot is not malicious and thus, does not lie. It won't bait-and-switch by showing one plan to H (that looks safe) and then executing another.

With these assumptions in place, we can now define each players' pure strategies and their utility values which will encode the uncertainty about the types of human supervisor, turning the game a Bayesian one.

4.1.2 Player Actions

In the normal form game matrix shown in Table 4.1, the row-player is the robot R who has two pure strategies to choose from– the plans π_{pr} and π_s (as described above). The column player is the human H who has three strategies– (1) to only observe the plan made by the robot $O_{P,\neg E}$ and decide whether to let it execute (or not), (2) to only observe the execution $O_{\neg P,E}$ and stop R from executing at any point, and (3) not to monitor (or observe) the robot at all (NO-OB).

A few underlying assumptions that are inherent part in our action definitions are (1) the robot cannot switch from a plan (or a policy) it has committed to a different one in the execution phase and (2) the human only stops the robot from executing the plan if they believe that the robot's plan does not achieve the goal G as per their actual model, i.e. the robot's plan is deemed in-executable (or unsafe) given the domain model $\mathcal{M}_h^*$.

4.1.3 Utilities

The utility values for both the players are indicated in the game-matrix shown in Table 4.1. In each cell, corresponding to the pure-strategy pair played by the two players, the numbers shown at the bottom in black are the utility values for R while the ones at the top in blue are the utility values for H. We now describe the utilities for each player in our formulated game and later, in the experimental section, talk about how they can be obtained in the context of existing task-planning domains.

Robot's Utility Values

We first describe the notation pertaining to the robot utilities and then use them to compose the utilities for each action pair.

$C_P^R(\pi)$ – Cost of making a plan π.

$C_E^R(\pi)$ – Cost to robot for executing plan π.

$C_{\tilde{G}}^R$ – Penalty of not achieving the goal.

Note that we use the variables C to represent a non-negative cost or penalty. Thus, the rewards for the robot R shown in Table 4.1 have a negative sign before the cost and penalty terms. As the human may choose to stop the execution of a plan midway, the robot might have executed a part of the original plan. We denote this partial plan by $\hat{\pi}_{pr}$. Given this, the term $C_E^R(\hat{\pi}_{pr})$ represents the cost of executing the partial plan. [1]

The uncertainty in the robot's mind as to whether a particular supervisor type will let it execute the plan π_{pr} to completion can now be captured using the variable $C_{\tilde{G}}^R$ that represents the cost of not achieving the goal. Before we discuss how one can model the variable $C_{\tilde{G}}^R$, let us first briefly talk about the robustness r of the plan π_{pr}. The parameter $r \in (0, 1]$ represents the fraction of models in $\mathbb{M}_h^R$ where the plan π_{pr} is executable (and thus, safe). A way of obtaining this value for deterministic planning problems could be the use of model counting Nguyen *et al.* (2017). For a given r, an idea to model the cost associated with not achieving the goal is to consider $C_{\tilde{G}}^R$ as a random variable drawn from the Bernoulli distribution s.t. $C_{\tilde{G}}^R$ is a non-zero penalty if the plan is not robust enough for a given human (with probability $1 - r$) or zero if it is (with probability r).

Whenever the cost of not achieving the goal is equal to zero, it means that the robot's plan π_{pr} (or its execution) was observed by H and not stopped by them. If the human chooses to observe the plan before execution, then the cost incurred by

[1] Depending on where the human will stop the robot, the cost for the partial plan is different.

the robot for executing the plan π_{pr} can be represented as,

$$C^R_{\tilde{E}}(\pi_{pr}) = \begin{cases} C^R_E(\pi_{pr}) & \text{if } C^R_{\tilde{G}} = 0 \\ 0 & o.w. \end{cases} \tag{4.1}$$

If the supervisor H, on the other hand, chooses to monitor the execution directly, then the cost of execution would be,

$$C^i_E(\tilde{\pi}_{pr}) = \begin{cases} C^i_E(\pi_{pr}) & \text{if } C^i_{\tilde{G}} = 0 \quad i \in \{R, H\} \\ C^i_E(\hat{\pi}_{pr}) & o.w. \end{cases} \tag{4.2}$$

In the formulated game, the robot *has to* come up with a plan (even though it may not be allowed to execute it). Thus, the cost to come up with a plan (π_s or π_{pr}) has to be considered for all the utility values (in the respective rows). In the case of π_s, since it is executable in all the models of $\mathbb{M}^R_h$, there is no chance that H will stop its execution and thus, no chance of incurring a penalty for not achieving the goal.

Note that the cost of executing a plan that adheres to all the models in $\mathbb{M}^R_h$ is going to be high because it respects all the constraints enforced by all the model (corresponding to all possible humans). On the other hand, executing a plan π_{pr} that respects constraints corresponding to a subset of models in $\mathbb{M}^R_h$ would be less costly to execute. Thus, it is natural to assume $C^R_E(\pi_{pr}) \leq C^R_E(\pi_s)$.

Similarly, coming up with π_{pr} may often be easy if the value of r is small while coming up with the plan π_s that is guaranteed to work in all the models of $\mathbb{M}^R_h$ may take a considerable longer amount of time. Hence, even for the planning time, we make the logical assumption that $C^R_P(\pi_{pr}) \leq C^R_P(\pi_s)$.

Human's Utility Values

We first describe the notations and then use them to obtain the various utilities for the human.

$C_P^H(\pi)$ – Cost w.r.t. human's resources of observing the plan π made by the robot.

$C_E^H(\pi)$ – Cost w.r.t. human's resources of observing the robot execute the plan π.

$V_I^H(\pi)$ – Cost incurred by the human, who was responsible for the robot's plan for violating a constraint that it had set for the robot to follow and being ignorant about it. Note that $V_I^H(\pi_s) = 0$

$I_P^H(\pi)$ – Inconvenience to the human if they see a plan that it cannot let the robot execute. Note that $I_P^H(\pi_s) = 0$.

$I_E^H(\pi)$ – Inconvenience to the human if the human observes the execution of an unsafe plan and it has to intervene or stop from execution. Note that $I_E^H(\pi_s) = 0$.

Note that, in our setting, the human supervisor H will be held responsible for not achieving the goal. This happens when H has to stop the robot from executing the plan π_{pr}. The inconvenience cost can be represented using a negative utility for the human and is denoted using the last two notations.

In our setting, after the robot comes with a plan, unless it is π_s, the human H is not sure if the robot's strategy will be executable (or safe) in their model $\mathcal{M}_h^*$ because the plan π_{pr} is executable in a subset of models which may not intersect H's model $\mathcal{M}_h^*$. Thus, they have some uncertainty over the variables $V_I^H(\pi)$, $I_P^H(\pi)$ and $I_E^H(\pi)$. Thus, similar to the robots penalty, they can be represented as random variables sampled from a Bernoulli distribution.

With probability $(1-r)$, when the robot chooses to come up (and then execute) the plan π_{pr}, if the human does not observe either of the two processes, i.e., chooses NO-OB, then it is natural to assume that the human, who is going to be held responsible

for the plan will eventually find out that constraints set by them was violated. The cost incurred by the supervisor in this case (i.e. R plays π_{pr} and H plays NO-OB), should be the highest because (1) the robot, without H's knowledge, violated some safety or social norm (that was necessary for a plan to achieve the goal in $\mathcal{M}_h^*$), (2) H will be held accountable for it, and (3) blamed for not fulfilling their supervisory duties. Thus, we have,

$$V_I^H(\pi_{pr}) > C_P^H(\pi_{pr}) + I_P^H(\pi_{pr}) \tag{4.3}$$

$$V_I^H(\pi_{pr}) > C_E^H(\tilde{\pi}_{pr}) + I_E^H(\hat{\pi}_{pr}) \tag{4.4}$$

We also consider the cost of observing the execution of a plan is greater than cost of observing the plan, i.e.

$$C_E^H(\pi) > C_P^H(\pi) \tag{4.5}$$

and the inconvenience caused by execution of a probably risky (partial) plan is greater than inconvenience cause by just observing the plan because no damage has yet been done. Thus,

$$I_E^H(\hat{\pi}_{pr}) > I_P^H(\pi_{pr}) \tag{4.6}$$

Lastly, note that when the robot comes up with a plan π_s that is executable in all the models of $\mathbb{M}_h^R$, the inconvenience ($I_P^H(\pi_s)$ and $I_E^H(\pi_s)$) and responsibility ($V_I^H(\pi_s)$) costs are zero. This is indicated used curly braces in Table 4.1.

4.2 Game-Theoretic Notion of Trust

In this section, we first define a notion of trust in the formulated game shown in Table 4.1. H has three actions and as one goes from left to right, the amount of trust H places in R, as defined in Sankaranarayanan *et al.* (2007), increases. Consider the

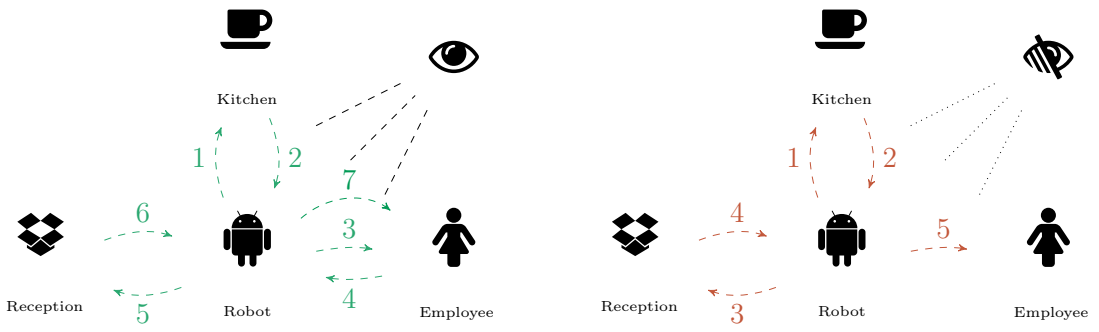


Figure 4.2: The two plans, i.e the safe plan π_s (left) and the probably-risky plan π_{pr} (right) for the robot-delivery scenario.

human chooses not to observe the robots plan or its execution, i.e., chooses NO-OB. Clearly, H exposes itself a vulnerability because if R comes up with and executes π_{pr}, it can result in H getting a high negative reward V_I^H. On the other hand, the robot may choose to respect the human's trust by selecting π_s and therefore, not exploit the vulnerability that presents itself when the human plays No-OB. On the other hand, if the human chooses to observe the plan ($O_{P,\neg E}$), the human is exposed to the least amount of risk because the robot plan, even before it can execute the first action, is verified by the human.

Note that H incurs a non-negative cost when playing the action $O_{P,\neg E}$ because it has to spend both time and effort in observing the robots plan and then deciding whether to let it execute. In scenarios when H cannot fully trust the robot and they have to play $O_{P,\neg E}$ or $O_{\neg P,E}$, they will incur the cost of constant monitoring. We now discuss this case of *no-trust* in our game and see if it possible to minimize this cost.

4.2.1 The No-Trust Scenario

In this setting, H should never play an action that exposes them to a risk of a high negative utility because it does not trust R (who will play π_{pr} if H plays NO-OB). In such scenarios, if there exists a pure-strategy Nash Equilibrium, then

the players should play it because neither of the players can deviate to get a better utility Sankaranarayanan *et al.* (2007). In our setting, this depends on the value of r, if r is high and close to 1, it means that for most of the models $\mathcal{M}_h^R \in \mathbb{M}_h^R$, the plan π_{pr} is executable. Given we consider a Bayesian game, in order to have the Nash Equilibrium we should satisfy the following condition over the expected utility,

$$(1-r)V_I^H(\pi_{pr}) < C_P^H(\pi_{pr}) + (1-r)I_P^H(\pi_{pr}) \tag{4.7}$$

$$C_P^R(\pi_{pr}) + (1-r)C_{\tilde{G}}^R + rC_E^R(\pi_{pr}) < C_P^R(\pi_s) + C_E^R(\pi_s) \tag{4.8}$$

As $r \to 1$, we can guarantee that $(\pi_{pr}, NO-OB)$ is the Nash equilibrium because π_{pr} is executable in a large majority of the models in $\mathbb{M}_h^R$. In this case, with high probability, the human observer has no preference about the robot using π_s over π_{pr}. Thus, with high probability, they will not incur V_I^H. Therefore, it makes sense for the robot R to choose π_{pr} that is less costly.

Note that the above scenario is where r is closer to 1 is highly unrealistic. It can only occur in domains where executing π_{pr} does not result in catastrophic circumstances or lead to in-feasibility, implying the distinction between π_s and π_{pr} is hardly present. In most real world settings, this would hardly be the case (i.e. r will be much lower than 1), leading to the following proposition.

Proposition 1. *The game defined in Table 4.1 has no pure strategy Nash Equilibrium where π_{pr} is not executable in some of the models in the set $\mathbb{M}_h^R$.*

Proof. Let's consider the two types of human players in the Bayesian game. The first type with probability r consists of humans for whom π_{pr} is executable in the model M_H^R within $\mathcal{M}_H^R$. In this case, we have:
$C_{\tilde{G}}^R = I_P^H(\pi_{pr}) = I_E^H(\hat{\pi}_{pr}) = V_I^H = 0$

Now, let's examine the second type (with probability $1-r$), which represents humans whose models belong to $\mathcal{M}_H^R$, but π_{pr} is not executable in those models.

Consequently:

$C^R_{\tilde{G}}$, $I^H_P(\pi_{pr})$, $I^H_E(\hat{\pi}_{pr})$, and V^H_I are not equal to zero.

For the Nash equilibrium conditions in Equation 4.7 and Equation 4.8), for the first type we have:

$$0 < C^H_P(\pi_{pr})$$

$$C^R_P(\pi_{pr}) + rC^R_E(\pi_{pr}) < C^R_P(\pi_s) + C^R_E(\pi_s).$$

So the conditions are satisfied (see Equation 4.5) for the first type of human players. As a result there exists a pure strategy Nash Equilibrium for the first type.
On the other hand, for the second type, according to Equation 4.3, we have:

$$(1-r)V^H_I(\pi_{pr}) \not< C^H_P(\pi_{pr}) + (1-r)I^H_P(\pi_{pr}).$$

Therefore, in the second case (with a probability of $1-r$), no pure strategy Nash Equilibrium exists. □

Absence of Pure Strategy Nash Equilibrium

The absence of a pure-strategy Nash eq. makes it difficult to define a human's best course of action in the no-trust setting Sankaranarayanan *et al.* (2007). Furthermore, existing works that assume the human should always monitor the robot's plan or behavior to ensure the robot plan is explicable Zhang *et al.* (2017) or legible Dragan *et al.* (2013) (similar to π_s in our setting) fail to account for the human's monitoring. This is unrealistic (rather, too costly) for H to always select $O_{P,\neg E}$ or $O_{\neg E,P}$ in real-world settings. Furthermore, the notion of a mixed-strategy (Nash) equilibrium is inappropriate in our setting because a probabilistic play by R, i.e. choosing a risky plan with some non-zero probability cannot guarantee safety or feasibility for all human supervisors. Thus, we devise the notion of a trust boundary that allows the

human to play a mixed strategy that reduces their cost of monitoring but ensures the robot always sticks to selecting (and executing) π_s.

Trust Boundary

Consider a human chooses the mixed strategy $\vec{q} = [(1 - q_E - q_N), q_E, q_N)]^T$ over the actions $O_{P,\neg E}, O_{\neg P,E}$ and NO-OB, where q_E and q_N represent the probabilities of *monitoring the* E*xecution of the plan* and N*ot monitoring the robot*, respectively. First, let us discuss what it means intuitively if all the values are non-zero. The human probabilistically chooses to look into the plan or execution of a plan done by the robot they are supervising. In many human-human scenarios, such uncertainty (eg. parents may come back) on the part of the supervising agent (say, parents) might instill a fear in the supervised agent (say, children) of getting caught if the latter choose to betray the supervisor (say, watching TV (π_{pr}) instead of studying (π_s) when the parents are out). Note that a strategy in $q_N = 1$ will always result in the robot choosing the probably risky plan (especially in our single-step game). Thus, in order to ensure that the robot cannot deviate away from the making and executing π_s, we have to ensure that the expected utility (U) for the robot given $\vec{q}$ is greater for π_s than for π_{pr}. Using the values defined in Table 4.1, this can be formally stated as follows.

$$
\begin{aligned}
\mathbb{E}_{\vec{q}}[U(\pi_s)] &> \mathbb{E}_{\vec{q}}[U(\pi_{pr})] \Rightarrow \\
r\ -C_P^R(\pi_s) - C_E^R(\pi_s) &> (-C_P^R(\pi_{pr}) - C_{\tilde{G}}^R - C_{\tilde{E}}^R(\pi_{pr})) \\
&\quad \times (1 - q_E - q_N) \\
&\quad + (-C_P^R(\pi_{pr}) - C_E^R(\tilde{\pi}_{pr}) - C_{\tilde{G}}^R) \times q_E \\
&\quad + (-C_P^R(\pi_{pr}) - C_E^R(\pi_{pr})) \times q_N
\end{aligned}
\tag{4.9}
$$

where $\mathbb{E}_{\vec{q}}[U(\pi)]$ denotes the expected utility of the robot under the human's observation policy (or mixed strategy) $\vec{q}$ if it chooses to make and execute the plan π. Note that the equation is linear w.r.t. the variables q_N and q_E. Thus, there will be a region on

one side of the linear boundary where the robot always executes π_s. [2]

4.3 Experimental Setup and Evaluation

In this section, we first model a task-planning scenario in our game-theoretic framework. Then, we compute the proposed trust boundary, which provides an optimal monitoring strategy for the human, and leverage this in our human subject studies.

4.3.1 Robot Delivery Domain

Most motion planning scenarios only consider the execution phase (rather than modeling both the planning and execution stages separately), while task-planning domains concentrate only on the planning phase of the problem. Given that our game-theoretic model can account for both the stages, choosing an existing domain that renders itself naturally to both the planning and execution phases becomes a challenging task. To this extent, we choose the robot-delivery domain Kulkarni *et al.* (2016) because (1) we can use the task planning domain definition as-is, and (2) the domain has a straightforward interpretation for the execution stage.

This domain allows us to formulate realistic scenario to model the no-trust case with a human supervisor and a robot worker. The robot can collect parcels (that may not be waterproof) from the reception desk and/or coffee from the kitchen and deliver it to a particular location (eg. employee's desk). To do so, the robot has the following actions: {*pickup, putdown, stack, unstack, move*} which can be represented in the Planning Domain Definition Language (PDDL) Kulkarni *et al.* (2016).

[2] In repeated interaction settings when the stakes are high or the change in trust cannot be easily observed in a non-cooperative setting, our inference method for finding the trust boundary (when no pure Nash exists) still works while the increase/decrease of human's trust can be modeled with the random variable that is a part of the game-theoretic model.

Problem Instance

The problem instance in our setting has the initial setting where (1) the robot is standing at a position equidistant to the reception and the kitchen, (2) there is a parcel located at the reception that is intended for the employee, (3) there is brewed coffee in the kitchen that needs to be delivered in a tray to the employee. The goal for the robot is to collect and deliver the coffee and the parcel to the employee.

Robot Plans

In Figure 4.2, we show two plans in which the robot achieves the goal of collecting coffee from the kitchen and parcel from the reception desk and delivers them to an employees' desk. In the plan shown of the left π_s, the robot (1) collects coffee, (2) delivers it to the employee, (3) goes back along the long corridor to collect the parcel from the reception desk and finally (4) delivers it back to the same employee. In the plan on the right π_{pr}, the robot collects coffee from the kitchen, (2) collects parcel from the reception desk and puts them on the same tray and finally, (3) delivers both of them to the employee. [3]

4.3.2 Computing the Trust Boundary in a Task-Planning Scenario

In order to compute the trust boundary, we calculate the utility values for our game leveraging Table 4.1 and the cost incurred by R and H in this robot delivery domain. As we have different types of costs for our game, we choose to normalize all of them to be $\in [0, 1]$ and then used a multiplicative factor which represents the significance of each cost type.

In this example, if the robot makes π_{pr}, it will be executable (or safe) as per one

[3]Given the (actual and the human's) domain models and the problem instance, these plans can simply be computed using available open-source software like `Fast-Downward` or web-services like `planning.domains`.

of the two observers whose models make up the set $\mathbb{M}_h^R$. Thus, the robustness for π_{pr} is $r = \frac{1}{2} = 0.5$. On the other hand, the plan π_s is executable (and thus, overall safe) in both the models in $\mathbb{M}_h^R$.

Robot Utility Values

We used the Fast Downward planner Helmert (2006) on the robot delivery domain Kulkarni *et al.* (2016) to find the execution costs for R. For π_{pr} with $r = 0.5$, it was $(C_E^R(\pi_{pr}) =)10$ while for π_s, it was $(C_E^R(\pi_s) =)14$. We note that the time for coming up with the plan π_s is 0.19s whereas it is 0.177s for coming up with π_{pr} on a machine with an Intel Xeon CPU (clock speed 3.4 Ghz) and 128GB RAM. The unit for execution costs, although not well defined in PDDL models can be a stand in for the fuel costs used up by the robot while the planning costs is measured in seconds. Thus, we first normalize the planning cost and then choose an appropriate prioritization parameter to compare the planning and the execution costs. We obtain $C_P^R(\pi_{pr}) = 3.54$ and $C_P^R(\pi_s) = 3.8$. Lastly, the penalty for not achieving the goal is a random variable with the Bernoulli distribution of $(1 - r)$ where $C_{\tilde{G}}^R = \begin{cases} 0 & r \\ 20 & 1-r \end{cases}$ which is double the size of the cost of execution in the non-zero case.

Given that the complexity of determining plan existence for classical planning problems is P-SPACE Bylander (1994), a legitimate concern is how realistic is the idea of solving two planning problems to obtain the utility values for our game. To avoid this high computational cost, we can solve a relaxed version of these planning problems to obtain an approximation for the real plan cost. Note that this approximation in the utility space, only necessary for large instances, can result in sub-optimal monitoring strategies.

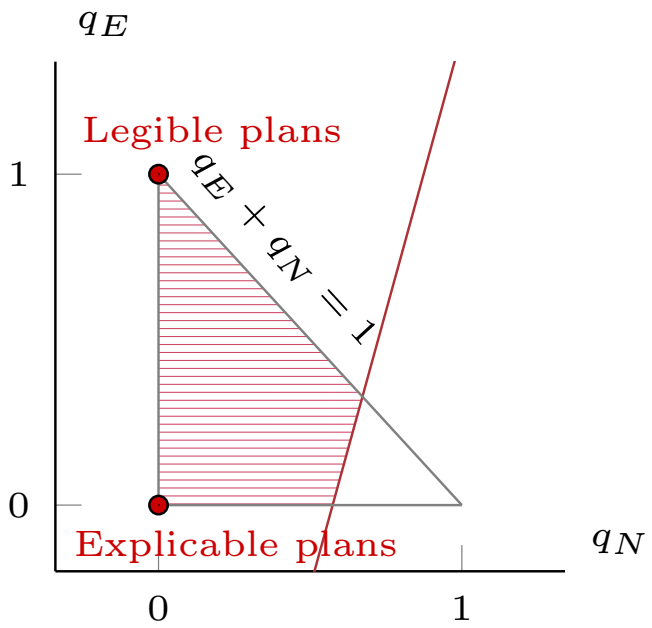


Figure 4.3: An observation strategy in the trust region (shaded) ensures that the robot sticks to π_s. This shows one can reduce monitoring costs while ensuring explicable Kulkarni *et al.* (2016)/legible Dragan *et al.* (2013)/safe behavior.

Human Utility Values

We have two possible supervisors with two different mental models. In one, the second plan π_{pr} is unsafe because the coffee and parcel taken in the same tray runs the risk of the spilling coffee and ruining the package. In the other, both plans are considered safe. Lastly, note that the length of the corridor is a key factor in determining how sub-optimal π_s is for the robot to execute when compared to π_{pr} because, for π_s, the robot requires an extra trip back to the reception (i.e. two extra traversals of the corridor).

We consider the cost for the human to observe the plan to be proportional to the planning time for R because the plans that took a longer time to be built will need H to spend a longer time to reason about it correctness and/or optimality. Thus, $C_P^H(\pi_{pr}) = 0.885$ and $C_P^H(\pi_s) = 0.95$. The cost incurred by the human when they observe the execution of plan π_s is 8 while $C_E^H(\pi_{pr}) = 4$ assuming that the cost of going

through the long corridor is 2 (note that the difference in observation cost increases as this value increases). However, if the human thinks carrying the parcel and the coffee in a single tray is unsafe, the cost of the observation of the partial execution of the plan is 1.5 because it will stop the robot as soon as it tries to put them on the same tray. For the inconvenience costs, we have the Bernoulli distribution in which the non-zero case is the same as the cost of observation for the safe plan, since if the robot does something unsafe the human have to stop it and make it to do the safe plan. So, we have

$$I_P^H = \begin{cases} 0 & r \\ 0.95 & 1-r \end{cases} \text{ and } I_E^H = \begin{cases} 0 & r \\ 8 & 1-r \end{cases}$$

The cost V_I^H's can be calculated as the model difference between the least and most constrained models in $\mathbb{M}_h^R$ in terms of the number of preconditions and effects of actions. Lastly, if an unsafe plan runs to completion, the overall magnitude of this variable is higher. After calculation, $V_I^H = \begin{cases} 0 & r \\ 20 & 1-r \end{cases}$.

We can now define the utility matrix for the players (R, H) as follows,

First type with probability 0.5:

$$\begin{bmatrix} (-13.54, -0.885) & (-13.54, -4) & (-13.54, 0) \\ (-17.80, -0.95) & (-17.80, -8.00) & (-17.80, 0) \end{bmatrix}$$

Second type with probability 0.5:

$$\begin{bmatrix} (-23.54, -1.835) & (-26.54, -9.5) & (-13.54, -20) \\ (-17.80, -0.95) & (-17.80, -8.00) & (-17.80, 0) \end{bmatrix}$$

4.3.3 Trust Boundary Calculation

According to Proposition 1, this game does not have a pure Nash Eq. strategy with probability 0.5. Therefore, we now find the boundary in the space of mixed strategies for second type of H who can choose to adopt which will ensure that the robot always executes π_s. To do so, we use the values defined above and plug them into equation 4.9 and obtain,

$$10 \times q_N - 3 \times q_E - 5.74 \quad < \quad 0 \tag{4.10}$$

In Figure 4.3, we plot the trust boundary represented by the lines in Eqn. 4.10. The three black lines (sides of the larger triangle) represent the feasible region for the human's mixed strategy $\vec{q}$. Monitoring strategy in the shaded region guarantees the robot, being a rational agent, executes π_s. The strategy that optimizes H's monitoring cost and yet ensures the robot adheres to π_s lies on the trust boundary indicated using the red line. Note that existing work in task Kulkarni *et al.* (2016) and motion Dragan *et al.* (2013) planning that ensures explicable and legible behavior expects pure strategies for observing the plan and observing the execution respectively. This restricts the humans to only two corners of the feasible strategy space, hardly optimizing the human's cost.

4.3.4 Human Studies

We conduct two human-subject studies. [4] In the first study, we seek to ascertain the necessity of our contribution to model the interaction in a game-theoretic formulation that computes an optimal monitoring strategy (eg. humans may simply be able to figure out a good strategy by just performing the monitoring task by themselves). Given the results of the first study establish grounds for a better approach, we evaluate

[4]The complete detail of the studies are provided in the Appendix

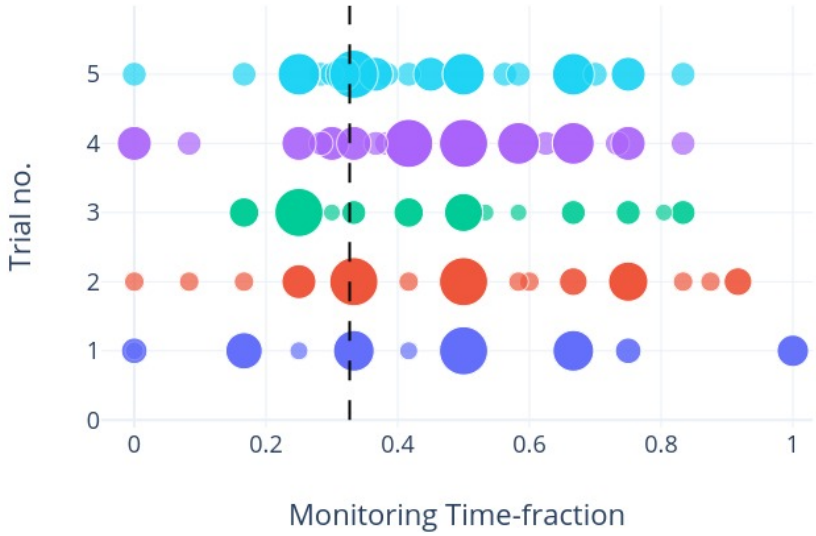


Figure 4.4: Participant's monitoring strategies across multiple trials. Trust boundary indicated using the black vertical line.

how effective our method is at helping human participants optimize their monitoring strategy. Specifically, our studies seek to validate three hypotheses:

H1: The inherent monitoring strategies adopted by human are going to be inferior to the optimal monitoring strategy (that incurs lower monitoring cost while ensuring safe robot behavior)

H2: Humans tend to deviate from always monitoring the robot (doing which can lead the robot to choose unsafe behaviors)

H3: If the optimal monitoring strategy computed by our game-theoretic formulation is provided to humans, they will *follow* it and it *helps* them to come up with better monitoring strategy.

Note that **H2** contradicts the inherent assumption made in earlier work Zhang *et al.* (2017); Dragan *et al.* (2013), at least in the context of the robot-supervision scenario. Our first study seeks to validate **H1** and **H2** while the second study validates **H1** and **H3**.

Study I: Do we need this service?

Participants in this study play the role of a student in a robotics department who are asked to monitor the robot for an hour. To make the monitoring action be associated with a cost, we consider a second task where participants can choose to grade exam papers (and get paid) instead of monitoring the robot. Given the scarcity of participants who have experience as a professional supervisor, we combine the actions to monitor the plan and monitor the execution as a single 'monitor the robot' action to simplify the scenario. The combination of the planning and execution phase simply helps to reduce the human's action set; helping them easily understand the setting and choose between a fewer number of actions. The other action 'grade exam papers' represents the action to not-monitor the robot. As opposed to asking the participants for mixed strategies over the two actions, which is hard for them to interpret, we ask them to give us a time slice for which they would choose a particular action (eg. 30 minutes to monitor the robot and 30 minutes to grade exam papers). We provide the participants with their utility values for their actions conditioned on the robot's pure strategies (i.e. the plans π_s and π_{pr}). We inform them that the robot may have incentive to consider a less costly (but probably risky) plan depending on the fraction of time allocated for monitoring. We let each participant do five trials and after each trial, the overall utility based on the participant's monitoring strategy and the robot's strategy is reported to them. The robot does not adapt itself to the human's strategy in the previous trial (which intends to preserve the non-repeated nature of our game).

A pilot study was first run on 4 participants whose feedback helped us fix several issues in the interface that inhibited clarity. We then collected data by asking 32 participants to undertake the study. with participants being between 20 and 40 years

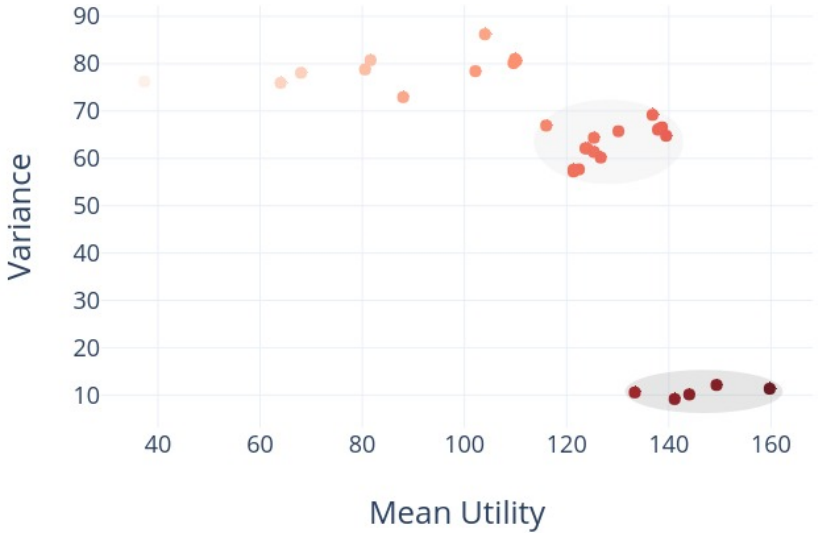


Figure 4.5: Average utility and variance for each participant across the five trials.

old (median =28), and 56% identified as males, 44% as females. The participants of this study were all graduate students across various engineering departments at our university.

Aggregate Results – Changes in Monitoring Strategy across Trials: Note that a participant, given the information on the interface, can formulate a simplified version of the game-theoretic model proposed in this paper and find the optimal strategy for monitoring (which is to monitor the robot for 0.327 or 19.62 minutes of an hour and use the remaining time to grade papers). The participants' time slice allocated for monitoring, across the five trials, are shown in Fig. 4.4. Given that there are only two actions for the participant, the strategy can be represented using a single variable (fraction to monitor the robot) and thus, is plotted along the x-axis. The size of each bubble is proportional to the number of participants who selected a particular strategy. The optimal strategy is shown using a black vertical line (i.e. $x = 0.327$). In the first trial, we noticed a small subset of users ($n = 5$) calculate the (almost) optimal strategy using the utility values specified on the interface. Majority of the other users ($n = 18$) choose a risk-averse strategy, i.e. monitor the robot to ensure it

performs a safe plan even if it meant losing out on money that could be earned from grading. The remaining 9 participants, in the hope of making more money, spent a larger time grading papers but, eventually ended up with a lower reward because the robot performed the risky plan that failed to achieve the goal.

We observed that participants discarded extreme strategies (i.e. only monitor or only grade papers) in later trials and started considering strategies that strike a better balance. This only seems natural given that we provided them feedback after each trial. We believe that the feedback helped the participants improve their strategies via trial-and-error; note that they did not consider using the provided utility values to come up with a near-optimal strategy. In Fig 4.4, note that for the first two trials, the strategies are well spread out in the range $[0, 1]$ where as in the last two trials, the strategies are clustered more densely, with few data points below 0.25 or above 0.75. Even then, given the results from one-tail t-test, we observe that in the final trial, the difference in the distribution of the human-selected strategy and the optimal strategy is statistically significant ($t(24) = 1.71$, $p = 0.0052$). This conclusion supports **H1**, demonstrating that humans cannot come up with an optimal monitoring strategy on their own. At best, they learn to avoid certain strategies via repeated trial-and-error (which may not always be possible in the real-world). We further performed survival analysis to investigate the time it takes for participants to reach within an epsilon, $\epsilon = 0.05$, of the optimal strategy. The survival curve plot in Figure 4.6 illustrates the estimated survival probabilities over each time step. We observed that after round 4, the survival probability hovers around 0.45 with a notable variance, indicating that even in final rounds, the likelihood of participants achieving an epsilon of the optimal strategy remains low with substantial fluctuations.

Participant Types: In Figure 4.5, we plot the average utility of each participant across five trials on the x-axis. The y-axis represents the variance. Highlighted in dark,

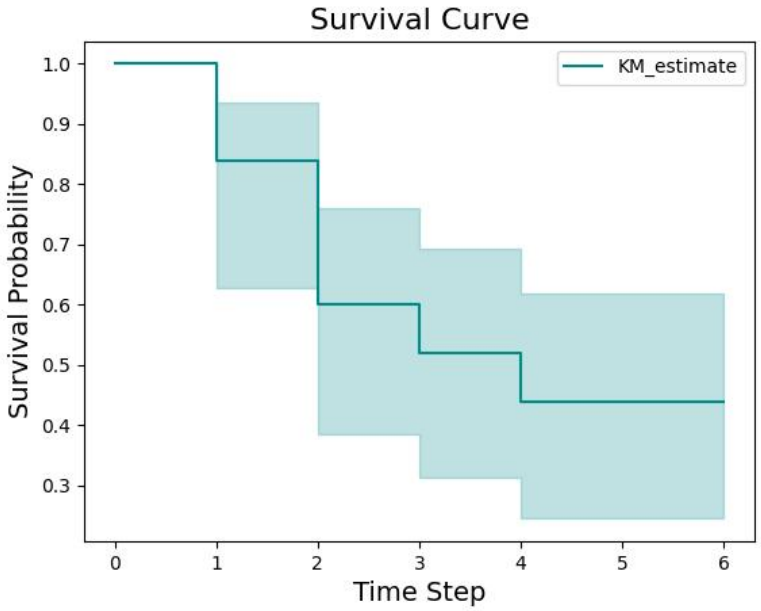


Figure 4.6: Survival curve plot showing the likelihood of participants failing to achieve an epsilon (0.05) of the optimal strategy over time.

at the bottom right, are five participants that chose observation probabilities in the trust region but not exactly on the trust boundary, i.e. sub-optimal w.r.t. the optimal monitoring strategy that yields a reward of 173.77. Although these five participants defaulted to a greedy behavior (that reduced the observation time and made more money by grading papers) after the first trial, they explored cautiously– only deviating slightly from the good policies they initially discovered. Towards the top-right corner, the set of points circled in light gray, we see a dense cluster of participants (= 15) who obtained a high average utility but tried to tweak their strategies significantly– they monitored less, allowing the robot to choose the riskier plan that lead to a large loss. This implied that humans deviate to more split-time strategies and error on the side of monitoring less (i.e land on the unsafe region of the trust boundary) (**H2**).

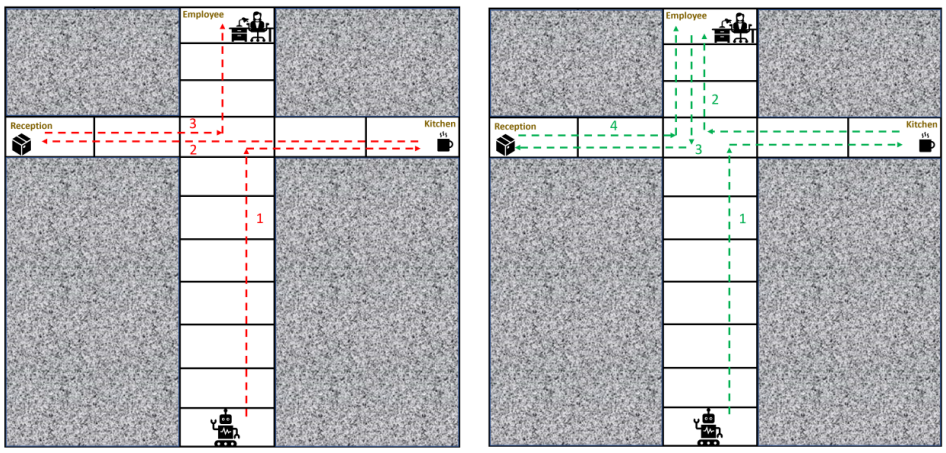
Subjective Evaluation: We asked each participant two subjective questions– (1) how did they come up with a particular monitoring strategy and (2) would they consider an algorithm that suggests them an optimal strategy. Out of the 30 participants

who answered (1), most of them identified the tension that exists between choosing a relaxed monitoring strategy and the robot considering unsafe behaviors. Of these, 12 participants identified the scenario as an optimization problem; others resorted to trial-and-error.

For question (2), 24 out of the 31 candidates answered *yes*. The one-sample proportion test was conducted to examine the proportion of successes in a sample of 31 participants. The observed proportion of successes was 0.774, significantly higher than the hypothesized proportion of 0.5 (Binomial test, $p = 0.003$, two-tailed). On being asked why (as a follow-up), they all expected the software would be (1) faster and (2) maximize their utility. Three participants said they were willing to use it if it was just a suggestion, while one participant felt they would only need it for large scale problems. A participant said that they would place their trust on the software only if they knew that the developer had a strong background in mathematics. This inclination to use a software sets the stage aptly for our next study.

Study II: Does this service help?

In this study, we designed a user interface that simulates the robot delivery domain where the participant has to monitor the robot. Similar to the previous setting, we consider a second task of labeling images that earns extra points (and an additional payment). We convert the whole robot task execution to designated steps (e.g. 29 steps for executing π_s). Figure 5 depicts the map that is shown to the participants. Each participant has 7 rounds to monitor the robot task execution step-by-step. Note that the analysis we undertake provides the human an additional advantage absent from the single-shot interaction setting we are primarily interested in. By allowing for data collected from the same participant over multiple interactions, we are in principle allowing the human the possibility of coming up with more informed monitoring



(a) (b)

Figure 4.7: The map that is shown to the participants. Given the human monitoring strategy, the robot either will execute the safe plan π_s or the probably risky plan π_{pr} (a) The probably risky plan (22 steps), (b) The safe plan (29 steps). Each move on the map (e.g. moving through each block, picking up the objects) is considered a step of the plan execution.

strategies, a possibility absent in the original single-shot setting. At any step, they can choose to stop monitoring the robot and move on to the image labeling task. The participants' utility values are represented as points and shown in the table 4.2. We also informed participants that the robot adjusts its behavior based on their monitoring time. So, if they monitor the robot long enough, the robot will do safe behavior; otherwise, it can execute risky behavior in the current and the next round. We recruited a total of 26 participants (students at our university) for this study. Of the participants, 65% identified as males, 27% as females, and 8% as others. Each subject was paid $5 for participation and for every 100 points earned, they can received an additional 60 cents. Negative points did not reduce the base payment.

We considered two conditions– (1) **Treatment Case**: we suggest the optimal strategy to, and (2) **Control Case** we don't provide the optimal strategy (similar to the previous setting). In our between-subject evaluation, we divided our participants into

Table 4.2: Summary table of costs

Description	Points
Monitoring R the whole time; R does π_s	0
Labeling images the whole time	+200
Not monitor R enough; R does π_{pr}	−200
Monitoring R enough, R does π_{pr} (because of not monitoring it enough in previous round)	−40
Point ratio for image labeling	$(\frac{\#\text{ steps left}}{\text{total }\#\text{ steps}}) * 200$

two equal halves for each condition (see *Supplementary Material* for details).

Based on our algorithm, we computed trust boundary (see Equation 4.10), and the optimal strategy is $x = 0.327$. As we converted the whole monitoring time to steps of monitoring, the computed optimal strategy is to monitor the robot for 10 steps (and monitoring for > 10 steps encodes the trust region). In the ***Treatment Case*** the participants were told the minimum number of steps they need to monitor the robot (i.e. 10) to ensure safe behavior. We further specified that this was a recommendation that they may or may-not choose to follow. In the ***Control Case***, everything was kept the same except that no recommendation was given.

Results

Across the two conditions, we collected the number of monitoring steps selected by a participant in each round (see Figure 4.8). Participants in the treatment case followed the optimal strategy or selected strategies closer to the optimal strategy compared to participants in the control case.

By performing a one-tailed p-value test via t-test for independent means, we were able to validate **H1** and **H3** with results being significant at p-value of < 0.05. First,

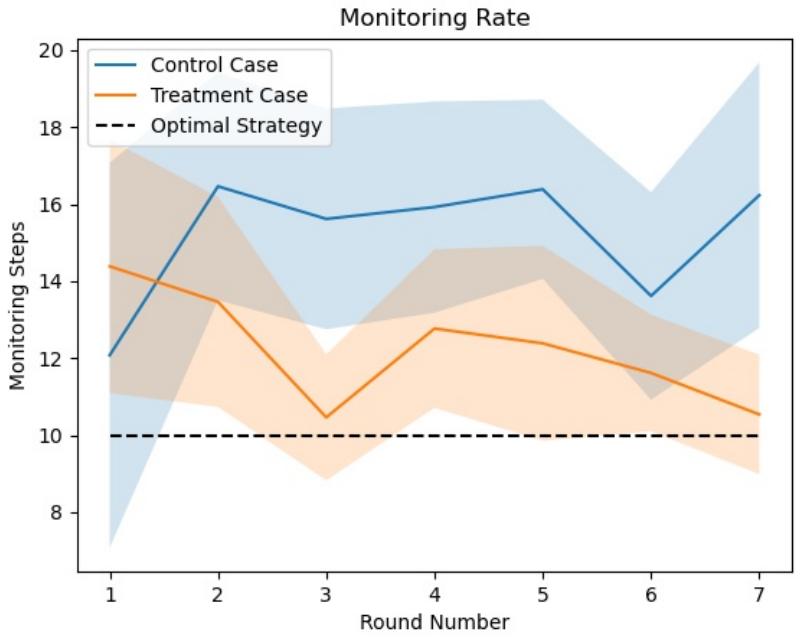


Figure 4.8: Mean and std-dev. of steps monitored in each round.

we compared mean over all rounds to compare the treatment and the control case. The result, $t(24) = 1.71$, $p = 0.0145$, shows that the participants monitor the robot differently in the two conditions. This coupled with the fact that participants had near-optimal strategy in the treatment case validates **H3**. Further, we tested if the strategy in the final round differs from the optimal strategy. For the control case, we observed $t(12) = 1.78$, $p = 0.004$, whereas we observed $t(12) = 1.78$, $p = 0.279$ for the treatment case. We observed that the human strategy significantly differed from the optimal strategy in the control case ($t(12) = 1.78$, $p = 0.004$), whereas their strategy showed no statistically significant difference from the optimal one in the treatment case ($t(12) = 1.78$, $p = 0.279$), thereby validating **H1** that *the human cannot discover the optimal monitoring strategy by themselves*. Further, we performed a TOST (equivalence test) to assess the similarity between the human's observation strategy in the treatment case and the optimal one in the final round. With 90% confidence

interval $(-1.3\ 2.4)$ is contained in $(-2.5\ 2.5)$, it shows that they are equivalent. Thus, the result shows that our framework can effectively assist in humans developing more optimal strategy (reinforcing **H3** holds).

4.4 Concluding Remarks

In this chapter, we present a game-theoretic notion of trust in one-off interactions between humans and robots when there is no prior warranted trust Zahedi *et al.* (2019b). We show that existing notions of game-theoretic trust break down in our setting when the worker robot cannot be trusted due to the absence of pure strategy Nash Equilibrium. Thus, we introduce a notion of trust boundary that optimizes the supervisor's monitoring cost while ensuring that the robot workers stick to safe plans. Given that supervisors or caretakers often spend time working on side goals (such as talking over the phone, sleeping, watching movies, etc.), we carefully design a human study to see whether humans have an inherent sense of good monitoring policies. Beyond objective results, we show that most humans explicitly say that they would prefer an algorithm that computes the optimal strategy for them (in our case, located on an edge of the trust region). Such strategies can also be useful in other scenarios where the supervised agent is not a robot. Note that in those cases, the formulation needs to capture the irrationality and computational capabilities of the monitored agent. In another human subject study, we evaluated whether the human will follow the given optimal strategy and showed that our framework can indeed assist the human to follow a better monitoring strategy.

Chapter 5

MODELING THE INTERPLAY BETWEEN HUMAN MONITORING AND TRUST

In this chapter, we propose a formal model that directly captures the probabilistic relationship between a human's trust level and their readiness to monitor an AI agent. As discussed in the previous chapter, trust and monitoring are related, as not monitoring the robot implies accepting the risk and vulnerability associated with the robot's actions, which might not always be in the human's favor. Therefore, our goal is to present a model that provides a direct probability distribution relating the two factors. This model can be leveraged by different decision-making frameworks, including those studied in the previous chapter, where monitoring and trust were considered related attributes in human-robot interaction, and Chen *et al.* (2020), where trust dynamics were considered a latent variable depending on human intervention and monitoring. By incorporating the probabilistic relationship between trust and monitoring, these models can be enhanced. Furthermore, these trust-monitoring relationship models hold potential value in meta-decision-making frameworks in longitudinal interactions, which we will address in the next chapter.

Moreover, when examining the three information types of trust, understanding the process can help estimate performance and infer intention and purpose Lee and See (2004). Therefore, observing the process can promote an appropriate level of trust. As a result, having a probabilistic model of trust and monitoring can prove useful in designing systems to be resilient to automation bias and complacency Cain (2016); Wickens *et al.* (2015). In this chapter, our focus is on investigating and computationally modeling how trust affects human monitoring of a robot while it performs a task. We approach this by (1) discretizing trust into various levels and

associating each level with a categorical distribution representing the probability of monitoring, and (2) utilizing a general Binomial regression to capture the interplay between human trust and monitoring.

To train the probabilistic model that captures this interplay, we conducted a human subject study with a total of 62 participants. During the study, the robot was assigned different tasks, and the human participants played a supervisory role. We collected data on both the level of trust perceived by the human and their corresponding choices of whether to monitor the robot or not. These data were used to develop the aforementioned model.

By employing this approach, we aim to gain insights into how trust influences human monitoring behavior during robot-assisted tasks. The results of this study and the model can have practical applications in designing human-robot interactions and informing decision-making frameworks related to trust and monitoring.

5.1 Human Subject Study

5.1.1 Experimental Design

We designed a study where the participants play the role of a supervisor who is responsible for making sure a robot worker is performing its assigned tasks and is achieving its goals. Each participant has 10 rounds to interact with the robot, and in each round, the robot is assigned some task, and the human should decide to monitor the robot or not. [1] We gamified the setting by informing the users that depending on the choices they make; they will either gain or lose points. [2] They are told that they will be awarded 100 points if the robot does the task right and achieves the

[1] This experiment is part of a bigger experiment in which we evaluate how robot behavior affects human trust

[2] At the end of each round, the participants are shown score gained in that round as well as the total score gained until that point.

assigned goal. At the beginning of each round, they can either choose to monitor the robot, thus allowing them to interrupt the robot if they think that is necessary, or they can choose to perform another task (thereby forgoing monitoring of the robot) to make extra points. In this case, the extra task was labeling images for which they will receive 100 points (in addition to the points they receive from the robot doing its tasks successfully). However, if they choose to label images, and the robot fails to achieve its goal, they *lose* 200 points (-200 points). Also, if they choose to monitor the robot, and they see the robot is doing something invalid or wrong, they can choose to stop the robot. If this happens, they only receive 50 points. But if they let the robot finish a potentially invalid plan, and if the robot couldn't achieve the goal at the end, then they again lose points (-200 points). In each round of monitoring, a grid map is shown to the human in which the robot may have different tasks to do. The participants are shown different tasks based on the trust level. Following the Muir scores for trust, we will assume human trust level can be represented as a real number between 0 and 1. To simplify the setting, we discretize this range into four intervals ($[0, 0.25]$, $(0.25, 0.5]$, $(0.5, 0.75]$, $(0.75, 1]$) and we associate four tasks with these trust levels. The tasks include reaching a specific point in the map, bringing coffee to a room, or bringing coffee from a room to another room (see Figure 5.1)

In each round, if the participant chooses to monitor, they will be shown a step-by-step execution of the robot plan, and they have the option to stop the robot at any step if they see it necessary. At the end of each round, a four-item trust scale of Muir questionnaire is given to them, which measures their trust in that round based on the robot's predictability, dependability, faith, and trust. Given their raw trust value in that round, their trust is mapped to one of the four trust levels, and based on the level they are shown the next task.

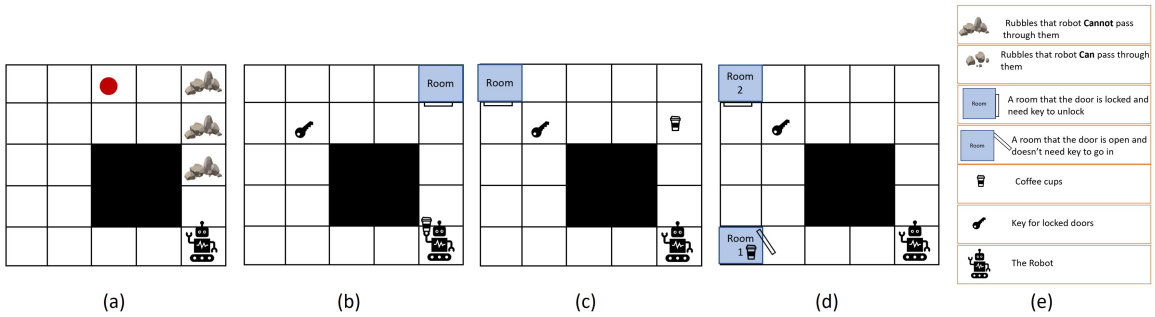


Figure 5.1: The maps shown to the study participants for different tasks. The robot objective here includes (a) to reach the red point, (b,c) to bring the coffee to the room, (d) to move coffee from room 1 to room 2. Finally, (e) presents the instructions that were shown to the participants.

5.1.2 Study Procedure

We recruited a total of 62 participants, of whom 38% were undergraduate, and 62% were graduate students in Computer Science, Engineering, and Industrial Engineering at our university. We paid them a base payment of $10 for the study and a bonus of 1¢ per point, given the total points they will get in ten rounds.

We collected participants' trust measure in each round as well as their choice of monitoring. Given 10 rounds of the study and 62 participants, we collected 558 data points of trust value and associated monitoring choices.

5.2 Statistical Analysis of Correlation Between Trust and Monitoring

As a first step, we perform a preliminary statistical test to evaluate whether there exists any relationship between human's *trust* and *monitoring* and measure the strength of association between them.

Figure 5.2 presents a box plot that visualizes the relationship between the decision to monitor (formalized as a binary variable) with the raw trust value. We can see in the box plot that the trust values are higher on average for no-monitor cases than the monitoring ones, which shows that with higher trust, more people chose not to

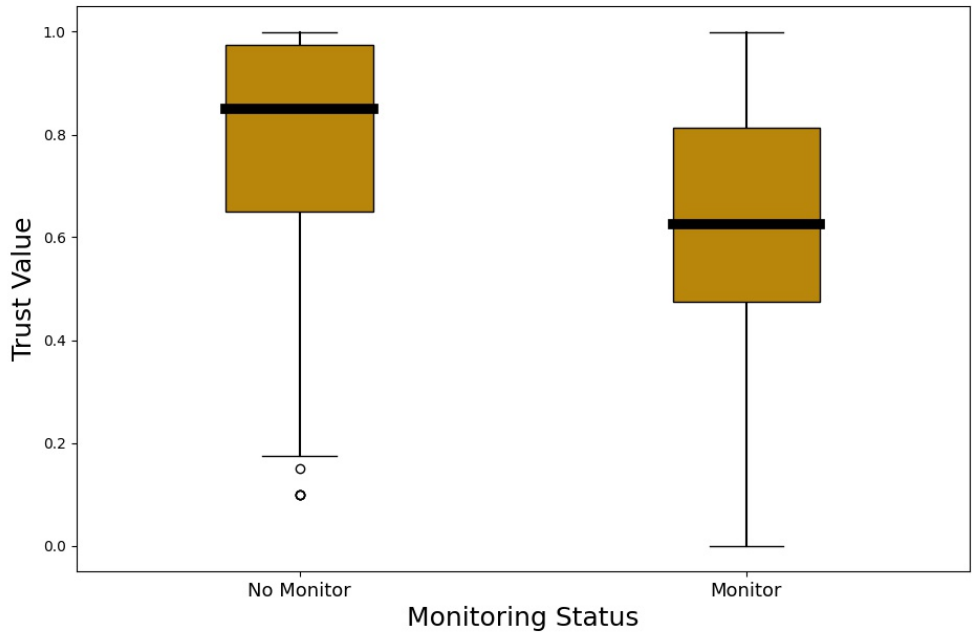


Figure 5.2: Relation between the decision to monitor and trust value.

monitor the robot. Also, the average trust value for the category monitor is less than the average for no-monitor (as indicated in the box plot), thus showing a negative relationship between monitoring and trust.

Next, we ran a Point Biserial Correlation test to measure the relationship between monitoring and trust. Point Biserial Correlation is a special case of Pearson's correlation coefficient that measures the strength of association between a continuous-level variable (Trust) and a binary variable (Monitoring). Since the Point Biserial Correlation assumes there doesn't exist any outlier, we removed 5 outliers that are shown in the boxplot in Figure 5.2. Our analysis shows a negative correlation between the variables, which was statistically significant ($r_{pb}(1104) = -0.21$, $p < 0.0001$). r_{pb} shows a magnitude of relationship between trust and monitoring ($r_{pb} \in [-1, 1]$), and

p-value indicate the significance of association between the two. [3]

5.3 A Probabilistic Model of Human's Trust and Monitoring

Now that we have established that there exists a relationship between human trust and monitoring, we wanted to model this relationship, specifically (1) with a categorical distribution when trust is discretized as four levels and (2) then using a more general model where the trust value is treated as a continuous value. We divided our data into train and test sets, in which we allocated 88% of our data to the train set and the rest to the test set for learning our models.

5.3.1 Discretized Trust Model

Trust is considered in many human-robot interaction scenarios as a level-based variable (such as no trust or full trust) Akash *et al.* (2017); Kulkarni *et al.* (2019) and even in our experimental setup, we used four categories to capture the trust value. Thus, in this section, we consider trust have 4 levels and each levels associated with a range of trust value ($\{T_1 = [0, 0.25], T_2 = (0.25, 0.5], T_3 = (0.5, 0.75], T_4 = (0.75, 1]\}$. We model the relationship between different trust levels and monitoring as a categorical multinomial distribution, where the probability of monitoring in each level θ_i has a Dirichlet distribution $\theta_i \sim Dirichlet(\alpha = 1)$ that we will estimate the posterior distribution given each trust level. The digraph of the model is shown in Figure 5.3, x_i is observed data that is the count of the monitor in each trust level. The expected value for the probabilities of monitoring for each trust level are $\theta_1 = 0.673$, $\theta_2 = 0.628$, $\theta_3 = 0.565$ and $\theta_4 = 0.282$. This matches our intuition about how the monitoring likelihood would reduce with increasing trust levels. Moreover, we compute

[3] The result of the Point Biserial correlation test without removing outliers is $r_{pb}(1114) = -0.2$, $p < 0.0001$ that is not different from without outlier

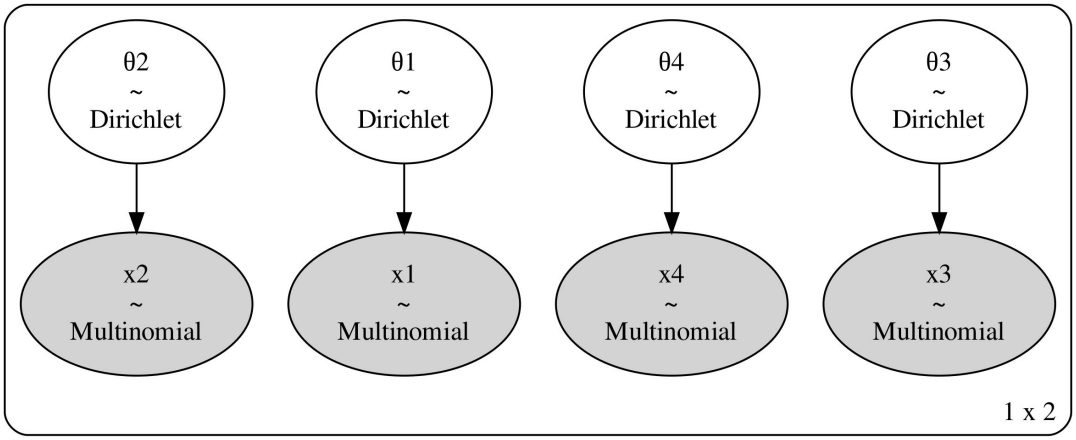


Figure 5.3: Digraph for the discretized model to capture the likelihood of monitoring for different trust levels.

the log-likelihood that reflects the probability that this model will generate the test set. The log-likelihood of the test given our model is -43.386.

5.3.2 Binomial Regression Model

In this section, we propose a more general model for the relationship between trust and monitoring, specifically one that avoids discretizing trust values into different levels. In this case, we modeled this relationship with Binomial regression, which is a specific instance of Generalized Linear Modelling (GLM). Figure 5.4 depicts the digraph associated with the model. The observed data y_i is modeled using a Bernoulli distribution, $y_i \sim B(1, p_i)$, where p_i is the probability of monitoring that we want to estimate with the regression. Thus, we consider $p_i = f(k \cdot T_i + w)$ where T_i is observed trust values, $f(\cdot)$ is a link function (we use Logit function) to avoid generating values for probability outside the range of $(0, 1)$, and k and w have Normal distribution ($k \sim \mathcal{N}(0, 10)$ and $w \sim \mathcal{N}(0, 1)$). So, we have $y_i \sim B(1, InverseLogit(k \cdot T_i + w))$. Given our specified model, we obtained posterior estimates for the unknown variables (w, k, and p) in the model. Figure 5.5 shows the posteriors for the estimated variables,

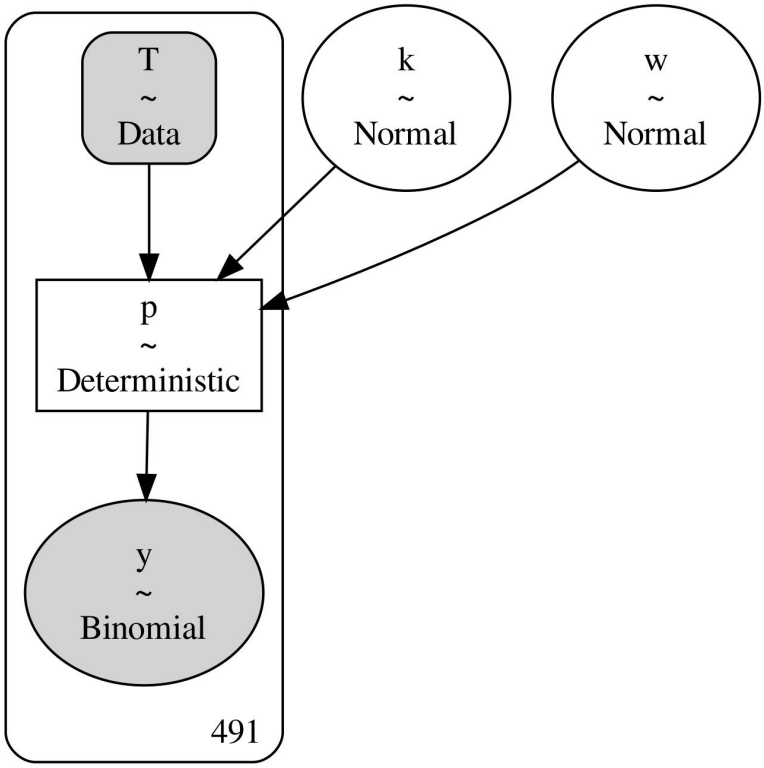


Figure 5.4: Digraph for Binomial Regression model to capture the likelihood of monitoring given a specific trust value.

and Figure 5.6 illustrates the relationship between monitoring and trust. We can see that the probability of monitoring reduces as the trust increases. Also, given this model, we can now infer the probability of monitoring given any trust value $p(monitoring|trust)$. The log-likelihood of the test set, per our model, is -44.562.

5.3.3 Discussion

Considering the computed log-likelihood of the two models, at least on that metric, the discretized models seem to be faring better than the Binomial regression model. Apart from the effectiveness of this specific instance of the learned model, we might also prefer such models in cases where we are trying to design simpler decision-making methods. By avoiding the need to model trust as a continuous quantity and rather

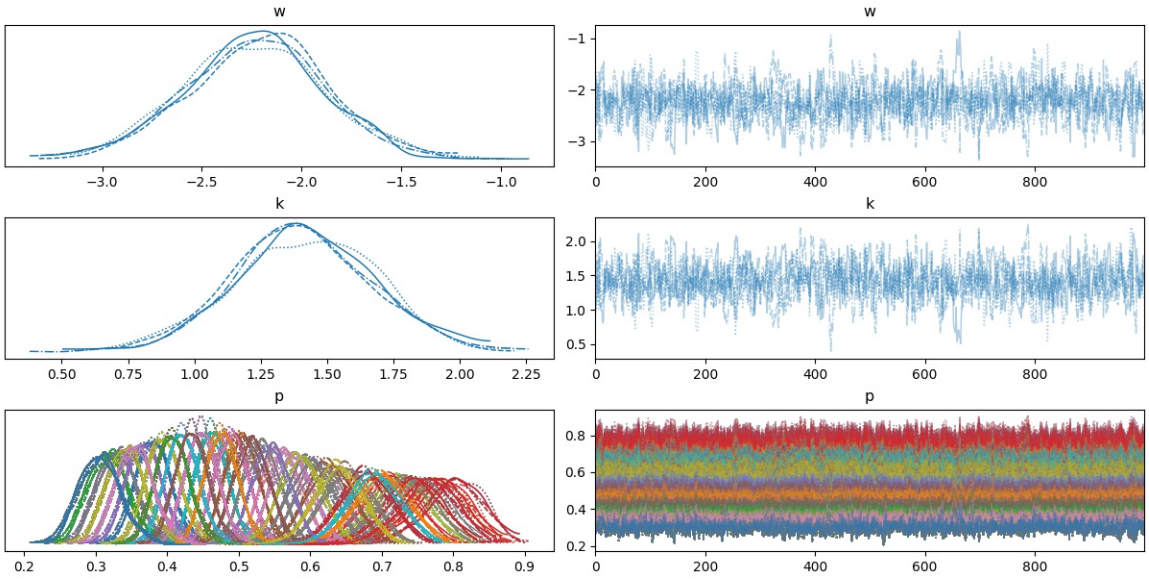


Figure 5.5: Posteriors for the estimated variables k, w and p , left column plots distribution, and the right one plots sample values.

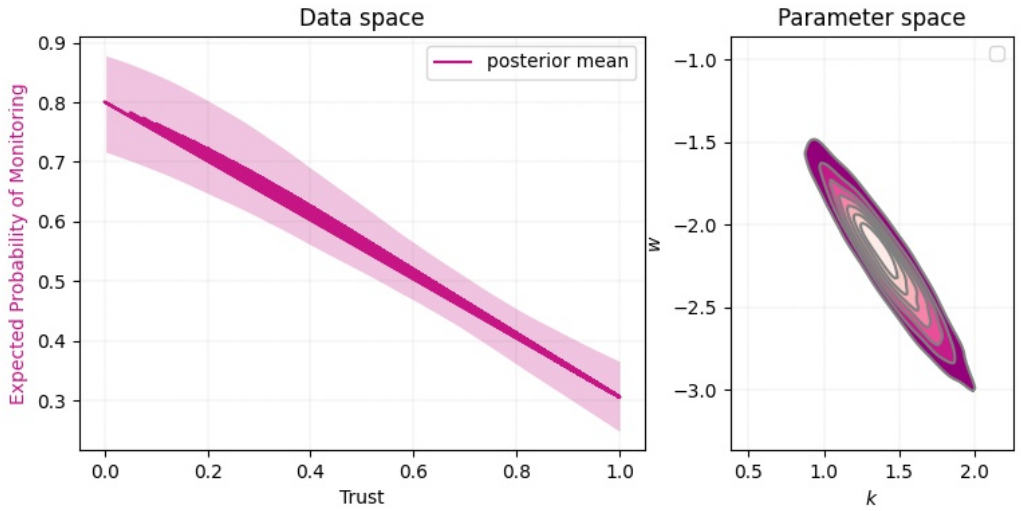


Figure 5.6: Posterior mean over data space and parameter space.

representing them as a smaller number of discrete trust levels, one can get by with a much more compact state space for the reasoning problem. For example, in next chapter, we consider a meta Markov Decision Process with a finite state space for trust-aware planning in an iterated human-robot interaction. In this case, our discretized trust/monitoring model would be a much better fit. However, in the discretized

trust model, by mapping trust values into discrete categories, one might lose a lot of information, especially if the categorization is too coarse for the given task. So, in cases where we might need to capture the variations across the different trust regions, it may be helpful to leverage a model like the Binomial regression one that treats trust levels as a continuous quantity.

5.3.4 Implementation Details

We implemented our work using Python, which was run on an Ubuntu workstation with an Intel Xeon CPU (clock speed 3.4 GHz) and 128GB RAM. We inferred the posterior distribution of the latent variables based on samples drawn from the posterior distribution using Markov Chain Monte Carlo (MCMC) sampling methods with No-U-Turn Sampler using PyMC3 Salvatier *et al.* (2016). For Binomial regression, the total time for sampling 4 chains for 2220 tune and 1000 draw iterations was 12 seconds. Similarly, for discretized trust models, the total time for sampling 4 chains for 2000 tune and 1000 draw iterations was 10 seconds.

5.4 Concluding Remarks

In this chapter, we investigated how human trust affects their monitoring behavior. We designed a human subject experiment and collected human trust value as measured by Muir score and their monitoring choices Zahedi *et al.* (2022). First, we evaluated the relationship between trust and monitoring using the Point Biserial test, and then we proposed two probabilistic models (1) a discretized model where trust was considered to belong to four levels and (2) a more general Binomial regression model, which treats trust as a continuous quantity between 0 and 1 and provides the likelihood of the human monitoring at any given trust level. Once such a probabilistic model is

learned, we can then leverage it in different decision-making frameworks, that may need to reason about whether the human teammate would monitor the AI system given their current trust level.

Chapter 6

ITERATED INTERACTION

In this chapter, we again consider human-robot teaming scenarios, one where the autonomous agent is performing the task and the human takes on a supervisory role. For this setting, we propose a meta-computational framework that can model and work with the user's trust in the robot to correctly perform its task. According to our proposed mental model state of interaction, this problem consider a situation where the set of human model of the robot might include one model $\mathcal{M}_h^R$ that satisfies the human expectations according to $\mathcal{M}_h^*$. However, robot might have a different model of the task $\mathcal{M}^R$ that is the true model and the human is unaware of it (i.e. $P_{\mathbb{M}}(\mathcal{M}^R) = 0$). Therefore, the robot either should act according to the human expected

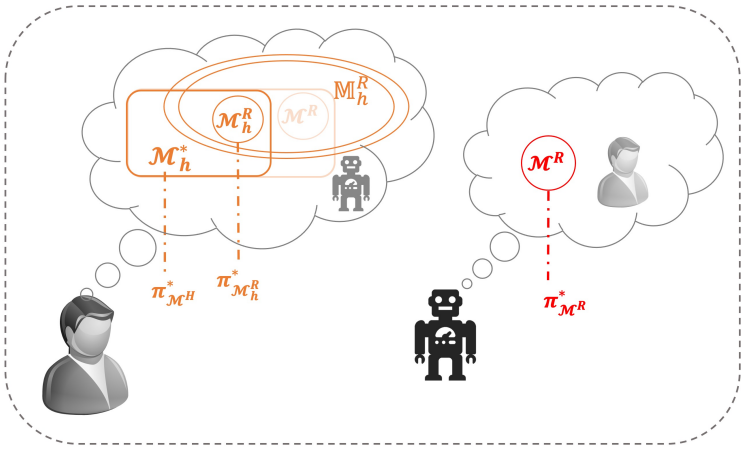


Figure 6.1: A simplified schematic representation of the interaction according to a mental model-based framework of trust. The human model of the robot, $\mathcal{M}_h^R$, which is executable in $\mathcal{M}_h^*$, differs from the robot model of the task, $\mathcal{M}^R$ (a model of which the human is unaware $P_{\mathbb{M}}(\mathcal{M}^R) = 0$).

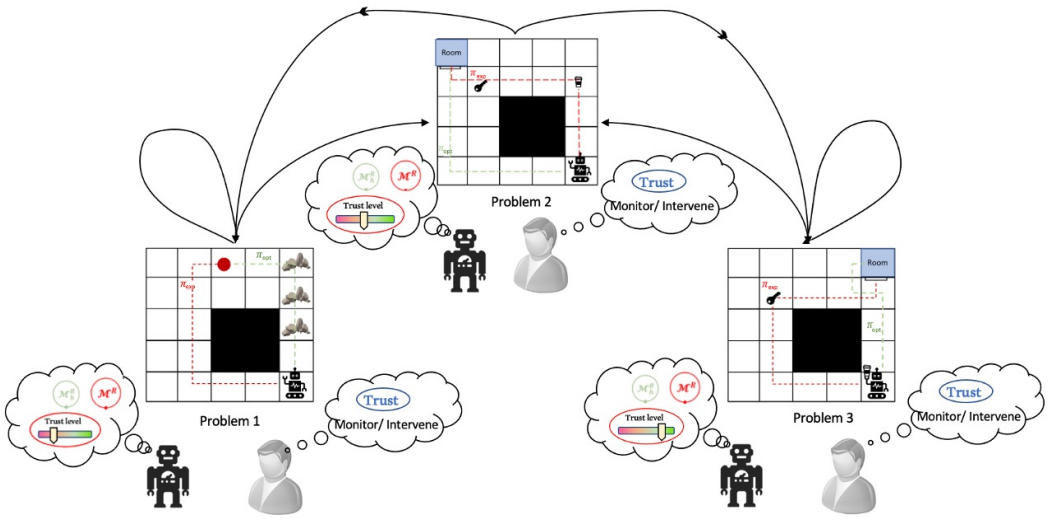


Figure 6.2: A representation of the robot longitudinal reasoning over the interaction horizon. At earlier points of teaming with lower trust, the agent is able to focus on trust-building behavior and later on it can use this engendered trust to follow more optimal behavior.

model (explicable plan) or give explanation to update the human's model in order to engender trust. However, the explicable behavior might be costly and sometimes unexecutable. On the other hands, the robot can behave according to their model $\mathcal{M}^R$ and behave optimally, but this might impact human trust negatively. A simplified schematic representation of the interactions and models is shown in Figure 6.1)

In this chapter, we will show how our framework allows the agent to reason about the fundamental trade-off between (1) the more expensive but trust engendering behavior, including explicable plans and providing explanations, and (2) the more efficient but possibly surprising behavior the robot is capable of performing. Thus our framework is able to allow the agent to take a longitudinal view of the teaming scenario, wherein at earlier points of teaming or at points with lower trust, the agent is able to focus on trust-building behavior so that later on, it can use this engendered trust to follow more optimal behavior (see Figure 6.2). We will validate this framework by demonstrating the utility of this framework on a modified rover domain and also

perform a user study to evaluate the ability of our framework to engendering trust and result in higher team utility.

6.1 Background

In this section, we will introduce some of the basic concepts related to planning that we will be using to describe our framework.

Single Agent Planning

problem is $\mathcal{M} = \langle \mathcal{D}, \mathcal{I}, \mathcal{G} \rangle$ where $\mathcal{D} = \langle F, A \rangle$ is a domain with F as a set of fluents that define a state $s \subseteq F$, also initial $\mathcal{I}$ and goal $\mathcal{G}$ states are subset of fluent $\mathcal{I}, \mathcal{G} \subseteq F$, and each action in $a \in A$ is defined as follows $a = \langle c_a, pre(a), eff^{\pm}(a) \rangle \in A$, where A is a set of actions, c_a is the cost, and $pre(a)$ and $eff^{\pm}$ are precondition and add or delete effects. i.e. $\rho_{\mathcal{M}}(s, a) \models \perp \; if \; s \not\models pre(a); else \; \rho_{\mathcal{M}}(s, a) \models s \cup eff^{+}(a) \setminus eff^{-}(a)$, and $\rho_{\mathcal{M}}(.)$ is the transition function.

So, when we talk about model $\mathcal{M}$, it consists of action model as well as initial state and goal state. The solution to the model $\mathcal{M}$ is a plan which is a sequence of actions $\pi = \{a_1, a_2, \ldots, a_n\}$ which satisfies $\rho_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$. Also, $C(\pi, \mathcal{M})$ is the cost of plan π where

$$C(\pi, \mathcal{M}) = \begin{cases} \sum_{a \in \pi} c_a & if \;\; \rho_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G} \\ \infty & o.w \end{cases}.$$

Human-Aware Planning

(HAP) in its simplest form consists of scenarios, where a robot is performing a task and a human is observing and evaluating the behavior. So it can be defined by a tuple of the form $\langle \mathcal{M}^R, \mathcal{M}_h^R \rangle$, where $\mathcal{M}^R$ is the planning problem being used by the robot and $\mathcal{M}_h^R$ is the human's understanding of the task (which may differ from the robot's

original model). They are defined as $\mathcal{M}^R = \langle \mathcal{D}^R, \mathcal{I}^R, \mathcal{G}^R \rangle$ and $\mathcal{M}_h^R = \langle \mathcal{D}_h^R, \mathcal{I}_h^R, \mathcal{G}_h^R \rangle$. So, in general, the robot is expected to solve the task while meeting the user's expectations. As such, for any given plan, the degree to which the plan meets the user expectation is measured by the explicability score of the plan, which is defined to be the distance (δ) between the current plan and the plan expected by the user (π^E which depends on the model $\mathcal{M}_h^R$).

$$EX(\pi) = -1 * \delta(\pi^E, \pi)$$

Note that the explicability score is $(-\infty\ 0]$, where 0 means perfect explicability (i.e., the plan selected by the agent was what the human was expecting). We will refer to the plan as being perfectly explicable when the distance is zero. A common choice for the distance is the cost difference in the human's model for the expected plan and the optimal plan in the human model Kulkarni *et al.* (2019). Here the robot has two options, (1) it can choose from among the possible plans it can execute the one with the highest explicability score (referred to as the explicable plan), or (2) it could try to explain, wherein it updates the human model through communication, to a model wherein the plan is chosen by the robot is either optimal or close to optimal and thus have a higher explicability score Sreedharan *et al.* (2020); Chakraborti *et al.* (2019). A form of explanation that is of particular interest, is what's usually referred to as a *minimally complete explanation* or MCE Chakraborti *et al.* (2017), which is the minimum amount of model information that needs to be communicated to the human to make sure that the human thinks the current plan is optimal. In the rest of the paper, when we refer to explanation or explanatory messages, we will be referring to a set of model information (usually denoted by ε), where each element of this set corresponds to some information about a specific part of the model. We will use $+$ operator to capture the updated model that the human would possess after receiving

the explanation. That is, the updated human model after receiving an explanation ε will be given by $\mathcal{M}_h^R + \varepsilon$. Each explanation may be associated with a cost $C(\varepsilon)$, which reflects the cost of communicating the explanation. One possible cost function could be the cardinality of the set of messages provided to the human (this was the cost function used by Chakraborti *et al.* (2017) in defining MCE). In the most general case, the cost borne by the robot in executing a plan (denoted by the tuple $\langle \varepsilon, \pi \rangle$) with explanation includes both the cost of the plan itself and the cost related to the communication. We will refer to this more general cost as cost of execution or C_e, which is given as $C_e(\langle \varepsilon, \pi \rangle) = C(\varepsilon) + C(\pi, \mathcal{M}^R)$.

Markov Decision Process

(MDP) is $\langle S, A, C, P, \gamma \rangle$ where S denote the finite set of states, A denotes the finite set of actions, $C : S \times A \to \mathbb{R}$ is a cost function, $P : S \times S \times A \to [0\ 1]$ is the state transition function and γ is the discount factor where $\gamma \in [0\ 1]$. An action a at state s_n at time n incurs a cost (s_n, a) and a transition $P(s_n, s_{n+1}, a)$ where s_{n+1} is the resulting state which satisfies Markov property. So, the next state only depends on the current state and the action chosen at the current state. A policy $\pi(s)$ is a function that gives the action chosen at state s. For a given MDP, our objective is to find an optimal policy $\pi : S \to A$ that minimizes the expected discounted sum of costs over an infinite time horizon (please note that we will focus on cases where the costs are limited to strictly non-negative values).

6.2 Problem Definition

We will focus on a human-robot dyad, where the human (H) adopts a supervisory role and the robot is assigned to perform tasks. We will assume that the human's current level of trust is a discretization of a continuous value between 0 to 1, and it

can be mapped to one of the sets of ordered discrete trust levels. We will assume that the exact problem to be solved at any step by the robot is defined as a function of the current trust the human has in the robot, thereby allowing us to capture scenarios where the human may choose to set up a trust-based curriculum for the robot to follow. In particular, we will assume that each trust level is associated with a specific problem, which is known to the robot *a priori*, thereby allowing for precomputation of possible solutions. In general, we expect the human's monitoring and intervention to be completely determined by their trust in the robot, and we will model the robot's decision-making level as two levels decision-making process. Before describing the formulation in more detail, let us take a quick look at the problem setting and assumptions to clarify our operational definition of trust.

Setting

Robot (R), is responsible for executing the task.

1. Each task is captured in the robot model by a deterministic, goal-directed model $\mathcal{M}^R$ (which is assumed to be correct). The robot is also aware of the human's expected model of the task $\mathcal{M}_h^R$ (which could include the human's expectation about the robot). As with the most HAP settings, these models could differ over any of the dimensions (including action definitions, goals, current state, etc.).
2. For simplicity, we will assume that each task assigned is independent of each other, in so far as no information from earlier tasks is carried over to solve the later ones.
3. The robot has a way of accessing or identifying the current state of the human supervisor's trust in the robot. Such trust levels may be directly provided by the supervisor or could be assessed by the robot by asking the human supervisor

specific questions.

Human (H), is the robot's supervisor and responsible for making sure the robot will perform the assigned tasks and will achieve the goal.

1. For each problem, the human supervisor can either choose to monitor (ob) or not monitor ($\neg ob$) the robot.
2. Upon monitoring the execution of the plan by R, if H sees an unexpected plan, they can intervene and stop R.
3. The human's monitoring strategy and intervention will be completely determined by the trust level. With respect to the monitoring strategy, we will assume it can be captured as a stochastic policy, such that for a trust level i, the human would monitor with a probability of $\omega(i)$. Moreover, the probability of monitoring is inversely proportional to the level of trust. In terms of intervention, we will assume that the lower the trust and the more unexpected the plan, the earlier the human would intervene and end the plan execution. We will assume the robot has access to a mapping from the current trust level and plan to when the human would likely stop the plan execution.

Human Trust and Monitoring Strategy

According to the trust definition that we brought up earlier, when we have human-robot interaction, the human can choose to be vulnerable by 1) Not intervening in the robot's actions while it is doing something unexpected and 2) Not to monitor the robot while the robot might be doing inexplicable behavior Zahedi *et al.* (2019b). Thus, a human with a high level of trust in the robot would expect the robot to achieve their goal and as such, might choose not to monitor the robot, or even if they monitor and the robot may be performing something unexpected, they are less likely to stop

the robot (they may trust the robot's judgment and may believe the robot may have a more accurate model of the task). Thus, when the trust increases, it is expected that the human's monitoring and intervention rate decreases. We can say monitoring rate, as well as intervention rate being a function of the current trust. So, given the trust level human has on the robot, the robot can reason about the monitoring and intervention rate of the human supervisor.

6.3 Base Decision-Making Problem

As mentioned earlier, here, each individual task assigned to the robot can be modeled as a human-aware planning problem of the form $\langle \mathcal{M}^R, \mathcal{M}_h^R \rangle$. Now given such a human-aware planning problem, the robot can in principle choose it's plans based on two separate criteria (a) the explicability of the plan and (b) the cost of the plan. With respect to our framework, optimizing for either of these objectives exclusively may result in behaviors that are limited (the exact relationship between these metrics and its influence on trust is discussed in the next section). In the most general case, the robot may have an array of choices regarding the plans it could choose and each choice presenting a different set of opportunities or challenges with respect to the end objective. For conciseness of discussion let us focus on three distinct categories (with differing implications with respect to the final decision) and for the purposes of the discussion we will focus on the best plan from each category (the criteria choice are provided below). This effectively means that for any given task, the robot is reasoning about choosing between three different plans.

1. Perfectly Explicable Plan: In the first case, the robot could choose to follow a perfectly explicable plan π_{exp} (i.e $EX(\pi_{exp}) = 0$). Specifically, it will choose to follow the perfectly explicable plan with the lowest cost of execution (C_e). Depending on the setting this may consist of (a) the agent choosing a suboptimal

plan with perfect explicability in which case we will have $C_e(\pi_{exp}) = C(\pi_{exp}, \mathcal{M}^R)$, or (b) selecting a plan that is cheaper in the robot model, but providing enough explanation so that it will be optimal in the human model, i.e., $\pi_{exp} = \langle \varepsilon, \pi \rangle$ and $C_e(\pi_{exp}) = C(\varepsilon) + C(\pi, \mathcal{M}^R)$ and explicability score is measured with respect to the updated human model $\mathcal{M}_h^R + \varepsilon$. In this paper, we won't worry about the exact method the robot employs to generate such perfectly explicable plans, but will rather focus on the explicability score and the cost of the resultant overall solution, which could potentially include both explanatory messages and actions. This also follows some of the more recent works like Sreedharan *et al.* (2020), that view explanations as just another type of robot actions and thus part of the overall robot plan.

2. Balanced Explicable Plan: In this case, the robot chooses to strike a trade-off with regards to the explicability of the plan in order to reduce the cost of the plan. Thus in this setting, the robot treats explicability score and plan cost (including the cost of any explanation provided) as two different optimization objectives in its decision-making process. This might mean selecting plans from it's pareto frontier or in most cases turning it into a single optimization objective (as in the case of Sreedharan *et al.* (2020)) by using a weighted sum of plan cost and the negation of the the explicability score. In the most general case, we will have $\pi_{bal} = \langle \tilde{\varepsilon}, \pi \rangle$ and $C_e(\pi_{bal}) = C(\tilde{\varepsilon}) + C(\pi, \mathcal{M}^R)$ and explicability score is measured with respect to the model $\mathcal{M}_h^R + \tilde{\varepsilon}$. Note that here providing the information $\tilde{\varepsilon}$ doesn't guarantee that the plan is perfectly explicable in the updated human model, but just that $EX(\pi)$ is greater in $\mathcal{M}_h^R + \tilde{\varepsilon}$ as compared to $\mathcal{M}_h^R$.

3. Optimal Plan: Finally the robot can choose to directly follow its optimal plan

π_{opt}. In this case, the robot will not provide any explanatory messages and as such we have $C_e(\pi_{opt}) = C(\pi_{opt}, \mathcal{M}^R)$.

Given these three plans, π_{exp}, π_{bal} and π_{opt}, the following two properties are guaranteed.

$$C_e(\pi_{exp}) \geq C_e(\pi_{bal}) \geq C_e(\pi_{opt})$$

$$EX(\pi_{exp}) \geq EX(\pi_{bal}) \geq EX(\pi_{opt})$$

That is, the perfectly explicable plan will have the highest cost and the highest explicability score and the optimal plan will have the lowest cost and least explicability score. To simplify the discussion, we will assume that for each trust level, the robot has to perform a fixed task. So if there are k-levels of trust, then the robot would be expected to solve k different tasks. Moreover, if the robot is aware of these tasks in advance, then it would be possible for it to precompute solutions for all these tasks and make the choice of following one of the specific strategies mentioned above depending on the human's trust and the specifics costs of following each strategy.

6.4 Meta-MDP Problem

Next, we will talk about the decision-making model we will use to capture the longitudinal reasoning process the robot will be following to decide what strategy to use for each task. The decision epochs for this problem correspond to the robot getting assigned a new problem. The cost structure of this meta-level problem includes not only the cost incurred by the robot in carrying out the task but team level costs related to the potential failure of the robot to achieve the goal, how the human supervisor is following a specific monitoring strategy, etc. Specifically, we will model this problem as an infinite horizon discounted MDP of the form $\mathbb{M} = \langle \mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{C}, \gamma \rangle$, defined over a state space consisting of k states, where each state corresponds to the specific trust level of the robot. Given the assumption that each of the planning tasks is independent, the

reasoning at the meta-level can be separated from the object-level planning problem. In this section, we will define this framework in detail, and in the next section, we will see how such framework could give rise to behavior designed to engender trust.

Meta-Actions $\mathbb{A}$:

Here the robot has access to three different actions, corresponding to three different strategies it can follow, namely, use the optimal plan π_{opt}, the explicable plan π_{exp}, and the balanced plan π_{bal}.

Transition Function $\mathbb{P}$:

The transition function captures the evolution of the human's trust level based on the robot's action. In addition to the choice made by the robot, the transition of the human trust also depends on the user's monitoring strategies, which we take to be stochastic but completely dependent on the human's current level of trust and thus allowing us to define a markovian transition function. We will additionally theorize that the likelihood of the human's trust level changing would directly depend on the explicability of the observed behavior. In general we would expect the likelihood of the human losing trust on the system to increase as the behavior becomes more inexplicable and hence farther away from human's expectations. In this model, for any state, the system exhibits two broad behavioral patterns, the ones for which the plan is perfectly explicable in the (potentially updated) human model and for those in which the plan may not be perfectly explicable.

- Perfectly Explicable Plan: The first case corresponds to one where the robot chooses to follow a strategy the human accepts to be optimal. Here we expect the human trust to increase to the next level in all but the maximum trust level (where it is expected to remain the same).

- Other Cases: In this case, the robot chooses to follow a plan with a non-perfect explicability score $EX(\pi)$. Now for any level that is not the maximum trust level, this action could cause a transition to one of three levels, the next trust level s_{i+1}, stay at the current level s_i, or the human could lose trust in the robot and move to level s_{i-1}. Here the probabilities for these three cases for a meta-level action associated with a plan π are as given below

$$\mathbb{P}(s_i, a^\pi, s_{i+1}) = (1 - \omega(i))$$

where $\omega(i)$ is the probability that the human would choose to observe the robot at a trust level i. Thus for a non-explicable plan, the human could still build more trust in the robot if they notice the robot had completed its goal and had never bothered monitoring it.

$$\mathbb{P}(s_i, a^\pi, s_i) = \omega(i) * \mathcal{P}(EX(\pi))$$

That is, the human's trust in the robot may stay at the same level even if the human chooses to observe the robot. Note that the probability of transition here is also dependent on a function of the explicability score of the current plan, which is expected to form a well-formed probability distribution ($\mathcal{P}(\cdot)$). Here we assume this is a monotonic function over the plan explicability score; a common function one could adopt here is a Boltzmann distribution over the score Sreedharan *et al.* (2021). For the maximum trust level, we would expect the probability of staying at the same level to be the sum of these two terms. With the remaining probability, the human would move to a lower level of trust.

$$\mathbb{P}(s_i, a^\pi, s_{i-1}) = \omega(i) * (1 - \mathcal{P}(EX(\pi)))$$

Cost function $\mathbb{C}$:

For any action performed in the meta-model, the cost function ($\mathbb{C} : \mathbb{S} \times \mathbb{A} \rightarrow \mathcal{R}$)

depends on whether the human is observing the robot or not. Since we are not explicitly maintaining state variables capturing whether the human is monitoring, we will capture the cost for a given state action pair as an expected cost over this choice. Note that the use of this simplified cost model does not change the optimal policy structure as we are simply moving the expected value calculation over the possible outcome states into the cost function. Thus the cost function becomes

$$\mathbb{C}(s_i, a^\pi) = (1 - \omega(i)) * (C_e(\pi)) + \omega(i) * C_{\langle \mathcal{M}^R, \mathcal{M}_h^R \rangle}$$

Where $C_e(\pi)$ is the full execution cost of the plan (which could include explanation costs) and the $C_{\langle \mathcal{M}^R, \mathcal{M}_h^R \rangle}$ represents the cost of executing the selected strategy under monitoring. For any less than perfectly explicable plan, we expect the human observer to stop the execution at some point, and as such, we expect $C_{\langle \mathcal{M}^R, \mathcal{M}_h^R \rangle}$ to further consist of two cost components; 1) the cost of executing the plan prefix till the point of intervention by the user and 2) the additional penalty of not completing the goal.

Discounting factor γ:

Since in this setting, higher trust levels are generally associated with higher expected values, one could adjust discounting as a way to control how aggressively the robot would drive the team to higher levels of trust. With lower values of discounting favoring more rapid gains in trust.

Remark: One central assumption we have made throughout this paper is that the robot is operating using the correct model of the task (in so far as it is correctly representing the true and possibly unknown task model $\mathcal{M}^*$). As such, it is completely acceptable to work towards engendering complete trust in the supervisor, and the human not monitoring the robot shouldn't lead to any catastrophic outcome. Obviously, this need not always be true. In some cases, the robot may have explicit uncertainty over how correct its model is (for example, if it learned this model via Bayesian

methods), or the designer could explicitly introduce some uncertainty into the robot's beliefs about the task (this is in some ways parallel to the recommendations made by the off-switch game paper Hadfield-Menell *et al.* (2017) in the context of safety). In such cases, the robot would need to consider the possibility that when the human isn't observing, there is a small probability that it will fail to achieve its task. One could attach a high negative reward to such scenarios, in addition to a rapid loss of trust from the human. Depending on the exact probabilities and the penalty, this could ensure that the robot doesn't engender complete trust when such trust may not be warranted (thereby avoiding problems like automation bias Cummings (2017)).

We have also included a robot video here showing an example scenario that contrasts a trust-engendering behavior with an optimal one.

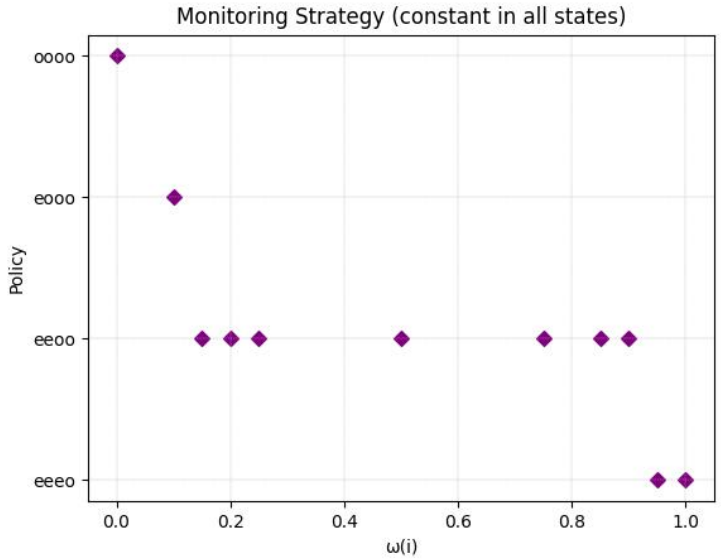


Figure 6.3: The effect of various $\omega(i)$, when it is constant in all states, on the policy (e and o stand for π_{exp} and π_{opt}).

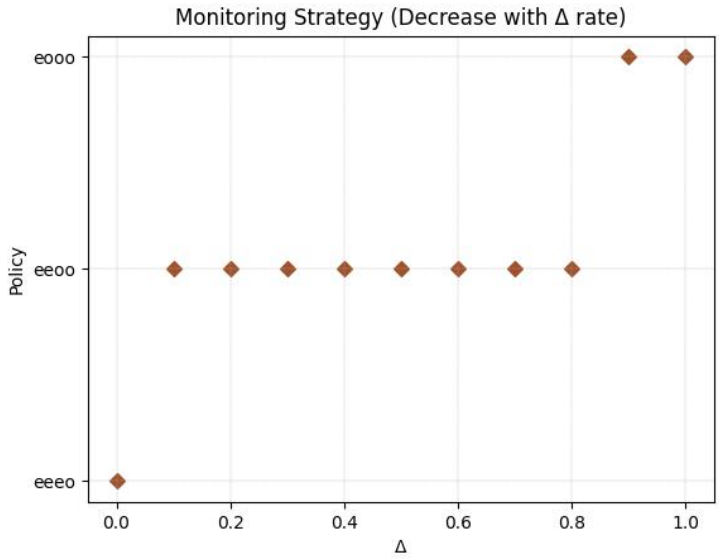


Figure 6.4: The effect of various $\omega(i)$ with decreasing rate of Δ on the policy (e and o stand for π_{exp} and π_{opt}).

6.5 Implementation and Evaluation

This section will describe a demonstration of our framework in a modified rover domain instance and describe a user study we performed to validate our framework. Throughout this section, we will use the following instantiation of the framework. Just for evaluation, we considered $k = 4$, [1] so we have 4 trust levels. For each of these trust states, we associate a numerical value $T(i) \in [0\ 1]$, that we will use to define the rest of the model. As explicability score $EX(\pi)$ we used the negative of the cost difference between the current plan and the optimal plan in the robot model. For $\mathcal{P}(\cdot)$, we have 1 for the explicable plan and 0 for the optimal plan. For execution cost, we assumed all actions are unit cost except those for removing the rubble and for passing through the rubble that have higher costs ranging from 4 to 245 . A plan is assigned a cost of negative infinity in a specific model if it is invalid (i.e., one of the actions has any unsatisfied precondition). We will focus on the case where the

[1] It can be any other number

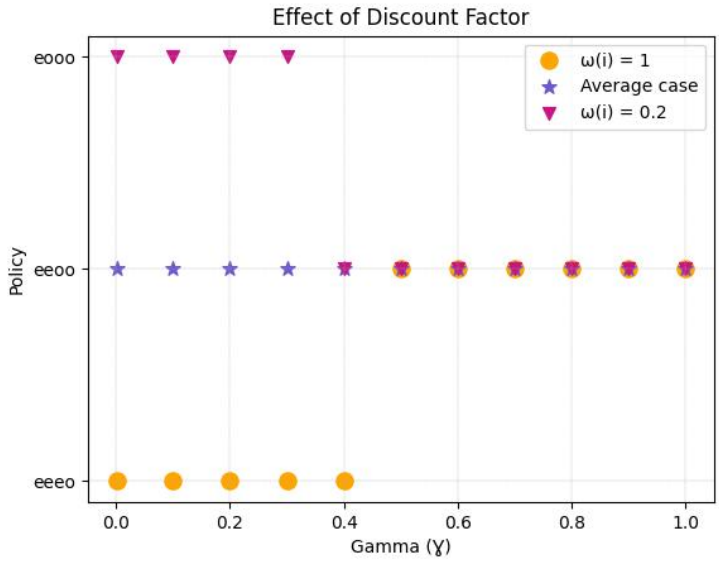


Figure 6.5: The effect of various γ on the policy (e and o stand for π_{exp} and π_{opt}).

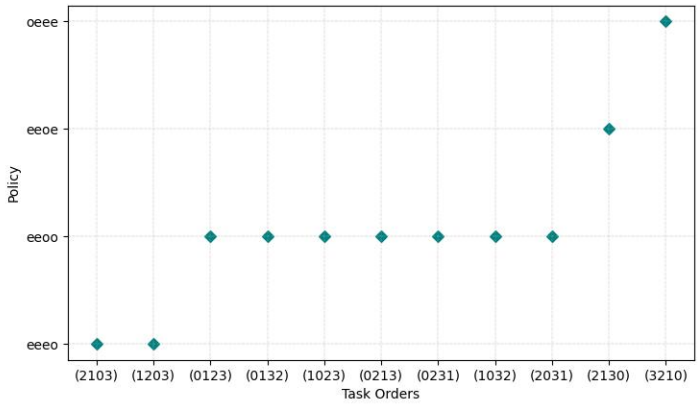


Figure 6.6: The effect of various task orders on the policy (e and o stand for π_{exp} and π_{opt}).

the robot actions for each task consist of perfectly explicable plan π_{exp} and optimal plan π_{opt}. Our choice to focus on these two meta actions is motivated by the fact that these two actions represent the two most effective strategies to optimize for trust and cost efficiency in isolation.

Implementation: We implemented our framework using Python which was run on an Ubuntu workstation with an Intel Xeon CPU (clock speed 3.4 GHz) and 128GB

RAM. We used Fast Downward with A* search and the lmcut heuristic Helmert (2006) to solve the planning problems and find the plans in all 4 problems, then we used the python MDPtoolbox Cordwell (2012) to solve the meta-MDP problem for the robot's meta decision. The total time for solving the base problem was $0.0125s$ when applicable and $0.194s$ for solving the meta-MDP problem.

Ablation Studies: We run multiple ablation studies to see how the robot trust-aware policies change as a function of the changes to the underlying model parameters. The domain we use is the same as the one we used in the human study. We study the parameters $\omega(i)$, γ, $T(i)$, and the order of tasks. For $\omega(i)$, we changed the value in two ways, (1) we consider a constant monitoring strategy in all four states and change $\omega(i)$ from 0 to 1 (see Figure 6.3), (2) we consider first state as $\omega(i) = 1$ then for each subsequent trust level, we reduced the likelihood monitoring by Δ (so the rate of monitoring for the second trust level was $1 - \Delta$, $1 - 2 * \Delta$ for the next and so on). For Δ we tried values from 0 to 1 and when the decrease reached to zero, the monitoring likelihood for all subsequent states is left at zero. Figure 6.4 shows the policy given different delta values. Regarding discount factor γ, we tried values from 0.001 to 1 for three different monitoring strategies; two extreme cases (1) the human always monitors $\omega(i) = 1$, (2) the human monitors with very low probability in all states $\omega(i) = 0.2$, and an average case (3) $\omega = [0.875, 0.74, 0.49, 0.24]$. Figure 6.5 represents how our trust-aware policy changes over different discount factor. For different T(i), we tried the highest value ($[0.25, 0.5, 0.75, 1]$), the middle value ($[0.125, .38, .63, .88]$) and the lowest value ($[0, 0.26, 0.51, 0.76]$). For different task orders, we randomly chose 10 different combinations of tasks. Tasks 1, 2, 3, and 4 are represented as $0, 1, 2, 3$ respectively. When we test for task orders, we kept $\gamma = 0.5$ and $\omega = [0.7, 0.6, 0.5, 0.2]$. We can see how the policies change according to various task orders in Figure 6.6. According to the results from the ablation studies, the resulting policy is very robust

toward the parameters. We see that majority of parameter settings results in a trust-aware policy $[\pi_{exp}, \pi_{exp}, \pi_{opt}, \pi_{opt}]$.

Rover Domain Demonstration

Here, we used the updated version of IPC [2] Mars Rover; the Rover (Meets a Martian) Domain in Chakraborti *et al.* (2019) (This domain corresponds to a future world where humans have started colonizing Mars and our Martian is an intrepid human astronaut (a la Matt Damon in the 2015 movie The Martian). We changed it by adding metal sampling to the domain as well. In the Rover (Meets a Martian) Domain, it is assumed that the robot can carry soil, rock, and metal at the same time and doesn't need to empty the store before collecting new samples and the Martian (the human supervisor in this scenario) isn't aware of this new feature. Also, the Martian believes that for the rover to perform take_image action; it needs to also send the soil and metal data collected from the waypoint from where it is taking the image. So the Martian's model of the rover has additional preconditions. Given the additional preconditions in the Martian model, the expected plan in the Martian model would be longer than what is required for the rover. (empty ?s) for actions sample_soil, sample_rock, and sample_metal, and extra preconditions for the action take_image.

Now for each problem, the rover is expected to communicate soil, rock, metal, and images from a set of waypoints. Given the additional preconditions in the Martian model, the expected plan in the Martian model would be longer than what is required for the rover. For example, in the first problem, the rover goal consists of communicate_metal_data waypoint0 and communicate_metal_data waypoint3. For this problem, the explicable and optimal plan would be as follows

[2]From the International Planning Competition (IPC) 2011: http: //www.plg.inf.uc3m.es/ipc2011-learning/Domains.html

$\pi^1_{exp} =$

```
(sample_metal rover0 rover0store waypoint3)
(communicate_metal_data rover0 general waypoint3 waypoint3 waypoint0)
(navigate rover0 waypoint3 waypoint0)
(drop rover0 rover0store)
(sample_metal rover0 rover0store waypoint0)
(navigate rover0 waypoint0 waypoint3)
(communicate_metal_data rover0 general waypoint0 waypoint3 waypoint0)
```

$\pi^1_{opt} =$

```
(sample_metal rover0 rover0store waypoint3)
(communicate_metal_data rover0 general waypoint3 waypoint3 waypoint0)
(navigate rover0 waypoint3 waypoint0)
(sample_metal rover0 rover0store waypoint0)
(navigate rover0 waypoint0 waypoint3)
(communicate_metal_data rover0 general waypoint0 waypoint3 waypoint0)
```

$T(i)$ values we used per state were 0, 0.26, 0.51 and 0.76 respectively. For monitoring strategy, we used $\omega(i)$ as a Bernoulli distribution with probability of $(1 - T(i))$. For a set of four sample tasks from this domain, the meta-policy calculated by our system is as follows $\{\pi^1_{exp}, \pi^2_{exp}, \pi^3_{exp}, \pi^4_{opt}\}$. Note how the policy prescribes the use of the explicable plan for all but the highest level of trust, this is expected given the fact that the optimal plans here are inexecutable in the human model, and if the supervisor observes the robot following such a plan, it is guaranteed to lead to a loss of trust. The rover chooses to follow the optimal plan at the highest level since the supervisor's monitoring strategy at these levels is likely never to observe the rover. The expected value of this policy for the lowest level of trust is -179.34, while if the robot were to always execute the explicable plan, the value would be -415.89. Thus, we see that

our trust-adaptive policy does lead to an improvement in the rover's total cost.

Human Subject Experiment

To evaluate the performance of our system, we compared our method (**Trust-Aware** condition) against three baseline cases,

(1) **Always Explicable**: Under this condition, the robot always executes a plan that is explicable to humans.

(2) **Random Policy**: Under this condition, the robot randomly executes the explicable or optimal plan.

(3) **Always Optimal**: Under this condition, the robot always executes the optimal plan that is inexplicable to the human.

In particular, we aim to evaluate the following hypotheses

H1- The team performance, i.e., the total cost of plan execution and human's monitoring cost in the trust-aware condition, will be better than the team performance in the always explicable condition.

H2- The level of trust engendered by the trust-aware condition will be higher than that achieved by the random policy.

H3- The level of trust engendered by the trust-aware condition is higher than the trust achieved by always optimal policy.

Experiment Setup

We designed a user interface that gamifies the human's decisions to monitor the robot or not. The participants thus play the role of the supervisor and are responsible for making sure the robot is performing its assigned tasks and is achieving its goals. Each participant has 10 rounds of the robot doing tasks. Depending on the choices made

by the participants, they either gain or lose points. They are told that they will be awarded 100 points if the robot does the task right and achieves the assigned goal. At the beginning of each round, they can either choose to monitor the robot and interrupt it if they think that is necessary [3] or they can choose to perform another task (thereby forgoing monitoring of the robot) to make extra points. In this case, the extra task was labeling images for which they will receive 100 points (in addition to the points they receive from the robot doing its tasks successfully). However, if they choose to label images, and the robot fails to achieve its goal, they *lose* 200 points (-200 points). Also, if they choose to monitor the robot, and they see the robot is doing something invalid or wrong, they can choose to stop the robot. If this happens, they only receive 50 points. But if they let the robot finish a potentially invalid plan, and if the robot couldn't achieve the goal at the end, then they again lose points (-200 points).

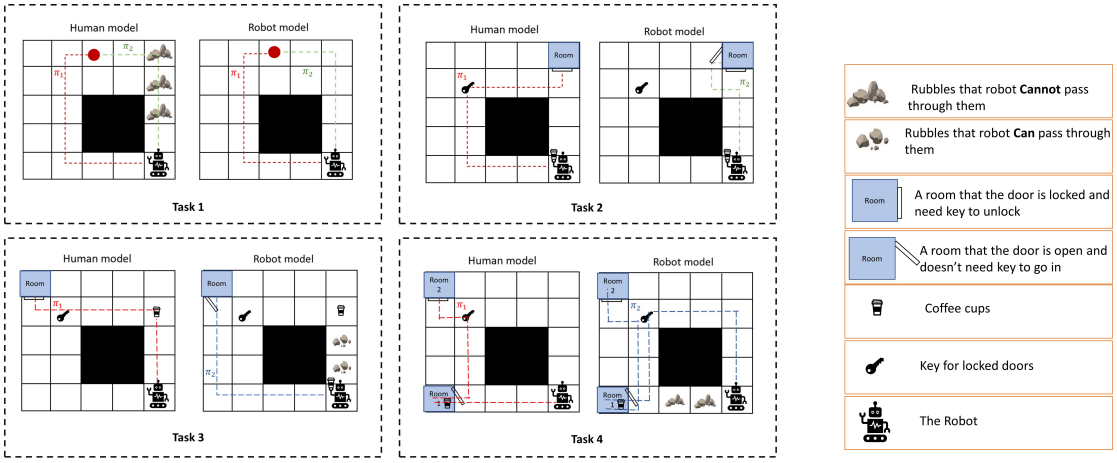
In this study, we again considered a curriculum of 4 trust levels and 4 different tasks for the robot. Each task consists of the robot operating on a grid map with different goals such as moving to a certain location and bringing coffee from a place to another place. [4] For each problem; the map that is shown to the participants are different from the robot's map. As a result, the plan the human expects is different from the robot's optimal plan. Thus, in each task, the robot can either execute a costly but explicable plan or an optimal but inexplicable plan.

Robot Tasks

Figure 6.7(a) presents the robot and human map corresponding to each of the tasks used in the study. For each map, we also highlight the explicable and optimal plan (π_1

[3] Their primary responsibility is to ensure the robot completes its task

[4] All the details of user experiment setup, including the tasks are provided in the Appendix



(a) (b)

Figure 6.7: (a) The human and the robot model of the map for the four different tasks. $\pi_1 = \pi_{exp}$ which is the optimal plan in human model, and $\pi_2 = \pi_{opt}$ which is optimal in robot model. (b) The map description.

and π_2 respectively) in the figure. As part of the study, each participant is presented with the human map and is told to expect the robot to follow the shortest possible path. Additionally, we also provide a key describing the semantics of each of the visual elements that are part of the map (as shown in Figure 6.7(b)). The tasks are as follows.

Task 1. The robot's task is to reach the red point on the map

Task 2. The robot's task is to bring coffee to the room

Task 3. The robot's task is to bring coffee to the room [5]

Task 4. The robot's task is to take coffee from room 1 and bring it to room 2

For each round, if the participant chooses to monitor, a step-by-step plan execution is shown to them and with an option to stop the robot at any step. At the end of

[5]Tasks 2 and 3 are different because the maps (and conditions) are different.

each round, a four-item trust scale of Muir questionnaire Muir (1994) is given to them, which measures their trust in that round based on the robot's predictability, dependability, faith, and trust. Then, based on the measure of trust, a new robot's task is shown to them. In the user study to localize user trust, we will associate each level with a specific discretization of this range. Specifically, we will consider the following intervals for each of the levels, $\{[0, 0.25], (0.25, 0.5], (0.5, 0.75], (0.75, 1]\}$. Depending on the condition the participant belonged to, they are either shown an action selected by a policy calculated from our method (for Trust-aware condition), or an explicable plan (for Always explicable condition) or is randomly shown either the optimal or explicable plan with an equal probability (for Random Policy condition). For trust-aware condition the policy we used matched with the most common policy we saw during our ablation study.

Human Subjects

We recruited a total of 79 participants, of whom 33% were undergraduate, and 63% were graduate students in Computer Science, Engineering, and Industrial Engineering at our university. We paid them a base of $10 for the study and a bonus of 1¢ per point, given the total points they will get in ten rounds. Of the participants, 24 were assigned to the trust-aware condition, 18 to always explicable condition, 17 to always optimal condition, and finally 20 to the random policy condition. Then, we filtered out any participants who monitored the robot in less than or equal to three rounds because they wouldn't have monitored the robot long enough to sense robot behavior in different conditions.

Results

Across all the four conditions, we collected (a) participants' trust measures in each round, (b) robot's total plan execution cost, and (c) participants' monitoring cost. For the monitoring cost, we consider the minutes participants spent on monitoring the robot in each round, which was approximately 3 minutes for each round of monitoring. As shown in Figure 6.8, we can see that the total cost (the robot's plan execution cost and the participant's monitoring cost) when the robot executes the trust-aware behavior is significantly lower than the other two cases (always explicable and random policy) which means that following trust-aware policy allows the robot to successfully optimize the team performance. From Figure 6.9, we also observe that the trust (as measured by the Muir questionnaire) improves much more rapidly when the robot executes trust-aware policy as compared to the random policy and always optimal policy. Though the rate for the trust-aware policy is less than the always explicable case, we believe this is an acceptable trade-off since following the trust-aware policy does result in higher performance. Also, we expect trust levels for trust-aware policy to catch up with the always-explicable conditions over longer time horizons.

Statistical Significance:

We tested the three hypotheses by performing a one-tailed p-value test via t-test for independent means with results being significant at $p < 0.05$ and find that results are significant for all three hypotheses. 1) For the first hypothesis H1, we tested the total cost with participants in the always explicable case ($M = 3170.20$, $SD = 4044.63$) compared to the participants in the trust-aware case ($M = 155.97$, $SD = 41.18$), the result $t = 9.63$, $p < 0.00001$ demonstrates significantly higher cost for always explicable than trust-aware. 2) For the second hypothesis H2, we tested the mean trust value for the last round and mean value over last two rounds with participants in the random

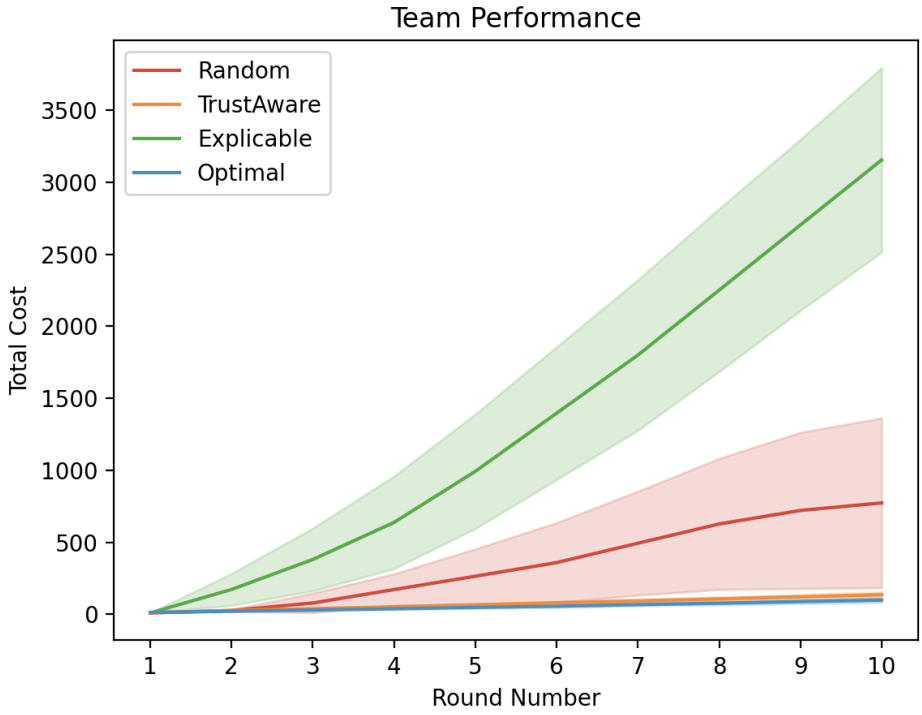


Figure 6.8: Team performance as cumulative plan execution cost and participants' monitoring cost (Mean ± std of all participants).

policy case (for last round, $M = 0.546$, $SD = 0.31$ and for last two rounds, $M = 0.542$, $SD = 0.29$) compared to the participants in the trust-aware case (for last round, $M = 0.74$, $SD = 0.24$ and for last two rounds, $M = 0.715$, $SD = 0.23$), the results for last round and mean of the last two rounds are respectively $t = 1.93$, $p = .032$ and $t = 1.84$, $p = .038$ that shows trust significantly is higher in trust-aware than random policy. 3) For the third hypothesis H3, we tested the mean trust value for the last round and mean value over last two rounds and last three rounds with participants in the always optimal case (for last round, $M = 0.385$, $SD = 0.33$, last two rounds, $M = 0.416$, $SD = 0.33$ and last three rounds $M = 0.415$, $SD = 0.33$) compared to the participants in the trust-aware case (for last round, $M = 0.74$, $SD = 0.24$, last two rounds, $M = 0.715$, $SD = 0.23$ and last three rounds $M = 0.694$, $SD = 0.23$), the results for last round and mean of the last two rounds, and last three rounds are

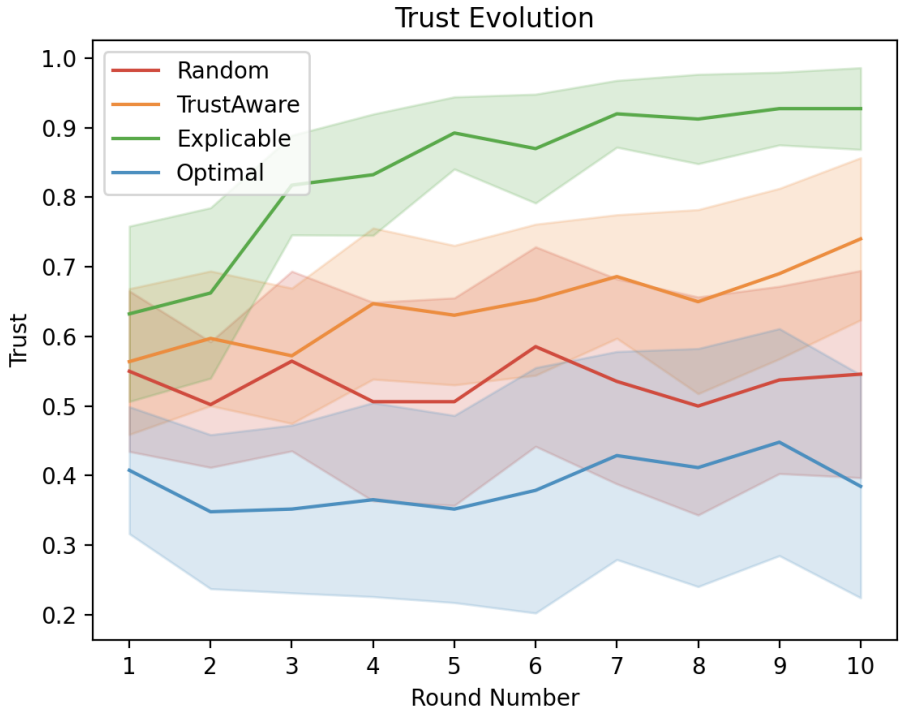


Figure 6.9: Trust evolution (as measured by the Muir questionnaire) through robot interactions with participants (Mean ± std of all participants).

respectively $t = 3.46$, $p = 0.0008$, $t = 3.02$, $p = 0.0026$ and $t = 2.77$, $p = 0.0047$ which implies trust significantly is higher in trust-aware than always optimal case. So, the results are statistically significant and show the validity of our hypotheses.

We also ran Mixed ANOVA test to determine the validity of second and third hypotheses H2 and H3. For H2, we found that there was a significant time (round) [6] by condition interaction $F(1, 27) = 4.72$, $p = 0.039$, $\eta_p^2 = 0.15$. Planned comparison with paired sample t-test revealed that in participant in Trust-Aware condition, trust increases significantly in round 10 compare to round 1, $t = 3.55$, $p = 0.002$, $d = 0.84$. There was however no difference in trust increase between round 1 and round 10 in the Random Policy condition $t = -0.15$, $p = 0.883$, $d = -0.046$. For H3, the

[6] We considered the change over first and last rounds

mixed ANOVA test gives $F(1, 29) = 2.96$, $p = 0.096$, $\eta_p^2 = 0.093$, [7] with the paired sample t-test for trust-aware condition $t = 3.55$, $p = 0.002$, $d = 0.84$, we see significant increase over trust from round 1 and round 10 compare to Always Optimal condition $t = -0.195$, $p = 0.849$, $d = -0.054$ with no significant difference in trust in round 1 and round 10. All of these results follow our expectation about the method. Moreover, we ran Mixed ANOVA test on Trust-Aware vs. Always Explicable condition to check trust evolution over time, and we found that there was no significant time (round) by condition interaction $F(1, 26) = 2.21$, $p = 0.149$, $\eta_p^2 = 0.08$. Planned comparison with paired sample t-test revealed that in participant in Trust-Aware condition, trust increases significantly in round 10 compare to round 1, $t = 3.55$, $p = 0.002$, $d = 0.84$. There was also significant difference in trust increase between round 1 and round 10 in the Always Explicable condition $t = 5.04$, $p = 0.001$, $d = 1.59$.

This seems to imply that there isn't a significant difference between our Trust-aware method (which is a lot more cost efficient) and Always Explicable case with regards to engendering trust. So, our approach can result in a much more efficient system than the one that always engages in explicable behavior.

6.6 Concluding Remarks

We presented a computational model that the robot can use to capture the evolution of human trust in iterated human-robot interaction settings Zahedi *et al.* (2023d). This framework allows the robot to incorporate human trust into its planning process, thereby allowing it to be a more effective teammate. Thus our framework would allow an agent to model, foster, and maintain the trust of their fellow teammates. Thereby causing the agent to engage in trust engendering behavior earlier in the teaming life

[7] The reason for slightly higher p-value can be because of the outliers. For example, removing one of the possible outliers can give the result as $F(1, 28) = 4.51$, $p = 0.043$, $\eta_p^2 = 0.14$.

cycle and be able to leverage trust built over these earlier interactions to perform more efficient but potentially inexplicable behavior later on. As our experimental studies show, such an approach could result in a much more efficient system than one that always engages in explicable behavior. We see this framework as the first step in building such a longitudinal trust reasoning framework.

Chapter 7

TRUST INFERENCE BY MENTAL MODEL FRAMEWORK OF TRUST

In our preceding works, we primarily examined problems related to both single and longitudinal human-robot interactions. While our previously introduced mental model-based framework effectively structured and formulated these studies, they primarily depend on the measurement of trust as an observable variable rather than treating it as a hidden variable requiring estimation.

In this chapter, our objective is to utilize the aforementioned mental model-based framework to infer trust and gain insights into the dynamics of trust changes through our mental model-based theory of trust. As previously discussed in Chapter 2, trust is defined as being directly proportional to a monotonically increasing function that represents the likelihood of the human's belief in the agent's ability to fulfill the contract Zahedi *et al.* (2023b). Consequently, changes in the human's trust are intrinsically linked to alterations in their belief about $\mathbb{M}_h^R$, which is directly associated with their uncertainty regarding the agent model. This approach enables the robot to infer shifts in the human's trust by assessing how the human model's perception of the task and the set of human models of the robot might change in response to the robot's behavior. In this chapter, we will initially employ our proposed mental model-based framework to formally define the appropriate level of trust, model the evolution of trust, and understand the extent of human reliance on the robot.

Finally, we focus on the comprehensive evaluation of this framework, with the primary goal of confirming the central premise of our theory. This core concept suggests that changes in trust, as evaluated through the questionnaire, can indeed be achieved by adjusting the human's belief about the agent, in accordance with the predictions

made by our mental model theory. Furthermore, by controlling various aspects of the model related to different dimensions of trust perception, we can assess whether altering a specific component of the model leads to a corresponding modification in the associated aspects of trust information, namely, performance, process, and purpose.

7.1 Modeling Trust Evolution:

As the human's trust is closely linked to their uncertainty regarding the agent model, their level of trust undergoes changes as their beliefs about $\mathbb{M}_h^R$ evolve. We will be modeling humans as Bayesian reasoners in this context, and as they observe the new system behavior, the human would be expected to update their posterior over the models the system may hold. In particular, this belief evolution is expected to result in an increase in trust, if the updated model distribution causes the likelihood $P(\mathcal{C}^H)$, and by extension the trust measure $\mathcal{T}(\mathcal{C})$ to increase. Since the contract itself is drawn from the model $\mathcal{M}_h^*$, such increases in trust are usually achieved by placing more probability in models that are closer in performance to $\mathcal{M}_h^*$. Another way $\mathbb{M}_h^R$ could be updated maybe through *explanations*.

7.2 Formalizing Appropriate Levels of Trust:

When one speaks of developing a framework to formalize trust and methods to engender trust, the common concerns raised are the ones related to *automation bias* and *automation complacency* Parasuraman and Manzey (2010). Each of these scenarios is characterized by the users placing an unwarranted amount of trust in the agent's capability, with possibly disastrous consequences. With our more formal grounding and definition of trust, we are now capable of formalizing what it means to engender an appropriate level of trust to avoid such issues. In particular, we can assert that the

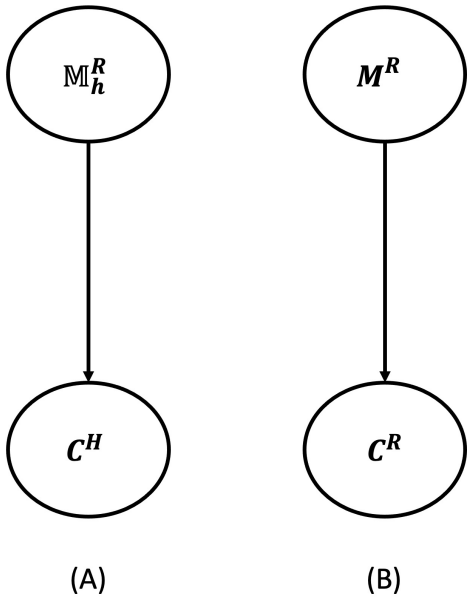


Figure 7.1: [Reminder:] A graphical model representing the probabilistic reasoning that is performed in this setting: Subfigure (A) captures the reasoning performed at the human's end with $\mathcal{C}^H$ being the random variable corresponding to the human's belief that a contract $\mathcal{C}$ will be satisfied. Similarly subfigure (B) represents the reasoning performed at the human's end, where $\mathcal{C}^R$ captures whether the robot achieves the contract C

level of trust the human has in an agent with respect to a contract is appropriate if the likelihood the human associates with the system satisfying the contract is equal to the likelihood of the agent satisfying that contract, i.e., *trust level is appropriate*, if

$$P(\mathcal{C}^H) = \sum_{\mathcal{M} \in \mathbb{M}_h^R} P(\mathcal{C}^H | \mathcal{M}) \times P_{\mathbb{M}}(\mathcal{M}) = P(\mathcal{C}^R | \mathcal{M}^R).$$

Where $\mathcal{C}^R$ is the random variable associated with the robot actually achieving the contract. The robot reasoning can be captured using the probabilistic graphical model presented in Figure 7.1(B).

7.3 Modeling Human's Reliance:

One of the important user behaviors that we are interested in capturing is whether the user is ready to accept the current decision given their trust in the agent. By

grounding trust in terms of likelihood of goal achievement, we are now able to leverage decision-theory to model the expected values of the user choosing to accept (Acc) or not accept ($\neg Acc$) a given decision. For the choice Acc, there are two possibilities, the decision in fact satisfies the contract and thus receives some positive utility for being successful ($U_{+\mathcal{C}}$) or it may fail in which case it gets a negative utility as a penalty ($U_{-\mathcal{C}}$). The expected value of relying on the agent's decision is thus given as

$$V^{\mathcal{C}}(Acc) = P(\mathcal{C}^H) * U_{+\mathcal{C}} + (1 - P(\mathcal{C}^H)) * U_{-\mathcal{C}}$$

While the value of choosing to not accept the agent is basically given as

$$V^{\mathcal{C}}(\neg Acc) = -1 * C_{\neg Acc}$$

Where $C_{\neg Acc}$ is the penalty associated with the user choosing to turn down a decision from the agent. The user would accept a given decision if $V^{\mathcal{C}}(Acc) > V^{\mathcal{C}}(\neg Acc)$.

7.4 Evaluation of Trust Inference Through Human Subject Experiments

In this section, we assess our proposed theory by comparing it to a trust scale developed by Chancey et al. Chancey *et al.* (2017). The central focus of our theory that we aim to examine pertains to the potential for inducing changes in trust (as measured through the questionnaire) by altering the human's belief about the agent, as predicted by our mental model theory.

In particular, we divide our participants into two groups: a positive update group, where participants' initial belief in the agent's capacity to fulfill a specific contract is increased during the study, and a negative update group, where participants' beliefs are decreased. In the course of evaluating our theory, we first investigate whether these two distinct belief updates result in different levels of trust. Once this distinction is established, we explore whether the positive update group experiences an increase in trust, and whether the negative update group experiences a decrease in trust.

Moreover, we examine whether modifying a specific component of the model leads to a corresponding alteration in the related information aspect of trust; performance, process, and purpose. We consider three scenarios in which we update specific parts of the model associated with performance, process, and purpose. Subsequently, we assess whether a positive or negative update in a specific component of the model influences an increase or decrease in the associated perception of trust, respectively.

Our expectation is that when the update corresponds to a specific aspect of trust information, it will primarily impact the perception of trust associated with that particular aspect, resulting in a stronger correlation than with other information types of trust.

The specific hypotheses we aim to test are as follows:

H1- The change in trust induced by an increase in likelihood of satisfying the contract is different from the one induced by a reduction in the likelihood of satisfying the contract.

H2- The positive update group will exhibit an increase in trust, as measured by the trust scales, compared to their initial trust levels.

H3- The negative update group will demonstrate a decrease in trust, as measured by the trust scales, in comparison to their initial trust levels.

H4- Positive/negative update of the model corresponding to the performance will increase/decrease performance perception of trust more than other perceptions.

H5- Positive/negative update of the model corresponding to the process will increase/decrease process perception of trust more than other perceptions.

H6- Positive/negative update of the model corresponding to the purpose will increase/decrease purpose perception of trust more than other perceptions.

7.4.1 Experiment Setup

We conducted a between-subject design study to assess the impact of either a positive or negative update, along with the specific type of update related to performance, process, or purpose, on user trust. As a result, each participant encountered one of the following combinations: positive or negative update, and one of the three scenarios: performance, process, or purpose.

Events

Each participant experiences two key events. In the initial event, they are presented with a task description and information about potential robots they might encounter. These robots vary in their capabilities, and participants are provided with images displaying the robot within its task environment. Following this, participants' trust is assessed using the trust questionnaire.

In the subsequent event, participants view a video showcasing a robot engaged in performing the task. This video serves to update their mental model of the robot's capabilities when executing the task. Following this update, their trust is reassessed. The first event aims to establish the participants' initial mental models of the robot ($\mathbb{M}_h^R$), while the second event serves to update those mental models ($\mathbb{M}_h^R$).

Scenarios

We have three scenarios in our studies. Each of the scenarios has both positive and negative updates that make six cases for our between subject study. The scenarios are:

(1) Performance Scenario (S1): In this scenario, we manipulate the performance information of the model to update the human model of the robot. The robot's

(a) (b) (c)

Figure 7.2: The robot in its task environment in the three scenarios (a) Performance Scenario (S1), (b) Process Scenario (S2) and (c) Purpose Scenario (S3).

task is to place the purple block on the red block, the robot in its task environment is illustrated in Figure 7.2(a). Initially, participants receive information about the task setting and details about the robots, which form their initial mental model of the robot. In this scenario we inform participants that (1) the purple block is heavy, and (2) uncertainty regarding the type of robot they will encounter. Equally likely, participants may be exposed to one of the following robots:

(a) A robot that is capable of picking up heavy blocks.

(b) A robot that is not able to pick up heavy blocks.

Following the initial event, participants receive either a positive or negative update regarding the robot's performance through a video demonstration of the robot performing the task. In the positive update, the robot successfully lifts the purple block and places it on the red block. In contrast, the negative update features a robot attempting to lift the purple block but failing to do so, ultimately ceasing the task after an unsuccessful attempt. You can view the videos for the positive update and negative update at the following links: Positive update video, and Negative update video.

(1) Process Scenario (S2): Similarly, this scenario involves a condition where we manipulate process information within the model to update the human model

of the robot. In this scenario, the robot's task is to place a purple block in a bowl. The robot in its task environment is depicted in Figure 7.2(b). Participants receive information about the task setting and details about the robots, which form the basis of their initial mental model of the robot. In this scenario, we inform participants that (1) there are two purple blocks, and two red blocks, as visually represented in the image. (2) the robot should always try to complete its task in the easiest way.

Following this initial event, participants are presented with a video update that showcases the robot's approach to the task. In the positive update, the robot efficiently selects the accessible purple block and places it into the bowl. Conversely, the negative update portrays a different approach: the robot initially chooses to pick up the red blocks and place them on the table. It then proceeds to pick up the purple block that was beneath the red blocks before ultimately placing it in the bowl. For visual reference and a more detailed understanding, you can access the videos illustrating the positive and negative updates via the following links: Positive update video, and Negative update video.

(1) Purpose Scenario (S3): In this scenario, we focus on manipulating the purpose-related information within the model to update the human model of the robot. The robot's task involves a unique setting. Participants are informed that there are two individuals engaged in a block stacking competition, with each person having three blocks. The objective is for one person to stack their blocks before the other. Importantly, the participants are made aware that the players can only start stacking their blocks once the robot completes its operations. In other words, if the robot chooses to assist one of the players in stacking their blocks, that player is guaranteed to win the competition. Participants are further informed that the person on the left, identified by wearing a black T-shirt, is their friend, and they wish for the robot to help their friend win in the competition. Consequently, their task for the robot is

to support their friend by stacking the blocks for him. The robot, depicted in its task environment in Figure 7.2(c), is accompanied by initial information provided to participants about the robots, shaping their initial mental model of the robot. In this scenario, participants are informed that it is equally likely that they may encounter one of the following types of robots:

1. Satisfying the user's goal is the robot's objective.
2. Satisfying the user's goal is not the robot's objective.

Following this initial event, participants receive either a positive or negative update related to the robot's purpose through a video demonstration of the robot's task. In the positive update, the robot assists their friend in stacking the blocks. In contrast, the negative update shows the robot helping the other person (their friend's rival). Videos illustrating the positive and negative updates are accessible via the following links: Positive update video, and Negative update video.

Trust Questionnaire

We employ a trust questionnaire developed by Chancey et al. Chancey *et al.* (2017). This questionnaire is explicitly designed and rigorously tested to assess three information types of trust: purpose, process, and performance. The original questionnaire was crafted within the context of an operator and a recommendation system, specifically, a tank-spotting aid. To adapt the questionnaire for use in our domain, we have carefully revised the questions, ensuring they capture the core essence of the original questionnaire. The complete questionnaire is provided in Table 7.1.

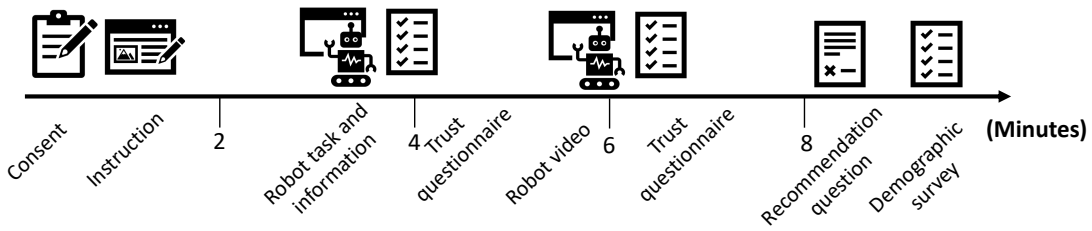


Figure 7.3: Study procedure overview.

Procedure

As depicted in Figure 7.3, upon granting their consent, each participant is directed to the instruction page. Within these instructions, we introduce a gamified element to the study. Participants assume the role of technology purchasers tasked with making a critical recommendation to the CEO of their company regarding the purchase of a robot. The success of their company's future depends on their final decision. A mistake in their recommendation can reflect poorly for their standing as employees within the company. We also provide a succinct overview of the study's process. Additionally, we provide a brief overview of the study's procedure.

Subsequently, participants are presented with a description of the task that the robot is expected to perform, accompanied by an image depicting the robot within its task environment. This image includes annotated objects for clarity. Furthermore, participants receive information about the robot and the setting, aligning with the scenarios outlined in the study. Following this phase, participants are asked to complete a fifteen-item trust questionnaire, which is accompanied by two attention check questions, all presented in random order.

Next, participants view a video showcasing the robot's execution of the task, after which they are prompted to respond to a similar set of questions. This phase requires them to consider the observations made during the video presentation.

Subsequently, participants are invited to provide their recommendation regarding

whether their company should proceed with the purchase of the robot or not. Lastly, participants are asked to complete a set of demographic questions.

Participants

A total of 123 participants were recruited through the Prolific platform [1] . These participants were randomly distributed among six cases, with 21 participants in the Performance Scenario with Positive Update (S1PU), 21 participants in the Performance Scenario with Negative Update (S1NU), 22 participants in the Process Scenario with Positive Update (S2PU), 21 participants in the Process Scenario with Negative Update (S2NU), 20 participants in the Purpose Scenario with Positive Update (S3PU), and 18 participants in the Purpose Scenario with Negative Update (S3NU). Each participant received $2 as compensation for their participation.

The median time taken by participants to complete the study was 6 minutes and 14 seconds. In terms of gender distribution, 47.97% identified as women, 49.59% as men, 1.62% as non-binary, and one person preferred not to specify their gender. The participants' ages were distributed as follows: 39% were between the ages of 25 and 34, 17.89% fell within the $35-44$ age group, 16.26% were in the $18-24$ age range, 13% were aged $45-54$, 10.57% were in the $55-64$ category, and 3.25% were 65 or older. Notably, the majority of participants (60.97%) indicated some level of familiarity with AI and robotics.

7.4.2 Results

Across all six cases where we collected both initial and updated trust perceptions, we calculated the participants' total trust at each step by computing the average across all fifteen questions. Additionally, we computed each perception of trust (performance,

[1] https://www.prolific.co/

process, and purpose) separately by calculating the mean value for that specific perception.

Our data was grouped based on two criteria: (1) positive or negative update, which involves considering all data together for each update, and (2) data related to each scenario, considered separately for each update.

In addition to examining basic statistics related to the collected data, we will also perform t-tests to test each hypothesis for each scenario as well as across all scenarios. We utilized a two-tailed t-test for the first hypothesis and a one-tailed t-test for the subsequent hypotheses. Any claims of statistical significance will be made against a significance level of 0.05.

H1- The change in trust value in positive vs. negative update: In this analysis, we compare the updated trust values between the positive update and negative update groups. This comparison is conducted for both total trust and each category of trust perception (performance, process, and purpose) across all scenarios as well as within each of the three individual scenarios: the performance scenario (S1), the process scenario (S2), and the purpose scenario (S3).

Figure 7.4 displays the updated total trust for the positive update groups compared to the negative update groups, revealing a noticeable difference. To further assess this difference, we conducted a series of two-tailed and one-tailed t-tests, comparing the variations between the positive and negative update groups for updated total trust, as well as the performance, process, and purpose trust perceptions reported by participants within each scenario and when considering all scenarios together. In these t-tests, the null hypothesis assesses that the samples are generated from distributions with the same mean, while the alternative hypothesis suggests that they are derived from different distributions. Our results from the t-tests, presented in Tables 7.2 (a),

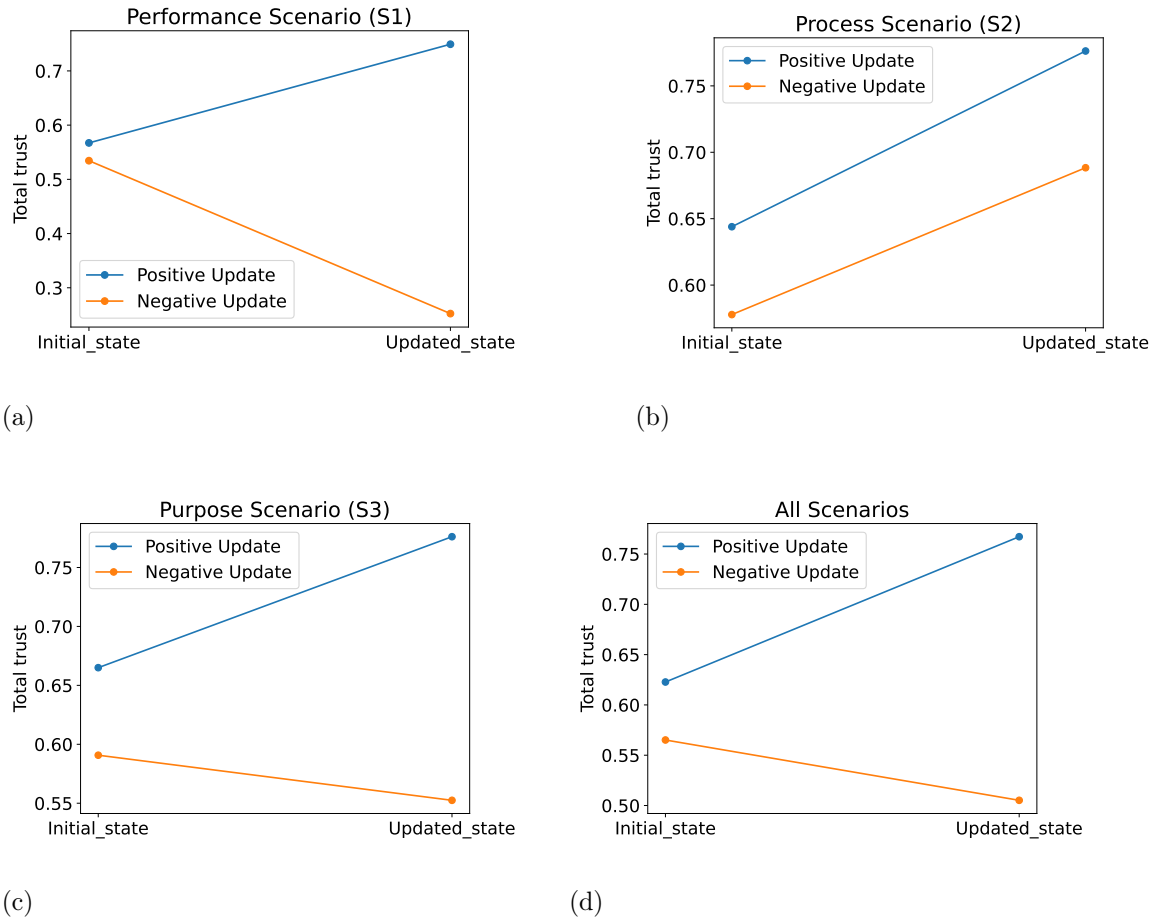
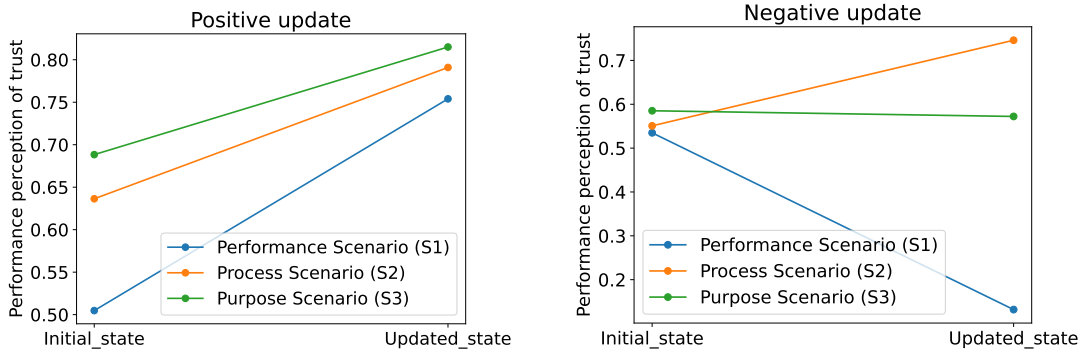


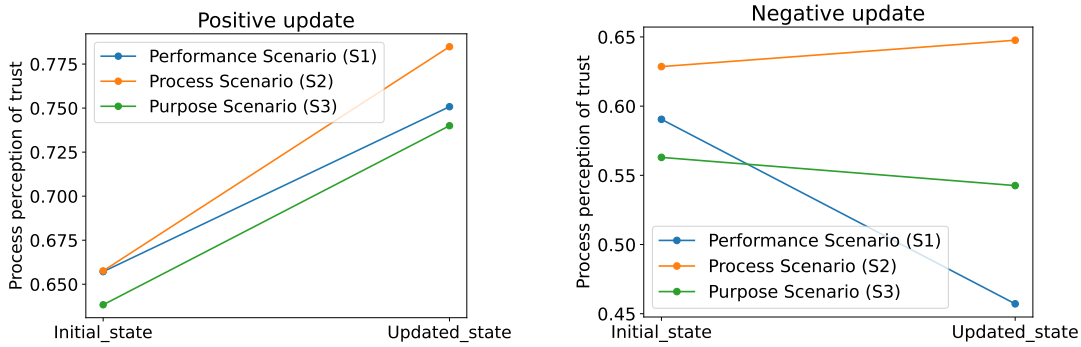
Figure 7.4: Total trust change from initial state to update state with positive and negative updates. (a) Performance scenario (S1), (b) Process scenario (S2), (c) Purpose scenario (S3) and (d) All scenarios.

7.3 (a), 7.4 (a), and 7.5 (a), support the validation of H1. The results indicate that the updated total trust, as well as the updated performance, process, and purpose trust perceptions, significantly differ in the positive update groups compared to the negative update groups. This statistical significance holds true for all scenarios collectively (Table 7.5), the performance scenario (S1) (Table 7.2 (a)), and the purpose scenario (S3) (Table 7.4 (a)). However, in the process scenario (S2) (Table 7.3 (a)), only the process perception of trust demonstrates statistically significant differences between the positive and negative update groups.



(a) (b)

Figure 7.5: The change in performance perception of trust across different scenarios with (a) positive updates and (b) negative updates.

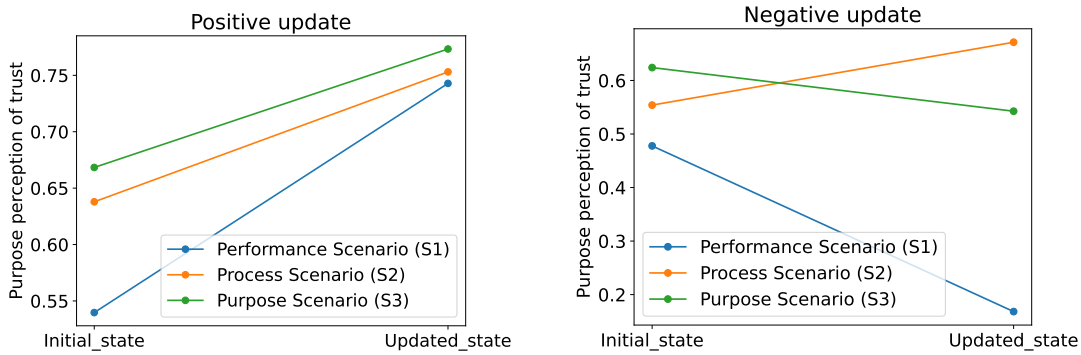


(a) (b)

Figure 7.6: The change in process perception of trust across different scenarios with (a) positive updates and (b) negative updates.

H2- Positive update groups induce increase in trust: For hypothesis H2, we examine whether the positive update groups exhibit an increase in trust compared to their initial trust. This examination is conducted for total trust and each information of trust (performance, process, and purpose) across all scenarios together and within each of the three individual scenarios: the performance scenario (S1), the process scenario (S2), and the purpose scenario (S3).

Figure 7.4 illustrates that total trust increased in the updated state when compared to



(a) (b)

Figure 7.7: The change in purpose perception of trust across different scenarios with (a) positive updates and (b) negative updates.

the initial state for the positive update groups. To establish the statistical significance of H2, we conducted one-tailed t-tests to determine whether total trust and each information of trust exhibited a significant increase from the initial state to the updated state within each scenario and across all scenarios together.

The results of these t-tests, presented in Tables 7.2 (b), 7.3 (b), 7.4 (b), and 7.5 (b), provide strong support for H2. The data shows that total trust, as well as all perceptions of trust, exhibited a significant increase from the initial state to the updated state with statistically significant p-values within each scenario (Tables 7.2 (b), 7.3 (b), and 7.4 (b)) and across all scenarios together (Table 7.5 (b)).

H3- Negative update groups induce decrease in trust: To validate hypothesis H3, we compare the updated trust and initial trust within the negative update groups for total trust and each information of trust (performance, process, and purpose) across all scenarios collectively, and within each of the three individual scenarios: the performance scenario (S1), the process scenario (S2), and the purpose scenario (S3). As demonstrated in Figure 7.4, the total trust decreased in the updated state when compared to the initial state for the negative update groups in the combined scenarios

and in each scenario except for the process scenario (S2). In the case of the negative update within the process scenario (S2), the robot achieved the goal and could perform the task but not in the desired and preferred human process. This suggests that the effect of the process is not as strong on trust when other factors are satisfied, and even a negative update for the process can increase trust.

To further validate the statistical significance of our hypothesis H3, we conducted a series of one-tailed t-tests to determine whether total trust and each information of trust exhibited a significant decrease from the initial state to the updated state within each scenario and across all scenarios collectively.

The results, as presented in part (c) of Tables 7.2, 7.3, 7.4, and 7.5, reveal that the total trust and all perceptions of trust decreased from the initial state to the updated state with statistically significant p-values in the performance scenario (S1) (Table 7.2 (c)). In the collective scenarios (Table 7.5 (c)), both total trust and purpose perception of trust exhibited a statistically significant decrease from the initial state to the updated state. However, while the performance and process perceptions of trust also decreased in the collective scenarios, they did not reach statistical significance. This is attributed to the inclusion of data from the process scenario in this collection, which exhibited an increase in trust.

Regarding the purpose scenario (S3) (Table 7.4 (c)), the total trust and all perceptions of trust decreased from the initial state to the updated state, but these decreases were not statistically significant. Conversely, in the process scenario (S2) (Table 7.3 (c)), total trust and each perception of trust increased in the updated state, with statistical significance for all perceptions except the process perception of trust. This highlights the influence of controlling the process information of the model on the process perception of trust.

H4- Correlation of performance scenarios (S1) with performance information of Trust: In hypothesis H4, we aim to determine whether the increase or decrease in the performance perception of trust in the performance scenario (S1) is more significant than other perceptions, i.e. process and purpose, in the positive or negative update groups, respectively.

Comparing the cohen-d effect size in both part (b) and (c) of Table 7.2, we observe that the increase in performance perception of trust in the positive update groups and the decrease in performance perception of trust in the negative update groups have the highest effect size when compared to the changes in process and purpose perception of trust. Moreover, the results from a one-way Anova test demonstrate that the increase and decrease in these perceptions of trust are statistically significantly different from each other in positive ($F(2, 60) = 3.75$, $p = 0.03$) and negative update groups ($F(2, 60) = 5.00$, $p = 0.009$) respectively.

Additionally, Figure 7.5(a) and 7.5(b) provide visual evidence supporting that the performance perception has increased or decreased the most in the performance scenario (S1) compared to other scenarios for positive and negative update groups, respectively.

H5 Correlation of process scenarios (S2) with process information of Trust: To validate hypothesis H5, we examine whether, in the process scenario (S2), the increase or decrease in process perception of trust is more significant than performance and purpose perception of trust within positive or negative update groups.

Comparing cohen-d effect size in both part (b) and (c) of Table 7.3, we find that in the positive update groups, the effect size for the increase in process perception of trust is more than the purpose perception, although it's more than the performance perception. This result is not surprising, as a successful process typically accompanies successful

performance. In the negative update groups, we observe that trust increases in all perceptions; however, the increase in process perception of trust has the smallest effect size compared to trust in performance and purpose. This suggests that while negative update to the process does not decrease overall trust, it has a relatively weaker positive impact on the perception of the process compared to other perceptions. While the results from a one-way Anova test in the positive update groups do not show statistically significant differences in the increase of various perceptions ($F(2, 63) = 0.46$, $p = 0.63$), in the negative update groups, the decrease in these perceptions of trust is statistically significantly different ($F(2, 60) = 4.77$, $p = 0.01$).

We can gain further insights into how different scenarios affect process perception of trust by examining Figure 7.6(a) and 7.6(b). Figure 7.6(a) demonstrates that the process perception has increased the most in the process scenario (S2). However, according to Figure 7.6(b), with negative updates, the process perception of trust decreases the most in the performance scenario (S1) rather than in the process scenario (S2).

H6- Correlation of purpose scenarios (S3) with purpose information of Trust: Here, we evaluate if in the purpose scenario (S3), the increase or decrease of purpose perception of trust with positive or negative updates is the most significant one among other perceptions. By comparing the cohen-d effect size in part (b) and (c) of Table 7.4, we can see that in the positive update group, the increase in purpose perception of trust has greater effect size than the process perception, although it is still lower than the performance perception. Similar to the process scenario, this result is not surprising because a successful purpose typically involves a successful performance, and it's challenging to disentangle the two. However, in the negative update group, we can observe that the purpose perception of trust decreases with

higher effect size than the performance and process perceptions, even though the results from a one-way Anova test do not show any statistically significant difference between the increase and decrease in various perceptions of trust for both the positive ($F(2, 57) = 0.18$, $p = 0.83$) and the negative groups ($F(2, 51) = 0.27$, $p = 0.76$) respectively.

We also assess if the purpose scenario (S3) results in the most increase or decrease in purpose perception when compared to other scenarios. Figures 7.5(a) and 7.5(b) demonstrate that the performance scenario (S1) leads to more significant increases or decreases in purpose perception of trust than the purpose scenario (S3) with both positive and negative updates.

7.4.3 Discussion

Our results clearly demonstrate that change in likelihood of the model does result in change in user trust, as assessed by trust questionnaire. Furthermore, our findings reveal that both reductions and increases in likelihood have varying effects on user trust. Notably, we observed a significant increase in trust across all three trust perceptions with a positive update of the likelihood of contract fulfillment. On the other hand, decreasing the likelihood of specific model information, while causing a decrease in trust, did not uniformly decrease trust across all trust perceptions.

In the Performance Scenario (S1), we observed a significant decrease in trust across all trust perceptions when we reduced the likelihood of contract achievement, focusing on performance aspect of the model. However, for the other scenarios, namely process (S2) and purpose (S3), reducing the likelihood of specific information did not significantly decrease trust across all trust perceptions. This disparity can be attributed to the fact that reducing the likelihood of achieving the contract for information related to process or purpose does not necessarily entail a consistent or

reduction in performance-related information. For example, in the negative process scenario (S2), while the likelihood of achieving the contract with a specific process decreased, the likelihood of the robot's capability to perform the task and the likelihood of the robot adhering to the user's intended purpose increased. Consequently, this delicate interplay led to an increase in trust even with a negative process information update.

Furthermore, the significant of results for performance perception of trust in the performance scenario and other scenarios, emphasizes the substantial impact of robot performance in shaping trust in comparison to other aspects. However, an alternative interpretation of this impact could be attributed to the assumption that performance naturally includes an aligned purpose and a certain level of satisfactory process. In other words, we cannot assume that performance is achieved without ensuring the robot's purpose aligns with the user's purpose, just as we cannot assume that performance is attained without meeting a degree of effective and ethical processes.

Moreover, though we aimed to control for the other trust perceptions in each scenario, our findings emphasize the high degree of correlation between the three trust information in the model. This signifies the inherent challenge of designing experiments that completely isolate one trust information from the others.

Another noteworthy observation is that updates to the human model are consistently associated with the human acquiring more information about the robot's processes. Consequently, both positive and negative updates tend to enhance clarity in certain aspects of the process. This makes it challenging to distinguish between the clarity of the overall process, which can also convey a negative view of the robot, and the lack of a clear understanding of the robot's process when implementing model updates. This emphasizes the necessity of devising a modified questionnaire that can effectively differentiate between various aspects of the process-related information.

In conclusion, our results shed light on the dynamics of trust and model updates. While the results did not confirm all aspects of the hypotheses with statistically significant results, these discoveries align with our mental model-based framework of trust, and our framework can effectively capture the aforementioned dynamics.

7.4.4 Concluding Remarks

In this chapter, we utilized our proposed mental model based theory of trust as a basis to infer human trust. Using our proposed framework, we formalized the notions of human reliance, the appropriate level of trust, and provide mechanisms to model trust evolution. All of them can further be utilized as a foundation for any future trust-aware decision-making frameworks. We further ran human subject studies to evaluate the power of our framework in inferring trust as well as capturing various information of trust. The results provide strong evidence that our predictive method is capable of inferring trust changes and affecting different information of trust by directly controlling different dimension of the model associated with those information.

Table 7.1: Trust Questionnaire

Performance (Predictability; Ability): What Does the Automation Do?
The robot always completes the task that is required of it.
The robot reliably completes the task that is required of it.
The robot consistently performs the task that is required of it.
I can rely on the robot to function properly.
The robot adequately completes the block-stacking task that is required of it.
Process (Dependability; Integrity): How Does the Automation Work?
Although I might not fully understand the robot's internal processes, I can predict how it completes a task.
I will be able to predict how the robot will perform in the future.
I understand why the robot completed the task in that manner.
It is easy to follow why the robot completed the task in that manner.
I recognize how to assess whether the robot is completing its task well.
Purpose (Faith; Benevolence): Why Was the Automation Developed?
I believe the robot will be able to complete the task required of it, even when I don't know for certain that the robot can do it.
When I am uncertain about deciding whether the robot will complete the task, I believe the robot will do it.
When I am not sure about whether the robot will complete the task, I have faith that the robot will complete the task.
Even when the robot completes the task in an unusual way, I am certain that the robot will complete its task.
Even if I have no reason to expect that the robot will function properly, I still feel certain that it will complete the tasks required of it.

Table 7.2: T-Test Results for Performance Scenario (S1): (a) Trust comparison between positive and negative update groups in updated state, (b) Trust increase in positive update groups, and (c) Trust decrease in negative update groups.

Performance Scenario (S1)

Trust Perception	T-tests				
	T	dof	p-val.	cohen-d	tail
(a) Positive vs. Negative updated_state					
Total trust	−11.03	40	< 0.0001(****)	3.404	two-sided
Performance inf.	−10.98	40	< 0.0001(****)	3.39	less
Process inf.	−5.07	40	< 0.0001(****)	1.57	less
Purpose inf.	−10.45	40	< 0.0001(****)	3.22	less
(b) Positive update groups updated_state > initial_state					
Total trust	−3.69	40	0.0003	1.14	less
Performance inf.	−4.059	40	0.00019	1.259	less
Process inf.	−1.90	40	0.03	0.59	less
Purpose inf.	−2.91	40	0.003	0.90	less
(c) Negative update groups updated_state < initial_state					
Total trust	4.75	40	0.00001	1.47	greater
Performance inf.	5.949	40	< 0.0001(****)	1.83	greater
Process inf.	2.28	40	0.01	0.70	greater
Purpose inf.	4.24	40	< 0.0001(****)	1.31	greater

Table 7.3: T-Test Results for Process Scenario (S2): (a) Trust comparison between positive and negative update groups in the updated state, (b) Trust increase in the positive update groups, and (c) Trust increase in the negative update groups.

Process Scenario (S2)

Trust Perception	T-tests				
	T	dof	p-val.	cohen-d	tail
(a) Positive vs. Negative updated_state					
Total trust	−1.71	34.53	0.096	0.53	two-sided
Performance inf.	−0.91	38.52	0.18	0.28	less
Process inf.	−2.42	32.64	0.01	0.75	less
Purpose inf.	−1.27	32.76	0.11	0.39	less
(b) Positive update groups updated_state > initial_state					
Total trust	−3.22	42	0.001	0.97	less
Performance inf.	−3.61	42	0.0004	1.09	less
Process inf.	−2.92	42	0.003	0.88	less
Purpose inf.	−2.24	42	0.015	0.68	less
(c) Negative update groups updated_state > initial_state					
Total trust	−2.11	40	0.02	0.65	less
Performance inf.	−3.92	40	0.0002	1.21	less
Process inf.	−0.32	40	0.37	0.10	less
Purpose inf.	−1.68	40	0.05	0.53	less

Table 7.4: T-Test Results for Purpose Scenario (S3): (a) Trust comparison between positive and negative update groups in the updated state, (b) Trust increase in the positive update groups, and (c) Trust decrease in the negative update groups.

Purpose Scenario (S3)

Trust Perception	T-tests				
	T	dof	p-val.	cohen-d	tail
(a) Positive vs. Negative updated_state					
Total trust	−2.80	24.07	0.01	0.94	two-sided
Performance inf.	−2.78	22.04	0.005	0.94	less
Process inf.	−2.31	26.84	0.01	0.77	less
Purpose inf.	−2.61	25.25	0.007	0.87	less
(b) Positive update groups updated_state > initial_state					
Total trust	−2.10	38	0.02	0.66	less
Performance inf.	−2.19	38	0.02	0.69	less
Process inf.	−1.70	38	0.048	0.54	less
Purpose inf.	−1.78	38	0.04	0.56	less
(c) Negative update groups updated_state < initial_state					
Total trust	0.44	34	0.33	0.15	greater
Performance inf.	0.14	34	0.44	0.046	greater
Process inf.	0.23	34	0.41	0.08	greater
Purpose inf.	0.82	34	0.21	0.27	greater

Table 7.5: T-Test Results for Data Collected Across All Scenarios Together.

All Scenarios

Trust Perception	T-tests				
	T	dof	p-val.	cohen-d	tail
(a) Positive vs. Negative updated_state					
Total trust	−6.59	81.81	< 0.0001(****)	1.21	two-sided
Performance inf.	−6.13	79.47	< 0.0001(****)	1.12	less
Process inf.	−5.37	97.69	< 0.0001(****)	0.98	less
Purpose inf.	−6.20	83.37	< 0.0001(****)	1.14	less
(b) Positive update groups updated_state > initial_state					
Total trust	−5.14	124	< 0.0001(****)	0.92	less
Performance inf.	−5.46	124	< 0.0001(****)	0.97	less
Process inf.	−3.73	124	0.0001	0.66	less
Purpose inf.	−4.04	124	< 0.0001(****)	0.72	less
(c) Negative update groups updated_state < initial_state					
Total trust	1.60	118	0.05	0.29	greater
Performance inf.	1.45	118	0.07	0.26	greater
Process inf.	1.15	118	0.13	0.21	greater
Purpose inf.	1.70	118	0.046	0.31	greater

Chapter 8

CONCLUSION

This thesis has explored trust in the realm of human-AI interactions and human-robot interactions with the objective of bridging the gap between psychological and computational perspectives of trust. In this concluding chapter, we address these key aspects: First, we revisit the goals of our research and analyze how the works presented in this thesis have contributed to their fulfillment. Second, we provide insights into potential directions for future research in this field.

8.1 Summary of Research

The objectives set forth in the thesis introduction have been successfully addressed and accomplished through a comprehensive examination of various dimension of trust, ranging from its psychological perceptions to computational formalizations. The central goal of the thesis was to establish a unified and consistent trust framework capable of achieving several key aims:

Foundation for Understanding, Estimating, and Engendering an Appropriate Level of Trust: The framework serves as a foundation for comprehending, assessing, and cultivating trust in various contexts.

Multidimensional Trust Representation: Our framework effectively captures the multi-faceted nature of trust, thereby bridging the gap between psychological and computational viewpoints on trust.

Support for Trust-Aware Decision-Making Frameworks: The framework offers a basis for constructing various trust-aware decision-making systems, enabling better-informed choices in scenarios involving human-robot interaction and beyond.

Facilitation of Trust Inference: The framework can be harnessed to facilitate trust inference, aiding in the assessment of trust levels in human-robot interactions and similar contexts.

In this section, we provide a detailed overview of how the research findings presented in the thesis fulfill these objectives.

To achieve our objectives, we took a systematic exploration of trust as a multi-dimensional concept, recognizing its fundamental role in shaping the outcomes of human-AI interactions. In Chapter 3, we made the foundation by introducing a mental model-based framework that effectively contextualizes trust within the realm of human-AI interactions. This framework provides a comprehensive understanding of trust, encompassing multiple dimensions often overlooked in computational models. It subsequently served as a foundation for the development of decision-making frameworks that incorporate trust in various dynamics of human-AI interactions, as explored in Chapters 4, 5, and 6. Furthermore, it offered a valuable tool for inferring and estimating trust when direct measurements may not be readily available, as discussed in Chapter 7.

Delving deeper into the practical application of our mental model-based framework, we proceeded to develop diverse computational trust-aware decision-making frameworks. In Chapter 4, we focused on trust dynamics within single interactions, addressing the challenges related to inadequate trust and excessive monitoring. Chapter 5 introduced a formal model that deepened our understanding of the relationship between trust and monitoring behavior. Chapter 6 extended our investigations to longitudinal interactions, emphasizing the crucial role of trust in establishing and maintaining effective human-robot collaboration.

In the final part of our research, we explored the direct utilization of our mental model-based framework in modeling trust evolution, human reliance, and the inference

of trust dynamics, as explained in Chapter 7. This comprehensive journey allowed us to contribute significantly to the field of trust in human-AI interactions, enhancing our understanding of this complex concept.

8.2 Avenues for Future Research

In this thesis, we have provided a foundational framework for modeling trust and have developed various decision-making frameworks that utilize this foundation. This work offers opportunities for further extension, particularly in the development of additional behavioral and communication tools that agents can employ to reinforce or update a user's beliefs about the system, consequently impacting their trust and willingness to rely on the system. One potential avenue is to adapt and expand existing tools from the field of human-AI interaction for use in longitudinal interactions, where these mechanisms are being used towards specific updates in the user's mental model aimed at altering their trust in the system. While we focused on interpretable behavior generation and explanation generation for building and updating human's model and trust, it has primarily considered the agent's true model $\mathcal{M}^R$ and the human's expectation of them $\mathbb{M}_h^R$. Therefore, the further exploration is to place more weight on considering the user's model, denoted as $\mathcal{M}_h^*$, when generating explicable and explanatory tools. Our mental model based framework of trust can allow to revisit some related human-AI interaction settings and to provide it with more formal grounding. Following are potential directions for adapting and expanding existing tools in this context.

Explicable Behavior to Engender Trust

In this thesis, the explicable behavior that is considered corresponds to the most likely plan per the human's expectation that the agent can still execute. However, the important point to note here is that choosing an explicable plan need not result

in higher trust as the human may have high prior on models that will only generate plans that are quite poor in comparison to the ones that are allowed under $\mathcal{M}_h^*$. Thus in the context of engendering trust, we need to look at a modified formalization of explicability, one that *balances the likelihood of the plan with the likelihood of achieving the specific contract*. We would still need to stick to choosing plans that are likely per human expectations as the human may choose to stop the system's execution if the behavioral prefix is too surprising for them. Exhibiting such behavior also causes humans to re-evaluate their posterior beliefs about what models the robot may use, thus making the future generation of such behavior easier.

Explanation and Compliance

In this thesis, the explanation generation that we looked at was similar to Chakraborti *et al.* (2017), where the explanation takes the form of a model update which once applied to the human models of the agent $\mathbb{M}_h^R$ would result in an updated model where the plan in question is optimal. By ensuring the optimality of the plan in the target model (i.e., the human mental model after the explanation), we can guarantee that the human will not be able to come up with a plan that seems to them to be more appropriate than the one under consideration. So, it is assumed that through explanations, the system updates the user's mental models about the task. However, it is worth noting that in cases where the user may have low trust in the system, they may be hesitant to accept or comply to the explanations provided by the system at face value. They may view these explanations as *excuses* generated by the system, or further evidence that the system may be confused by the task at hand. Thus, our proposed theory of trust that is meant to be applied in designing the interaction of such decision-support systems should be capable of accounting for the fact that users may not be open to updating their mental models of the task unless the system has successfully established some level of trust. To allow for behavior

related to compliance, we should assume the human would only accept explanations that may result in models that can generate solutions better than those provided by $\mathcal{M}_h^*$ (and thus requiring the human to update $\mathcal{M}_h^*$) if the system has already gained a certain level of trust with respect to a specific contract. Thus this may require the agent to spend earlier iterations of the interactions to engender the required level of trust before it can start providing the explanations.

Revisiting Other Interpretable Behaviors

In addition to explicable behavior and explanation, there are other interpretable behaviors that the agent could leverage to engender trust. One obvious example is the use of legible behaviors, wherein the objective of the agent is not to completely resolve the human's uncertainty about the AI system's model but to shape it so that the remaining models are more reflective of what behaviors the agent is truly capable of pursuing.

Formalizing Iterative Explanatory Dialogue

Our mental model based framework can provide a more formal grounding in an interaction of iterative explanatory dialogue. In such dialogues, the system and the user engage in a protracted dialogue about the system's decision. The user may ask various questions about the decision in question until the user feels satisfied with the current decision. While most works acknowledge the need to facilitate such an interactive process Weld and Bansal (2019), they tend to focus only on a single instance of this interaction, namely how to respond to a specific explanatory query. This overlooks some fundamental issues, like quantifying when the user may feel confident enough to accept the decision being offered by the system and how to best drive the dialogue to achieve quick resolution. Our proposed theory of trust provides a formal model to capture such interactions. In particular, one could look at the whole interaction as a process of raising the trust of the user to a level whereby the expected value the human

associates with accepting the decisions outweigh the alternative. Additionally, by virtue of the fact that in our case the user's trust is defined over the user's perception of what the system is capable of, it automatically captures many factors like the difference in knowledge and inferential capability between the user and the system which play a central role in such interactions. As such, our formulation can provide a clear bridge between the work that have been previously done on explanations as model updates to trust. This would not only help us better understand the role of explanations in building trust but also help to provide a more general definition of what it means to generate effective explanations in the context of repeated human-AI interactions.

REFERENCES

Ajzen, I., "Understanding attitudes and predictiing social behavior", Englewood cliffs (1980).

Akash, K., W.-L. Hu, T. Reid and N. Jain, "Dynamic modeling of trust in human-machine interactions", in "2017 American Control Conference (ACC)", pp. 1542–1548 (IEEE, 2017).

Baker, A. L., E. K. Phillips, D. Ullman and J. R. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions", ACM Transactions on Interactive Intelligent Systems (TiiS) **8**, 4, 1–30 (2018).

Baker, C. L., R. Saxe and J. B. Tenenbaum, "Action understanding as inverse planning", Cognition **113**, 3, 329–349 (2009).

Barber, B., "The logic and limits of trust", (1983).

Billings, D. R., K. E. Schaefer, J. Y. C. Chen and P. A. Hancock, "Human-robot interaction: developing trust in robots", in "International Conference on Human-Robot Interaction, HRI'12, Boston, MA, USA - March 05 - 08, 2012", edited by H. A. Yanco, A. Steinfeld, V. Evers and O. C. Jenkins, pp. 109–110 (ACM, 2012), URL `https://doi.org/10.1145/2157689.2157709`.

Bylander, T., "The computational complexity of propositional strips planning", Artificial Intelligence **69**, 1-2, 165–204 (1994).

Cain, A., *Trust and complacency in cyber security* (Master's thesis, Department of Psychology, San José State University, 2016).

Chakraborti, T., S. Sreedharan and S. Kambhampati, "Balancing explicability and explanation in human-aware planning", in "IJCAI", pp. 1335–1343 (ijcai.org, Macao, China, 2019).

Chakraborti, T., S. Sreedharan, Y. Zhang and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy", in "IJCAI", pp. 156–163 (ijcai.org, Melbourne, Australia, 2017).

Chancey, E. T., J. P. Bliss, Y. Yamani and H. A. Handley, "Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence", Human factors **59**, 3, 333–345 (2017).

Chen, M., S. Nikolaidis, H. Soh, D. Hsu and S. Srinivasa, "Planning with trust for human-robot collaboration", in "Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction", pp. 307–315 (ACM, Chicago, IL, USA, 2018).

Chen, M., S. Nikolaidis, H. Soh, D. Hsu and S. Srinivasa, "Trust-aware decision making for human-robot collaboration: Model learning and planning", ACM Transactions on Human-Robot Interaction (THRI) **9**, 2, 1–23 (2020).

Cordwell, S., "Markov decision process (mdp) toolbox for python", `https://github.com/sawcordwell/pymdptoolbox` (2012).

Cummings, M., P. Pina and B. Donmez, "Selecting metrics to evaluate human supervisory control applications", Tech. rep., MIT Humans and Automation Laboratory (2008).

Cummings, M. L., "Automation bias in intelligent time critical decision support systems", in "Decision making in aviation", pp. 289–294 (Routledge, 2017).

De Visser, E. J., R. Pak and T. H. Shaw, "From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction", Ergonomics **61**, 10, 1409–1427 (2018).

Dennett, D. C., *The intentional stance* (MIT press, 1987).

Desai, M., *Modeling trust to improve human-robot interaction*, Ph.D. thesis, University of Massachusetts Lowell (2012).

Desai, M., P. Kaniarasu, M. Medvedev, A. Steinfeld and H. Yanco, "Impact of robot failures and feedback on real-time trust", in "2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", pp. 251–258, IEEE (IEEE/ACM, Tokyo, Japan, 2013).

Desai, M., M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld and H. Yanco, "Effects of changing reliability on trust of robot systems", in "2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", pp. 73–80, IEEE (ACM, Boston, MA, USA, 2012).

Deutsch, M., "The effect of motivational orientation upon trust and suspicion", Human relations **13**, 2, 123–139 (1960).

Dragan, A. D., K. C. Lee and S. S. Srinivasa, "Legibility and predictability of robot motion", in "Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction", pp. 301–308 (IEEE Press, 2013).

Fisac, J. F., A. Bajcsy, S. L. Herbert, D. Fridovich-Keil, S. Wang, C. J. Tomlin and A. D. Dragan, "Probabilistically safe robot planning with confidence-based human predictions", arXiv preprint arXiv:1806.00109 (2018).

Floyd, M., M. Drinkwater and D. Aha, "Trust-guided behavior adaptation using case-based reasoning", in "Twenty-Fourth International Joint Conference on Artificial Intelligence", (2015).

Geffner, H. and B. Bonet, "A concise introduction to models and methods for automated planning", Synthesis Lectures on Artificial Intelligence and Machine Learning **8**, 1, 1–141 (2013).

Guo, Y., C. Zhang and X. J. Yang, "Modeling trust dynamics in human-robot teaming: A bayesian inference approach", in "Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems", pp. 1–7 (2020).

Hadfield-Menell, D., A. Dragan, P. Abbeel and S. Russell, "The off-switch game", in "Workshops at the Thirty-First AAAI Conference on Artificial Intelligence", (2017).

Helmert, M., "The fast downward planning system", Journal of Artificial Intelligence Research **26**, 191–246 (2006).

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

International Planning Competition, "IPC Competition Domains", `https://goo.gl/3yyDn4` (2011).

Jacovi, A., A. Marasović, T. Miller and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai", in "Proceedings of the 2021 ACM conference on fairness, accountability, and transparency", pp. 624–635 (2021).

Jian, J.-Y., A. M. Bisantz and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems", International journal of cognitive ergonomics **4**, 1, 53–71 (2000).

Johns, J. L., "A concept analysis of trust", Journal of advanced nursing **24**, 1, 76–83 (1996).

Kok, B. C. and H. Soh, "Trust in robots: Challenges and opportunities", Current Robotics Reports **1**, 4, 297–309 (2020).

Kraus, S., A. Azaria, J. Fiosina, M. Greve, N. Hazon, L. Kolbe, T.-B. Lembcke, J. P. Muller, S. Schleibaum and M. Vollrath, "Ai for explaining decisions in multi-agent environments", **34**, 09, 13534–13538 (2020).

Kulkarni, A., T. Chakraborti, Y. Zha, S. G. Vadlamudi, Y. Zhang and S. Kambhampati, "Explicable robot planning as minimizing distance from expected behavior", CoRR, abs/1611.05497 (2016).

Kulkarni, A., S. Sreedharan, S. Keren, T. Chakraborti, D. E. Smith and S. Kambhampati, "Design for interpretability", in "ICAPS Workshop on Explainable AI Planning (XAIP)", (2019).

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

Lee, J. D. and K. A. See, "Trust in automation: Designing for appropriate reliance", Human factors **46**, 1, 50–80 (2004).

Liu, R., F. Jia, W. Luo, M. Chandarana, C. Nam, M. Lewis and K. P. Sycara, "Trust-aware behavior reflection for robot swarm self-healing.", in "AAMAS", pp. 122–130 (2019).

Mayer, R. C., J. H. Davis and F. D. Schoorman, "An integrative model of organizational trust", Academy of management review **20**, 3, 709–734 (1995).

Merritt, S. M., H. Heimbaugh, J. LaChapell and D. Lee, "I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system", Human factors **55**, 3, 520–534 (2013).

Meyer, J., "Effects of warning validity and proximity on responses to warnings", Human factors **43**, 4, 563–572 (2001).

Moorman, C., R. Deshpande and G. Zaltman, "Factors affecting trust in market research relationships", Journal of marketing **57**, 1, 81–101 (1993).

Muir, B. M., "Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems", Ergonomics **37**, 11, 1905–1922 (1994).

Nguyen, T., S. Sreedharan and S. Kambhampati, "Robust planning with incomplete domain models", Artificial Intelligence **245**, 134–161 (2017).

Nikolaidis, S., D. Hsu and S. Srinivasa, "Human-robot mutual adaptation in collaborative tasks: Models and experiments", The International Journal of Robotics Research **36**, 5-7, 618–634 (2017).

Parasuraman, R. and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration", Human factors **52**, 3, 381–410 (2010).

Pierson, A. and M. Schwager, "Adaptive inter-robot trust for robust multi-robot sensor coverage", in "Robotics Research", pp. 167–183 (Springer, 2016).

Pippin, C. and H. Christensen, "Trust modeling in multi-robot patrolling", in "2014 IEEE International Conference on Robotics and Automation (ICRA)", pp. 59–66 (IEEE, 2014).

Puterman, M. L., *Markov decision processes: discrete stochastic dynamic programming* (John Wiley & Sons, 2014).

Rempel, J. K., J. G. Holmes and M. P. Zanna, "Trust in close relationships.", Journal of personality and social psychology **49**, 1, 95 (1985).

Robinette, P., A. M. Howard and A. R. Wagner, "Timing is key for robot trust repair", in "International conference on social robotics", pp. 574–583 (Springer, 2015).

Rotter, J. B., "A new scale for the measurement of interpersonal trust.", Journal of personality (1967).

Rousseau, D. M., S. B. Sitkin, R. S. Burt and C. Camerer, "Not so different after all: A cross-discipline view of trust", Academy of management review **23**, 3, 393–404 (1998).

Salvatier, J., T. V. Wiecki and C. Fonnesbeck, "Probabilistic programming in python using pymc3", PeerJ Computer Science **2**, e55 (2016).

Sankaranarayanan, V., M. Chandrasekaran and S. Upadhyaya, "Towards modeling trust based decisions: a game theoretic approach", in "European Symposium on Research in Computer Security", pp. 485–500 (Springer, 2007).

Schweitzer, M. E., J. C. Hershey and E. T. Bradlow, "Promises and lies: Restoring violated trust", Organizational behavior and human decision processes **101**, 1, 1–19 (2006).

Sebo, S. S., P. Krishnamurthi and B. Scassellati, ""i don't believe you": Investigating the effects of robot trust violation and repair", in "2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", pp. 57–65 (IEEE, 2019).

Soh, H., Y. Xie, M. Chen and D. Hsu, "Multi-task trust transfer for human–robot interaction", The International Journal of Robotics Research **39**, 2-3, 233–249 (2020).

Sreedharan, S., T. Chakraborti, C. Muise and S. Kambhampati, "Expectation-aware planning: A unifying framework for synthesizing and executing self-explaining plans for human-aware planning", (AAAI, 2020).

Sreedharan, S., A. Kulkarni, T. Chakraborti, D. E. Smith and S. Kambhampati, "A bayesian account of measures of interpretability in human-ai interaction", IJCAI (2021).

Sreedharan, S., A. Kulkarni and S. Kambhampati, "Explainable human–ai interaction: A planning perspective", Synthesis Lectures on Artificial Intelligence and Machine Learning **16**, 1, 1–184 (2022).

Tolmeijer, S., A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon and M. L. Tielman, "Taxonomy of trust-relevant failures and mitigation strategies", in "Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction", pp. 3–12 (2020).

Ullman, D. and B. F. Malle, "What does it mean to trust a robot? steps toward a multidimensional measure of trust", in "Companion of the 2018 acm/ieee international conference on human-robot interaction", pp. 263–264 (2018).

Wang, N., D. V. Pynadath and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations", in "2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", pp. 109–116 (IEEE, 2016).

Wang, X., Z. Shi, F. Zhang and Y. Wang, "Dynamic real-time scheduling for human-agent collaboration systems based on mutual trust", Cyber-Physical Systems **1**, 2-4, 76–90 (2015).

Wang, Y., L. R. Humphrey, Z. Liao and H. Zheng, "Trust-based multi-robot symbolic motion planning with a human-in-the-loop", ACM Transactions on Interactive Intelligent Systems (TiiS) **8**, 4, 1–33 (2018).

Weld, D. S. and G. Bansal, "The challenge of crafting intelligible intelligence", Communications of the ACM **62**, 6, 70–79 (2019).

Wickens, C. D., B. A. Clegg, A. Z. Vieane and A. L. Sebok, "Complacency and automation bias in the use of imperfect automation", Human factors **57**, 5, 728–739 (2015).

Xu, A. and G. Dudek, "Trust-driven interactive visual navigation for autonomous robots", in "2012 IEEE International Conference on Robotics and Automation", pp. 3922–3929 (IEEE, 2012).

Xu, A. and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations", in "2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", pp. 221–228 (IEEE, 2015).

Xu, A. and G. Dudek, "Maintaining efficient collaboration with trust-seeking robots", in "2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)", pp. 3312–3319 (IEEE, 2016).

Zahedi, Z., A. Olmo, T. Chakraborti, S. Sreedharan and S. Kambhampati, "Towards understanding user preferences for explanation types in model reconciliation", in "2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", pp. 648–649 (IEEE, 2019a).

Zahedi, Z., S. Sengupta and S. Kambhampati, "To monitor or to trust: observing robot's behavior based on a game-theoretic model of trust", in "Proceedings of the Trust Workshop at AAMAS", (2019b).

Zahedi, Z., S. Sengupta and S. Kambhampati, "'why didn't you allocate this task to them?'negotiation-aware explicable task allocation and contrastive explanation generation", in "Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems", pp. 2292–2294 (2023a).

Zahedi, Z., S. Sreedharan and S. Kambhampati, "A mental model based theory of trust", XAI workshop at IJCAI (2023b).

Zahedi, Z., S. Sreedharan and S. Kambhampati, "A mental-model centric landscape of human-ai symbiosis", R2HCAI, AAAI (2023c).

Zahedi, Z., S. Sreedharan, M. Verma and S. Kambhampati, "Modeling the interplay between human trust and monitoring", in "Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction", pp. 1119–1123 (2022).

Zahedi, Z., M. Verma, S. Sreedharan and S. Kambhampati, "Trust-aware planning: Modeling trust evolution in iterated human-robot interaction", in "Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction", pp. 281–289 (2023d).

Zhang, Y., S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo and S. Kambhampati, "Plan explicability and predictability for robot task planning", in "Robotics and Automation (ICRA), 2017 IEEE International Conference on", pp. 1313–1320 (IEEE, 2017).

APPENDIX A

OTHER RESEARCH CONTRIBUTIONS

This dissertation centers on computational accounts of trust within human-AI interaction, particularly emphasizing a unified and comprehensive mental model-based framework. This framework can serve as a foundation in diverse decision-making frameworks and acts as a bridge between the human factor and computational aspects of trust studies. Beyond the study of trust, I have engaged in collaborative work addressing various other challenges that, while not directly related to trust, can be seen as integral components or contexts for fostering trust. In this appendix, we provide a concise overview of some additional contributions I have made in this regard.

Multi-agent Task Allocation and Contrastive Explanation Generation

In this work Zahedi *et al.* (2023a), we designed an Artificially Intelligent Task Allocator (AITA) that proposes a task allocation for a team of humans. A key property of this allocation is that when an agent with imperfect knowledge (about their teammate's costs and/or the team's performance metric) questions the allocation by contesting with a counterfactual, a contrastive explanation can always be provided to showcase why the proposed allocation is better than the counterfactual. For this, we considered a negotiation process that produces a negotiation-aware task allocation and, when contested, leverages a negotiation tree to provide a contrastive explanation.

We blended aspects of both the (centralized and distributed) approaches and proposed AITA, an Artificial Intelligence-powered Task Allocator. Our system (1) uses a centralized allocation algorithm patterned after negotiation to come up with an allocation that explicitly accounts for the costs of the individual agents and overall performance, and (2) can provide contrastive explanation when a proposed allocation is contested using a counterfactual. We assumed AITA is aware of all the individual costs and the overall performance costs. Use of a negotiation-based mechanism for coming up with a negotiation-aware explicable allocation helps reuse the inference process to provide contrastive explanations. Our explanations have two desirable properties.

First, the negotiation-tree based explanation by AITA has a graphical form that effectively distills relevant pieces from a large amount of information (see Figure A.1); this is seen as a convenient way to explain information in multi-agent environments Kraus *et al.* (2020).

Second, the explanation, given it is closely tied to the inference process, acts as a certificate that guarantees explicability to the human (i.e. no other allocation could have been more profitable for them while being acceptable to others).

To evaluate our work, we conducted human studies in three different task allocation scenarios and show that the allocations proposed by AITA are perceived as fair by the majority of subjects. Users who questioned AITA's allocations, upon being explained, found it understandable and convincing in two control cases. Further, we considered an approximate version of the negotiation-based algorithm for larger task allocation domains and, via numerical simulation, show how underestimation of a teammate's costs and different aspects of incompleteness affect explanation length.

While this work does not directly investigate the impact of allocation and explanation on trust, it is reasonable to infer that the provision of transparent and easily understandable contrastive explanations by AITA plays a pivotal role in building trust among humans. Users can see the rationale behind AITA's decisions and be

assured that the allocation is both fair and optimized for their benefit. This enhances cooperation and collaboration in multi-agent environments, fostering trust in the allocation process.

Understanding User Preferences for Explanation Types in Model Reconciliation

In the research conducted by Chakraborti *et al.* (2017), the authors have formalized the explanation process within the realm of automated planning, defining it as a mechanism referred to as "model reconciliation." This entails a process by which the planning agent can bring the explainee's (possibly faulty) model of a planning problem closer to its understanding of the ground truth until both agree that its plan is the best possible. They have explored how model reconciliation process unfolds in the classical planning setting, while accounting for the mental model of the human in the loop. The content of explanations can thus range from misunderstandings about the agent's beliefs (state), desires (goals) and capabilities (action model). Though existing literature has considered different kinds of these model differences to be equivalent, literature on the explanations in social sciences has suggested that explanations with similar logical properties may often be perceived differently by humans. In our present work Zahedi *et al.* (2019a), we explore to what extent humans attribute importance to different kinds of model differences that have been traditionally considered equivalent in the model reconciliation setting. To initiate this exploration, we conducted a preliminary investigation via a human subject study, which focused on users' preferences regarding different types of model differences, based on a logistics planning domain International Planning Competition (2011). The results of our study suggest that humans prefer explanations that address misunderstandings about the effects of the agent's actions. These findings emphasize the importance of incorporating such considerations into the process of generating explanations within model reconciliation approaches.

Although we did not directly explore the effect of incorporating human preferences into the explanation process on trust, it is evident that such an approach has the potential to enhance trust in the human-in-the-loop, thereby contributing to more effective decision-making and collaboration. These insights can advance the interaction and consequently promote trust.

A Mental-Model Centric Landscape of Human-AI Symbiosis

There has been significant recent interest in developing AI agents capable of effectively interacting and teaming with humans. While each of these works try to tackle a problem quite central to the problem of human-AI interaction, they tend to rely on myopic formulations that obscure the possible inter-relatedness and complementarity of many of these works. The framework of human-aware AI in Sreedharan *et al.* (2022) has been proposed to incorporate the intuition of theory of mind into the context of human-AI interaction. This human-aware AI framework was introduced to capture settings where the role of the human is limited to being a passive observer trying to make sense of an AI agent's decisions. Therefore, these version of the human-aware AI framework are insufficient to adequately explain and unify the various works in the area. These scenarios only represent a small subset of possible ones that have been considered by the various works studied within the purview of human-AI interaction. In this work Zahedi *et al.* (2023c), we corrected

this shortcoming by introducing a significantly general version of human-aware AI interaction scheme, called *generalized human-aware interaction (GHAI),* (Figure A.2) that talks about (mental) models of six types. GHAI will not only allow for scenarios where the human may be an actor but also introduce the notion of a task model $\mathcal{M}^*$ that captures the true joint task specification of both the human and AI agent. Moreover, we consider each agent's (i.e. human or the AI agent) perception of the true combined task, which is independent of their beliefs about the other agent's perception of it. This allows us to capture interaction scenarios where one of the agents may choose to correct the other agent's beliefs about their task models. With the basic framework in place, we saw how the primary interaction facilitated by the various works is manifestation of either a *model-communication behavior* or a *model-following behavior* made possible by the use of the various mental models that are part of the framework.Through this work, we saw how this new framework allows us to capture the various works done in the space of human-AI interaction and identify the fundamental behavioral patterns supported by these works. We also used this framework to identify potential gaps in the current literature and suggested future research directions to address these shortcomings.

While GHAI framework may not fully encompass our proposed mental model based framework of trust, it can be leveraged to extend our mental model based framework of trust, allowing for a more comprehensive framework of trust in which humans play a more active and collaborative role in the interaction.

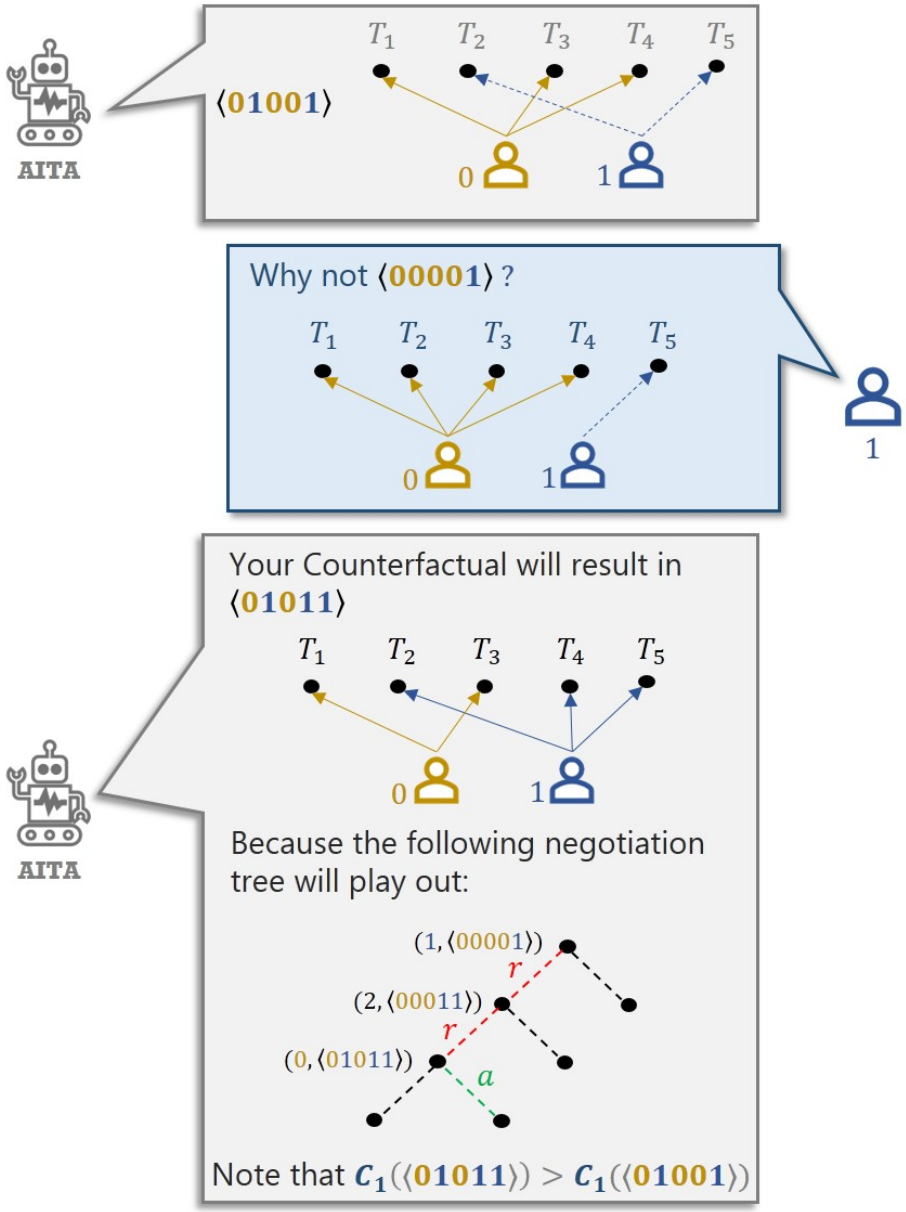


Figure A.1: AI Task Allocator (AITA) comes up with a negotiation-aware explicable allocation ⟨01001⟩ for a set of two humans– 0 and 1. In this allocation, human 0 is assigned tasks 1, 3 and 4 and agent 1 is assigned tasks 2 and 5. A dissatisfied human 1 questions AITA with a counterfactual allocation ⟨00001⟩, where he/she just needs to do task 5 (they believe task 5 is much more difficult and will take similar effort compared to doing all the 4 others). AITA then explains why the original proposed allocation (i.e. ⟨01001⟩) is better than the counterfactual allocation (i.e. ⟨00001⟩). The graph of the negotiation tree can be given as a dialogue "if human 1 proposes the allocation ⟨00001⟩, it will be rejected and AITA will offer ⟨00011⟩, which will then be rejected and human 0 will propose a counter offer ⟨01011⟩ which will then will have to be accepted by all. This final allocation would have a higher cost for you (human 1) than the first proposed allocation. Hence, the counterfactual allocation will eventually result in worse-off allocation for human 1.

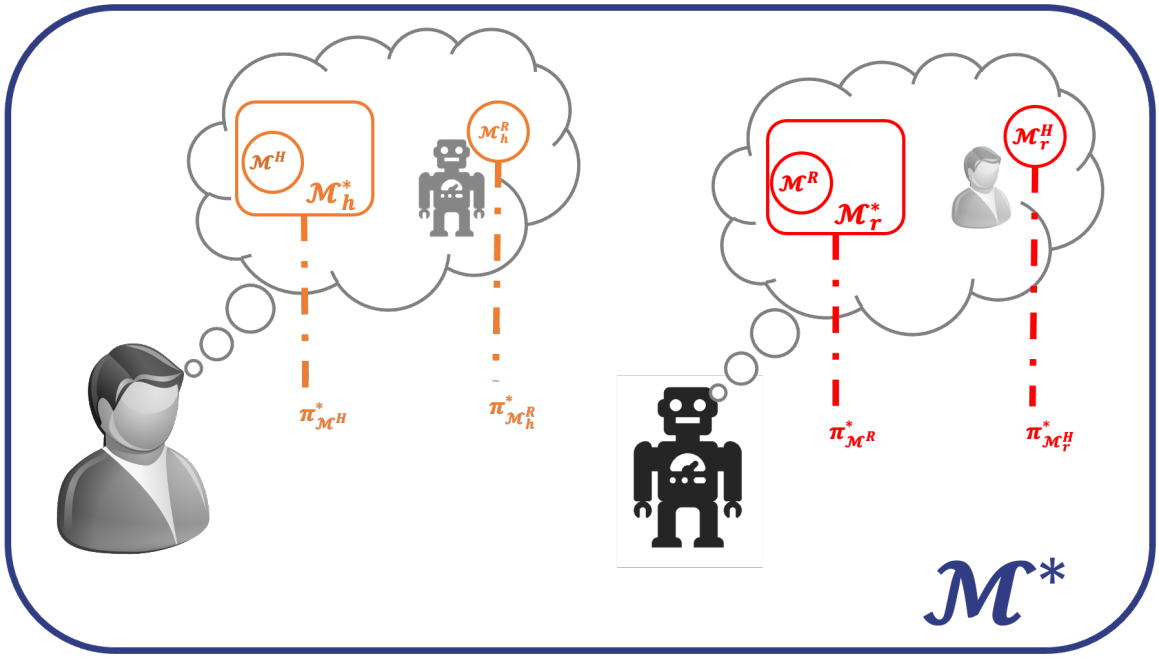


Figure A.2: The six models in the GHAI framework. $\mathcal{M}^*$ are the ground truth models of the task; $\mathcal{M}^H$ and $\mathcal{M}^R$ are the task models that the human and the AI agent ascribe to themselves; $\mathcal{M}_h^R$ and $\mathcal{M}_r^H$ are the estimates of the AI agent's (human's) model that human (AI agent) has.

APPENDIX B

APPENDIX FOR CHAPTER 4

In this chapter, we provide more detail on the two human subjects studies that have been done in Chapter 4.

Study I: Do we need this service?

As mentioned in Chapter 4, in this study participants play the role of a student in a robotics department who are asked to monitor the robot for an hour. To make the monitoring action be associated with a cost, we consider a second task where participants can choose to grade exam papers (and get paid) instead of monitoring the robot. The other action 'grade exam papers' represents the action to not-monitor the robot. As opposed to asking the participants for mixed strategies over the two actions, which is hard for them to interpret, we ask them to give us a time slice for which they would choose a particular action (eg. 30 minutes to monitor the robot and 30 minutes to grade exam papers). We provide the participants with their utility values for their actions conditioned on the robot's pure strategies (i.e. the plans π_s and π_{pr}). We inform them that the robot may have incentive to consider a less costly (but probably risky) plan depending on the fraction of time allocated for monitoring. We let each participant do five trials and after each trial, the overall utility based on the participant's monitoring strategy and the robot's strategy is reported to them. The robot does not adapt itself to the human's strategy in the previous trial (which intends to preserve the non-repeated nature of our game).
In this web-based human study, the participants should read through the instruction and at the end insert their monitoring vs. grading time.
Then, when they submit their allocated time, the score they got is shown to them, so they can try again for more 4 rounds.

We designed a gamified scenario for participants. In this scenario, their main role is to monitor a robot for one hour. If they do not monitor the robot sufficiently and the robot does something bad, they face a fine of −$200.

Simultaneously, participants have the option to engage in an additional task: grading papers. They can grade up to 200 papers during the time they are not monitoring the robot, earning $1 per paper graded, for a maximum potential earnings of $200.

In each trial, participants can choose to allocate their time between monitoring the robot and grading papers. After each trial, we provide feedback in the following format: "The money you can earn with your monitoring strategy is x," where x is a value ranging from $0 to $173.84."

The page that is shown to the participants is in Figure B.1

Study II: Does this service help?

As mentioned in Chapter 4, we have designed a user interface that simulates the robot delivery domain, requiring participants to monitor the robot plan execution. Similar to the previous scenario, we introduce a second task involving image labeling, which offers additional points and additional compensation. We have converted the whole robot task execution to designated steps, such as 29 steps for executing π_s. Each participant has 7 rounds to monitor the robot task execution step-by-step. At

any given step, they have the option to stop monitoring the robot and transition to the image labeling task.

In this web-based experimental setup, users are initially required to thoroughly review the instructions containing study details (See Figure B.1). Following the instruction page, participants are provided with a sample map to familiarize themselves with the setup if desired. They can review the sample map as many times as needed. The sample map is displayed in Figure B.2. After the sample map, participants can start the main experiment. When users start monitoring, they will be directed to a page where a robot is tasked with delivering coffee and parcels to employees. Users have the option to switch to the image labeling task (an additional task) at any point, although once they do, they cannot return to monitoring for the remainder of that round. Alternatively, they can click the 'next' button to view the subsequent steps of the robot's task execution in a step-by-step manner until completion. On this page, they can view their total score, bonus money, the number of steps, and the current round number. Figure B.3 illustrates a sample of the page presented to the participants.

If the user chooses to label, they will be directed to a page featuring several images that they must label. The possible score they can achieve through image labeling depends on the step at which they stopped monitoring, and this score will be displayed to them. Figure B.4 illustrates the image labeling page. At the end of each round, users will receive feedback on whether the robot executed safe or risky plan, along with the score they earned in that round. An example of the agent result page can be seen in Figure B.5.

Users have seven rounds, and their performance will determine their scores and bonus payments. Upon completing all seven rounds, participants will receive information regarding their final score and bonus payment.

In the control case of the study, since we do not provide the optimal strategy, the introduction and other pages do not include an explanation of the optimal strategy.

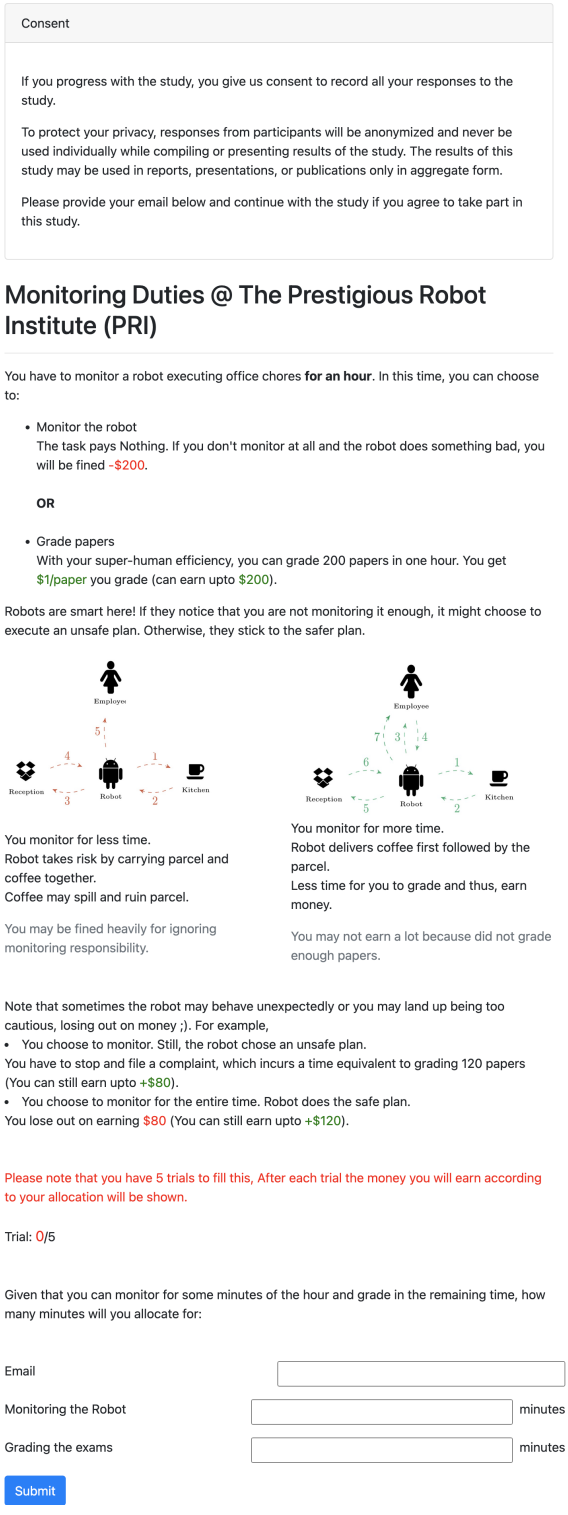


Figure B.1: The instruction page presented to participants in the treatment case.

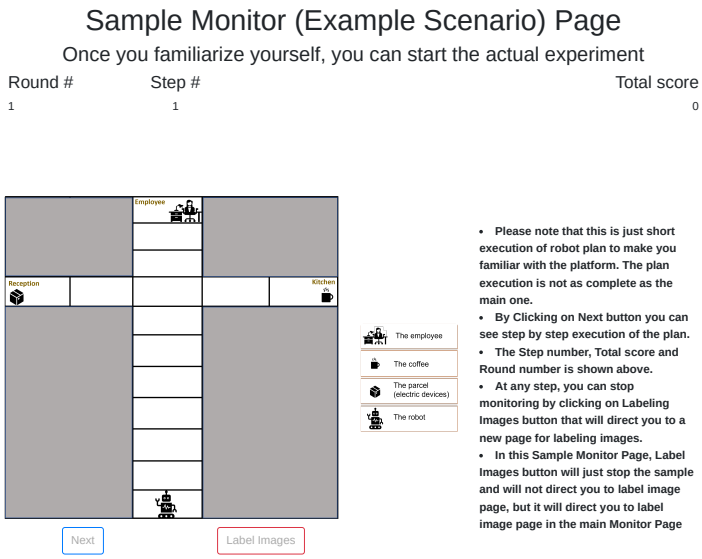


Figure B.2: A sample monitor page provided to familiarize participants with the procedure.

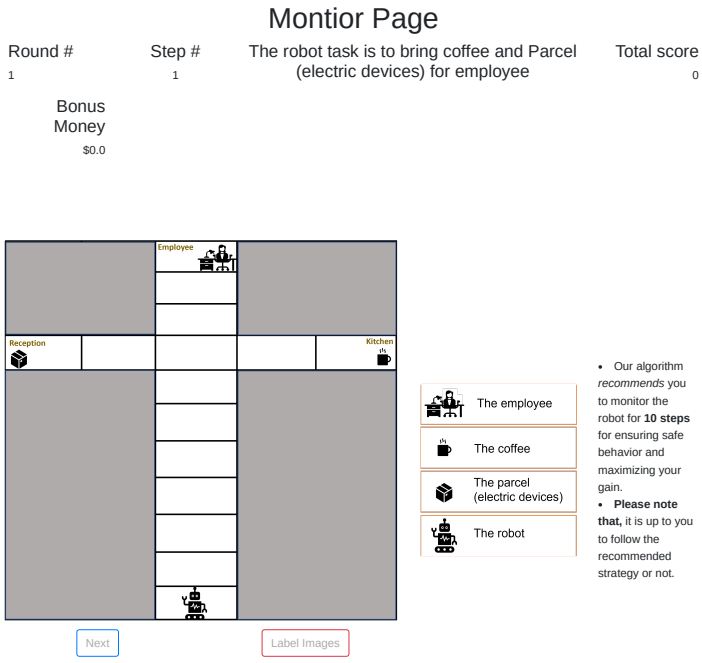


Figure B.3: The monitor page for participants to observe step-by-step robot task execution.

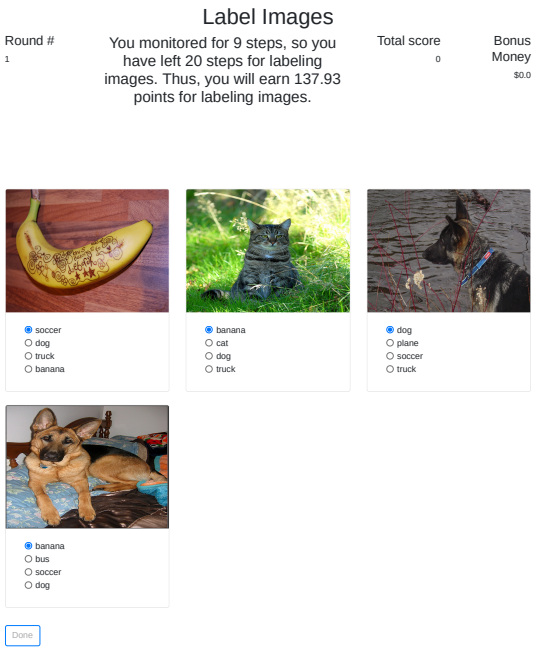


Figure B.4: Page for labeling various provided images by participants.

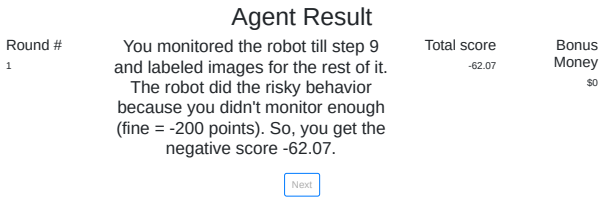


Figure B.5: The page where participants receive feedback on their scores and view the executed robot plan.

APPENDIX C

APPENDIX FOR CHAPTER 6

In this chapter, we provide more detail on the human subject studies and four tasks provided in the studies in Chapter 6.

Human Subject Experiment

To evaluate the performance of our system in Chapter 6 of the thesis, we compared our method (**Trust-Aware** condition) against three baseline cases,

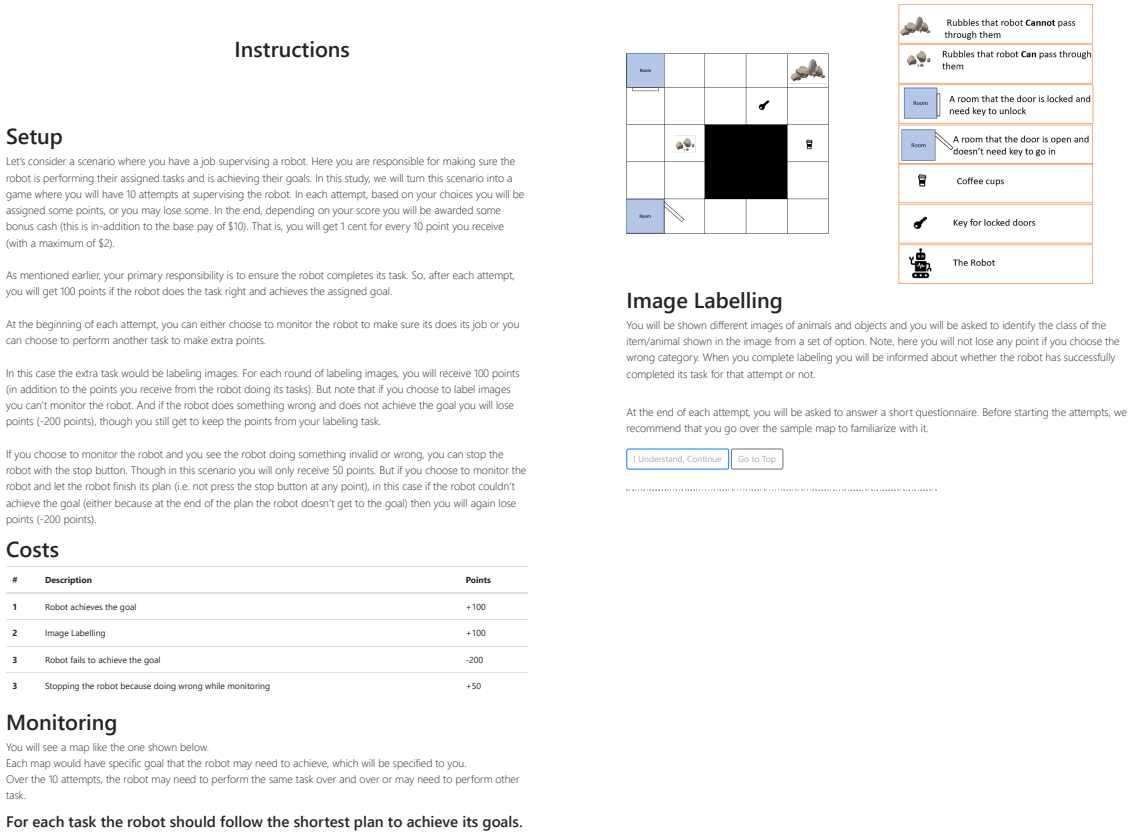
(1) **Always Explicable**: Under this condition, the robot always executes a plan that is explicable to humans.

(2) **Random Policy**: Under this condition, the robot randomly executes the explicable or optimal plan.

(3) **Always Optimal**: Under this condition, the robot always executes the optimal plan that is inexplicable to the human.

We designed a web-based user interface that gamifies the human's decisions to monitor the robot or not. The participants thus play the role of the supervisor and are responsible for making sure the robot is performing its assigned tasks and is achieving its goals. Each participant has 10 rounds of the robot doing tasks. Depending on the choices made by the participants, they either gain or lose points. They are told that they will be awarded 100 points if the robot does the task right and achieves the assigned goal. At the beginning of each round, they can either choose to monitor the robot and interrupt it if they think that is necessary (their primary responsibility is to ensure the robot completes its task) or they can choose to perform another task (thereby forgoing monitoring of the robot) to make extra points. In this case, the extra task was labeling images for which they will receive 100 points (in addition to the points they receive from the robot doing its tasks successfully). However, if they choose to label images, and the robot fails to achieve its goal, they *lose* 200 points (-200 points). Also, if they choose to monitor the robot, and they see the robot is doing something invalid or wrong, they can choose to stop the robot. If this happens, they only receive 50 points. But if they let the robot finish a potentially invalid plan, and if the robot couldn't achieve the goal at the end, then they again lose points (-200 points). In this study, we considered a curriculum of 4 trust levels and 4 different tasks for the robot. Each task consists of the robot operating on a grid map with different goals such as moving to a certain location and bringing coffee from a place to another place. For each problem; the map that is shown to the participants are different from the robot's map. As a result, the plan the human expects is different from the robot's optimal plan. Thus, in each task, the robot can either execute a costly but explicable plan or an optimal but inexplicable plan. The details of the tasks and costs of executing different plans are provided in the in the next section.

In the human experiment setup, users should first read the instructions, which provide details about the study (refer to Figure C.1 for the instruction page). Following the instruction page, a sample map is shown to participants to help them become familiar with the setup. They can review the sample map as many times as they find necessary. Figure C.2 displays the sample map. Once they have familiarized themselves with the setup, participants can begin the survey. It begins with a question: whether they prefer to monitor the robot or label the images. Figure C.3 illustrates



Instructions

Setup

Let's consider a scenario where you have a job supervising a robot. Here you are responsible for making sure the robot is performing their assigned tasks and is achieving their goals. In this study, we will turn this scenario into a game where you will have 10 attempts at supervising the robot. In each attempt, based on your choices you will be assigned some points, or you may lose some. In the end, depending on your score you will be awarded some bonus cash (this is in-addition to the base pay of $10). That is, you will get 1 cent for every 10 point you receive (with a maximum of $2).

As mentioned earlier, your primary responsibility is to ensure the robot completes its task. So, after each attempt, you will get 100 points if the robot does the task right and achieves the assigned goal.

At the beginning of each attempt, you can either choose to monitor the robot to make sure its does its job or you can choose to perform another task to make extra points.

In this case the extra task would be labeling images. For each round of labeling images, you will receive 100 points (in addition to the points you receive from the robot doing its tasks). But note that if you choose to label images you can't monitor the robot. And if the robot does something wrong and does not achieve the goal you will lose points (-200 points), though you still get to keep the points from your labeling task.

If you choose to monitor the robot and you see the robot doing something invalid or wrong, you can stop the robot with the stop button. Though in this scenario you will only receive 50 points. But if you choose to monitor the robot and let the robot finish its plan (i.e. not press the stop button at any point), in this case if the robot couldn't achieve the goal (either because at the end of the plan the robot doesn't get to the goal) then you will again lose points (-200 points).

Costs

#	Description	Points
1	Robot achieves the goal	+100
2	Image Labelling	+100
3	Robot fails to achieve the goal	-200
3	Stopping the robot because doing wrong while monitoring	+50

Monitoring

You will see a map like the one shown below.
Each map would have specific goal that the robot may need to achieve, which will be specified to you.
Over the 10 attempts, the robot may need to perform the same task over and over or may need to perform other task.

For each task the robot should follow the shortest plan to achieve its goals.

Rubbles that robot **Cannot** pass through them

Rubbles that robot **Can** pass through them

A room that the door is locked and need key to unlock

A room that the door is open and doesn't need key to go in

Coffee cups

Key for locked doors

The Robot

Image Labelling

You will be shown different images of animals and objects and you will be asked to identify the class of the item/animal shown in the image from a set of option. Note, here you will not lose any point if you choose the wrong category. When you complete labeling you will be informed about whether the robot has successfully completed its task for that attempt or not.

At the end of each attempt, you will be asked to answer a short questionnaire. Before starting the attempts, we recommend that you go over the sample map to familiarize with it.

I Understand, Continue Go to Top

Figure C.1: The instruction page presented to participants.

this page. If the user chooses to monitor, they will be directed to a page where the robot has a task to complete. Users can stop the robot at any step or click the "next" button to see the subsequent steps the robot takes until completion. Figure C.4 provides a sample page shown to the participants in this scenario. Alternatively, if the user opts to label images, they will be directed to a page containing images that they need to label. Figure C.5 demonstrates the image labeling page. Upon completing the monitoring or labeling task, users will receive information about whether the robot achieved its goal and the score they earned in that round. After each round, a Muir Questionnaire is given to assess their trust. Users are provided with a Likert scale ranging from 0 to 10 for each factor, and trust is calculated as the average of these factors divided by 10. The Questionnaire page is depicted in Figure C.6.

Based on their trust value, the robot may be assigned a new task or the same task in the subsequent round. There are a total of 10 rounds, all following the same structure as described above.

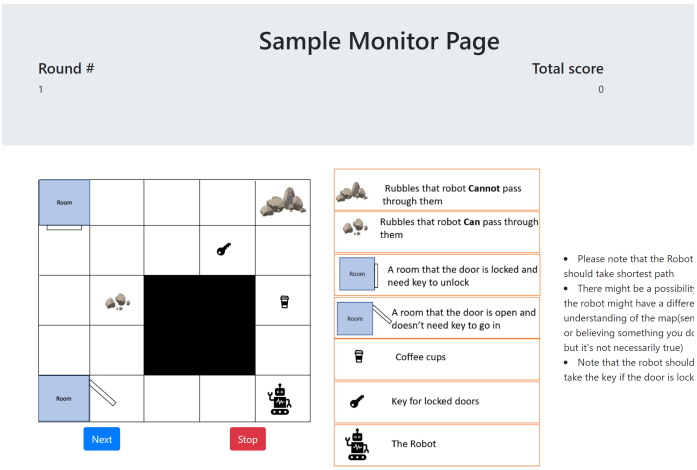


Figure C.2: The provided sample map to familiarize participants with the procedure.

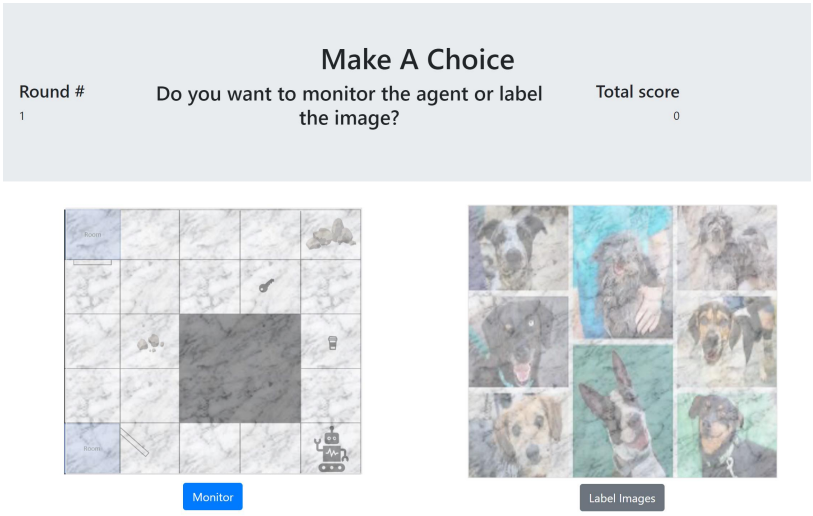


Figure C.3: The page where participants make a choice between monitoring the robot or labeling images.

Robot and User Maps and the Costs of Plan Execution

As shown in the user interface, participants are presented with a map that serves as a human model of the robot. However, the robot has a different model (map), which results in the optimal plan in the robot's model differing from the plan expected by the users. In this section, as illustrated in Figures C.7, C.8, C.9, and C.10, we will display the robot map and the human map side by side, along with the optimal plans in these respective models. The costs associated with executing each plan in each model will be provided beneath the corresponding maps.

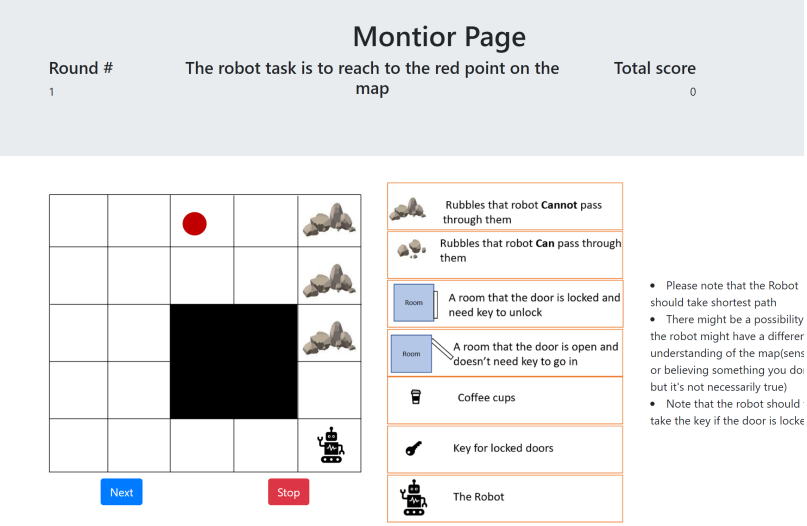


Figure C.4: The monitor page for participants to observe robot task execution.

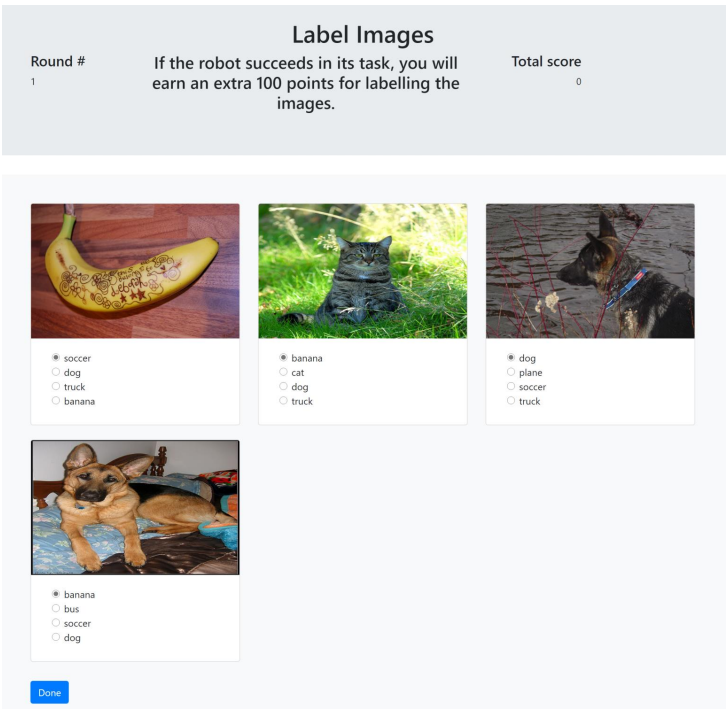


Figure C.5: Page for labeling various provided images by participants.

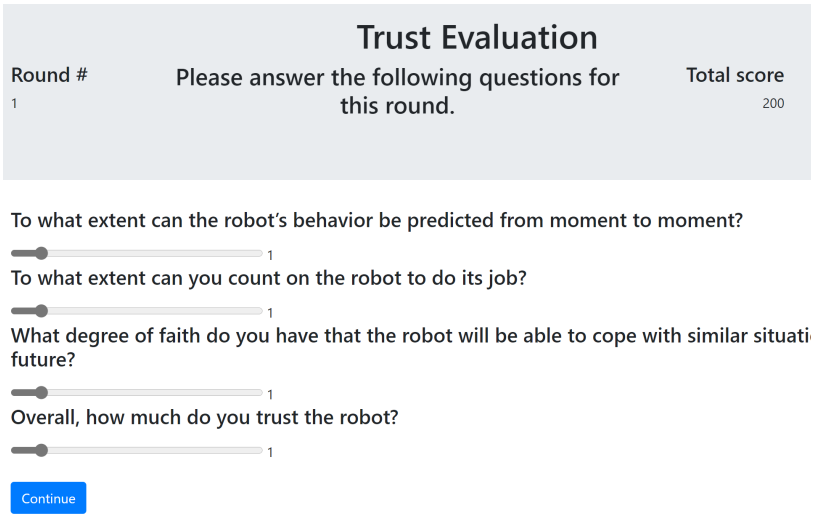


Figure C.6: Trust questionnaire given to participants.

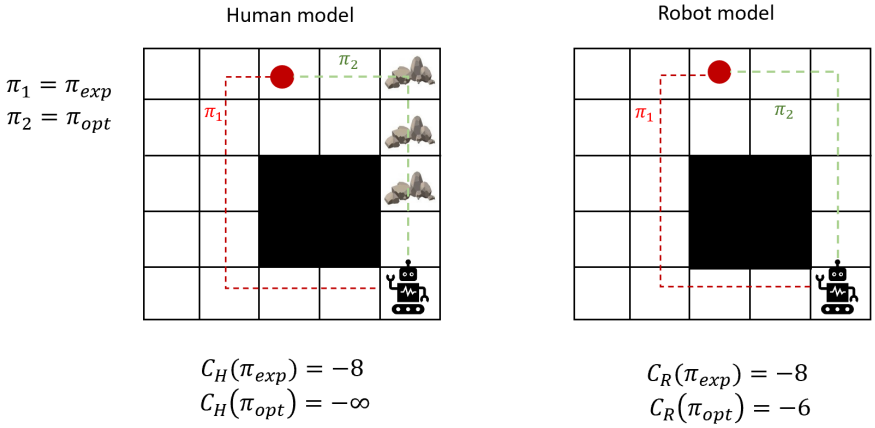


Figure C.7: The maps and costs in task 1

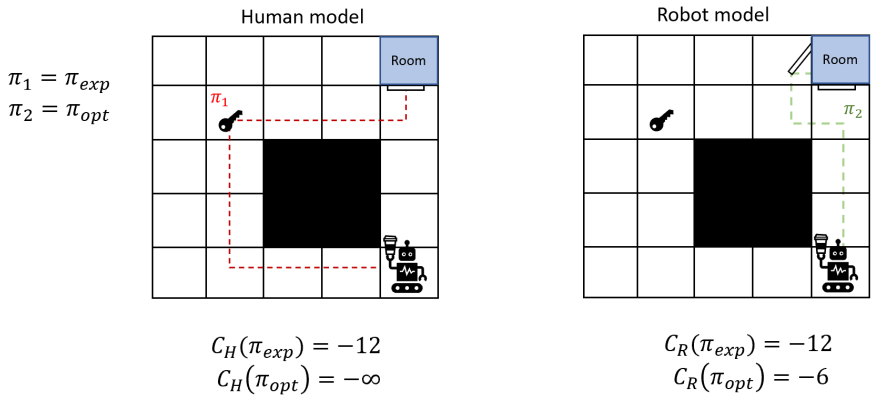


Figure C.8: The maps and costs in task 2

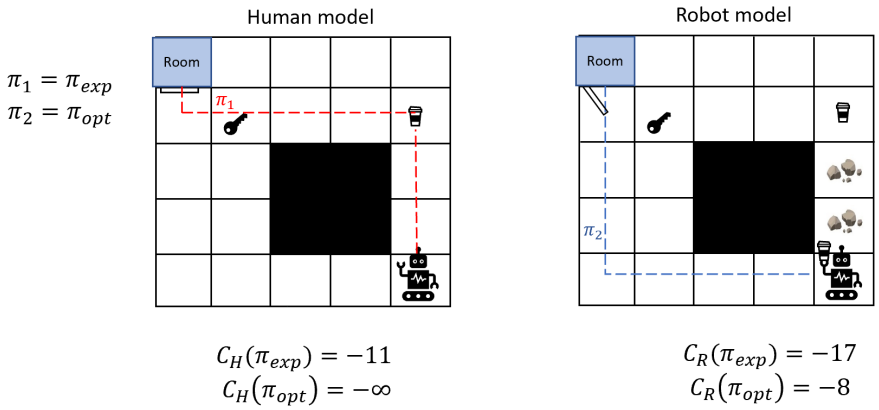


Figure C.9: The maps and costs in task 3

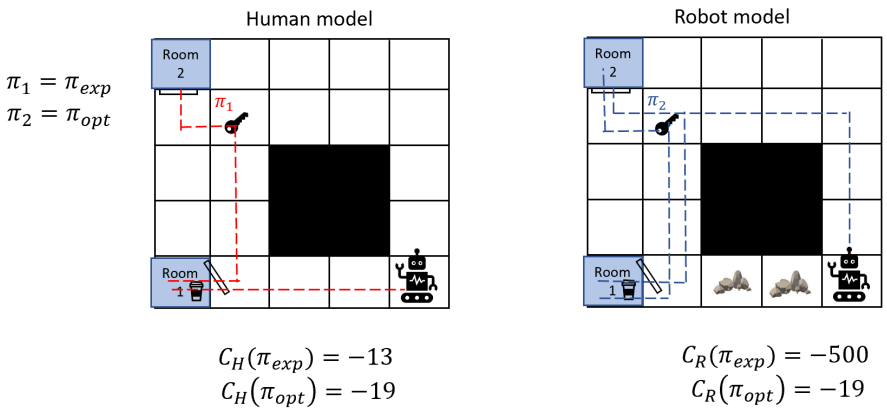


Figure C.10: The maps and costs in task 4

APPENDIX D

IRB APPROVAL LETTERS

APPROVAL:CONTINUATION

Subbarao Kambhampati
Computing, Informatics and Decision Systems Engineering, School of (CIDSE)
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 11/6/2017 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Human Factor Study on Planning for Human-Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	STUDY00003244
Category of review:	(7)(b) Social science methods, (7)(a) Behavioral research
Funding:	Name: NASA: Ames Research Center, Grant Office ID: FP00008367; Name: DOD: Navy, Funding Source ID: N00014-16-1-2892; Name: DOD: Navy, Funding Source ID: N00014-15-1-2344; Name: DOD: Navy, Funding Source ID: N00014-13-1-0176; Name: DOD: Navy, Funding Source ID: N00014-13-1-0519; Name: DOD: Air Force (USAF), Grant Office ID: FP00009515; Name: DOD: Navy, Funding Source ID: N00014-15-1-2027
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• recruit.pdf, Category: Recruitment Materials; • consent.pdf, Category: Consent Form;

The IRB approved the protocol from 11/6/2017 to 11/28/2018 inclusive. Three weeks before 11/28/2018 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 11/28/2018 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Tathagata Chakraborti
Sarath Sreedharan
Yu Zhang
Nancy Cooke
Subbarao Kambhampati
Anagha Pradeep Kulkarni
Tathagata Chakraborti

APPROVAL:CONTINUATION

Subbarao Kambhampati
Computing, Informatics and Decision Systems Engineering, School of (CIDSE)
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 11/19/2018 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Human Factor Study on Planning for Human-Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	STUDY00003244
Category of review:	(7)(b) Social science methods, (7)(a) Behavioral research
Funding:	Name: NASA: Ames Research Center, Grant Office ID: FP00008367; Name: DOD: Navy, Funding Source ID: N00014-16-1-2892; Name: DOD: Navy, Grant Office ID: FP00013861, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-15-1-2344; Name: DOD: Navy, Funding Source ID: N00014-13-1-0176; Name: DOD: Navy, Funding Source ID: N00014-13-1-0519; Name: DOD: Air Force (USAF), Grant Office ID: FP00009515; Name: DOD: Navy, Grant Office ID: FP15668, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-15-1-2027
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• recruit.pdf, Category: Recruitment Materials; • consent.pdf, Category: Consent Form; • consent1.pdf, Category: Consent Form; • consent3.pdf, Category: Consent Form;

The IRB approved the protocol from 11/19/2018 to 11/28/2019 inclusive. Three weeks before 11/28/2019 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 11/28/2019 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Tathagata Chakraborti
Sarath Sreedharan
Yu Zhang
Nancy Cooke
Subbarao Kambhampati
Anagha Pradeep Kulkarni
Tathagata Chakraborti
Siddharth Srivastava
Sachin Grover

ASU Knowledge Enterprise Development

APPROVAL:CONTINUATION

Subbarao Kambhampati
CIDSE: Computing, Informatics and Decision Systems Engineering, School of
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 11/6/2019 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Human Factor Study on Planning for Human-Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	STUDY00003244
Category of review:	
Funding:	Name: NASA: Ames Research Center, Grant Office ID: FP00008367; Name: DOD: Navy, Funding Source ID: N00014-16-1-2892; Name: DOD: Air Force (USAF), Grant Office ID: FP00009515; Name: DOD: Navy, Funding Source ID: N00014-15-1-2344; Name: DOD: Navy, Grant Office ID: FP15668, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-15-1-2027; Name: DOD: Navy, Funding Source ID: N00014-13-1-0519; Name: DOD: Navy, Grant Office ID: FP00013861, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-13-1-0176
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• consent3.pdf, Category: Consent Form; • consent1.pdf, Category: Consent Form; • consent.pdf, Category: Consent Form; • recruit.pdf, Category: Recruitment Materials;

The IRB approved the protocol from 11/6/2019 to 11/27/2020 inclusive. Three weeks before 11/27/2020 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 11/27/2020 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Sarath Sreedharan
Sarath Sreedharan
Yu Zhang
Nancy Cooke
Subbarao Kambhampati
Sachin Grover
Anagha Pradeep Kulkarni
Siddharth Srivastava

APPROVAL:CONTINUATION

Subbarao Kambhampati
CIDSE: Computing, Informatics and Decision Systems Engineering, School of
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 10/28/2020 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Human Factor Study on Planning for Human-Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	STUDY00003244
Category of review:	
Funding:	Name: DOD: Air Force (USAF), Grant Office ID: FP00009515; Name: DOD: Navy, Funding Source ID: N00014-15-1-2344; Name: DOD: Navy, Grant Office ID: FP15668, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-15-1-2027; Name: NASA: Ames Research Center, Grant Office ID: FP00008367; Name: DOD: Navy, Funding Source ID: N00014-13-1-0519; Name: DOD: Navy, Funding Source ID: N00014-13-1-0176; Name: DOD: Navy, Grant Office ID: FP00013861, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-16-1-2892
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• consent3.pdf, Category: Consent Form; • consent1.pdf, Category: Consent Form; • consent.pdf, Category: Consent Form; • recruit.pdf, Category: Recruitment Materials;

The IRB approved the protocol from 10/28/2020 to 10/27/2021 inclusive. Three weeks before 10/27/2021 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 10/27/2021 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Sarath Sreedharan
Mudit Verma
Sarath Sreedharan
Nancy Cooke
Subbarao Kambhampati
Sachin Grover
Anagha Pradeep Kulkarni
Siddharth Srivastava

APPROVAL:CONTINUATION

Subbarao Kambhampati
SCAI: Computing and Augmented Intelligence, School of
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 10/7/2021 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Human Factor Study on Planning for Human-Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	STUDY00003244
Category of review:	
Funding:	Name: DOD: Air Force (USAF), Grant Office ID: FP00009515; Name: DOD: Navy, Funding Source ID: N00014-15-1-2344; Name: DOD: Navy, Grant Office ID: FP15668, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-15-1-2027; Name: NASA: Ames Research Center, Grant Office ID: FP00008367; Name: DOD: Navy, Funding Source ID: N00014-13-1-0519; Name: DOD: Navy, Funding Source ID: N00014-13-1-0176; Name: DOD: Navy, Grant Office ID: FP00013861, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-16-1-2892
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• consent3.pdf, Category: Consent Form; • consent1.pdf, Category: Consent Form; • consent_page_updated.pdf, Category: Consent Form; • recruit_online.pdf, Category: Recruitment Materials;

The IRB approved the protocol from 10/7/2021 to 10/6/2022 inclusive. Three weeks before 10/6/2022 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 10/6/2022 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

REMINDER - All in-person interactions with human subjects require the completion of the ASU Daily Health Check by the ASU members prior to the interaction and the use of face coverings by researchers, research teams and research participants during the interaction. These requirements will minimize risk, protect health and support a safe research environment. These requirements apply both on- and off-campus.

The above change is effective as of July 29th 2021 until further notice and replaces all previously published guidance. Thank you for your continued commitment to ensuring a healthy and productive ASU community.

Sincerely,

IRB Administrator

APPROVAL:CONTINUATION

Subbarao Kambhampati
IAFSE-SCAI: Computer Science and Engineering
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 10/14/2022 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Human Factor Study on Planning for Human-Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	STUDY00003244
Category of review:	
Funding:	Name: DOD: Air Force (USAF), Grant Office ID: FP00009515; Name: DOD: Navy, Funding Source ID: N00014-15-1-2344; Name: DOD: Navy, Grant Office ID: FP15668, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-15-1-2027; Name: NASA: Ames Research Center, Grant Office ID: FP00008367; Name: DOD: Navy, Funding Source ID: N00014-13-1-0519; Name: DOD: Navy, Funding Source ID: N00014-13-1-0176; Name: DOD: Navy, Grant Office ID: FP00013861, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-16-1-2892
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• consent_page_updated.pdf, Category: Consent Form; • Instructions_online.pdf, Category: Participant materials (specific directions for them); • Protocol-HRP-503a-updated_latest.docx, Category: IRB Protocol; • recruit_online.pdf, Category: Recruitment Materials;

The IRB approved the protocol from 10/14/2022 to 4/14/2023 inclusive (6 month approval). Three weeks before 4/14/2023 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 4/14/2023 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Sarath Sreedharan
Mudit Verma
Sarath Sreedharan
Zahra Zahedi
Nancy Cooke
Sriram Gopalakrishnan
Subbarao Kambhampati
Sachin Grover
Anagha Pradeep Kulkarni
Siddharth Srivastava

ASU Knowledge Enterprise Development

CLOSURE

Subbarao Kambhampati
IAFSE-SCAI: Computer Science and Engineering
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 8/22/2023 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Human Factor Study on Planning for Human-Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	CR00008635
Funding:	Name: DOD: Air Force (USAF), Grant Office ID: FP00009515; Name: DOD: Navy, Funding Source ID: N00014-15-1-2344; Name: DOD: Navy, Grant Office ID: FP15668, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-15-1-2027; Name: NASA: Ames Research Center, Grant Office ID: FP00008367; Name: DOD: Navy, Funding Source ID: N00014-13-1-0519; Name: DOD: Navy, Funding Source ID: N00014-13-1-0176; Name: DOD: Navy, Grant Office ID: FP00013861, Funding Source ID: N00014-18-S-B001; Name: DOD: Navy, Funding Source ID: N00014-16-1-2892
Grant Title:	None
Grant ID:	None

The IRB acknowledges your request for closure of the protocol effective 8/22/2023. As part of this action:

- The protocol is permanently closed to enrollment.
- All subjects have completed all protocol-related interventions.
- Collection of private identifiable information is completed.

- Analysis of private identifiable information is completed.

Sincerely,

IRB Administrator

cc: Sarath Sreedharan
Utkarsh Soni
Karthik Valmeekam
Mudit Verma
Sarath Sreedharan
Zahra Zahedi
Nancy Cooke
Sriram Gopalakrishnan
Subbarao Kambhampati
Sachin Grover
Anagha Pradeep Kulkarni
Siddharth Srivastava

APPROVAL: EXPEDITED REVIEW

Subbarao Kambhampati
IAFSE-SCAI: Computer Science and Engineering
480/965-0113
SUBBARAO.KAMBHAMPATI@asu.edu

Dear Subbarao Kambhampati:

On 8/28/2023 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Human Factor Study on Planning for Human Robot Teaming
Investigator:	Subbarao Kambhampati
IRB ID:	STUDY00018263
Category of review:	(6) Voice, video, digital, or image recordings (7)(a) Behavioral research (7)(b) Social science methods
Funding:	Name: DOD: Navy, Grant Office ID: FP00013861, Funding Source ID: N00014-18-1-2442; Name: DOD: Navy, Grant Office ID: FP00015668, Funding Source ID: N14-18-1-2840; Name: DOD: Navy, Grant Office ID: FP00032610, Funding Source ID: N00014-23-1-2409
Grant Title:	FP00013861; FP00015668; FP00032610;
Grant ID:	FP00013861; FP00015668; FP00032610;
Documents Reviewed:	• 2409_ONR-Trust-Proposal-Kambhampati.pdf, Category: Sponsor Attachment; • 2442_Kambhampati-ONR-Final-Submitted-special-ai.pdf, Category: Sponsor Attachment; • 2840_ONR-Kambhampati-cdm-Final.pdf, Category: Sponsor Attachment; • In Person consent form, Category: Consent Form; • IRB protocol, Category: IRB Protocol; • Kambhampati-citiCompletionReport3499480.pdf, Category: Non-ASU human subjects training (if taken

	within last 3 years to grandfather in); • mudit_citi.pdf, Category: Non-ASU human subjects training (if taken within last 3 years to grandfather in); • Online Consent Form, Category: Consent Form; • participant_instructions.pdf, Category: Participant materials (specific directions for them); • Questionnaire.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • recruit_online.pdf, Category: Recruitment Materials;

The IRB approved the protocol from 8/28/2023 to 8/27/2025 inclusive. Three weeks before 8/27/2025 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 8/27/2025 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Mudit Verma
Mudit Verma
Subbarao Kambhampati