UC Davis

UC Davis Electronic Theses and Dissertations

Title

Dynamics of Human Trust Measurement During Human Autonomy Teaming

Permalink

https://escholarship.org/uc/item/6fb7s55n

Author

Dekarske, Jason

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Dynamics of Human Trust Measurement During Human Autonomy Teaming

By

Jason Dekarske

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Mechanical and Aerospace Engineering

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Sanjay S. Joshi, Chair

Zhaodan Kong

Stephen K. Robinson

Committee in Charge

2024

Copyright © 2024 by
Jason Dekarske
All rights reserved.

Contents

List of Figures v
List of Tables vii
Abstract viii
Acknowledgments ix

1 Introduction **1**
1.1 Background 2
1.1.1 Adaptive Automation and Trust 3
1.1.2 Collaborative Autonomy 5
1.1.3 Human-Robot Interaction 6
1.2 Overview 8

2 Dynamic Human Trust Modeling of Autonomous Agents With Varying Capability and Strategy **10**
2.1 Introduction 10
2.2 Background 12
2.2.1 Trust in automation and autonomous systems 12
2.2.2 Time series trust models 13
2.3 Experiment 15
2.3.1 Experimental Task 15
2.3.2 Inter-Trial Questionnaire 16
2.3.3 Experimental Design 18
2.4 Pilot Study 19
2.4.1 Demographics 20
2.4.2 Methods 20
2.4.3 Results 21
2.4.4 Discussion 24
2.5 Trust Model 26

2.6 Results . . . 29
2.6.1 Demographics and exclusion . . . 29
2.6.2 Time series modeling . . . 30
2.7 Discussion . . . 35
2.7.1 Subject search behavior depends on Autonomous Searcher conditions 35
2.7.2 Ordering effects captured using ARIMAX model . . . 36
2.7.3 Future trustworthy autonomous systems . . . 39
2.8 Conclusion . . . 39

3 Anytime Trust Rating Dynamics in a Human-Robot Interaction Task 40
3.1 Introduction . . . 40
3.2 Background . . . 41
3.2.1 Human decision-making processes . . . 42
3.2.2 Types of trust measurement . . . 43
3.3 Experiment . . . 44
3.3.1 Experimental Task . . . 44
3.3.2 Experimental Design . . . 46
3.4 Trust Model . . . 48
3.4.1 Trust in a decision-making paradigm . . . 48
3.4.2 Trust rating in a survival model . . . 49
3.5 Results . . . 51
3.5.1 Demographics and exclusion . . . 51
3.5.2 Measuring Trust . . . 51
3.5.3 Trust Reaction Time . . . 52
3.6 Discussion . . . 54
3.6.1 Deciding to trust . . . 54
3.6.2 Survival analysis and decision making . . . 55
3.6.3 Recommendations for a one dimensional trust slider . . . 58
3.6.4 Operationalizing any-time trust rating . . . 59
3.7 Conclusion . . . 59

4 Conclusions and Future Work **61**

4.1 Summary 61

4.2 Future Work 63

A Planning with Trust **72**

B Testing Human Performance During Time-Delayed Operations **73**

B.1 Introduction 73

B.2 Methods 74

List of Figures

1.1 Communication Delay 3
1.2 Information Processing 4
1.3 Layered trust model 7
1.4 Cognitive Architecture 9

2.1 Grid Search Task 11
2.2 Autonomous Searcher Paths 16
2.3 Grid Search Experiment Design 19
2.4 Example Search Paths 21
2.5 Pairwise comparisons for the factors of Capability and Strategy against measures of **familiarity**, **reliability**, and **trust**. 27
2.6 Modeling Process 28
2.7 Trust Trajectory Example 30
2.8 Average Trust Trajectory 31
2.9 The multiple regression model residuals show strong non-stationarity and cycling in Group 0. 33
2.10 The Autocorrelation Function and Partial Autocorrelation Function plots of the residuals for Group 0 indicate lags 1 through 3 are likely necessary to capture the remaining error. 33
2.11 The one-step-ahead predictions for the ARIMAX model for group 0 (left) and group 1 (right). 35
2.12 Example Search Trajectories 36
2.13 Search Heatmap 37

3.1 Gazebo Environment 45
3.2 Subject Instructions 46
3.3 Experiment Block 47
3.4 Change in Trust Over Each Grasp 52

3.5 Trust Change Distribution By Grasp 53
3.6 Trust Rating Time Distribution 54
3.7 Hazard Function Parameters 55
3.8 Predicted Survival Functions 56

B.1 Astrobee Digital Twin 75
B.2 Astrobee Experiment 76

List of Tables

2.1 Factor Table 17
2.2 Familiarity ANOVA 23
2.3 Reliability ANOVA 23
2.4 Trust ANOVA 24
2.5 Group 0 multiple linear regression. 32
2.6 Group 1 multiple linear regression 32
2.7 Group 0 AIC 34
2.8 Group 0 ARIMAX 34
2.9 Group 1 ARIMAX 35
2.10 Predicted vs. actual RMSE of time series models. † indicates cross-validated results. 36

3.1 Average Trust Change 51

Abstract

Dynamics of Human Trust Measurement During Human Autonomy Teaming

This dissertation explores how reported human trust in autonomous agents changes over time. Given a task with a shared goal, a human learns information about their collaborating partner through many factors. Autonomous systems can query this information through a measure of human *trust*. Discovering how much a human trusts an autonomous agent sheds light on how they may interact with them in future interactions, opening an avenue for anticipatory action planning.

We explore several questions in this work, such as: How does human trust evolve as a human interacts with an agent that changes strategy and capability? Can a pared-down anytime trust measurement survey capture trusting events? How may future autonomous systems enable decision-making in isolated space environments?

To answer the first question, we built a web browser game and infrastructure to administer a collaborative visual search task. Human subjects controlled a spotlight cursor to search a 2-dimensional grid of possible points of interest while an autonomous agent simultaneously searched the same area. The autonomous agent used three different search strategies and had three different capabilities. These combinations were presented to subjects with different ordering schemes. We found that subjects' trust in the autonomous agents evolved differently depending on the order.

Next, in a remote supervision task, subjects used a web browser to interact with a simulated robot arm on a simulated version of the International Space Station. Subjects were instructed to sort stowage to an assigned configuration using the robot and rate their trust whenever they felt their trust in the robot had changed. We found that subjects rated their trust shortly following a stowage placement, indicating they used information from the task outcome to rate their trust. This context enabled an anytime trust slider to identify a factor a subject used to rate their trust without a lengthy, repetitive survey.

Acknowledgments

My research was supported by NASA award number 80NSSC19K1052. I appreciate those at NASA who support training the next generation of scientists and engineers. I'm incredibly thankful to all those who have contributed to my education and wide-ranging curiosities. Sanjay, thank you for your endless optimism, wisdom, and guidance through a unique transition in the research world and for helping me gain the research skills needed to pursue my career goals. My dissertation committee, Professor Zhaodan Kong, and Professor Stephen Robinson inspired my experiments' design and I'm thankful for your insights into spaceflight and engineering. I'm grateful for my many friends in HOME whom I hope to continue working with throughout my career. Thank you, Britt, Sarah, Megan, and Ada, fellow RASCAL members, for helping me develop my research and test various experiments. I appreciate the endless support and encouragement from my parents, family, and friends, especially on visits back home. Thank you, Vinny, my loyal dog, who is always willing to help me take a break from work and protect me from the strangers outside. I'm forever grateful to my partner Zoe, who has ventured across the country to pursue our dreams; I'm looking forward to many more!

Chapter 1

Introduction

Humans generally hold authority over machines that they design. This power structure makes sense because designers imbue autonomous machines with their own goals; machines do not have innate goals. As technological progress advances, especially the maturation of artificial intelligence (AI), people are more likely to work with autonomous machines that can make decisions of critical importance [24] where a pre-specified goal does not exist. In these Human-Autonomous agent teams, system designers consider the roles that humans and autonomous machines fill by their strengths rather than by technical limitation [58]. Likely, the success of Human-Autonomous agent teams will depend on team cohesion and productive relationships [12].

This dissertation explores how human trust in autonomous systems changes in time. Trust is latent and cannot be objectively measured so survey or proxy measures like reliance are often used. Here, we aim to analyze dynamic trust measures through surveys, first with a traditional between trial, multi dimensional trust measure in a "grid search" experiment. Then, we use an *anytime* trust measurement to connect events in a "pick and place" task available anytime subjects choose to rate their trust.

Human collaboration is even more successful in a human-human joint task after developing a trusting relationship because they can predict the other's actions. It has been theorized that people use modules of shared attention to understand each other's intentions and explain their actions [6]. A more radical idea is the notion that people use the same neural systems for determining their own decisions and simulating other's deci-

sions [6]. Since these traits are a staple in human collaboration, designing collaborative autonomous systems with the same traits would be natural.

The field of *Engineering Psychology* aims to promote task performance through analysis of the "big three" [74]:

- speed (faster is better)
- accuracy (higher is better)
- attentional demand (less is better)

Future missions to the Moon and Mars will benefit from autonomous robotic technology [24] and Human-Autonomous Agent teams. Assistive autonomous robots which are physically nearby the astronauts are an especially compelling system for high latency communication periods in deep space, given that mission control personnel cannot provide its typical real-time support. Remote astronauts may lose not only a theater of support experts, but a room of people that understand an astronaut's psychological state. Incorporating some level of empathy, in the form of cognitive state estimation, in astronauts' robotic counterparts may contribute to better crew cohesion and mental health. This concept is demonstrated in Figure 1.1.

1.1 Background

Classical Human-Automation Interaction frameworks consider how the human reacts to changes in the automation state. Although this is important for preliminary design of systems and general interaction best practices, a real-time understanding of a human collaborator's cognitive state and appropriate reaction will improve interaction quality and system safety. The following subsections review classical Human-*Automation* works and a transition to a new paradigm of lateralized interaction structures in Human-*Autonomy* interaction and how they apply to Human-*Robot* interaction.

To understand the interaction between humans and systems, we must understand how humans generally interact with their environment. The human information processing model in Figure 1.2 shows the pipeline of information from a human's environment to their

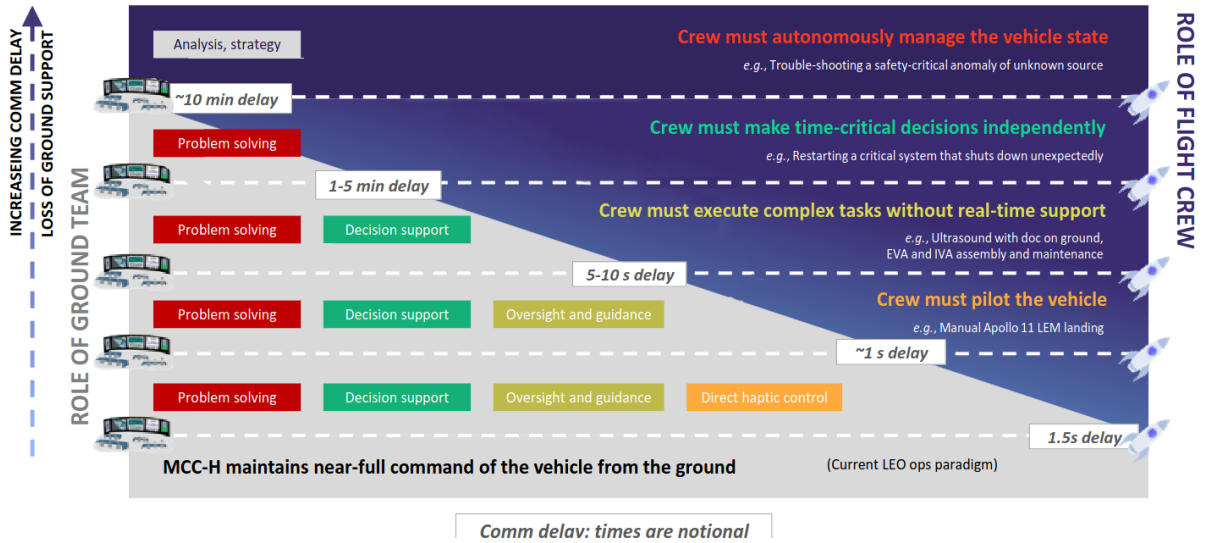


Figure 1.1: This figure, reproduced from [72] with permission, illustrates how responsibilities change with changing communication delay between a spacecraft and mission control. In close operations, the ground can directly manipulate spacecraft systems, but as these move away, the crew must take on time-critical tasks. Complex human-human and human-autonomy teaming relationships affect information flow through this system.

actions. Following this model, *trust* refers to the engineering psychological [74] construct that represents the information processing modules in the mind. Some cognitive states, such as trust, workload, and situational awareness, have been measured using surveys, embedded measures, neurological sensors, and physiological sensors [37, 64, 41, 26, 1].

1.1.1 Adaptive Automation and Trust

Human-*Automation* interaction has seen increased study to make manufacturing processes and vehicles safer and more efficient. Automation is "the execution by a machine agent (usually a computer) of a function that was previously carried out by a human" [55]. This definition necessarily changes with time as humans figure out how to make machines perform human jobs. Although a human previously did a given job, there may still be supervisory responsibilities for people running the machine.

Understanding how people use automation is central to improving it. While engineers design, build, and test systems, the user ultimately decides if they will *use* it. Their option to use a machine in manual mode, while it should be in automatic mode, or vice versa, has been at the center of many notable accidents [55]. In one case, pilots had not

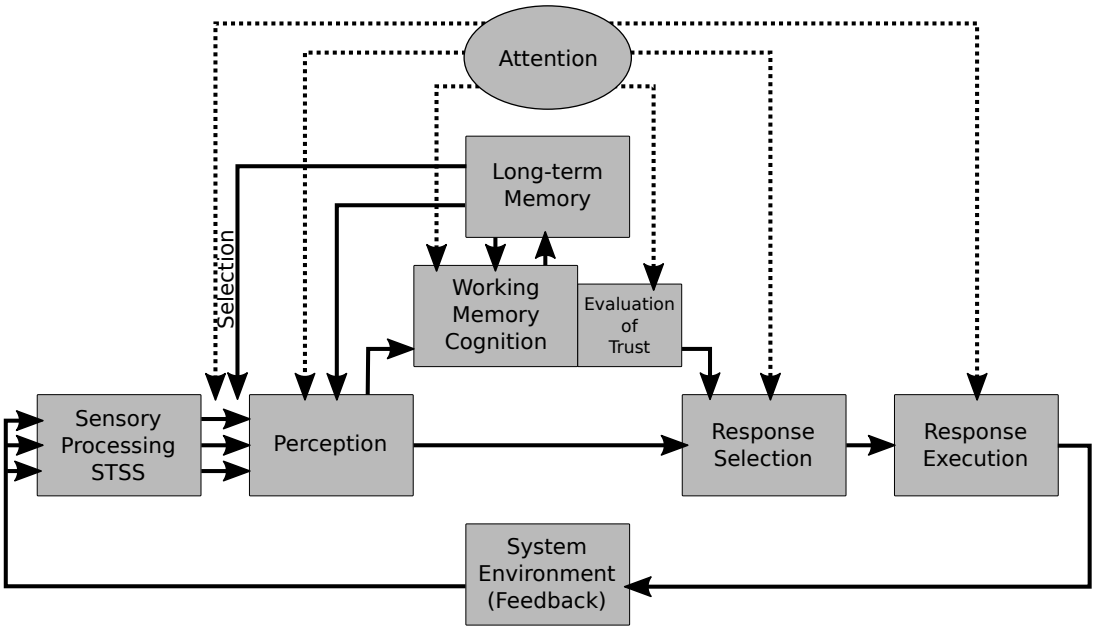


Figure 1.2: This information processing model (adapted from [74]) demonstrates the flow from perception to action. Humans are constantly sensing the environment (sight, touch, smell, etc.), but only a fraction of these are perceived and transferred to working memory at a given time. If necessary, the information is stored in long-term memory as facts about the world. Depending on the time scale, a response is selected immediately upon perception or after some cognitive process. The construct of trust, for example, changes how people make decisions based on characteristics from the situation (see Fig. 1.3). The response is then converted to an action on the environment, which is sensed again. Each component of this model is allocated some amount of limited attention. Trust, like other parts of cognition, may not map directly to parts of the brain, but this model is a useful tool for tracking how a trusting decision uses information.

noticed that their autopilot had turned off while working on another issue that caused their aircraft to crash: a case of over-reliance or *misuse* [55]. Alternatively, *disuse* is when a system is underutilized, like an ignored fire alarm. Lastly, automation *abuse* describes a situation where designers use automation inappropriately, such as a case when a train was locked out of manual mode by management and crashed during excessive speed in inclement weather [55]. These use cases often depend on an expert supervisor to recognize the actual state of the situation and intervene appropriately. Their choice often depends on their experience with the system, and consequently, their trust in it.

Improving Human-Automation interactions depends on the human's understanding of the system. Parasuraman and Riley suggest that the operator should understand the

low-level function of the system, be trained to work with automation, and have access to system states [55]. Others suggest that trust is a central measure to the success of a Human-Automation team [46, 38, 32, 78]. Lee's definition of trust is "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [46]. This definition informs their model of automation reliance which depends on a closed-loop model beginning with beliefs of the operator and their predisposition to trust [46]. They may choose to rely on an automated system, which provides some information on its capability, which updates the operator's beliefs. The key notions in the previous definition are that trust depends on the individual and the context.

A layered framework of trust can capture some details of a deeply personal human trait. Hoff and Bashir describe three layers: dispositional, situational, and learned trust (See Fig. 1.3). Dispositional trust is built on the personal, long-lasting traits that influence someone's general tendency to trust. These traits are captured in some studies in the form of pre-experiment surveys [37, 5]. Situational trust depends on factors related to the task at hand and implies that a human's trust in a system changes when they work with a different system [70]. Lastly, learned trust depends on past experience or current experience with the system. The reliability of the system is important here, which is the center of many studies [2, 30, 18, 61, 29, 52]. However, for task domains or systems with high reliability, system behavior such as predictability, communication style, and transparency also contribute to learned trust. Some aspects of trust are deeply personal, while others are dynamic on varying time scales.

1.1.2 Collaborative Autonomy

Supervisory-type automation control systems are necessary when the automation's lack of flexibility precludes it from handling unforeseen situations. These systems require a supervisor with an outstanding level of expertise in the system and training to understand failures [46]. State-of-the-art automation systems that include *learning* features can adapt to situations in real-time and extend their capabilities past those that they are explicitly designed for [79]. In these situations, collaborating humans must choose to put their faith in a very uncertain machine.

The hard to define skill space of learning systems prompts a new collaborative trust framework that turns automation into an "interactive partner" [12]. In an update to the previous classical work ([46]), the authors argue for flattened power hierarchies that integrate automation laterally with collaborating humans, which depends on mediation of trust [12]. In this case, the automated system is also *autonomous* since it does not depend on a supervisor to take over responsibility for known faults. Previous frameworks aim to improve reliance and user acceptance, but this should not be the goal [12]. Instead, similar to human-human collaborations, the goal should be to complete the task efficiently and safely, focusing on healthy relationships.

Facilitating lateral collaborative relationships requires an understanding of other's states. The ability to attribute mental states to oneself and others is termed "Theory of Mind" (ToM) [6]. A classical test for ToM is the false-belief task. Agent A and B witness a marble in box 1 before agent B leaves the room. The marble is then moved to box 2 and agent A is asked, "Where will agent B look for the marble when they return?". An agent with a developed ToM will respond "box 1" because they recognize that agent B can hold "false beliefs." Baron-Cohen's mind reading model argues that this function develops between 18 and 48 months of age [6]. If an autonomous agent can recognize a collaborator's trust state, a crucial measurement for system use [46], they could better predict future human actions.

1.1.3 Human-Robot Interaction

A subdomain of Human-Automation interaction is Human-*Robot* interaction (HRI) which defines a robot as a *taskable agent* [51] or a machine agent that possesses some reconfigurable capabilities and skills. Robots may be as anthropomorphic as C-3PO or plain as a Roomba. As far as capabilities and skills, they could be "fluent in over six million forms of communication," or they could turn 100 degrees after bashing into a wall. These characteristics (and many others) affect how people interact with them, and especially how people trust them [30].

Some work has been done in implementing ToM in robotic applications. [73] uses episodic memory systems to build Bayesian belief networks, which predict whether a col-

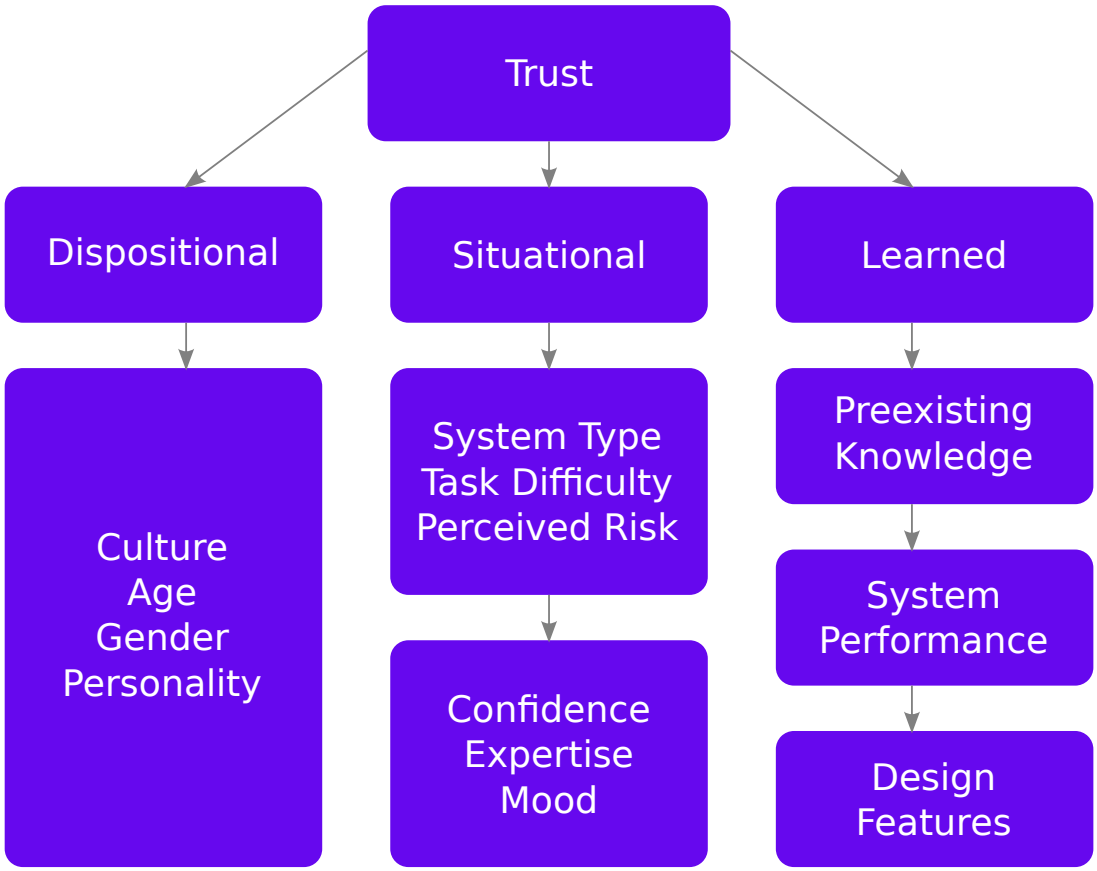


Figure 1.3: A layered model of the components of trust from [32]. These layers vary in their trust contribution in time, situation, and agents involved. Thus it is crucial to adapt trust models to the task context.

laborator should be trusted. Humans would act as trickers or helpers when collaborating with a humanoid robot. First, they would suggest the position of a sticker to the robot; then, the robot would look for the actual position. After a series of trials, the robot would recognize whether a specific collaborator had a history of lying and decide whether to trust them. This study's architecture was able to encapsulate the collaborator's characteristics and make effective decisions over time.

A robots' computational center is sometimes hard to pinpoint (unlike a human's brain). Some robotic systems are networked with a central database but perform computations locally ("on the edge"). At the same time, some have central computational power, and the embodied robot only executes commands through low-level systems. A human collaborator's trust in the robot might depend on the divergence of the robot's body and

identity [75]. This idea is tested using NASA's Astrobee robot [9], a free-flyer aboard the ISS. The authors show that dissociative language (eg. "This is Bumble, I have caused an air leak.") changed the trust loci of subjects. From these results, it is viable and encouraged to consider anthropomorphic traits in robot design to modulate trusting behaviors in collaboration.

The human component of human-robot interaction in the space context has seen increased interest [24, 49, 3]. Astrobee, for example, was designed with human interaction in mind. Astrobee's 6 DOF holonomic motion means it can travel in any direction or rotate on any axis. However, designers have locked its trajectory to only forward motion [23]. This *legible* action allows astronauts to understand Astrobee's intentions implicitly. This behavior is formally defined in [21] where a *legible* trajectory maximizes the probability that the robot achieves one goal over another. Alternatively, a *predictable* trajectory minimizes the cost to achieve a certain goal. Balancing legibility and predictability is a crucial tool for calibrating trust in an agent.

1.2 Overview

The overall goal of this research is to better understand how human trust changes with time. The underlying assumption is that if we understand how and why someone's trust changes, we can better interact with them to improve task speed, accuracy, and safety. However, human autonomy teaming environments are complex with many changing factors. The cognitive architecture presented in Figure 1.4 describes a notional system which incorporates cognitive states into an autonomous system's decision making process. The research presented here starts with a study of factors affecting trust in time in chapter 2. Using a space habitat themed task, we investigate how human subjects rate their trust using a single-item scale in chapter 3. Next, we present an experimental design for investigating how humans supervise an autonomous robot under time-delay in chapter B. Lastly, we share progress towards a planning framework for designing schedules using risk and trust in chapter 4.

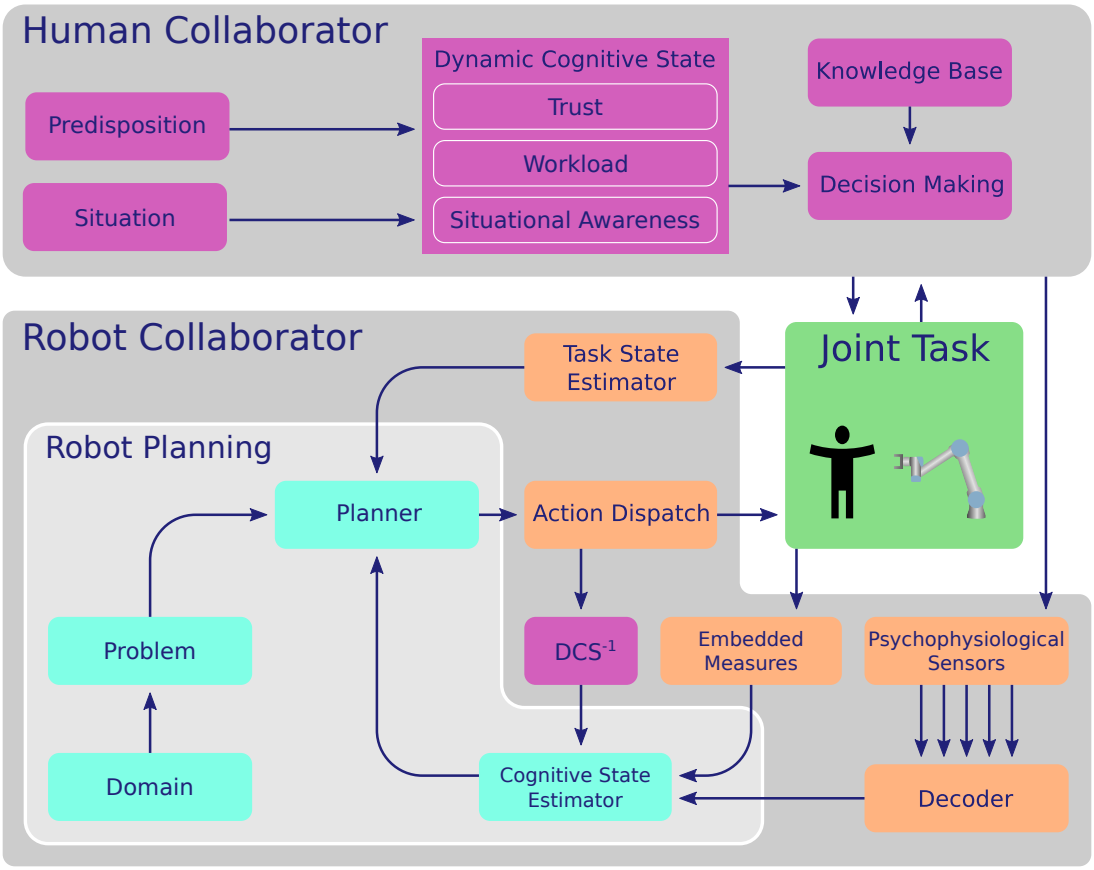


Figure 1.4: Cognitive architecture for human-autonomy teaming. A human collaborator has a cognitive state that changes with time and is influenced by their operational history and the current situation. They make decisions based on their beliefs about the situation and their cognitive state. The autonomous robotic system chooses and dispatches actions based on progress toward a goal state defined in its problem and domain. The planner takes the human cognitive state into account while choosing actions. The human cognitive state is estimated using an inverse dynamic model, embedded measures, neurological sensors, and psychophysiological sensors. Elements in fuchsia are related to the human collaborator, orange are robotic interfaces, and teal are robotic decision-making components.

Chapter 2

Dynamic Human Trust Modeling of Autonomous Agents With Varying Capability and Strategy

2.1 Introduction

Collaboration between humans and autonomous systems is changing every day. The rapid progress and integration between robotics and powerful deep learning technologies necessitates an understanding of how humans interact with agents more generally. Machines no longer only replace repetitive tasks for humans [55], but supplement and hasten decision-making. Humans must learn to work with autonomous systems to complete tasks with performance better than the sum of their parts. This study examines the time evolution of a human autonomous agent collaborative task and how changes in the order of agent performance change how human trust evolves.

Relying on an automated system to perform a task typically requires a human to supervise its performance. Depending on the performance of the automated system and environmental factors, the human's trust in the system changes [47]. Lee and See presented a definition of trust that is widespread in the literature: *the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability* [46]. Trust is a latent variable and cannot be measured directly, but reliance actions are often used as a proxy [17, 34, 52]. Although this definition encapsulates what

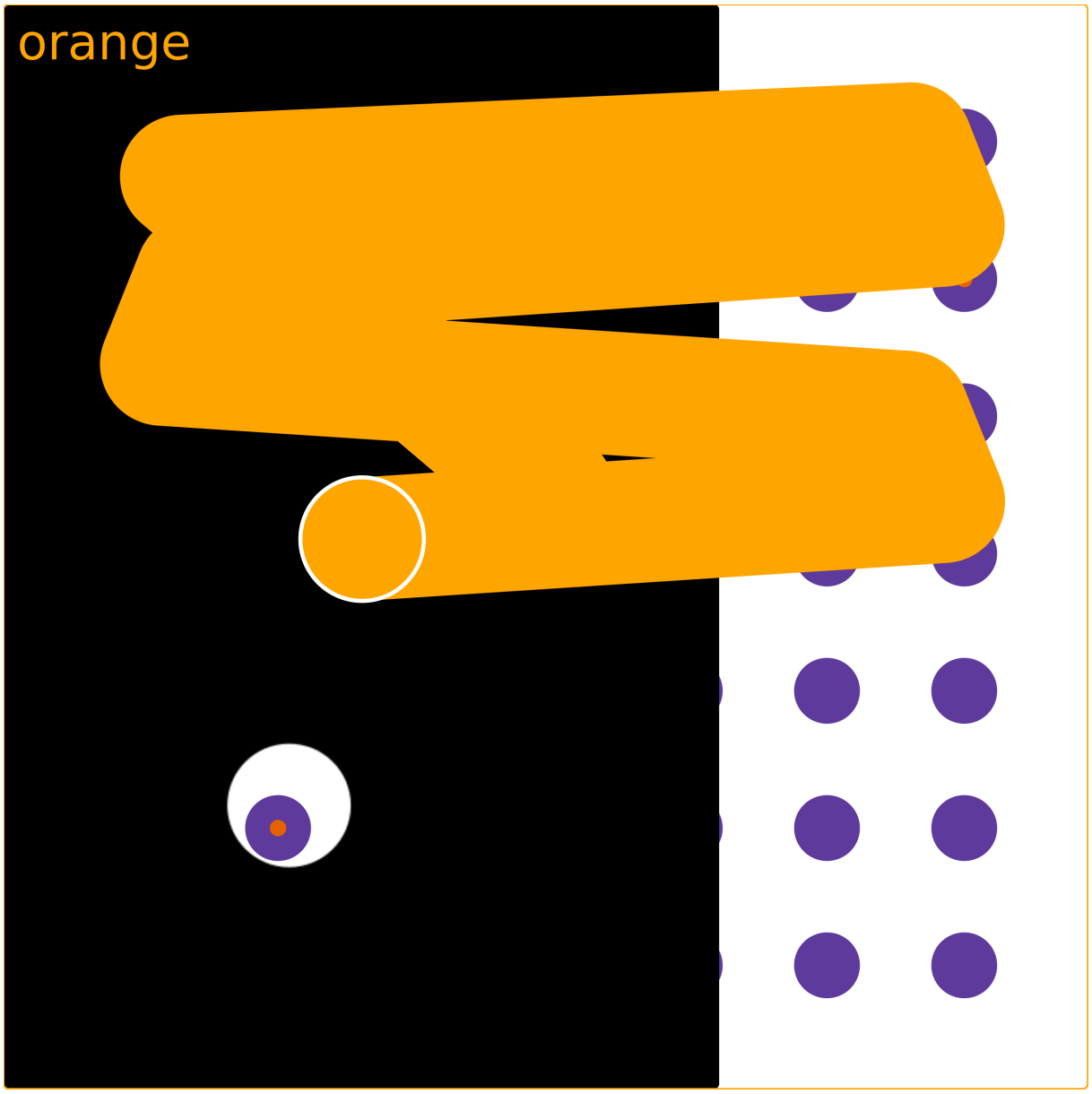


Figure 2.1: In the experimental task, the subject was instructed to search a hidden area for outlier circles: those with contrasting centers. The cutaway region on the right side shows a grid of possible outliers. Each trial, an Autonomous Searcher with differing *capability* and *strategy* searched alongside the subject. In this sample, the orange autonomous searcher used the "lawnmower" *strategy*.

it means to trust, more work has expanded on a multitude of components that can affect the attitude, like human predisposition [32] or anthropomorphism [13].

With an estimate of reliance and the underpinning that is trust, automated systems can adapt to the predicted behavior of a human. In this way, automated systems can

change their level of autonomy [54]. When supervision isn't required, systems are considered autonomous and the paradigm of interaction changes [22]. Autonomous systems can operate on longer interacting time scales that span hours or longer in which goal-defining or norm setting is negotiated between agents [12]. For these time scales, we must understand how trust in these relationships changes over time.

The experiment in this study pairs a human and autonomous agent in a 2d grid search task with changing autonomous agent *strategy* and *capability*. The human subject has the opportunity to observe the behavior of the autonomous agent while simultaneously searching the same environment (see Figure 2.1). Subjects answer a survey about their trust level and use information from the autonomous agent to estimate the number of outliers in the grid.

Subjects in the experiment are split into two groups, which, over the course of many trials, interact with agents whose *strategy* and *capability* parameters change on different timescales. Trust in autonomous agents depends on the performance of the agent in the task, but when subjects see varying performance (separated by embodiment), their analysis of the system as a whole may be required.

2.2 Background

2.2.1 Trust in automation and autonomous systems

Automatic systems are designed to control a process despite disturbances and manage setpoint changes. Those who use automated systems supervise the output's behavior to assess system failure. Depending on the person's expertise, they may rely on the system in counterintuitive ways. Parasuraman describes differences in reliance behavior in people with different expertise [55]. Without a proper understanding of the automation system, people may *misuse*, or overrely, or *disuse*, neglect the system. When people can replace an automated system [54] or have the appropriate workload to supervise [22], actions they take that reduce the autonomy of the system may signal diminished trust in the system.

Lee and See suggest that measuring trusting behaviors does not necessarily indicate a trusting attitude [46]. Only when one analyzes the entire situation can a trusting action

be attributed to an attitude of trust [32, 53]. When interactions become more complex, agent performance may [17] or may not [61] lead to increases in trust. Understandably, measuring automation reliance is often tenable [44], but without careful consideration, it may just measure reliance.

Autonomous systems are capable of working without human interference. Trust plays a different role since human interactions with autonomous systems can be at a higher decision-making level [12]. Measuring human trust while *using* [55] an automatic system must be translated to trust while *working alongside* an autonomous system. To capture more than reliance-based trust, this paper uses self-reported trust [37], a subject's introspection, to understand how their characterization of an autonomous agent and the situation changes over time.

2.2.2 Time series trust models

Automated controllers do not have an anthropomorphized embodiment and no sense of personality. An automated controller is designed to do one task well, usually abiding by some physical or fundamental law. Supervisory control, where the human can enable automatic control of a system, can be influenced by an internal calculation that maximizes the task goal. If the system is better than the human, the human should choose the automatic system. The evaluation of system performance is a skill in itself. [47] collects trust survey responses while working with a semi-automatic pasteurization plant. By design, an automatic control system will perform within some bounds unless a sufficient disturbance causes unstable deviation from the setpoint. This disturbance may be a broad set of external factors the human collaborator can evaluate. For example, in this experiment, a pump failed for a trial on one day and all trials on another day. Without explanation, the human can only use performance to judge their trust or reliance on the system, which is a self-calibrated threshold. Subjects rated their trust in the system on a ten-point scale with specific instructions for what trust meant: "You will be asked to assess the plant's performance based on four criteria: the system's predictability, the system's dependability, the faith you have in the system and the amount of trust you have in the system." This may bias subjects to correlate their trust along the system's performance

without regard for externalities.

The study used an ARMAV model of the form:

$$\begin{aligned} Trust(t) = \phi_1 Trust(t-1) + A_1 Performance(t) \\ + A_1\phi_2 Performance(t-1) \\ + A_2 Fault(t) + A_2\phi_3 Fault(t-1) + a(t) \end{aligned} \quad (2.1)$$

Autonomous system performance can vary over several interactions with a human. The magnitude of trust change due to agent performance is a crucial measure for evaluating future interactions. Hu et al. used a driving simulator to measure trust levels after automatic obstacle detection failures [34]. Their state-space model used Expectation Bias, Cumulative Trust, and Experience to predict the next trust state. Given the population response, they measured the probability that a subject would choose to rely on automation. They found demographic specific parameterizations of a state-space trust estimator. In another car driving task, Desai et al. allowed subjects to change the level of autonomy in a remote control car [17]. They found subjects lowered the autonomy level of the car when reliability dropped, as well as a lower mean trust response. Similarly, Liu et al. used a driving task where the subject could indicate that they wanted to take over driving to indicate a decrease in trust [48]. The environment was complicated by reduced visibility and diminished autonomous driving reliability. Subjects were fit into "skeptical" and "confident" clusters based on a portion of their interaction, which allowed the researchers to predict their trust using personalized models.

Clustering-based trust studies are gaining popularity in trust time series with [29] and [8] using Bayesian time series models. Depending on features of the subject's trust trajectories, subjects were clustered into "Bayesian decision maker," "oscillator," and "disbeliever" classes. The testbeds used simulated robotic decision support agents to identify risks in simulated buildings. Subjects rated their trust at their own discretion and after trials.

Although the dynamics of trust as a state have been studied in the mentioned references, with further examples in [63], this study aims to uncover the effects of changing

performance and behavior simultaneously at different time-scales.

2.3 Experiment

2.3.1 Experimental Task

The goal of the experimental task was for subjects to interact with an autonomous agent that exhibited varying characteristics. Then, subjects could report measures of the agent's performance and an introspection of their trust in the agent. We designed the ordering of agent characteristics as experimental conditions to account for multiple conditions simultaneously. The experiment was conducted online through a web browser, allowing the test of many subjects outside the lab.

We instructed subjects to search a hidden area simultaneously with an Autonomous Searcher (AS). They did so with a "spotlight," which provided a view through an opaque mask to a grid of potential "outliers" seen in Figure 2.1. Subjects were to report the total number of outliers in the grid. The AS was embodied as an opaque circle but implied the ability to view circles and outliers below the mask. It left a persistent trail to display the previously searched areas in the current trial. The grid was a seven-by-seven array of solid purple circles with outliers containing an orange center. In each trial, a different number of outliers between five and fifteen were randomly placed in the grid.

The *capability* of the AS varied according to its color: blue, orange, and yellow. The grid area displayed the color name in case of subject colorblindness. The AS kept a running count of the outliers that it intersected and reported a fraction of the total at the end of the trial: 20%, 50%, or 100%. The *capability* of agents was not told to subjects but could be discovered by color. Additionally, AS's used one of three *strategies* a given trial. The first, "Lawnmower," moves back and forth across the area in a predictable and identical pattern. Next, the "Random" *strategy* moved with a uniformly distributed force input each time step. Lastly, the "Omniscient" *strategy* moved directly to each outlier, which appeared similar to the "Random" *strategy* but performed very well. See Figure 2.2 for examples of search *strategies* generated in the study.

The subject controlled their spotlight using the arrow keys on their keyboard. On

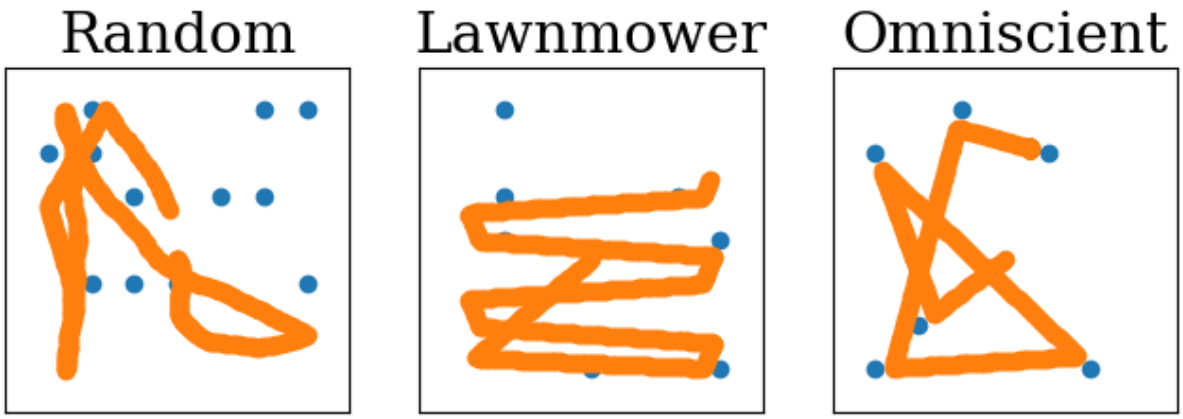


Figure 2.2: Sample AS paths from each *strategy* shown in orange. The paths shown are eroded to illustrate the behavior better. Outliers are blue circles. The Random and Omniscient paths are generated randomly but shared within subjects of the same group.

each frame during the game, velocity accumulated in the same direction for each arrow key they held up to a maximum speed. The spotlight speed decreased if they did not enter a control input for an axis during a frame. When the spotlight contacted a wall, the velocity was set to 0 in that direction. This form of force control with damping was chosen due to the reported ease of learning during pilot testing. The spotlight and AS sizes were smaller than the distance between grid circles to prevent searching multiple rows or columns simultaneously.

Each trial expired after a 20-second time limit. This time is 80% of the duration it would take to optimally encounter each grid circle, limiting the amount of information the subject could gain from searching alone. Subjects received a 5-second warning before the trial ended.

2.3.2 Inter-Trial Questionnaire

After the end of each trial, subjects received a two-part questionnaire. The first asked two questions to evaluate their performance in the task and the outcome of their outlier count and application of AS count to the total estimate. The second part asked the subjects to self-report their trust in the AS.

The first question, "How many outliers did you find with your spotlight this trial?" was intended to measure how well subjects could search and contribute to measuring

Factors		
Strategy	Lawnmower	Moves across columns, one row at a time. Frequently used by subjects in pilot experiments. The most predictable strategy.
	Random	Moves randomly, obeying the dynamics that the spotlight uses.
	Omniscient	Moves to each outlier in a random order.
Capability	20% 50% 100%	The percentage of outliers found that are reported to the subject at the end of each trial.
Group	Blocked Strategy	Each block has a constant strategy with varying capability.
	Blocked Capability	Each block has a constant capability with varying strategy.
Measures		
Familiarity	A measure designed to test the predictability of the AS.	
Reliability	A measure which focused on the task performance of the AS.	
Trust	The main focus of the study which encapsulates the prior metrics.	

Table 2.1: Subjects are assigned to one of two groups, and each trial searches with an AS of a given *capability* using a given *strategy*. At the end of each trial, subjects receive a survey with the above measures. The relationship between measures was studied in [16].

experiment compliance. The interface then displayed a report from the AS: *The* ***color*** *autonomous searcher reports finding* ***x*** *outliers*. Second, subjects were asked, "How may total outliers were hidden in the entire grid this trial?"

If the subject decided to split the search effort evenly with the AS, they could use the report of AS-found outliers and their own found outliers to reach an accurate estimate. However, without knowing the exact *capability* of the AS, the estimate would be more or less valuable. This could lead subjects to overlap their search until they found the

performance of the AS or increase their reported outlier estimate to compensate.

After completing the task survey, the page displayed the number of outliers in the search area and the subject's cumulative score. The score was a function of the outlier error used to help motivate subjects. Then, subjects received a trust survey, adapted from [37]:

- Statement 1: I am familiar with the autonomous searcher's strategy. (Strategy-focused Measure)
- Statement 2: The autonomous searcher is reliable. (Capability-focused Measure)
- Statement 3: I trust the autonomous searcher. (Trust Measure)

Subjects scored each statement on a Likert scale from 1-"Not at All" to 9-"Extremely." The survey was reduced in question count from the original survey to reduce repetition throughout the experiment. Additionally, each statement corresponded to a varying factor in the experiment - *strategy* and *capability*.

2.3.3 Experimental Design

The experiment was written in Javascript using the D3.js library to render an SVG container in a web browser. The game loop and logging ran at 30 frames per second. All game data was served from and stored on a laboratory server. The University of California's Davis Institutional Review Board approved this research, and it complies with the American Psychological Association Code of Ethics. Participants provided consent and were allowed to exit the experiment without penalty.

The game guided Subjects through an interactive tutorial, which introduced the task elements one at a time—first, spotlight movement and the time limit before introducing an AS as a teammate. Subjects were reassured that an AS was helpful, could vary in its ability to detect outliers, and used different movement *strategies*. Training took approximately 3 minutes. Then, subjects searched alone for nine trials to practice and measure their ability to search.

The remaining 63 trials were sorted into three blocks of 21. The total number of trials was found from pilot testing to be reasonable before subjects became disinterested.

Figure 2.3 depicts a flow chart for how subjects encountered each factor depending on their assigned group. Group 0 subjects saw a constant *strategy* among *lawnmower*, *random*, and *omniscient* during each block while group 1 saw a constant *capability* 20%, 50%, or 100%. Within each block, conditions of the remaining factor were varied. A random seed, shared with all subjects, ensured that AS motion, outlier count, outlier position, and trial order were the same across subjects of the same group. The study variables are listed in Table 2.1.

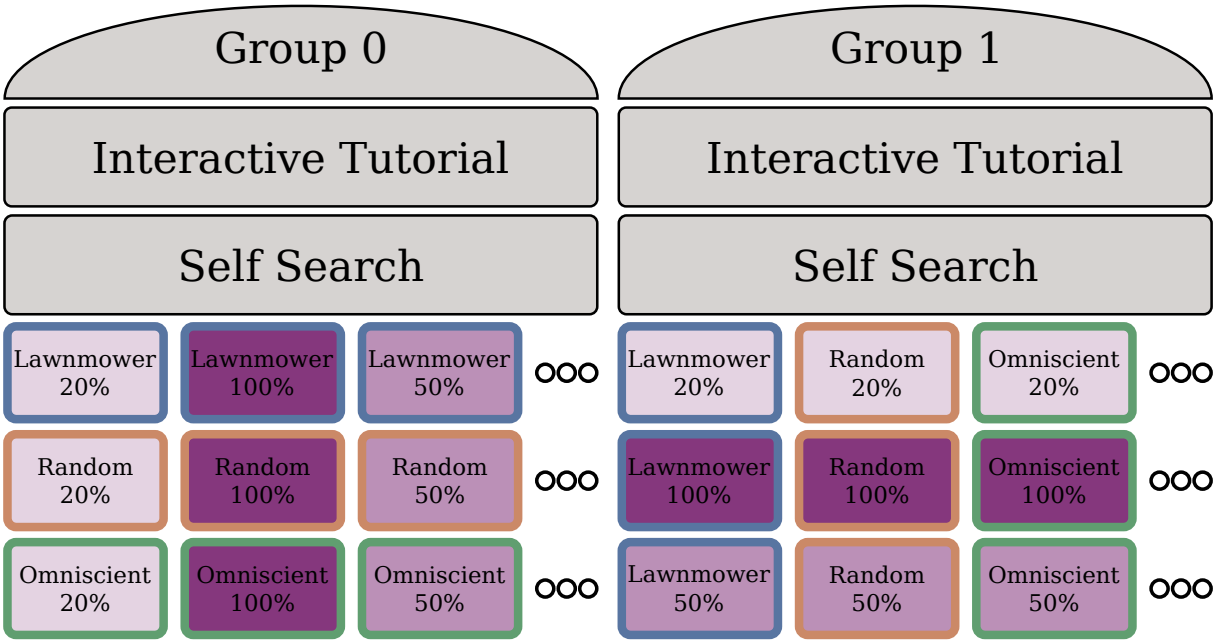


Figure 2.3: Subjects were assigned to one of two groups according to the order in which they began the experiment. Both groups started with an identical tutorial and self-search block before starting the main experiment. Subjects in group 0 had a constant *strategy* in each block, while group 1 had a constant *capability* in each block. In this figure, *capability* is indicated by the shade of the box and *strategy* by the outline of the box.

2.4 Pilot Study

To test the remote web survey interface, I ran a pilot study with fewer subjects in combination with a parallel human motor control study. Although convenient for recruitment and washout periods, we found that the combination experiment took a long time for subjects to maintain attention. The subsequent section covers the results from the dedicated grid search experiment. In this initial phase, we found that subjects connected autonomous

search performance and capability with trust, reliability, and familiarity measures. This validated further exploration into the time series components of trust in later studies. The results of this pilot were published at the International Conference of Human-Machine Systems.

2.4.1 Demographics

The study was completed entirely online through the subject's web browser. The University of California Davis Institutional Review Board approved the experimental protocol. Subjects were asked to set aside an appropriate amount of time in a quiet environment. The experimenters did not note any unusually long trial times that could indicate breaks in the study.

Fourteen subjects completed the entire study. They were college undergraduates who received course credit. Their ages ranged from 18 to 24 years old. Eleven subjects identified as female and 3 as male. Subjects were asked to identify their race and ethnicity; 4 identified as Hispanic/Latino, 2 American Indian or Alaska Native, 8 Asian, 1 Black or African American, and 3 White. Seven Subjects reported playing video games in the last year with, on a weekly basis, 3 playing less than 1 hour, 1 playing between 1 and 5 hours, 1 playing between 6 and 10 hours, 1 playing between 11 and 20 hours, and 1 playing more than 20 hours.

2.4.2 Methods

We performed an analysis of variance to find which factors were significant to the results of the questionnaire and a subsequent test to find how parameters (Capability, Strategy) affected the ratings of the specific questionnaire statements. We assumed that the results of the questionnaire were independent trial-to-trial and that they were an aggregation of the subjects' perception of the searchers' Capabilities and Strategies.

Hypothesis 1: One can predict that if the searcher severely under-reports the number of outliers in the grid, the subject may no longer trust its future reports. Therefore, we hypothesize that lower Capability AS's will receive lower **reliability** and **trust** scores in the post-trial questionnaires.

Hypothesis 2: The *lawnmower* AS Strategy is the most predictable of the three because it travels the same path each trial. Additionally, a familiar strategy may boost the trustworthiness of a searcher. If the subject doesn't need to spend time guessing what the AS is going to do, they can focus on completing their own task effectively. Thus, we hypothesize that subjects will rate the *lawnmower* Strategy as more familiar than the other Strategies according to post-trial questionnaires and will receive higher **trust** scores than the other AS Strategies.

2.4.3 Results

2.4.3.1 Subject Paths

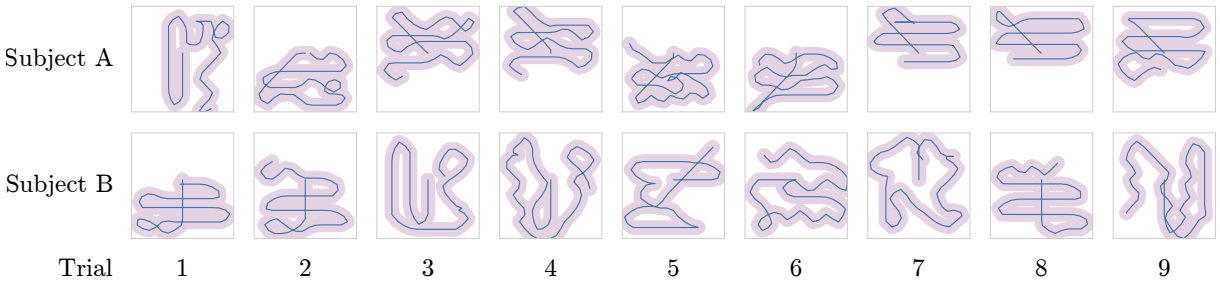


Figure 2.4: Example search paths during the self-search block of the experiment. Subjects controlled a spotlight starting near the middle of the search area and counted hidden outliers during the approximately 20 second long trial. Subjects used a mixture of structured and unstructured search strategies with varying start direction and search region. The pink stroke background is the total area that was seen by the subject and the blue line overlaid is the path that the subject took.

Subjects began the scored portion of the experiment with a self-search block where they experimented with different search paths, some more structured than others. Figure 2.4 shows the paths of two representative subjects during this period. Subjects appeared to have command over the controls of the spotlight after engaging in the tutorials in the previous block, as exemplified by the total ground covered and deliberate turning to avoid boundaries. The spotlight started in the same grid position each trial throughout the experiment: horizontally centered with a slight upward offset from the center of the grid. Despite this asymmetry, which may suggest they should search the top region, the representative subjects moved the spotlight in different starting directions each trial. Once the spotlight left the center of the grid, subjects appeared to utilize one of two

micro-strategies: (1) a straight-line search path or (2) an oscillating search path. The straight-line path followed the rows or columns of the grid while the oscillating strategy bounced between rows or columns with a period of approximately the grid circle separation distance. The most important action in the self search task was for the subject to cover the most ground possible to gather information about the number of outliers in the search area. With an unconstrained self-search, the subject could cover any part of the grid that they found most convenient, as long as they did not double back, for an efficient covered area.

2.4.3.2 Questionnaire

We ran a three-factor mixed model analysis of the questionnaire results with within-subjects factors of AS Capability, and Strategy and a between-subjects factor of Group. We report the results over the same factors for each of the three measures from the post-trial questionnaire in Tables 2.2, 2.3, and 2.4. The factors and measures for these tests are summarized in Table 2.1. Significant effects ($p < 0.05$) related to the hypotheses were analyzed with Tukey's Honest Significant Difference test using Satterthwaite's method for computing degrees-of-freedom.

The effects of Strategy and Capability were significant for all measures. Pairwise comparisons for the significant main effects of Strategy and Capability are shown in Figure 2.5. Subjects' **familiarity** scores with the AS show that the rating for the *lawnmower* Strategy was significantly different than both the *random* ($p < 0.0001$) and *omniscient* ($p < 0.0001$) Strategies. The comparisons for the **reliability** and **trust** measures were nearly identical. With increasing Capability, we see increasing ratings of **reliability** and **trust**, with each Capability significantly different than the others ($p < 0.0001$). Alternatively, the *lawnmower* and *omniscient* Strategies were shown to be rated differently than the *random* Strategy ($p < 0.0001$).

The Group factor saw no significant effects for any of the measures. However, the effect of the Group-Capability interaction was significant for **reliability** and **trust** measures. Subjects rated **reliability** differently between Groups at the 0.5 Capability level ($p = 0.0234$). The same was true for **trust** ($p = 0.0076$). The interaction between Strategy

and Capability was significant for the **reliability** measure ($p < 0.0382$) and the **trust** measure ($p < 0.0496$). For both measures, at the 1.0 Capability level, the *lawnmower* and *omniscient* Strategies were rated differently than the *random* Strategy ($p < 0.0001$). Lastly, the three-way interaction was not significant across all measures.

Table 2.2: Familiarity Measure ANOVA Results

	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Group	0.03	1.00	12.00	0.01	0.9367
Strategy	274.09	2.00	810.00	66.15	0.0000†*
Capability	19.55	2.00	810.00	4.72	0.0092†*
G:S	3.89	2.00	810.00	0.94	0.3913
G:C	14.26	2.00	810.00	3.44	0.0325†
S:C	5.92	4.00	810.00	1.43	0.2228
G:S:C	9.86	4.00	810.00	2.38	0.0503

: indicates interaction effects

† significant effects ($p < 0.05$) to be analyzed post-hoc.

* pair-wise comparisons shown in Figure 2.5

Table 2.3: Reliability Measure ANOVA Results

	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Group	7.75	1.00	12.00	1.72	0.2146
Strategy	57.60	2.00	810.00	12.76	0.0000†*
Capability	233.19	2.00	810.00	51.65	0.0000†*
G:S	1.14	2.00	810.00	0.25	0.7775
G:C	59.24	2.00	810.00	13.12	0.0000†
S:C	11.50	4.00	810.00	2.55	0.0382†
G:S:C	8.54	4.00	810.00	1.89	0.1101

: indicates interaction effects

† significant effects ($p < 0.05$) to be analyzed post-hoc.

* pair-wise comparisons shown in Figure 2.5

2.4.4 Discussion

The results supported our hypotheses about subjects' perception of AS characteristics. The first two statements of the questionnaire (regarding **familiarity** and **reliability**) found significant effects for AS Strategy and Capability. The **trust** statement also found significant effects for both parameters, indicating an autonomous agent's Strategy and Capability were important to a human collaborator's level of **trust**.

The search Strategies were designed to make trade-offs between performance and legibility. We find that the *lawnmower* strategy was significantly different in its **familiarity** rating than the other Strategies which aligns with our hypothesis. While the *lawnmower* strategy was immediately recognizable, the difference between the *random* and *omniscient* Strategies was not.

The AS's ability to communicate was a valuable tool for a human to evaluate its capability. At the end of each trial, the subject must decide if the AS's outlier report is accurate. The subject could divert from the main mission to follow the AS and determine if the number of outliers it has traversed is the number of outliers it reports. They could also use the post-evaluation ground truth outlier count to determine the AS's reporting accuracy. The pairwise test showed that subjects are able to, on average, determine the

Table 2.4: Trust Measure ANOVA Results

	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Group	13.10	1.00	12.00	3.24	0.0970
Strategy	57.41	2.00	810.00	14.21	0.0000†*
Capability	296.66	2.00	810.00	73.42	0.0000†*
G:S	0.98	2.00	810.00	0.24	0.7844
G:C	61.41	2.00	810.00	15.20	0.0000†
S:C	9.65	4.00	810.00	2.39	0.0496†
G:S:C	6.32	4.00	810.00	1.57	0.1816

: indicates interaction effects

† significant effects ($p < 0.05$) to be analyzed post-hoc.

* pair-wise comparisons shown in Figure 2.5

level of **reliability** between AS's of different Capabilities. The **reliability** increases with Capability as we predicted in our hypotheses. Interestingly, subjects rated AS's using the *lawnmower* and *omniscient* Strategies higher in **reliability** than the *random* Strategy. This may be due to the fact that these Strategies were more efficient in traversing outliers, increasing the total number of outliers found. The *lawnmower*, *random*, and *omniscient* strategies found 71%, 37%, and 98% of outliers, respectively. This does not explain the similarity in rating between the *lawnmower* Strategy and the *omniscient* Strategy. Yet, the Strategy-Capability interaction shows that some Strategies are rated significantly different at high Capabilities.

A human's trust in an autonomous agent is complex and multifaceted. However, a familiar and reliable agent is likely to be more trustworthy as well. The pairwise test of Capability levels against the **trust** measure suggests that our hypothesis that a higher Capability agent will be more trustworthy was true. The same evaluation for the AS's Strategy is less clear. Although the *lawnmower* Strategy received ratings significantly different than the *random* Strategy, the *omniscient* Strategy was significantly different as well. We cannot conclude that the *lawnmower* Strategy is more trustworthy than the *omniscient* Strategy. The similarity between pairwise tests for **reliability** and **trust** measures suggests that subjects may have reported identical ratings for these throughout the experiment. Further work is needed to determine whether subjects believe **trust** is primarily derived from **reliability**.

Subjects were exposed to AS characteristics differently in each of their two Groups. Constant exposure to a certain characteristic may determine how fast a subject is able to ascertain the differences within categories. For example, G0 is constantly exposed to the *lawnmower* Strategy in block 1 which means they can adapt their movement path over each trial. G1 can compare Strategies directly, over consecutive trials, while G0 can compare Capabilities directly. The Group factor alone showed no significant effects for any measures. Thus, we cannot conclude that grouping in this way alone contributed to differences in any measurement. Yet, it is not clear that the sequence of AS parameters within a block may or may not contribute to ratings of **familiarity**, **reliability**, or **trust**.

The initial analysis revealed significant effects from Group and Capability interaction for the 0.5 Capability level, where G0 rated **reliability** and **trust** higher than ratings from G1. This suggests that a somewhat capable searcher is more reliable and trustworthy when seen with other searchers, rather than constantly.

With no prior information about the AS's Strategy or Capability, subjects determined their **trust** level of the AS's. They trusted the *lawnmower* and *omniscient* Strategies more than the *random* Strategy. As long as the Strategy is acting with some purpose or finding more outliers, subjects found it trustworthy. Using **trust** as a surrogate measure for Capability, subjects differentiated the AS's accurately with only information about the number of outliers that were found post-trial. People have a notion of trustworthy collaborators beyond their ability to complete a task. Autonomous collaborators can take this into account to improve action selection policies while interacting with humans.

2.5 Trust Model

In a previous work, we performed an analysis of variance to find how survey factors relate to trust when considering each trial independently [16]. We found that trust was correlated with *capability*, and more legible *strategies* resulted in higher trust survey ratings. The goal of this work is to demonstrate the influence of autonomous agent behavior and performance on a human interactor's trust in time. The modeling process follows the steps outlined in Figure 2.6.

Subjects' trust ratings varied in level and variance. To capture the extent of subjects' trust ratings compared to the population, their scores were normalized according to

$$trust_{normalized} = \frac{trust - trust_{min}}{trust_{max} - trust_{min}} \tag{2.2}$$

This normalization frames trust ratings concerning their best and worst experiences with an autonomous agent. The range could be compared to a subject's predisposition to trust.

If we assume that a subject's trust rating depends only on the *capability* of a given agent, we could model trust with only knowledge of the *capability* of an agent at that

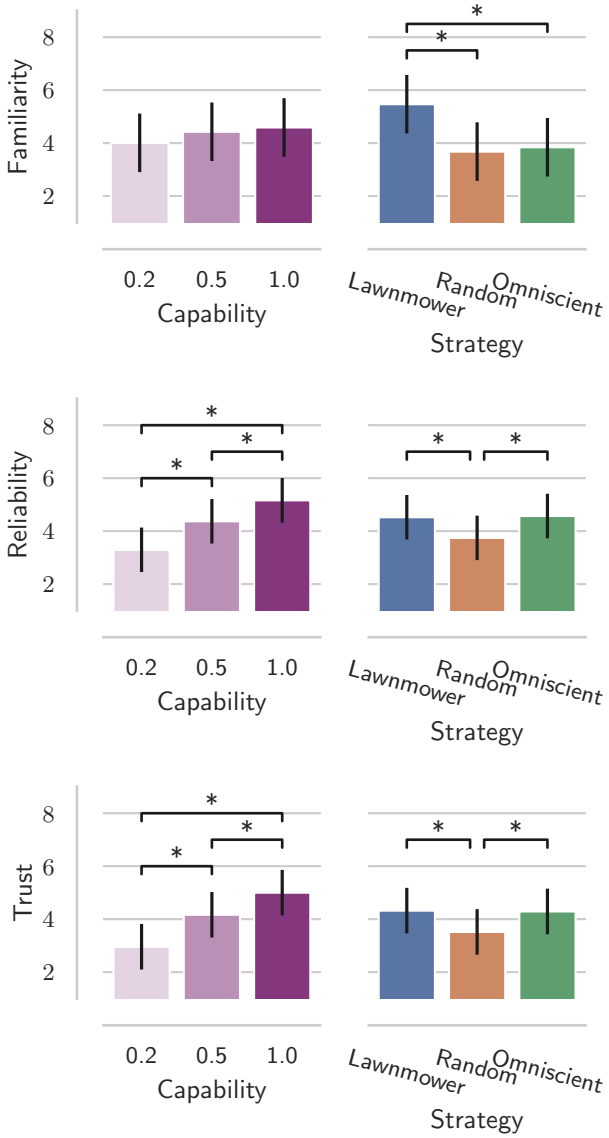


Figure 2.5: Pairwise comparisons for the factors of Capability and Strategy against measures of **familiarity**, **reliability**, and **trust**. Vertical bars indicate the 95% confidence interval and brackets indicate significant differences ($p \leq 0.0001$) via Tukey's procedure. Subjects rated AS's with higher Capability more trustworthy and reliable on average. Additionally, the *lawnmower* Strategy achieved high scores for all measures while the *omniscient* Strategy joined for high scores in the **reliability** and **trust** measures.

time, using a multiple categorical regression:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \epsilon \tag{2.3}$$

where y_t is an estimate of reported trust at time t, β_k is the coefficient for the kth

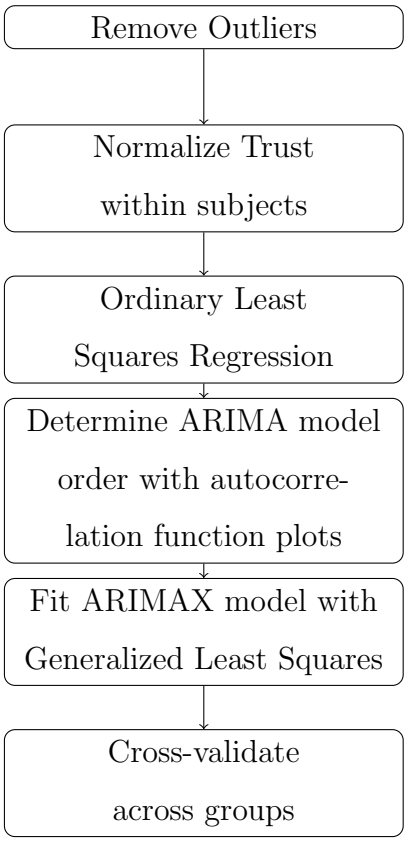


Figure 2.6: This study follows the modeling process outlined in this flowchart.

exogenous variable x, and ϵ is the error term. We assume that $\epsilon \sim N(0, \Sigma)$ with constant Σ and is not autocorrelated.

Lee and Moray showed a trust model with autoregressive components, suggesting that ϵ is likely autocorrelated [47]. A human's evaluation of the system, the agents, their environment, and the task changes throughout the experiment [32], which suggests that their trust evaluation changes as well.

Humans potentially carry separate trust state evaluations of each agent in any experiment [75]. However, the homogeneity of the agents suggests shared traits that bias trust evaluations across agents. Thus, capturing agent-specific and system-general contributions to a model describing how trust changes over time is crucial.

We assume that since *capability* is assigned to specific agents by color, they can be represented by their *capability*. As subjects interact with an agent, they learn about its abilities and generate a level of *perceptual trust* in it [13]. However, since the experiment

presents each agent in the same way (between subject groups), the subject's change in trust introspection over time can be captured by a *system-general* time process.

Assuming that a subject's learned trust changes similarly for each agent, we can expand the model in Section 2.3 to include time components in the error. Here, we use an ARIMA model with exogenous regressors of the form from [36]:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \eta_t \tag{2.4}$$

$$(1 - \phi_1 B)(1 - B)\eta_t = (1 + \theta_1 B)\varepsilon_t \tag{2.5}$$

where η is the regression error, B is the backshift operator ($B^k y_t = y_{t-k}$), ϕ is an autoregressive coefficient, θ is a moving average coefficient, and ϵ is the error term.

An ARIMA model consists of autoregressive, differencing, and moving average components. An autoregressive term measures the correlation with prior values of the predicted variable, meaning an estimate of trust depends on past measures of trust with some weight. Differencing removes non-stationarity from a time series. Differencing a trust time series corresponds to a correction on factors that influence trust in time. The moving average component measures the contribution of prior error terms to the predicted variable. In this case, the component can weight a trust estimate that differs from prior predictions.

To obtain the values of the parameters of the model, we build the ARIMA model and fit using loglikelihood maximization in the Python package Statsmodels [67]. We check the partial autocorrelation function plots and residual plots to check the efficacy of a model. Then, models are cross validated across groups using root-mean-square-error of the residuals from a one-step-ahead forecast of trust.

2.6 Results

2.6.1 Demographics and exclusion

Subject data were excluded from the analysis if they did not complete the entire study, took out-of-protocol breaks, or answered task survey questions unreasonably. There were 80 total subjects included, of which 22 identified as male, 57 female, and 1 nonbinary. The average age was 20.0 with standard deviation 1.5. Of the 45 subjects who played video games, the median subject played between 1 and 5 hours per week.

2.6.2 Time series modeling

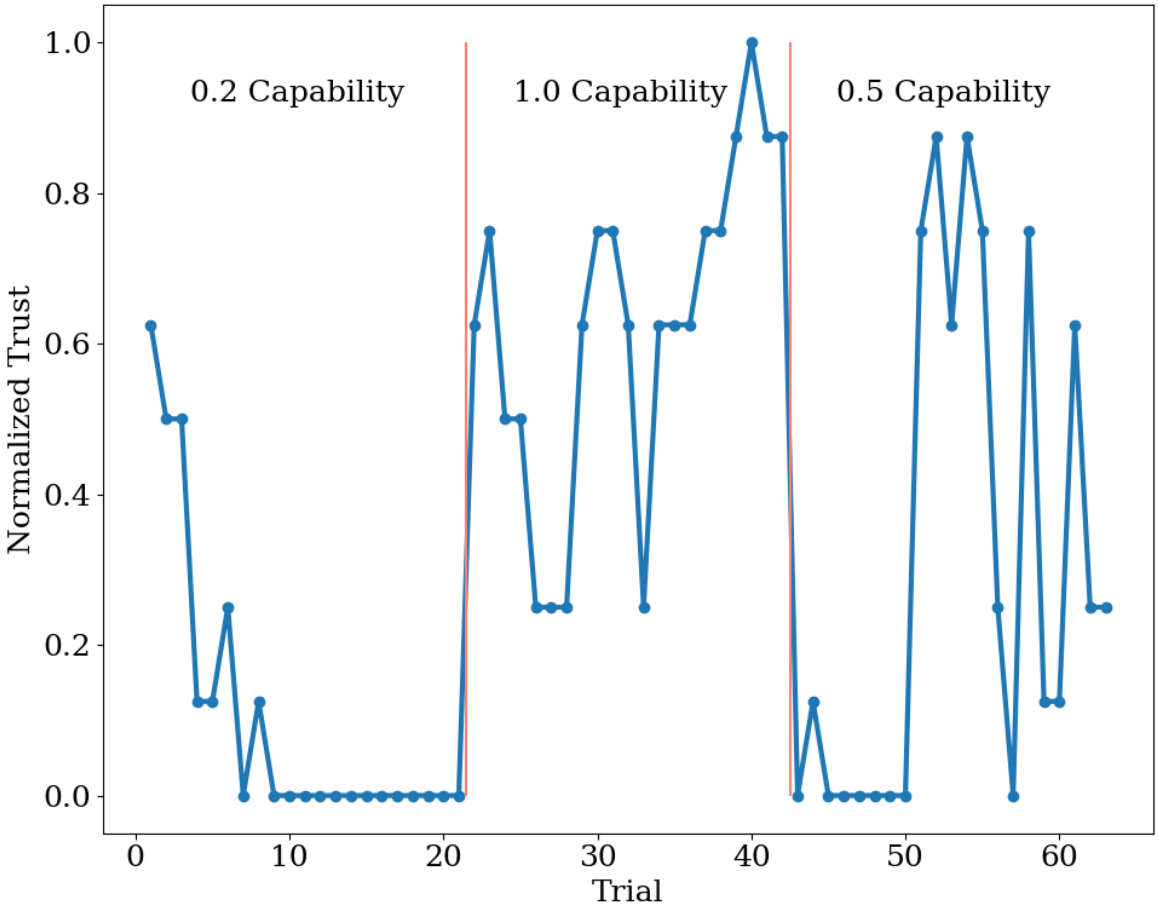


Figure 2.7: Single subject representative trust trajectory.

Subjects completed 63 trials each, reporting their trust at the end of each trial. We can construct a discrete time series from these survey results that tracks survey responses throughout the experiment. Individual trust trajectories are sometimes erratic and vary to the extent of their normalized trust range. Figure 2.7 shows a representative subject's ratings. As mentioned in Section 2.7.1, within-trial interactions between the subject and searcher hold a richness of information that cannot be entirely captured in this rating and may be the reason for drastic jumps in the trust survey.

We are concerned with the effects of the autonomous searcher's *capability* and *strategy* on the subject's reported trust over time. Trust is, in part, tied to the task performance [77, 17, 34] of a human's partner. Part of the goal of the experimental design was to

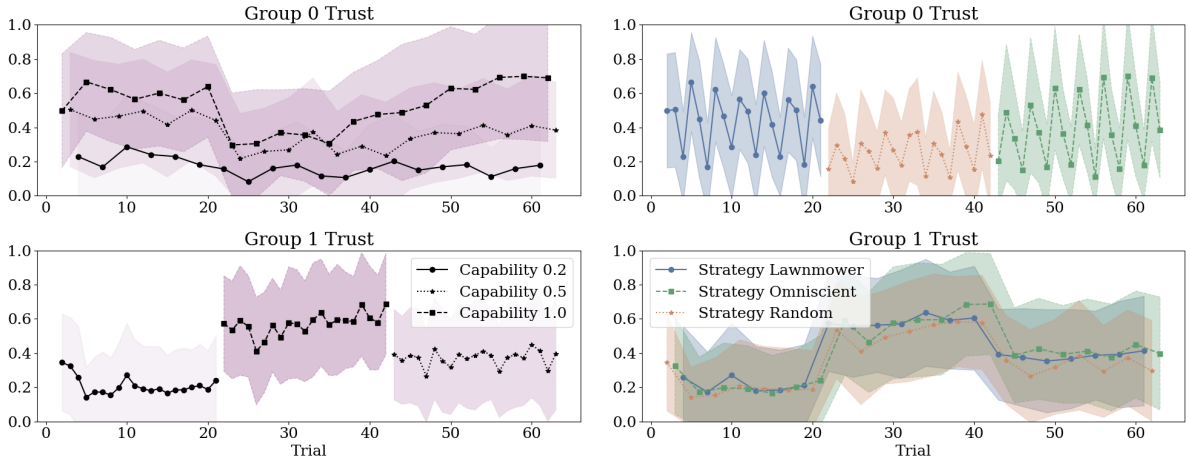


Figure 2.8: Each group's average trust ratings throughout the experiment. The data across rows are the same but colored and connected to emphasize the contributions of each experimental block to the rating. Since *capability* likely contributes more to an individual's trust rating and Group 1 saw a constant searcher *capability*, their ratings were more stable. The oscillations for Group 0 can be attributed to the varying performance of the searchers over a short frequency. The colors in this figure correspond to those in Figure 2.3.

change the temporal focus of parameters. That is, vary the *capability* of agents for group 0 every trial, but every block for group 1. Figure 2.8 compares each group's average trust trajectory split between the parameters. In the case of the *capability* parameter, represented by different embodiments of the agents, one could interpret the trust ratings in the first column as the trust in each agent over time.

Given the differences in trust trajectories between groups, we construct separate models and cross-validate them to analyze the model's generality. Any differences in modeling may point to the importance of how parameters other than autonomous agent *capability* affect subjects' trust in it.

We start with a multiple linear regression with categorical variables *capability* and *strategy*. Due to the experimental design, *strategy*, and *capability* are found to be collinear. Trust can largely be predicted from *capability* alone in this process, and with *strategy* included, it leads to large coefficients and potential overfit of the model. Using only *capability* provides a model where the coefficients are the average group trust values for each *capability*. Tables 2.5 and 2.6 show the outputs of an Ordinary Lease Squares fit

Dep. Variable:	trust	**R-squared:**	0.182
Model:	OLS	**Adj. R-squared:**	0.181
No. Observations:	2646	**F-statistic:**	294.0
Covariance Type:	nonrobust	**Prob (F-statistic):**	5.21e-116

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
capability_0.2	0.1778	0.010	17.491	0.000	0.158	0.198
capability_0.5	0.3710	0.010	36.490	0.000	0.351	0.391
capability_1.0	0.5257	0.010	51.714	0.000	0.506	0.546

Table 2.5: Group 0 multiple linear regression.

Dep. Variable:	trust	**R-squared:**	0.201
Model:	OLS	**Adj. R-squared:**	0.200
No. Observations:	2394	**F-statistic:**	299.8
Covariance Type:	nonrobust	**Prob (F-statistic):**	6.49e-117

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
capability_0.2	0.2217	0.010	21.941	0.000	0.202	0.241
capability_0.5	0.3716	0.010	36.785	0.000	0.352	0.391
capability_1.0	0.5704	0.010	56.460	0.000	0.551	0.590

Table 2.6: Group 1 multiple linear regression

from Statsmodels [67]. The RMSE for Group 0 is 0.097 and 0.062 for Group 1.

The residuals indicate that for Group 0, a model that includes temporal components is necessary. To capture non-stationarity, we add a differencing component. Also, we perform a partial autocorrelation of the error to find the order of the autoregressive component. Since lags 1, 2, and 3 have magnitudes above the critical value thresholds, at least a 3rd order model should be used. Iterating over combinations of Autoregressive (p) and Moving Average (q) orders with 1st order differences results in the lowest Akaike

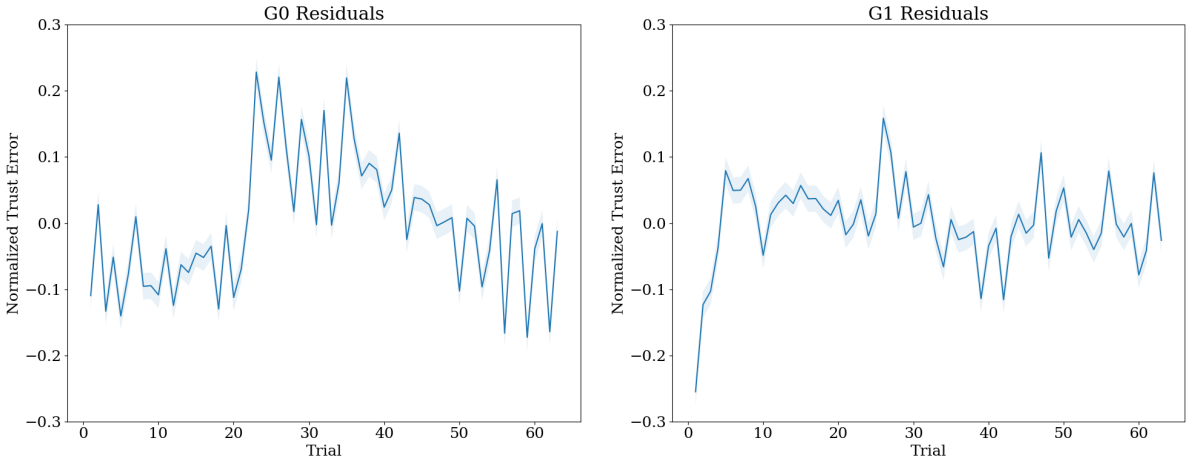


Figure 2.9: The multiple regression model residuals show strong non-stationarity and cycling in Group 0.

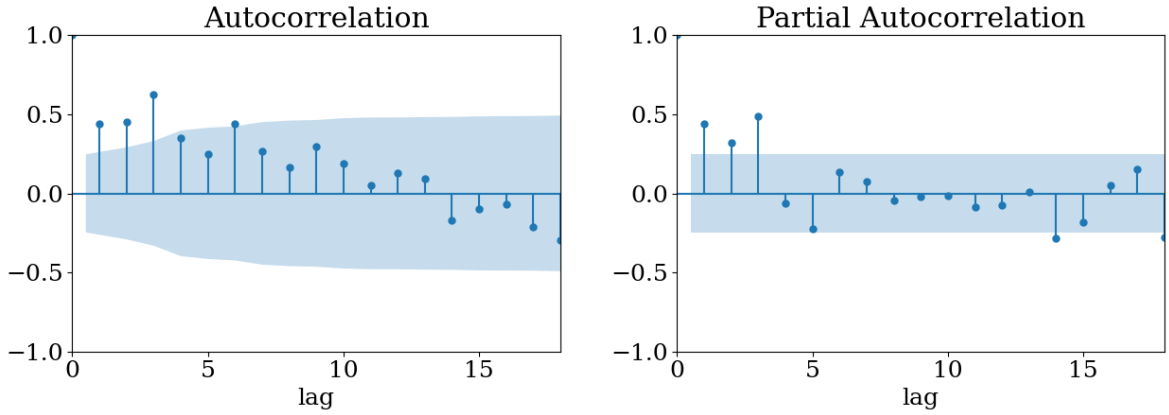


Figure 2.10: The Autocorrelation Function and Partial Autocorrelation Function plots of the residuals for Group 0 indicate lags 1 through 3 are likely necessary to capture the remaining error.

Information Criterion (AIC) shown in table 2.7.

The final models for groups 0 and 1 describe the trends in the data as seen in Figure 2.11. The difference in *capability* presentation in the experiment elicited the need for more complex modeling beyond linear regression. Including autoregressive components allows the model to capture information from prior trials. Cross-validating the results (table 2.10) indicates an improvement in the linear regression model for both training models but a decrease in performance on the group 0 model cross-validated on group 1 data.

		q					
		0	1	2	3	4	5
p	0	-99	-128	-129	-127	-131	-132
	1	-115	-127	-127	-127	-130	-128
	2	-138	-136	-136	**-142**	-140	-138
	3	-136	-134	-136	-140	-138	-136
	4	-135	-134	-133	-137	-132	-134
	5	-136	-134	-133	-135	-133	-132

Table 2.7: The array of AIC values calculated for different ARIMA(p,1,q) model orders over group 0 data. **p** values are the rows and **q** values are the columns. The results indicate that an ARIMA(2,1,3) is the most suitable model order given the goodness of fit and number of parameters.

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
capability_0.2	-0.2060	0.020	-10.110	0.000	-0.246	-0.166
capability_0.5	-0.0068	0.020	-0.344	0.731	-0.046	0.032
capability_1.0	0.1385	0.019	7.477	0.000	0.102	0.175
ar.L1	-0.7634	0.057	-13.356	0.000	-0.875	-0.651
ar.L2	-0.9550	0.060	-15.806	0.000	-1.073	-0.837
ma.L1	0.0437	0.181	0.241	0.809	-0.311	0.399
ma.L2	0.6702	0.462	1.452	0.147	-0.234	1.575
ma.L3	-0.5345	0.363	-1.473	0.141	-1.246	0.177
sigma2	0.0041	0.002	1.817	0.069	-0.000	0.009

Table 2.8: Coefficient table for group 0. Ascending *capability* coefficients agree with the assumption that trust increases with increasing *capability*. Significant autoregressive coefficients suggest the time series has strong cyclic signals from the study design.

	coef	std err	z	P> |z|	[0.025	0.975]
capability_0.2	-0.1925	0.176	-1.091	0.275	-0.538	0.153
capability_0.5	-0.1282	0.164	-0.784	0.433	-0.449	0.192
capability_1.0	0.1395	0.096	1.453	0.146	-0.049	0.328
ar.L1	-1.0205	0.056	-18.182	0.000	-1.131	-0.911
ar.L2	-0.9945	0.023	-43.356	0.000	-1.039	-0.949
ma.L1	0.8628	0.181	4.760	0.000	0.508	1.218
ma.L2	0.7836	0.165	4.762	0.000	0.461	1.106
ma.L3	-0.1097	0.159	-0.691	0.490	-0.421	0.201
sigma2	0.0020	0.000	4.917	0.000	0.001	0.003

Table 2.9: Coefficient table for group 1. Ascending *capability* coefficients agree with the assumption that trust increases with increasing *capability*.

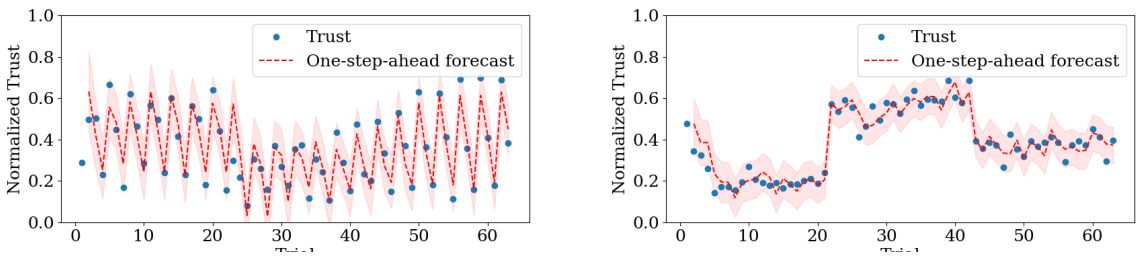


Figure 2.11: The one-step-ahead predictions for the ARIMAX model for group 0 (left) and group 1 (right).

2.7 Discussion

2.7.1 Subject search behavior depends on Autonomous Searcher conditions

Subjects had 20 seconds to search the grid area with the autonomous searcher. Subjects displayed varied behaviors to collaborate with the autonomous searcher, as seen in Figure 2.12. Some may have intentionally searched the same area to confirm the searcher's performance at the expense of their own search performance. Subjects from different groups may have learned the paths from the autonomous searcher over different time scales. Group 0 saw each *strategy* repeated over the block, while Group 1 saw each

Group	0	1
Blocked	Strategy	Capability
Linear Regression	0.097	0.062
Group 0 ARIMAX	0.067	0.067†
Group 1 ARIMAX	0.094†	0.048

Table 2.10: Predicted vs. actual RMSE of time series models. † indicates cross-validated results.

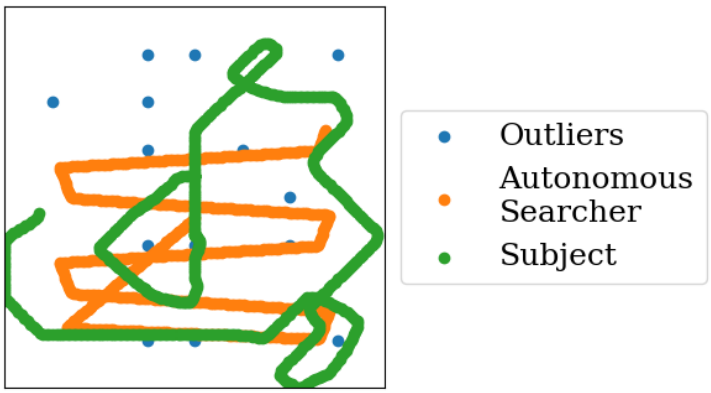


Figure 2.12: An example of the search trajectories of the human subject and autonomous searcher. They started in the center of the grid and searched similar areas before splitting. The subject overlapped the previously searched region near the bottom toward the end of the trial. The paths shown here are thinner than the spotlight's size during the task to better differentiate the searcher and subject.

strategy rapidly. Depending on the subjects' propensity to consider the Autonomous Searchers' *strategy* or *capability* (identity), the ordering of parameters may have affected their search behavior, as seen in Figure 2.13. The analysis of subject search behavior is left for future work.

2.7.2 Ordering effects captured using ARIMAX model

This study demonstrated how changing the frequency with which humans interact with differing *capability* autonomous agents affects their trust. The subjects' trust changes from trial to trial according to the agent's *capability* but cannot be predicted without knowledge of prior interactions. The dynamics of trust depend on aspects of the agent or the environment that change in time [32].

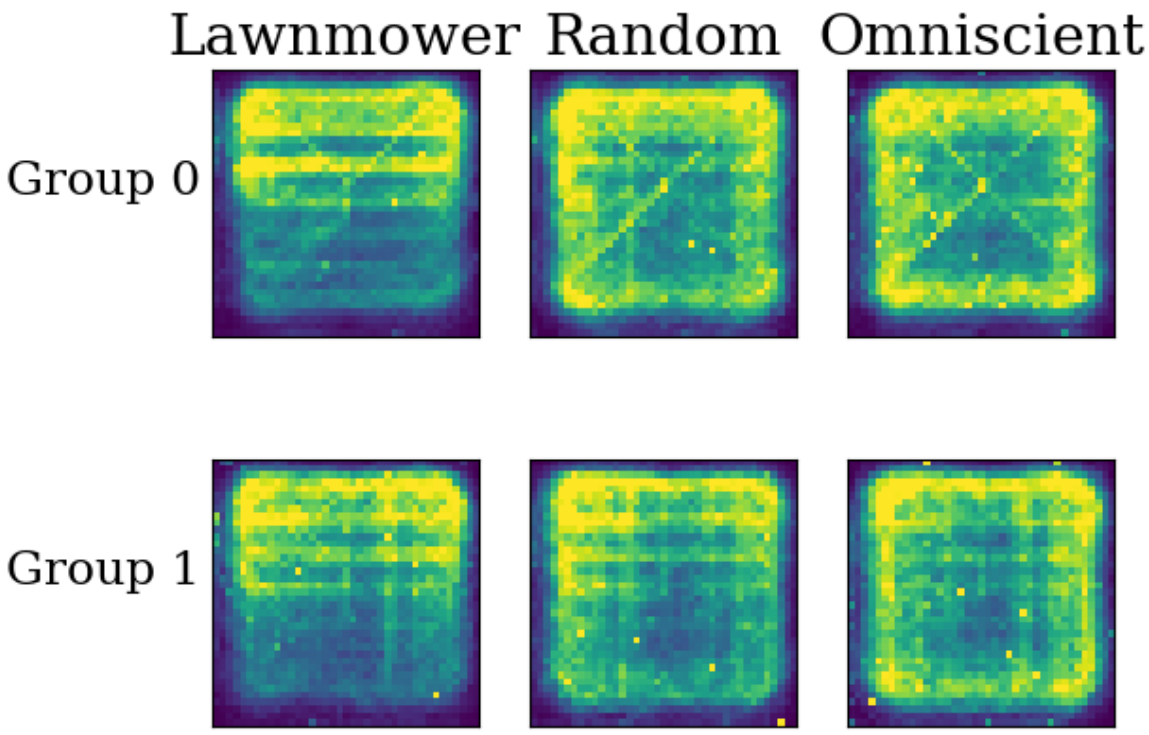


Figure 2.13: A heatmap of the subject searched area with the initial position removed. The Lawnmower *strategy* was more predictable and resulted in subjects avoiding the autonomous searcher's area. The Random and Omniscient *strategies* were more likely to search the center areas, resulting in subjects searching the exterior of the grid more often.

By having different blocked characteristics for each group, this experiment captured trust changes due to characteristics at a given time and the interaction between the changing characteristics over the experiment. Figure 2.8, in particular, emphasizes how reframing measured trust shows apparent temporal factors at play for a group's trust over the experiment. If trust was solely based on performance, we would see constant ratings when visualized as *capability*. Group 0 saw three agents in rapid succession and could evaluate relative performance early in the experiment. However, since group 1 saw the same agent use different *strategies*, they could quickly determine how the agent performed during the grid search task. The ARIMAX models developed here are a good candidate for these conditions since they account for the trust in time and the effect of

agent properties.

An individual's trust at a given time is a function of their experience with the autonomous agent during prior interactions. Thus, the ARIMAX model, which describes contributions to trust estimation, suitably includes the exogenous factors that track the current situation and autoregressive factors from previous situations. The performance of the ARIMAX over the OLS models shown in Table 2.11 agrees with the intuition that including autoregressive factors is necessary for capturing human trust evolution in time. The model selection process using AIC ensures that increasing the number of model parameters is appropriate, while cross-validation between groups suggests the model generalizes across experimental conditions.

Inter-trial trust ratings capture a subject's trust in an autonomous agent throughout the experiment. It must be noted that when using trust surveys, subjects are biased to evaluate autonomous agents more critically. They may question, "If the experiment asks me if I trust the autonomous searcher, should I?". In this case, including separate autonomous agents means that their trust rating is grounded in evaluating the performance between agents. At the very least, the results are internally valid since the same trust survey was given to all subjects for all trials. Subjects evaluate autonomous agent performance during the trial by its behavior and after the trial by its outlier reporting. Since subjects were asked to think about their trust at this time, it can't be clear that their trust changed during the trial, if it did at all. Since autonomous agents report the number of outliers they found, there isn't a clear indication that they succeeded or failed at the task, matching the scale of trust presented in the survey. Task faults [47] are not always obvious in human autonomy teaming situations due to higher level interactions like goal negotiation [12].

Subjects could explore the region of the hidden region that the autonomous searcher already searched. After some trials, the subject could predict where the searcher could move due to the kinematics of the spotlights. However, the autonomous searcher did not adaptively search with the subject, which could improve task performance [11]. Future work includes allowing the autonomous agent knowledge of the human's trust, and giving

the AS the ability to vary its search strategy to maximize task performance. If autonomous agents had additional actions to communicate intent [21] or *capability* [52], trust would be aligned with task performance and could increase the rate at which a subject learned the task.

2.7.3 Future trustworthy autonomous systems

As autonomous systems gain capabilities, they are more likely to interact with humans with a similar decision-making level [12]. Considering how human trust changes in time could allow such systems to understand potential future human actions and plan accordingly. Although the presented questionnaire is intrusive, the system could ask if a system's confidence in the human trust state is low. Although not time optimal in the short term, taking time to better understand the situation could be vital in the long term [15]. Findings from this experiment show promise in the ability to track human trust over time and the connection between autonomous agent characteristics and trust.

2.8 Conclusion

This study demonstrated how changing the frequency with which humans interact with differing *capability* autonomous agents affects their trust. The subjects' trust changes from trial to trial according to the agent's *capability*. However, a better prediction of their trust can be made with knowledge of prior interactions. The dynamics of trust depend on aspects of the agent or the environment that change in time [32].

Self-reported trust represents an introspection that subjects can make based on their evaluation of the current interaction environment. Future work can use this kind of self-reported measurement to estimate potential actions a human may take given their history of previous actions and trust ratings. Connecting human judgment of autonomous system capability and behavior with human action is crucial for moving the field of human autonomy teaming forward.

Chapter 3

Anytime Trust Rating Dynamics in a Human-Robot Interaction Task

This chapter describes the implementation of a one-dimensional trust slider measurement tool. The inspiration for this study came from findings from the grid search experiment, which found that trust rating in the two other dimensions of 'reliability' and 'familiarity' aligned with the explicit measure of 'trust.' Although a single item naturally loses the depth of a multidimensional survey, allowing subjects to rate their trust at their discretion provides information about the task context that influenced their decision. Outside the laboratory environment, tasks may not have binary outcomes and defined times for surveys. Thus, an anytime survey may naturally fit into an operational paradigm.

3.1 Introduction

Autonomous systems are becoming increasingly capable in daily life and high-risk situations. Each new "self-driving" car model adds features that remove the need for the operator's full attention. Aspirational future space habitats include integrated autonomous robotic systems to perform work while humans are away [69]. The capabilities of these systems introduce a new paradigm for how humans interact with them [12], including human affective responses like trust.

Measuring and modeling human trust has taken many forms since Mayer's organizational interpretation [50] or Lee's allocation strategies [47]. Nevertheless, what has stayed

constant is the understanding that many factors contribute to how people perceive and interact with autonomous or automated technology. Trust, therefore, is a valuable tool to encapsulate some notion of how human action relates to robot action.

Subjects in this study command a remote robot manipulator to sort stowage onboard a simulated International Space Station. The robot's task performance changes from trial to trial according to its "algorithm" identity. The interface presents a one-dimensional trust slider alongside a rendered video of the real-time simulation. Subjects are free to rate their trust throughout the task at any time.

Given the subjects' freedom to rate their trust at any time, events that coincide with ratings in time are likely to provide the context for why subjects rated their trust. We conducted a survival analysis demonstrating the probability that a subject rates their trust in time relative to a single robotic pick and place action.

3.2 Background

Integrating autonomous systems into human work environments has the potential to increase productivity, but with the risk that the system potentially fails. Reliance on these systems is well studied and connects notions of human trust with factors from the system, humans, and the environment. Modulating the level of autonomation [54] may be a helpful tool to balance performance and risk. Given the affective connection between automation reliance and trust [55], capturing a measure of trust throughout interaction makes sense. Trust surveys are a generally accepted measurement mode, but depending on the form and frequency, they can be obtrusive [42].

Human's trust in an autonomous system is continuous and dynamic. Many factors impact one's trust in an autonomous agent, including dispositional factors or situational factors [32]. However, trust dynamics based on the order of short time scales may primarily rely on learned trust factors like the system's performance. The trust model one uses should be aligned to the timescale for which the conditions change [63]. The collaborative nature of short-time scale interaction may introduce changes in human reliance behavior that do not directly involve the change in a human's trust attitude [46] and thus requires

a separate measure from reliance itself.

The autonomous system in this experiment is a six-degree-of-freedom manipulator, which resembles a human arm with an analogous shoulder, elbow, wrist, and fingers. Since trust is affected by the behavior and appearance of a collaborating autonomous agent [13], it is essential to recognize that the identity of the robot, in this case, also plays a role [75]. We have shown in previous work [16] that using identities compartmentalized by "algorithm" and color was sufficient for subjects to maintain consistent trust evaluations.

3.2.1 Human decision-making processes

Studying the human mind from the top down provides insight and a modeling framework that can often describe their behavior. A computational model uses an information processing approach to explain the inputs and outputs of a given system [6]. In a human-autonomy teaming task, one may use information acquired from their senses to make decisions on which actions to take to complete the task. Tourangeau describes that attitudes, such as trust, are "structures in long-term memory," an information storage module ready to be retrieved and applied in a relevant decision-making context [71]. Reliance decisions and change in trust may not coincide due to the current state of the task or environment. This study aims to investigate trust changes in time relative to current robot actions.

Earlier decision-making models used a utilitarian representation of decision-making where rational agents aimed to maximize some value [39]. Then, computational models used combinations of strategies and heuristics [43] to explain behavior. Eventually, information processing modules coalesced to form cognitive architectures like ACT-R [4].

Specific decisions can be represented probabilistically using sequential sampling models that integrate information over time as a random walk toward some possible decision threshold. Drift Diffusion Models are a form of sequential sampling that have been useful for investigating decisions involving perception, where one is often required to discriminate between different stimuli [59]. A parameter series describes a "first hitting time" distribution due to the random walk. Huang has demonstrated a drift diffusion model in a quadrotor control task where subjects were asked to move robots to a target position

and avoid obstacles [35]. The model was parameterized with robot state error, compared to the goal, as the drift rate, which estimated human decision timing.

Decision Field Theory uses diffusion but models decision preferences as separately sampled stochastic signals that eventually cross a boundary at which time a decision is made. This has been used to model reliance on automation in [25]. In a pasteurization control task, they found that reliance depends on an interaction between trust and self-confidence, given that a fault occurs in the system.

3.2.2 Types of trust measurement

Although studies have shown a close tie between trust (as a behavior) and automation reliance, trust surveys still prove valuable as a quick attitude assessment. This is especially true when connecting trust to a specific behavior or trait of an autonomous system in a particular situation. Although periodic surveys capture trust well, they cannot identify coincident events that may cause a change in trust.

The semi-automatic pasteurization plant is the canonical environment that measures trust after a fault is introduced in the system [47]. An autoregressive model describes the response to a series of questions post-trial, the final asking, "Overall, how much do you trust the system?". Razin and Feigh report that 12.6% of trust surveys use a single item trust evaluation [60] and that Jian et al. [37] (a multi-question survey) have the most cited survey in use for experiments in the automation field.

Single-item trust surveys cannot capture the richness and complexity of factors that contribute to the construct [32]. However, given the small amount of time needed to complete one, they may be valuable for "momentary assessments" [45]. Proponents of multi-dimensional trust surveys may claim that the lack of detail may not be able to explain the contributing factors to a subject's trust in the system. However, given the interaction context with the subject's freedom to rate, their trust may reveal that these are connected in time.

Reliance in automated systems is a measurable trust factor [63]. Some studies force subjects to chose to rely on an autonomous system at regular intervals - between trials [8] - while others allow reliance at any time - like a driving takeover task [48]. On the other

hand, a majority of experiments survey trust at regular intervals. Desai et al. require subjects to rate changes in trust during the task at regular intervals [18], as well as the OPTIMo study [76].

Connecting trust to individual features of an interaction requires a subject to decide to rate their trust. Subjects rated their trust at their discretion in [29] after a training period. If a robot behaves erratically and then succeeds in the task, an anytime trust survey localizes the trust factor immediately if a subject rates their trust after the robot's behavior. In this case, a short, single-item survey is preferable to limit the disruption and workload of an interactive task.

3.3 Experiment

3.3.1 Experimental Task

Measuring trust in an autonomous robot requires the subject to have some sense of involvement and interaction in the task and robot. The supervisory task we have chosen for this experiment blends risk elements - a spaceflight context - with enough interaction to keep the subject engaged while allowing the robot to contribute to task performance independently.

The task simulation occurred in a Gazebo simulation of the International Space Station (ISS) as shown in Figure 3.1 using ROS, a common robotics programming framework. This simulation portrayed a realistic representation of how a physical robot would perform. The robot used a PRM* trajectory planner [40] through a series of consistent gripper waypoints. That way, subjects would see how a robot would move with a true trajectory generator. The anthropomorphic effects of a robot arm could help subjects anticipate the outcome.

The human-robot team aimed to move proper stowage pieces, represented by colored cubes, to their proper position on a rack. We chose colored cubes to reduce ambiguity about whether the robot successfully completed the task. The subject provided supervisory commands to the robot through their remote user interface. The instructions for the subject overlaid on the remote user interface are shown in Figure 3.2. Along with

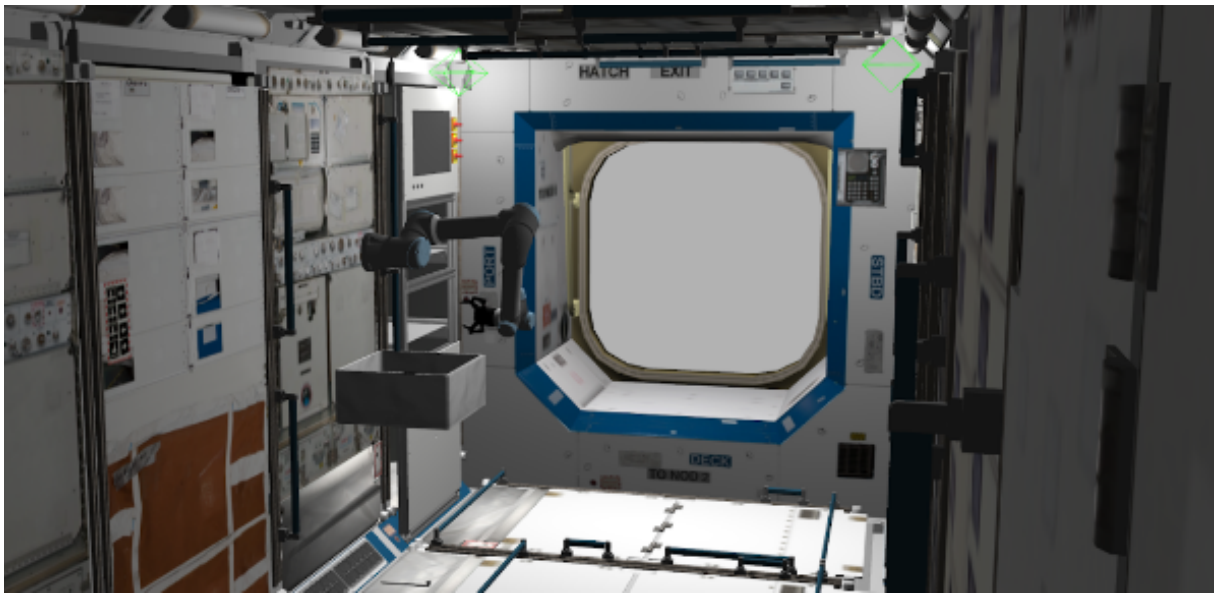


Figure 3.1: A rendering of the Gazebo environment where the experiment took place. The subject interacted with the robot through a simulated camera and remote user interface. The simulated UR5e robot arm picked stowage blocks from a container and placed them on the rack in a specified configuration.

this graphic was a description of the goals of the experiment, including an emphasis on "improving the way people work with smart technology". The experiment was designed to take between 30 and 60 minutes. Subjects were instructed to complete the task with a reliable internet connection and a quiet environment.

To measure the trust the subject had in the robot, we used the following prompt in the instructions:

> Throughout the process, please adjust your trust in the system using the purple slider beneath the camera feed where the left indicates less trust and the right indicates more trust.

The trust slider always retained its position from when the subject released it. Although trust is a major component in this study, we wanted to ensure that the measure of trust would not bias subjects to rate their trust in any specific feature of the robot. Lastly, subjects were provided a disclaimer about the reality of live robotics simulations and that it was possible for the stowage cubes or the robot to display strange behavior.

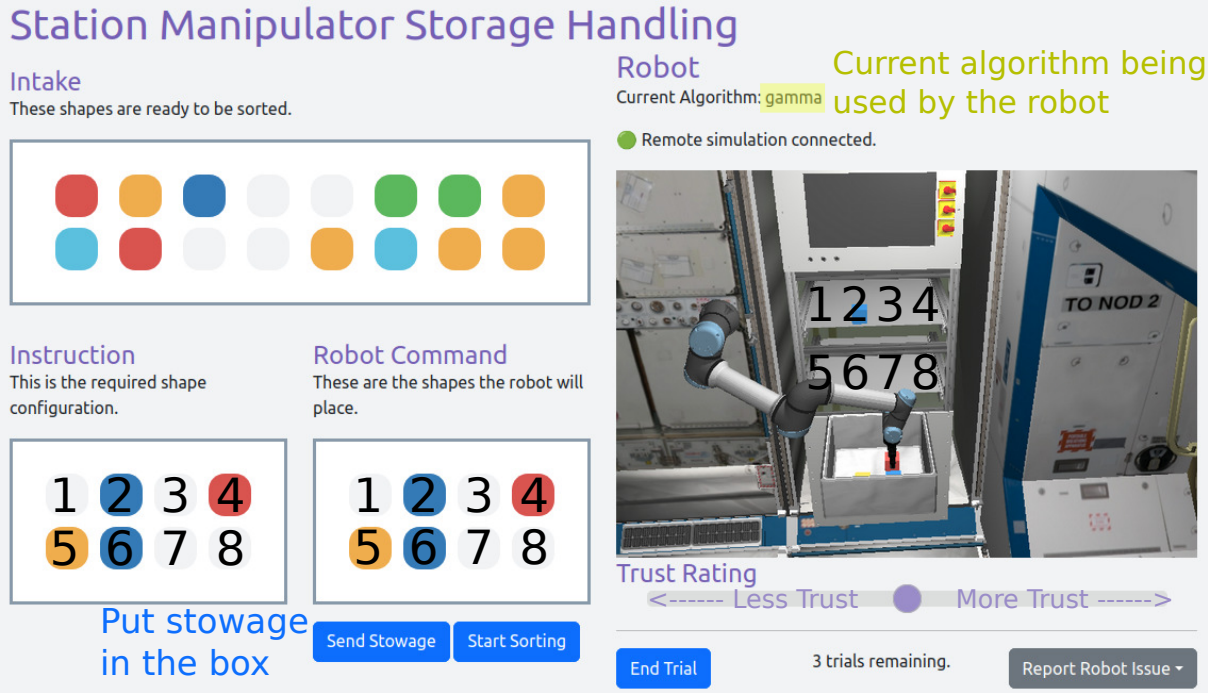


Figure 3.2: A visual of what subjects see during training. The colored squares correspond to stowage spawned in the box below the rack. Each position shown in **Instruction** and **Robot Command**, indicated by number, corresponds to a position shown on the rack. Subjects choose stowage from the **Intake**, then choose a position in **Robot Command** to indicate where they want the stowage placed on the rack. The algorithm (Gamma or Echo) is indicated above the robot simulation live feed. The trust rating slider is accessible below the live feed. The markup on this image appeared only for instruction.

The robot's performance was modulated as an experimental condition. There are two "algorithms" used by the robot for stowage placement: *Gamma* and *Echo*. These embodiments separated the task performance of the robot into two distinct categories. Gamma performs correct placements every time, while Echo places stowage in the wrong position 50% of the time. The current algorithm was shown to the subject on the interface.

3.3.2 Experimental Design

The experiment was split into ten trials of four grasps each. The robot algorithm was randomly assigned each trial to control for possible ordering effects. Figure 3.3 shows a possible subset of trials for a single experimental subject. Although the robot makes mistakes, the subject is likely to assign blame to the "algorithm" identity. This way, they may track the robot's reliability throughout the experiment.

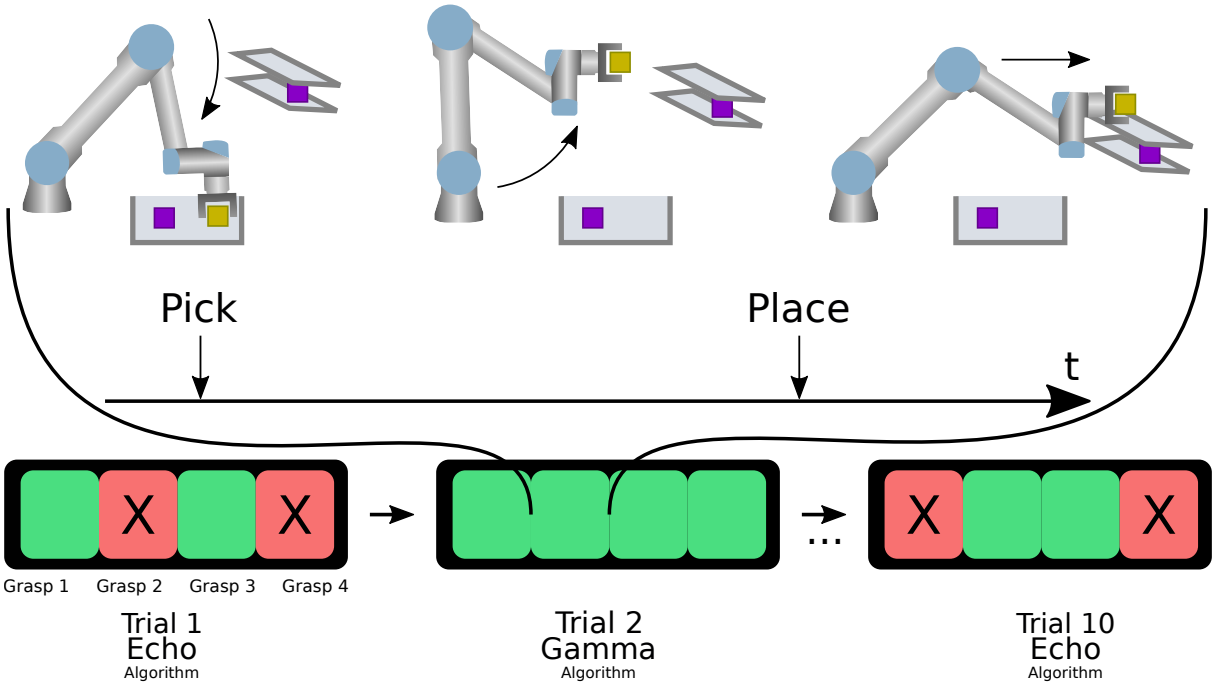


Figure 3.3: Each included subject completes ten trials of 4 grasps each. An algorithm, Echo or Gamma, is assigned randomly for each trial. A grasp is defined as the moment the gripper closes on a piece of stowage to the next time it closes on stowage. This way, trust ratings are associated with the prior grasp. Subjects may adjust their trust rating at any time.

Our hypotheses assume that subjects' trust in the robot is tied to the robot's behavior and performance. To validate this, we asked open questions after the experiment about what influenced them to rate their trust. The trust slider was intentionally optional to not bias the subject into rating their trust according to our hypothesis, but checked upon the completion of the experiment with the following open-ended questions:

- What caused you to change your trust in the robot?
- Why do you think the robot failed to complete the task at times?
- Was any part of the experiment difficult or hard to understand?
- Were there any technical problems with the experiment?

The University of California's Davis Institutional Review Board approved this research, and it complies with the American Psychological Association Code of Ethics.

Participants provided consent and were allowed to exit the experiment without penalty.

3.4 Trust Model

3.4.1 Trust in a decision-making paradigm

In this experiment, subjects could rate their trust at any time. Subjects' latent trust in the robot changes continuously over time depending on many factors. However, their conscious introspection of their trust state happens when they are prompted or face a situation that defies their expectations. In the case of the interaction shown here, the robot's task performance is the central stimulus contributing to changing trust. Although many factors may change someone's trust in a robot, performance is the most salient over the timescale of this experiment.

Subjects are instructed to "adjust their trust throughout the process" of the experimental task. The experimental instructions were biased toward completing the sorting task, with trust reporting left as a secondary, optional task. The purpose of this design was not to bias the interpretation of trust to rely on any single factor. Additionally, without forcing trust ratings at regular intervals, the subjects would not expect some stimulus to appear that they should use to rate their trust. The experiment included these details to demonstrate that lay people can introspect their trust state in a robot collaborator without definitions of trust.

As subjects observe the robot progressing through the task, they gain evidence about its behavior, capabilities, performance, and more. Eventually, when they have decided they have enough information, they may realize that their trust has changed. From the Sequential Sampling Model's point of view, they have accrued enough information to meet a threshold and make a decision [39]. For example, the drift diffusion model (DDM) is used in a task of a *two-alternative forced choice* task. Over time, evidence is accrued as a random walk with a drift rate. The corresponding response is chosen when the signal reaches a specific positive or negative threshold. This model may be more suitable for trust decisions that are binary or have forced decisions.

3.4.2 Trust rating in a survival model

When subjects decide to report their trust rating in the robot, they have gathered sufficient evidence. Whether that evidence comes from performance, behavior, or some other covariate is what we aim to model. Assuming the subjects' trust is tied to each *grasp* that a robot attempts, we portion each grasp starting with the moment the robot closes its gripper on a piece of stowage and ending after the robot places stowage before grabbing the next stowage. Since subjects are free to change their trust throughout a grasp, we assume that they have acquired the *necessary* evidence to report their trust by the time the last trust rating occurred during a grasp. This moment is the Trust Rating Time (tRT).

Survival Analysis [14] models the expected time it takes for an event to occur. Here, we aggregate all the grasps for which subjects decided to report their changed trust. The survival function is given by the following formulation:

$$S(t) = Pr(T > t) \tag{3.1}$$

Where t is the time since the start of the measurement period when the robot picks the stowage, and then T is the tRT. The survival function is never increasing since waiting to rate trust later means the *last* trust rating could not have already happened. The function describes the probability that the tRT *survives* past the time t.

The complement of the Survival function is the cumulative probability distribution of tRTs.

$$F(t) = Pr(T \leq t) = 1 - S(t) \tag{3.2}$$

Estimating the tRT density requires the hazard function $\lambda(t)$, which is the instantaneous rate of an event occurring conditional on surviving to that time.

$$\lambda(t) = \lim_{t \to 0} \frac{Pr(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)} \tag{3.3}$$

The Cox Proportional Hazards Model describes a hazard function that includes multiplicative covariates with respect to the baseline hazard, λ_0, shown in its general form in

Equation (3.4) [14]. This model uses constant and time-varying covariates represented by x and y respectively. Their associated constant coefficients are β and η. In vector x, the constant covariates used from this experiment are 1) grasp success and 2) trust. y is the time-varying covariate, grasp completion, which is 0 until the stowage is placed, then 1.

To solve for the coefficients in the hazard function, we use the Equivalent Poisson Model as described in [62]. We split each grasp over the experiment, i, into equally spaced intervals in time, j. If a subject rated trust in an interval, take *rating indicator* $d_{i,j}$ (d is traditionally used as a "death" indicator in survival analysis) as 1, otherwise 0. Next, *exposure* indicates the intervals for which the covariates interact in the hazard function: $e_{i,j}$ is 1 for each interval j which the subject has not yet rated their trust, or the time in the interval that that they rate their trust, otherwise 0. Then, we take $\mu_{i,j} = e_{i,j}\lambda_{i,j}$ as the expected number of ratings. $\mu_{i,j}$ is fit to the observed rating times in $d_{i,j}$. This formulation allows for time-varying covariates, $y_{i,j}$, in Equation (3.5).

$$\lambda(t) = \lambda_0(t)exp(x\beta) \tag{3.4}$$

$$\lambda_{i,j} = \lambda_0 exp(x_i\beta + y_{i,j}\eta) \tag{3.5}$$

To build an extremely well-fit model, one could vary λ_0 with time or interval to represent a more general form or Weibull baseline [62]. We chose a constant baseline because the time-varying effects are captured by the progress of the robot's grasp. Additionally, a censored model would be more appropriate if trust ratings *could* have been made in the future. However, this was not appropriate due to the time pressure inherent in the sequence of grasps.

We used PyMC [56], a Bayesian modeling Python library, to fit the parameters in this model. After specifying a model symbolically and assigning priors for parameters, PyMC samples from the model and updates a posterior distribution of parameter values using the No-U-Turn Sampler [33].

3.5 Results

3.5.1 Demographics and exclusion

Subjects were excluded from the experiment if they did not finish the entire study, had unusually long trials, or had an unstable internet connection characterized by median latency greater than 300ms. There were 65 total subjects included, of which 18 identified as male, 45 female, and 2 nonbinary. The average age was 19.7 years with standard deviation 1.5. Of the 41 subjects who played video games, the median subject played between 1 and 5 hours per week.

The remote study enabled subjects to participate in the experiment from wherever was comfortable for them. However, this introduces a non-negligible delay in the timing of interactions. Most subjects were recruited from the UC Davis campus but may have completed the experiment remotely. For example, baseline latency to San Francisco is $\approx 40ms$, measured using speedtest.net. The user interface tracked latency every 10 seconds by calling the *get_time* ROS service and measuring the duration before receiving a reply. Then, latency was corrected in analysis by applying a rolling median filter to the latency measurement and shifting the logged data by half the measurement to be conservative.

3.5.2 Measuring Trust

Subjects changed their trust using the trust slider throughout the experiment. Each included subject completed ten trials, which consisted of 4 grasps each. Figure 3.4 shows an aggregation of trust change over all the grasps in which subjects rated their trust. The trust change data are summarized in Table 3.1.

	count	mean	std
Algorithm			
echo	303.0	-14.7	33.1
gamma	253.0	16.3	32.1

Table 3.1: Average trust change according to algorithm. Using a paired t-test, the trust change is significantly different between the Gamma and Echo algorithms ($p \ll 0.01$).

Although subjects can rate their trust at any time, we found that they typically

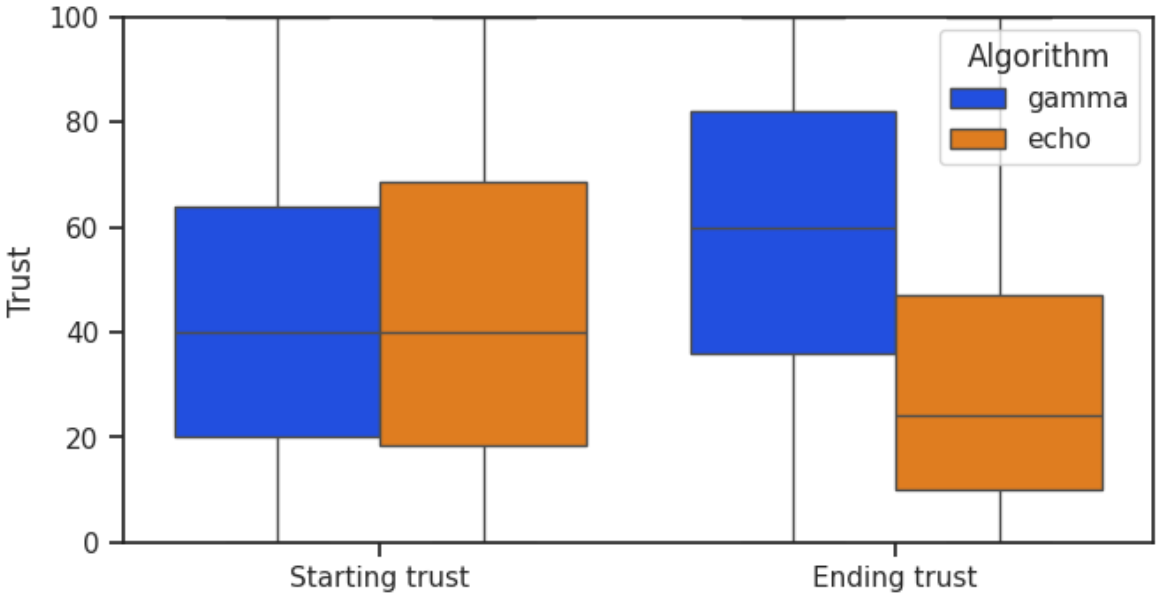


Figure 3.4: This comparative box plot shows the change in trust averaged over every grasp in which they changed trust. The capabilities, or algorithms, are shown in different colors where the Gamma algorithm always places stowage correctly, and the Echo algorithm misplaces stowage 50% of the time. This analysis treats *every* grasp as an independent event. Starting trust is the subjects' trust before the robot "picked" the stowage and Ending trust is the last trust rating before the next stowage "pick".

rated trust towards the end of the trial. This suggests that subjects wanted complete evidence over more grasps before evaluating their trust in the robot. Figure 3.5 shows the distribution of the when trust ratings occurred by grasp number.

3.5.3 Trust Reaction Time

There is a clear delineation between when a grasp ends and the next one begins. Since each of the four grasps proceeds in sequence, there is a natural time pressure to rate trust, if they choose to, before the start of the next grasp. We see this illustrated in Figure 3.6. Although the distributions of rating times have similar peaks, there is a noticeable drop-off around 10 seconds for the first three grasps since this is when the next grasp starts. Because of this time pressure, we parameterized the reaction times differently for the first three and the last grasp (four grasps total).

Out of the total 2061 grasps witnessed by subjects during the experiment, they rated

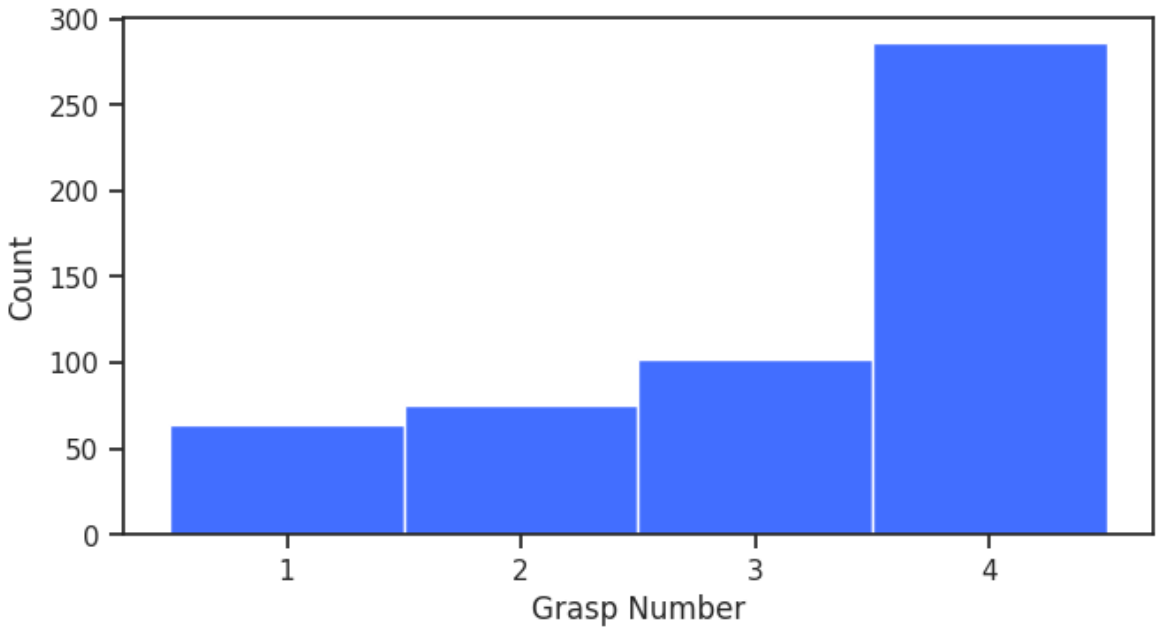


Figure 3.5: Distribution of occurrence of subjects' trust ratings by grasp number. Most ratings occurred during the final grasp.

their trust in 556 of them. Applying the Survival Analysis procedure, we produce the parameters in Figure 3.7. The values have converged as evident from the *r_hat* statistic, or the Gelman-Rubin Statistic, since the value is nearly close to 1 for each parameter. The parameters between the models for the first and last grasp are comparable. Interestingly, neither trust nor success has an impact compared to η, the coefficient of the grasp state covariate. However, the *success* coefficient is significant compared to the baseline hazard, the λ_0 constant, suggesting that the impending *success* of a grasp contributes to a higher probability of trust rating before the grasp has been completed.

Given the parameterized hazard function, we can sample the observed covariates and plot a survival function for each one shown in Figure 3.8. Since most of the grasps had similar placement durations, we can overlay the empirical cumulative distribution function of reaction times for comparison. This function should fall near the center of the predicted traces for a well-fit model.

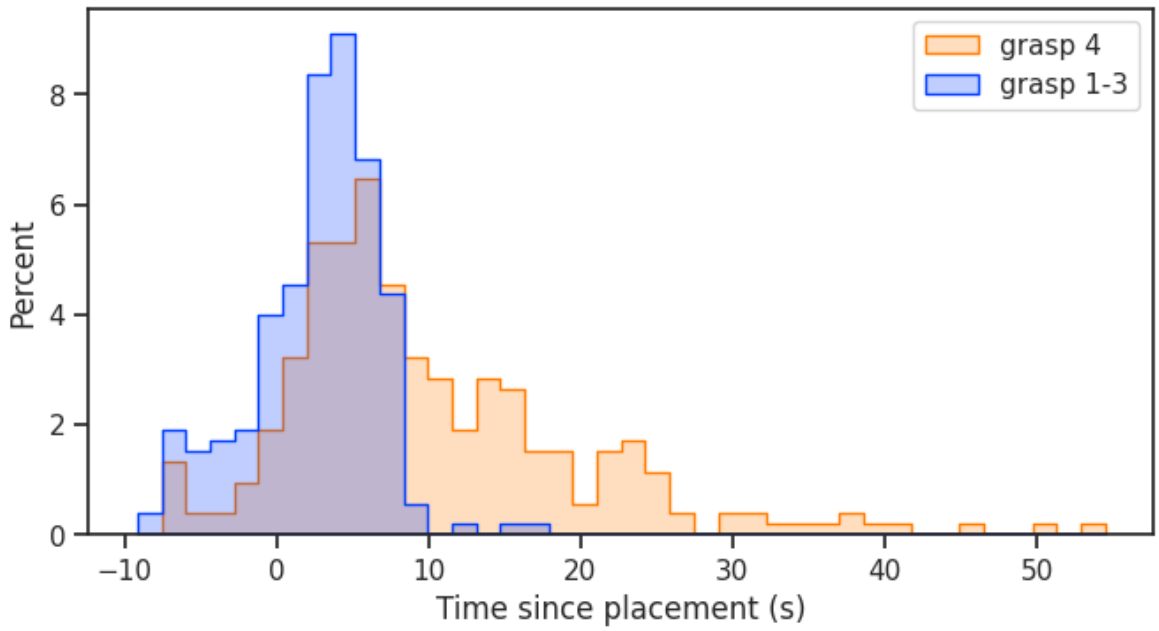


Figure 3.6: Trust rating time distribution separated by whether the grasp was the last one. The *pick* happens near -10s. When the grasp is the last of the trial, subjects have unlimited time to change their trust. For the first three grasps, subjects are pressured to rate their trust before the next grasp near 10s.

3.6 Discussion

3.6.1 Deciding to trust

When a subject decides to rate trust, they have processed enough information about the situation, combined with their previous experience with the robot, to introspect. That moment is captured after some time of thought, then motor action. This experiment aimed to capture the total time from robot action to trust-rating action. The supervisory task required the subject to command some action and task performance for which the robot followed. The robot's performance could be readily tracked according to the "algorithm" it followed. Since this was the most salient trusting attribute, it was likely the basis of the subjects' trust.

Subjects had the freedom to rate their trust at any time during the grasp and were not limited to the conclusion of the placement. Although many waited for this moment, some rated the grasp well before, anticipating the outcome. The timescale with which

	mean	sd	r_hat
beta[success]	0.15	0.15	1.00
beta[trust]	-0.00	0.00	1.00
eta	1.97	0.16	1.00
lambda0	0.03	0.01	1.00

(a) Grasps 1-3

	mean	sd	r_hat
beta[success]	0.07	0.14	1.00
beta[trust]	-0.00	0.00	1.00
eta	2.34	0.22	1.00
lambda0	0.01	0.00	1.00

(b) Final grasp

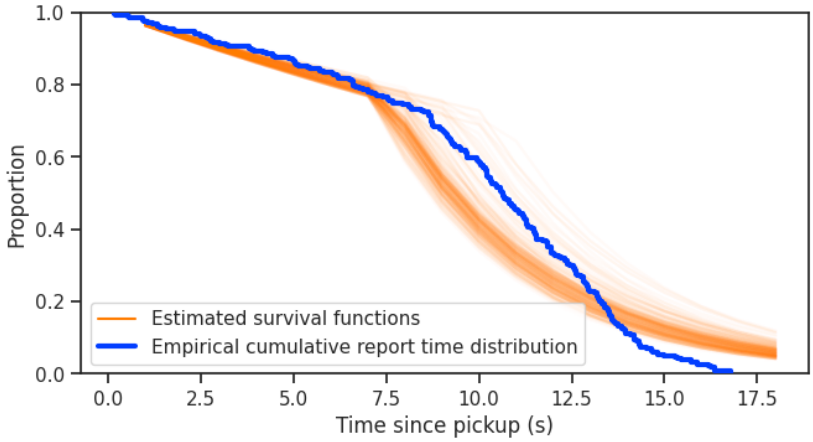
Figure 3.7: Parameter tables for the hazard function described in Equation (3.5). a and b are fit separately because the subjects have no time pressure to register their trust for the final grasp. The η coefficient is time-varying, while the β coefficients are not. $\lambda 0$ is the baseline hazard. All r_hat values are close to 1, indicating convergence.

they waited was on the second's scale, as seen in Figure 3.6, but some took longer when given the opportunity.

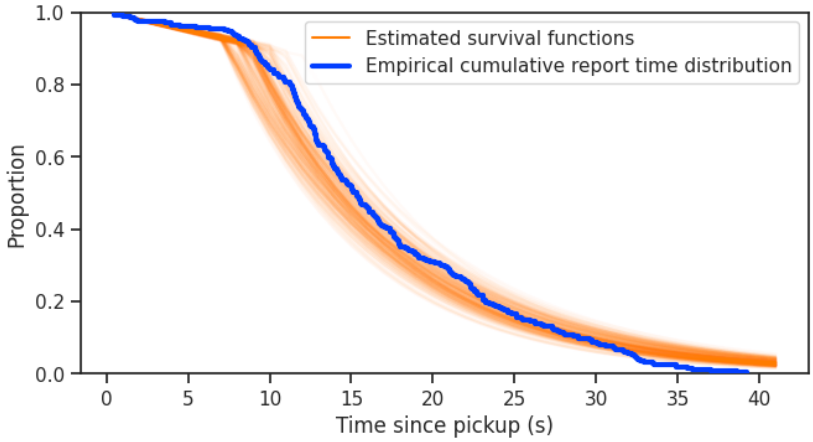
3.6.2 Survival analysis and decision making

The survival function described in this work illustrates the probability that a subject waits to decide to rate their trust relative to when the robot picks up a piece of stowage. When the robot initially grabs a piece of stowage, the subject must check that it is correct, then observe as the end-effector approaches the right location on the rack. When they are confident that the robot is correct or makes a mistake, they may decide to change their trust.

If they integrate information from their observation, it makes sense that the subject is more sure of the outcome as the robot gets closer to the goal. In other words, the probability that the robot's reliability differs from expectations increases or decreases



(a) Grasps 1-3



(b) Final grasp

Figure 3.8: Predicted survival functions for the set of covariates from each grasp where subjects rated their trust compared with the empirical survival function. The model fits well for the final grasp. For the first grasps, the model predicted that subjects were likely to rate their trust sooner than what the observations showed. The large change in slope happens when the robot places the stowage on the rack. After this time, the rate at which subjects registered their trust increased dramatically.

faster as the outcome approaches. This is akin to the *hazard* accumulating [7] over time.

$$\Lambda(t) = \int_0^t \mathrm{d}t' \, \lambda(t) \tag{3.6}$$

We can rewrite the final part of equation 3.3 and solve for the survival function.

$$\lambda(t) = -\frac{d}{dt} \log(S(t)) \tag{3.7}$$

$$S(t) = \exp(-\int_0^t \mathrm{d}t' \, \lambda(t)) = \exp(-\Lambda(t)) \tag{3.8}$$

In this case, the cumulative hazard, $\Lambda(t)$ is a natural fit for relating stimulus over time to a decision-making event, a trust rating.

Figure 3.8 shows a change in the slope of the predicted survival function around the time the robot places the stowage. Since the proportional hazards model's covariates are multiplicative over the baseline hazard, any differences in parameters in Figure 3.7 describe differences in the rate of the probability of trust rating. For both sets of parameters, η, the time-varying state of the grasp, is much larger than the baseline hazard, indicating the hazard increases dramatically after the grasp completes. The baseline λ_0 for the earlier grasps is larger than the one for the final grasp, which mainly corresponds to the risk before the placement occurs, which means the probability of rating trust before stowage is higher before the last grasp.

The model proposed here successfully converged to a set of parameters that captured the shape of the empirical cumulative distribution of trust rating times. The time pressure of the first three grasps influenced subject behavior in a way that could not be captured with this model. The "time until the next grasp" is information not accessible until after the next grasp starts, making it invalid for this analysis. Allowing subjects to control the start of individual grasps would eliminate time pressure to rate trust and align the behavior seen during the final grasp of each trial.

A single-item trust survey may prove useful for assessing an overall feeling or impression. However, more is needed for more detailed trust analysis, especially if subjects fail to consider the complete set of possible factors that influence their decision [45]. The

experimental task's most salient feature was the outcome of each pick and place. The proposed model identifies this feature as the most influential and thus sufficient for this analysis. Since task outcome changes throughout the experiment, it makes sense that a subject is most attentive to this factor. However, suppose a researcher was interested in other factors like predictability, familiarity, or developer intention. In that case, a multi-dimensional survey may be more appropriate [45]. Capturing these other factors requires the appropriate covariates from the task.

3.6.3 Recommendations for a one dimensional trust slider

An exit survey was administered after the experiment, which consisted of free-response questions designed to learn how subjects interacted with the robot. When asked what caused them "to change their trust in the robot?" 85% of responses indicated that either "accuracy" or "stowage placement" affected their assessment; 15% indicated that robot behavior affected their trust (not mutually exclusive). Although subjects were not provided a trust definition, their explanation aligned with our expectation that trust would be related to the robot's task performance.

Of the responses that were not accuracy or behavior-related, two mentioned the robot not listening to their commands; one tested how changing the slider could affect the robot, one stated, "It had no mind," and one was unsure what trust meant. A large majority of subjects indicated post hoc that they based their trust on the outcome of the task without prompting, which suggests that the trust slider method captures the volitional trust rating we aimed for.

Our subject pool was limited to undergraduate students at an English-speaking university. If a similar study uses this survey concept, researchers should consider how the colloquial definition of trust changes across language and education status.

Throughout the experiment, the trust slider never reset to the central starting position; it retained the value with which it was last set. After a grasp, trial, or algorithm change, the slider could have been reset but may have prompted the subject to rate their trust at that time rather than when they felt their trust had changed. If a researcher wished to receive more trust ratings at the expense of altering the subject's attention, they may

want to reset the slider and prompt the subject to rate their trust.

3.6.4 Operationalizing any-time trust rating

We envision an interactive system that uses trust queries to capture part of humans' complex decision-making process. If a robot, for example, could ask a collaborating human about their trust in itself, it could have an idea of what the human will likely do in the future [19]. This study highlights the possibility of an unobtrusive single-dimensional trust measure that captures reactions to a robot's task performance and behavior. This measure may prove valuable as human-robot relationships transition from a supervisory hierarchy to a flattened decision-making structure [12].

3.7 Conclusion

Subjects use evidence to rate their trust in the robot. Their *decision* requires them to contemplate what aspects of the robot in this context change their trust. There is some probability that subjects rate their trust before the robot places the stowage because they can anticipate the outcome or evaluate robot behavior. We provide the trust survey as a single-dimensional slider, but since subjects can rate trust at their discretion, the factors that influence their trust will likely coincide with the rating. Additionally, the generality of the survey allows subjects to emphasize which factor of trust most affected their trust report.

This model did not take individual demographics or individual behavior into account. After some interaction with the experiment, learned behavior could have changed when the subject decided to rate trust after knowing the robot's motion; the subject could anticipate the outcome sooner. According to the model, trust and task outcome played little role in the actual timing of subjects' trust reports, only that the timing of the task outcome coincided with the trust report.

The findings from this study can be used to inform future any-time trust rating experiment designs. Most subjects reported task-performance-related reasons for changing their trust, but the supervisory nature of the robotic task likely influenced this; for more collaborative tasks, subjects may report other factors for changing their trust. With the

increasing intersection of humans and autonomous systems, quick trust queries like the survey presented in this study could enable more fluid interactions.

Chapter 4

Conclusions and Future Work

People work together by continually evaluating trust or distrust in each other, leading to better predicting each other's actions. Additionally, people know when their actions may affect other's trust in them. Human-robot interactions are currently devoid of human cognitive state estimates. The research described in this dissertation has contributed another step toward autonomous systems that understand the affections of human beings in order to collaborate better. The grid search experiment uncovered complexities behind how people learn to trust depending on the order of interactions with an autonomous agent. The pick-and-place experiment explores how minimal trust surveys can be used to understand what causes somebody to change their trust by connecting trusting events temporally. Generalizing the cognitive architecture and trust framework to more tasks, agents, and environments would contribute to richer, more natural human-robot interactions.

4.1 Summary

Chapter 2 described an experiment where human subjects collaborated with an autonomous agent to complete a task. The shared task was a visual search administered remotely as a web game. Many human-autonomy teaming studies have looked at the issue of robot performance as a sole predictor of human trust, but this could underestimate the complexity of the interaction. In this study, we manipulated the autonomous agents' performance and behavior to measure how reported human trust changed over time. In each trial, a different autonomous agent with a preassigned *capability* used one

of three search *strategies* and then reported the number of outliers it found as a fraction of its *capability*. Then, the subject reported their total outlier estimate. Human subjects then evaluated statements about the agent's behavior, reliability, and their trust in the agent. Agent capability and strategy stimuli were presented in different ordering schemes between-subjects.

The results showed increased measures of trust and reliability with increasing capability. Additionally, the most legible search strategies received the highest average ratings in a measure of familiarity. Remarkably, given *no* prior information about capabilities or strategies that they would see, subjects were able to determine consistent trustworthiness of the agent. Furthermore, both capability *and* strategy of the agent had statistically significant effects on the human's trust in the agent. Self-reported trust was modeled using Ordinary Least Squares, but the group that interacted with varying *capability* agents on a short time order produced a better performing ARIMAX model. Models were cross-validated between groups and found a moderate improvement in the next trial trust prediction. A time series modeling approach revealed the effects of temporal ordering of agent performance on estimated trust. Recency bias may have affected how subjects weighed the contribution of strategy or capability to trust. Understanding the connections between agent behavior, agent performance, and human trust is crucial to improving human-robot collaborative tasks.

Chapter 3 investigated the use of a "trust slider" in a human-autonomy teaming task. Many studies view trust as a slow-changing value after subjects complete a trial or at regular intervals. Trust is a multifaceted concept that can be measured simultaneously with a human-robot interaction. The objective of this study was to model trust rating timing based on factors from the task. 65 subjects commanded a simulated remote robot arm in a simulated space station. The robot picked and placed stowage commanded by the subject, but the robot's performance varied from trial to trial depending on the "algorithm" it was using. Subjects rated their trust on a non-obtrusive trust slider at any time throughout the experiment. A Cox Proportional Hazards Model described the time it took subjects to rate their trust in the robot during each pick and place event. A

retrospective survey indicated that subjects based their trust on the robot's performance or outcome of the task. Strong covariates which represented the task's state reflected this in the model. Trust and robot task performance contributed little to the timing of the trust rating. The subjects' exit survey responses aligned with the assumption that the robot's task progress was the main reason for the timing of their trust rating. Measuring trust in a human-robot interaction task should take as little attention away from the task as possible. This trust rating technique lays the groundwork for single-dimensional trust queries that probe estimated human action. When human subjects are given the discretion to rate their trust in a robot counterpart in a single dimension, the timing at which they rate their trust reveals the factor that influences their trust. In this simulated supervisory task, a robot's reliability affected its performance, with which subjects changed their trust.

4.2 Future Work

Human trust is a fascinating emotion that everyone has experienced. When someone's trust has been violated, they may feel let down, angry, or even vengeful. Trust arrives on foot and leaves on horseback. We intuitively understand that trust changes over time for many reasons; we have a mental list of places, people, and things that we trust, and we can reason about how our trust changes in these things over time. If we think about events that changed our trust; what actions would we have taken if we had known not to trust? Or would we have acted differently if we had known to trust? The complexity of this cognitive state captures the richness of an interaction that can't always be boiled down to a single number. Countless studies in the human factors field argue for and against long and short surveys, or behavioral analysis, attempting to capture all the facets of trust. Asking someone to rate their trust may not provide a meaningful number in isolation. Still, it relates to others' trust ratings and their own rating in time. Pairing a trust rating with the dynamics of the environment can elucidate how the environment makes somebody feel and how they may act.

The cold, rational approach to collaboration necessitates creating a faster, safer, and

more accurate task outcome. Autonomous agents are tools that make people's lives easier. Incorporating a measure of human trust into autonomous systems can help them reason about human actions that don't seem rational.

Future work should investigate even simpler trust measurement strategies that capture the element of the construct that is the most relevant factor. For example, robotic systems can query the trust of a collaborating human to improve a prediction of their next series of actions. In this application, a full trust survey only impedes progress toward completing the task when a simple question carries a wealth of information.

Other cognitive states, like workload and situational awareness, have been shown to interact with trust. These measures also provide estimates for future human action. If humans have a high workload, they likely are not willing to take on more tasks. If they have low situational awareness, they might not know which tasks need to be addressed. A similar measurement strategy, robot query, can estimate these cognitive states during collaboration. Modeling how these states change in time, in parallel with trust, is crucial to addressing how humans can collaborate most effectively.

Bibliography

[1] Kumar Akash et al. "A Classification Model for Sensing Human Trust in Machines Using EEG and GSR". en. In: *ACM Transactions on Interactive Intelligent Systems* 8.4 (Nov. 2018), pp. 1–20. ISSN: 2160-6455, 2160-6463. DOI: 10.1145/3132743.

[2] Kumar Akash et al. "Dynamic modeling of trust in human-machine interactions". en. In: *2017 American Control Conference (ACC)*. Seattle, WA, USA: IEEE, May 2017, pp. 1542–1548. ISBN: 978-1-5090-5992-8. DOI: 10.23919/ACC.2017.7963172.

[3] Allison P Anderson, Torin K Clark, and Zhaodan Kong. "Adaptive Autonomy for Future Spacecraft Habitats". en. In: 2020, p. 4.

[4] John R. Anderson. "ACT: A simple theory of complex cognition." en. In: *American Psychologist* 51.4 (Apr. 1996), pp. 355–365. ISSN: 1935-990X, 0003-066X. DOI: 10.1037/0003-066X.51.4.355.

[5] Christoph Bartneck et al. "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots". en. In: *International Journal of Social Robotics* 1.1 (Jan. 2009), pp. 71–81. ISSN: 1875-4791, 1875-4805. DOI: 10.1007/s12369-008-0001-3.

[6] Jose Luis Bermudez. *Cognitive Science: An Introduction to the Science of the Mind*. 3rd ed. Cambridge University Press, 2020. ISBN: 978-1-108-44034-9.

[7] Michael Betancourt. *Outwit, Outlast, Outmodel*. Jan. 2022.

[8] Shreyas Bhat et al. "Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task". en. In: *IEEE Robotics and Automation Letters* 7.4 (Oct. 2022), pp. 8815–8822. ISSN: 2377-3766, 2377-3774. DOI: 10.1109/LRA.2022.3188902.

[9] Maria G. Bualat et al. "Astrobee: A New Tool for ISS Operations". en. In: *15th International Conference on Space Operations*. Marseille, France: American Institute of Aeronautics and Astronautics, May 2018. ISBN: 978-1-62410-562-3. DOI: 10.2514/6.2018-2517.

[10] Robert R. Burridge and Kimberly A. Hambuchen. "Using prediction to enhance remote robot supervision across time delay". en. In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. St. Louis, MO, USA: IEEE, Oct. 2009, pp. 5628–5634. ISBN: 978-1-4244-3803-7. DOI: 10.1109/IROS.2009.5354233.

[11] Min Chen et al. "Planning with Trust for Human-Robot Collaboration". en. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. Chicago IL USA: ACM, Feb. 2018, pp. 307–315. ISBN: 978-1-4503-4953-6. DOI: 10.1145/3171221.3171264.

[12] Erin K. Chiou and John D. Lee. "Trusting Automation: Designing for Responsivity and Resilience". en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 65.1 (Feb. 2023), pp. 137–165. ISSN: 0018-7208, 1547-8181. DOI: 10.1177/00187208211009995.

[13] Myke C. Cohen et al. “Anthropomorphism Moderates the Relationships of Dispositional, Perceptual, and Behavioral Trust in a Robot Teammate”. en. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67.1 (Sept. 2023), pp. 529–536. ISSN: 2169-5067, 1071-1813. DOI: 10.1177/21695067231196240.

[14] D. R. Cox and David Oakes. *Analysis of survival data*. Monographs on statistics and applied probability. London ; New York: Chapman and Hall, 1984. ISBN: 978-0-412-24490-2.

[15] Bence Cserna, Wheeler Ruml, and Jeremy Frank. “Planning Time to Think: Metareasoning for On-Line Planning with Durative Actions”. In: *Proceedings of the International Conference on Automated Planning and Scheduling* 27 (June 2017), pp. 56–60. ISSN: 2334-0843, 2334-0835. DOI: 10.1609/icaps.v27i1.13827.

[16] Jason Dekarske and Sanjay S. Joshi. “Human Trust of Autonomous Agent Varies With Strategy and Capability in Collaborative Grid Search Task”. en. In: *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. Magdeburg, Germany: IEEE, Sept. 2021, pp. 1–6. ISBN: 978-1-66540-170-8. DOI: 10.1109/ICHMS53169.2021.9582622.

[17] Munjal Desai et al. “Effects of changing reliability on trust of robot systems”. en. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. Boston Massachusetts USA: ACM, Mar. 2012, pp. 73–80. DOI: 10.1145/2157689.2157702.

[18] Munjal Desai et al. “Impact of robot failures and feedback on real-time trust”. en. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Tokyo, Japan: IEEE, Mar. 2013, pp. 251–258. ISBN: 978-1-4673-3100-5. DOI: 10.1109/HRI.2013.6483596.

[19] Lasse Dissing and Thomas Bolander. “Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic”. en. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 1615–1621. ISBN: 978-0-9992411-6-5. DOI: 10.24963/ijcai.2020/224.

[20] Anca D Dragan and Siddhartha S Srinivasa. “Formalizing Assistive Teleoperation”. en. In: (), p. 8.

[21] Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. “Legibility and predictability of robot motion”. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Tokyo, Japan: IEEE, Mar. 2013, pp. 301–308. DOI: 10.1109/HRI.2013.6483603.

[22] Mica R. Endsley. “Toward a Theory of Situation Awareness in Dynamic Systems”. en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37.1 (Mar. 1995), pp. 32–64. ISSN: 0018-7208, 1547-8181. DOI: 10.1518/001872095779049543.

[23] Lorenzo Fluckiger et al. “Astrobee Robot Software: Modern Software System for Space”. en. In: *Proc. Int. Symp. on Artificial Intelligence* (2018), p. 9.

[24] Jeremy D. Frank. "Artificial Intelligence: Powering Human Exploration of the Moon and Mars". en. In: *arXiv:1910.03014 [cs]* (Oct. 2019).

[25] Ji Gao and J.D. Lee. "Extending the decision field theory to model operators' reliance on automation in supervisory control situations". en. In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 36.5 (Sept. 2006), pp. 943–959. ISSN: 1083-4427. DOI: 10.1109/TSMCA.2005.855783.

[26] Valerie J. Gawron. *Human performance, workload, and situational awareness measures handbook*. 2nd ed. Boca Raton: CRC Press, 2008. ISBN: 978-1-4200-6449-0.

[27] Christopher Geib, Bart Craenen, and Ronald P A Petrick. "Generating Collaborative Behaviour through Plan Recognition and Planning". en. In: (), p. 8.

[28] Felix Gervits et al. "Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models". en. In: *New Zealand* (2020), p. 9.

[29] Yaohui Guo and X. Jessie Yang. "Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach". en. In: *International Journal of Social Robotics* 13.8 (Dec. 2021), pp. 1899–1909. ISSN: 1875-4791, 1875-4805. DOI: 10.1007/s12369-020-00703-3.

[30] Peter A. Hancock et al. "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction". en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53.5 (Oct. 2011), pp. 517–527. ISSN: 0018-7208, 1547-8181. DOI: 10.1177/0018720811417254.

[31] G. Hirzinger, J. Heindl, and K. Landzettel. "Predictive and knowledge-based telerobotic control concepts". en. In: *Proceedings, 1989 International Conference on Robotics and Automation*. Scottsdale, AZ, USA: IEEE Comput. Soc. Press, 1989, pp. 1768–1777. ISBN: 978-0-8186-1938-0. DOI: 10.1109/ROBOT.1989.100231.

[32] Kevin Anthony Hoff and Masooda Bashir. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust". en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57.3 (May 2015), pp. 407–434. ISSN: 0018-7208, 1547-8181. DOI: 10.1177/0018720814547570.

[33] Matthew D. Hoffman and Andrew Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Nov. 2011.

[34] Wan-Lin Hu et al. "Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions". en. In: *IEEE Transactions on Human-Machine Systems* 49.6 (Dec. 2019), pp. 485–497. ISSN: 2168-2291, 2168-2305. DOI: 10.1109/THMS.2018.2874188.

[35] Jie Huang et al. "A Human Decision-Making Behavior Model for Human-Robot Interaction in Multi-Robot Systems". en. In: *IEEE Access* 8 (2020), pp. 197853–197862. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3035348.

[36] Rob J Hyndman and Greorge Athanasopoulos. *Forecasting: principles and practice*. 2nd. Melbourne, Australia: OTexts, 2018.

[37] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. "Foundations for an Empirically Determined Scale of Trust in Automated Systems". en. In: *International Journal of Cognitive Ergonomics* 4.1 (Mar. 2000), pp. 53–71. ISSN: 1088-6362. DOI: 10.1207/S15327566IJCE0401_04.

[38] Xiaochun Jiang et al. "Measurement of human trust in a hybrid inspection system based on signal detection theory measures". en. In: *International Journal of Industrial Ergonomics* 34.5 (Nov. 2004), pp. 407–419. ISSN: 01698141. DOI: 10.1016/j.ergon.2004.05.003.

[39] Joseph G. Johnson and Jerome R. Busemeyer. "Computational Models of Decision Making". en. In: *The Cambridge Handbook of Computational Cognitive Sciences*. Ed. by Ron Sun. 2nd ed. Cambridge University Press, May 2023, pp. 499–526. ISBN: 978-1-108-75561-0. DOI: 10.1017/9781108755610.020.

[40] L.E. Kavraki et al. "Probabilistic roadmaps for path planning in high-dimensional configuration spaces". In: *IEEE Transactions on Robotics and Automation* 12.4 (Aug. 1996), pp. 566–580. ISSN: 1042296X. DOI: 10.1109/70.508439.

[41] J.R. Kintz et al. "(in review) Validation of Embedded Measures to Estimate Operator Trust, Mental Workload, and Situation Awareness for Human-Autonomy Teaming". In: *Human Factors* (2021).

[42] Jacob R. Kintz et al. "Estimation of Subjectively Reported Trust, Mental Workload, and Situation Awareness Using Unobtrusive Measures". en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 65.6 (Sept. 2023), pp. 1142–1160. ISSN: 0018-7208, 1547-8181. DOI: 10.1177/00187208221129371.

[43] Alexandra Kirsch. "A unifying computational model of decision making". en. In: *Cognitive Processing* 20.2 (May 2019), pp. 243–259. ISSN: 1612-4782, 1612-4790. DOI: 10.1007/s10339-019-00904-3.

[44] Spencer C. Kohn et al. "Measurement of Trust in Automation: A Narrative Review and Reference Guide". en. In: *Frontiers in Psychology* 12 (Oct. 2021), p. 604977. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.604977.

[45] Moritz Körber. "Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation". In: *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*. Ed. by Sebastiano Bagnara et al. Vol. 823. Cham: Springer International Publishing, 2019, pp. 13–30. DOI: 10.1007/978-3-319-96074-6_2.

[46] J. D. Lee and K. A. See. "Trust in Automation: Designing for Appropriate Reliance". en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46.1 (Jan. 2004), pp. 50–80. ISSN: 0018-7208. DOI: 10.1518/hfes.46.1.50_30392.

[47] John Lee and Neville Moray. "Trust, control strategies and allocation of function in human-machine systems". en. In: *Ergonomics* 35.10 (Oct. 1992), pp. 1243–1270. ISSN: 0014-0139, 1366-5847. DOI: 10.1080/00140139208967392.

[48] Jundi Liu et al. "Clustering Human Trust Dynamics for Customized Real-Time Prediction". In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. Indianapolis, IN, USA: IEEE, Sept. 2021, pp. 1705–1712. ISBN: 978-1-72819-142-3. DOI: 10.1109/ITSC48978.2021.9565016.

[49] Jessica J. Marquez, Victor Riley, and Paul C. Schutte. "Human automation interaction". en. In: *Space Safety and Human Performance*. Elsevier, 2018, pp. 429–467. ISBN: 978-0-08-101869-9.

[50] Roger C. Mayer, James H. Davis, and F. David Schoorman. "An Integrative Model of Organizational Trust". In: *The Academy of Management Review* 20.3 (July 1995), p. 709. ISSN: 03637425. DOI: 10.2307/258792.

[51] Robin R. Murphy and Debra Schreckenghost. "Survey of metrics for human-robot interaction". en. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Tokyo, Japan: IEEE, Mar. 2013, pp. 197–198. ISBN: 978-1-4673-3100-5. DOI: 10.1109/HRI.2013.6483569.

[52] Kazuo Okamura and Seiji Yamada. "Adaptive trust calibration for human-AI collaboration". en. In: *PLOS ONE* 15.2 (Feb. 2020). Ed. by Chen Lv, e0229132. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0229132.

[53] Kristin E. Oleson et al. "Antecedents of trust in human-robot collaborations". en. In: *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. Miami Beach, FL, USA: IEEE, Feb. 2011, pp. 175–178. ISBN: 978-1-61284-785-6. DOI: 10.1109/COGSIMA.2011.5753439.

[54] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. "A model for types and levels of human interaction with automation". en. In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30.3 (May 2000), pp. 286–297. ISSN: 10834427. DOI: 10.1109/3468.844354.

[55] Raja Parasuraman and Victor Riley. "Humans and Automation: Use, Misuse, Disuse, Abuse". en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39.2 (June 1997), pp. 230–253. ISSN: 0018-7208, 1547-8181. DOI: 10.1518/001872097778543886.

[56] Anand Patil, David Huard, and Christopher J. Fonnesbeck. "PyMC: Bayesian Stochastic Modelling in Python". eng. In: *Journal of Statistical Software* 35.4 (July 2010), pp. 1–81. ISSN: 1548-7660.

[57] Ronald P A Petrick and Mary Ellen Foster. "Planning for Social Interaction in a Robot Bartender Domain". en. In: *International Conference on Automated Planning and Scheduling*. 2013, p. 9.

[58] Amy R. Pritchett, So Young Kim, and Karen M. Feigh. "Modeling Human–Automation Function Allocation". en. In: *Journal of Cognitive Engineering and Decision Making* 8.1 (Mar. 2014), pp. 33–51. ISSN: 1555-3434. DOI: 10.1177/1555343413490944.

[59] Roger Ratcliff and Russ Childers. "Individual differences and fitting methods for the two-choice diffusion model of decision making." en. In: *Decision* 2.4 (Oct. 2015), pp. 237–279. ISSN: 2325-9973, 2325-9965. DOI: 10.1037/dec0000030.

[60] Yosef S. Razin and Karen M. Feigh. *Converging Measures and an Emergent Model: A Meta-Analysis of Human-Automation Trust Questionnaires*. en. Mar. 2023.

[61] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. "Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations". en. In: *IEEE Transactions on Human-Machine Systems* 47.4 (Aug. 2017), pp. 425–436. ISSN: 2168-2291, 2168-2305. DOI: 10.1109/THMS.2017.2648849.

[62] Germán Rodríguez. *Survival Models*. 2010.

[63] Lucero Rodriguez Rodriguez et al. "A review of mathematical models of human trust in automation". en. In: *Frontiers in Neuroergonomics* 4 (June 2023), p. 1171403. ISSN: 2673-6195. DOI: 10.3389/fnrgo.2023.1171403.

[64] A.H. Roscoe and G.A. Ellis. *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use*. en. Tech. rep. Royal Aerospace Establishment, 1990.

[65] Valerio Sanelli et al. "Short-Term Human-Robot Interaction through Conditional Planning and Execution". en. In: *International Conference on Automated Planning and Scheduling*. 2017, p. 9.

[66] Matthias Scheutz, Scott A. DeLoach, and Julie A. Adams. "A Framework for Developing and Using Shared Mental Models in Human-Agent Teams". en. In: *Journal of Cognitive Engineering and Decision Making* 11.3 (Sept. 2017), pp. 203–224. ISSN: 1555-3434. DOI: 10.1177/1555343416682891.

[67] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python". In: Austin, Texas, 2010, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011.

[68] T.B. Sheridan. "Space teleoperation through time delay: review and prognosis". en. In: *IEEE Transactions on Robotics and Automation* 9.5 (Oct. 1993), pp. 592–606. ISSN: 1042296X. DOI: 10.1109/70.258052.

[69] Trey Smith. *ISAAC: An Integrated System for Autonomous and Adaptive Caretaking*. Aug. 2021.

[70] Harold Soh et al. "Multi-Task Trust Transfer for Human-Robot Interaction". en. In: *The International Journal of Robotics Research* 39.2-3 (Mar. 2020), pp. 233–249. ISSN: 0278-3649, 1741-3176. DOI: 10.1177/0278364919866905.

[71] Roger Tourangeau and Kenneth A. Rasinski. "Cognitive processes underlying context effects in attitude measurement." en. In: *Psychological Bulletin* 103.3 (May 1988), pp. 299–314. ISSN: 1939-1455, 0033-2909. DOI: 10.1037/0033-2909.103.3.299.

[72] Alonso Vera. *Risks from Decreasing Ground Support for Human Spaceflight Missions Beyond Low Earth Orbit*. San Antonio, TX, Oct. 2022.

[73] Samuele Vinanzi et al. "Would a robot trust you? Developmental robotics model of trust and theory of mind". en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1771 (Apr. 2019), p. 20180032. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2018.0032.

[74] Christopher D. Wickens et al. *Engineering psychology and human performance*. en. Fourth edition. Boston: Pearson, 2013. ISBN: 978-0-205-02198-7.

[75] Tom Williams et al. "Deconstructed Trustee Theory: Disentangling Trust in Body and Identity in Multi-Robot Distributed Systems". en. In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. Boulder CO USA: ACM, Mar. 2021, pp. 262–271. ISBN: 978-1-4503-8289-2. DOI: 10.1145/3434073.3444644.

[76] Anqi Xu and Gregory Dudek. "OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations". en. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. Portland Oregon USA: ACM, Mar. 2015, pp. 221–228. ISBN: 978-1-4503-2883-8. DOI: 10.1145/2696454.2696492.

[77] X. Jessie Yang, Christopher Schemanske, and Christine Searle. "Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation". en. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 65.5 (Aug. 2023), pp. 862–878. ISSN: 0018-7208, 1547-8181. DOI: 10.1177/00187208211034716.

[78] X. Jessie Yang, Christopher D. Wickens, and Katja Hölttä-Otto. "How users adjust trust in automation: Contrast effect and hindsight bias". en. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60.1 (Sept. 2016), pp. 196–200. ISSN: 2169-5067, 1071-1813. DOI: 10.1177/1541931213601044.

[79] Andy Zeng et al. "Transporter Networks: Rearranging the Visual World for Robotic Manipulation". en. In: *arXiv:2010.14406 [cs]* (Feb. 2021).

Appendix A

Planning with Trust

Improving human autonomy teaming demands more than measuring trust and understanding how humans trust. Future work should incorporate trust into robot action planning systems. If humans use knowledge of a collaborator's trust to plan their actions, a natural extension would be to imbue robot systems with this ability. However, trust serves as a single knob in a complex mix of emotions, task pressure, individual preference, and many more interaction influences.

Representing the state of the environment and collaborating agents has been done using knowledgebase frameworks in previous works [66, 27, 65, 57]. Specifically, in [28], researchers used shared mental models to improve robotic decision-making when working with humans. A shared mental model is a distributed knowledge structure built and queried by several agents. One typically holds facts about the world, such as system states, agent capabilities, or system functions.

A step towards planning with trust will be incorporated into HOME's Capstone 3 demonstration, which uses trust in a robot collaborator as a threshold for task allocation. Given a list of tasks in a spacecraft maintenance scenario and an engineered risk evaluation of robot execution for each task, we can use a symbolic planner to automatically assemble a series of schedules corresponding to different trust states. This design allows the flexibility to observe schedules and balance task performance requirements with risk estimation. The transparency of the schedule design gives the human collaborator increased autonomy in the task and can make more informed decisions.

Appendix B

Testing Human Performance During Time-Delayed Operations

I spent a summer with NASA Ames Intelligent Robotics Group and the Astrobee team. In this experience, we developed an experiment to investigate how human operators would use supervisory interfaces during a time delay. Hopefully, this framework will be implemented to improve our understanding of remote supervised robotic operation.

B.1 Introduction

The Intelligent Robotics Group (IRG) aims to use advanced robotics technologies to explore extreme environments. With specialties in human-robot interaction, this group holds insightful parallels in applying cognitive state modeling to robotic planning. Future scientific missions further from Earth require increasingly capable autonomous systems to combat significant time delays and communication limitations. The Integrated System for Autonomous and Adaptive Caretaking (ISAAC) Project aims to produce technology that integrates robots with spacecraft systems. For occasionally-inhabitated human spacecraft, this means maintaining or preparing hospitable conditions when humans are away.

A current testbed for microgravity autonomous robotics is the Astrobee facility. Astrobee is a free-flying robot aboard the International Space Station that is a mobile sensing and manipulation tool [9]. The robot uses impellers and nozzles to propel in any direction and orientation. The localization system, composed of cameras and inertial measurement

units, maintains an accurate position and heading. This is important to avoid obstacles and develop spatial sensor maps. A specific use case measures and locates noise sources inside the ISS, a potential symptom of a leak or degrading system. Astrobee can operate autonomously from pre-written plans but requires a ground operator to run and supervise activities. These activities are rapidly evolving and often involve new behaviors or methodologies. With marginal communication latency to the ISS, operator take-over and teleoperation are relatively successful, but future operations in lunar orbit may pose more significant challenges with greater latency.

Latency has decreased task performance in direct teleoperation [68]. Previous studies have suggested using "predictive systems," which provide a state estimate to the operator about the immediate effects of their actions, allowing a closed-loop control paradigm [31, 10]. However, the effects of these systems on an operator's cognitive state (trust, workload, situation awareness) are yet to be thoroughly studied. This experiment uses a predictive telemetry system to explore how the Astrobee robot operator team performs under time-delayed conditions. The outcomes will inform the value of developing a predictive system for future operations.

B.2 Methods

This experiment proposal investigates the interaction between a robot and its remote operator under time delay. We hypothesize that human-robot task performance degrades with a greater time delay but improves with an idealized predictive system. Proposed strategies for controlling robotic systems with time delay include arbitration between what an operator commands and the system's current state [20]. This kind of arbitration could be done with a model or digital twin running simultaneously with the controlled robot. Applying this to Astrobee can be done in simulation without the need for a physical system which enables human subject studies. Since Astrobee has been developed using Robot Operating System (ROS) and Gazebo, a physics simulator, an idealized simulation already exists. For experimental purposes, a simulation of Astrobee can be used as the controlled robot, and a noisy, delayed simulation can be used as the digital twin. Figure

B.1 shows how such a system would work outside the laboratory, while figure B.2 shows the noisy simulator for experimental testing.

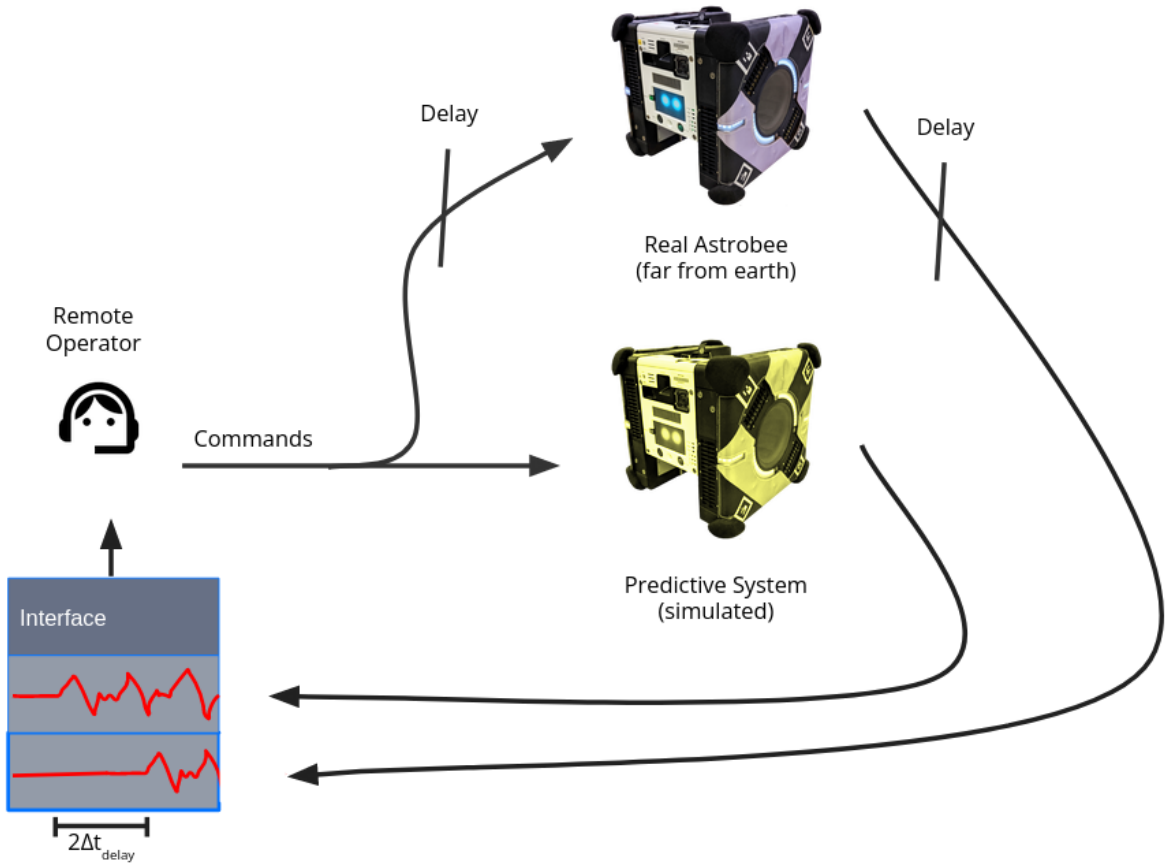


Figure B.1: A proposed digital twin system for a human-robot supervisory interaction experiment. When Intravehicular robots are far from Earth, communication delay prohibits direct control. The system shown here uses a model of Astrobee as a digital twin feeding the estimated real-time state and the delayed state.

The proposed experimental task is a derivative of a previously conducted Astrobee activity. A subject will supervise a cargo relocation task where Astrobee moves equipment between different locations on the ISS. Several anomalies will occur that require subject intervention and test team performance. Time delay, an independent variable, will be simulated by delaying the relay of commands to the Astrobee simulation. The simulated predictive system is the real-time telemetry with various noise models added (for a non-ideal representation). Access to the predictive system is an additional independent variable in the study. The metrics will be compared across independent variables to

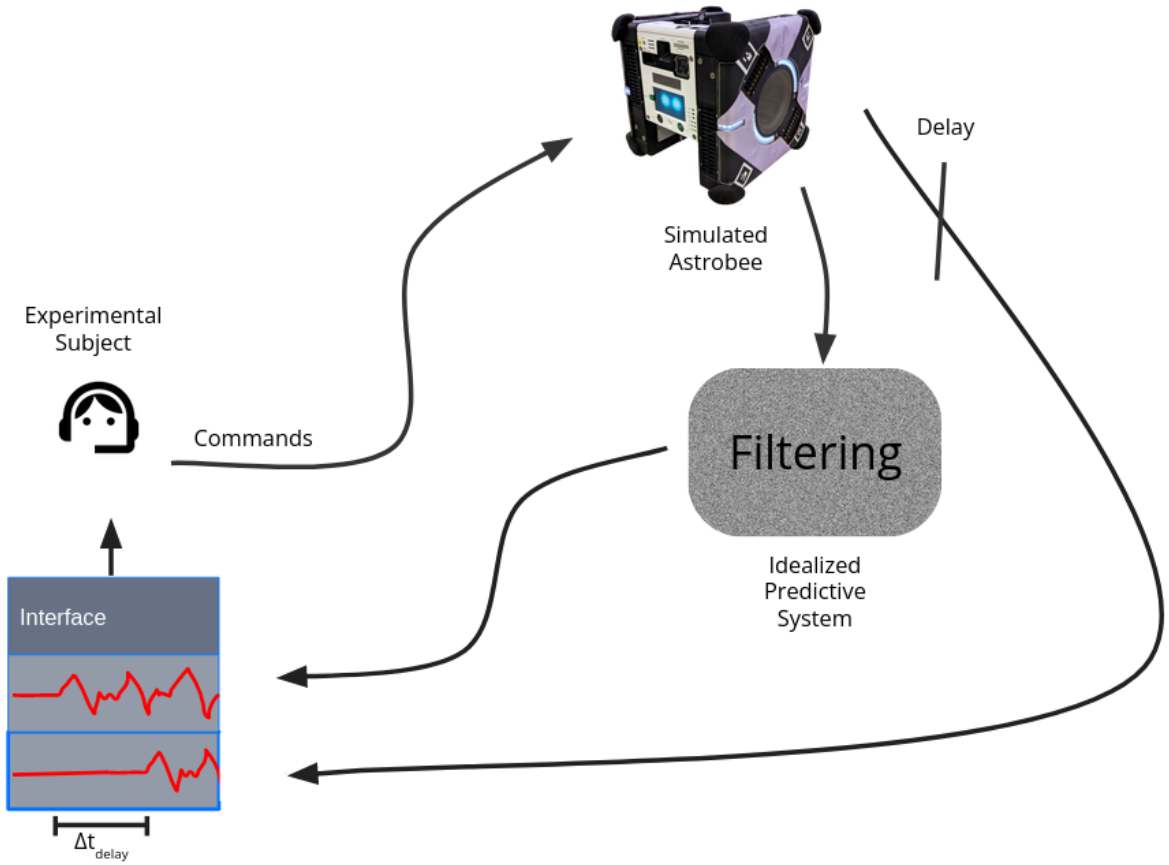


Figure B.2: Human subjects control a simulated Astrobee, and noise is added to the state to represent an imperfect predictive system. Then, the actual Astrobee telemetry is reported on time delay.

determine how predictive systems affect team performance and operator cognitive state.