"GOOD ENOUGH" AGENTS: INVESTIGATING RELIABILITY IMPERFECTIONS IN HUMAN-AI INTERACTIONS ACROSS PARALLEL TASK DOMAINS

BY

SEBASTIAN SAMUEL RODRIGUEZ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Emeritus Alex Kirlik, Chair
Professor Karrie Karahalios
Associate Professor H. Chad Lane
Assistant Professor Jessie Chin
Dr. James Austin Schaffer, Realtor.com

ABSTRACT

Advances in technology have resulted in the development of automation, which has facilitated various difficult tasks by extending the human's capabilities. However, the rapid adoption of automated systems begin to present issues as we continue interacting with technology. Automation is subject to the implicit social contracts we have with other entities in our lives (such as other humans, organizations, and groups), one of them being trust. It is important for us to understand the purpose of automation and its capabilities to set proper expectations into what the system can handle, and set an appropriate amount of trust in them. When we trust automation excessively or insufficiently, it may lead to either misuse (over-trust) or disuse (under-trust) of the automation, which may lead to sub-optimal, harmful, or in the worst of cases, fatal outcomes. The goal for trust within human-AI interactions is to reach calibrated trust.

Prior research has investigated approaches and alternatives to addressing under-trust and over-trust – with under-trust receiving the brunt of the work in order to increase technology adoption of new systems. Over-trust is researched in the context of supervisory control, automation interaction, and human-agent teaming, but it has limited resolution in traditional human-computer interaction research. The most viable approaches are repeated exposure and training of automated systems to prevent users to being lulled into a state of complacency, hampering performance. Addressing over-trust then becomes challenging due to various individual, task, situation, automation, and prior factors that affect the cognitive investment that the human sets in the situation at hand.

This dissertation focuses on designing the reliability of a system to promote calibrated trust, and show this across varied task domains. We inquire whether an agent with less than ideal reliability can promote better calibrated trust by presenting itself as imperfect, much alike human-human interactions where skills and capabilities are assessed and calibrated. Furthermore, we present how this manipulation in reliability can affect AI systems in multiple domains, as to demonstrate that the trust dynamics between humans and AI is not only

restricted to agents that live behind a screen (e.g., automation support, machine learning models, recommender systems), but also in physical and tangible systems much like we see in robotics today (e.g., drone swarms, robotic assembly). The approach of this dissertation is divided into 3 studies.

Recommender systems are a type of decision support system used to provide personalized recommendations to users, and are often the archetype of human-AI interactions (for instance, the plethora of applications that recommend content to us in our smartphones). We delve into recommender systems and compare how features commonly used in decision support system design (i.e., explanations, control settings, reliability) can affect the acquisition of domain knowledge. We discuss 2 sub-studies (n = 526 and n = 529) with a recommendation system each, where we vary the presence of explanation, amount of control over the system, and reliability (i.e., quality of recommendations). We find that features often used to increase trust (e.g., explanation of outputs, control over the system) can lead to over-trust, which is mitigated by a lowered reliability to allow humans to exercise their own judgment.

Since recommender systems are not the only type of AI systems we can interact with, we next focus on a physical domain where collaboration can be tangible (such as humans and robots). We investigate a simulated physical task with a pursuit-style objective, where the human is tasked to collaborate with 2 AI agents to capture a singular moving target. In this study (n = 104), we manipulate the reliability of the agent teammates, and measure both individual differences, perceptions of the agents, and task outcomes. Using mediation modeling, we demonstrate how reliability and performance is mediated by trust, situation awareness, and user individual differences. We additionally show how reducing reliability can have interaction effects with the domain and the environment, sometimes presenting unintended benefits.

Finally, we explore the simulated physical domain of human-robot interaction in a collaborative decision-making task. We control reliability in a signal detection theory-based task in with distinct robot representations to explore human perception of reliability thresholds, and how robot embodiment affects decision-making. In this study (n = 119), we ascertain that embodied interactions point to higher perceived workload and self-reported trust, and a lower reliability can facilitate trust calibration by allowing users to recognize multiple

erroneous cues.

The findings in this dissertation contribute to the general knowledge in trust calibration, reliability, and human-AI interaction across virtual and physical domains, which serves for engineers and designers to be cognizant of these effects to build AI systems that are able to cue their users on how to improve the amount of trust that should be allocated. This process then may become more akin to how humans calibrate their trust with other humans, a small step towards improved human-AI integration.

To my family, this doctorate is as mine as it is yours.

A mi familia, este doctorado es tan mio como es de ustedes.

ACKNOWLEDGMENTS

This dissertation is the largest endeavor I have taken up to this point in my life, and its completion is only due to the unending support of many important people who I have had the privilege to interact with for the past several years.

First and foremost, I thank my advisor, Professor Emeritus Alex Kirlik, for guiding this journey. The academic wisdom that you exude in each of our conversations (from day one) is something I can aspire to reach for the rest of my career. I cannot thank you enough for taking me under your wing, supporting me morally and academically, and pushing me to be a little bit more curious every day.

To Dr. James Schaffer, I am continually grateful for both your mentorship and friendship. From late night weightlifting advice, statistical modeling discussions, to playing Nioh 2 and speedrunning Monster Hunter World, your presence in my doctoral training has been pivotal to my growth and success.

To my dissertation committee: Professor Karrie Karahalios, Professor Chad Lane, Professor Jessie Chin; thank you for all the suggestions, notes, feedback, and conversations throughout the formulation and development of this body of work throughout the past 6 years. The individual expertise of each of you made this dissertation unique, interesting, and fun to research. I complete my dissertation knowing that I still have much to learn, and the breadth knowledge each of you have motivate me to continue honing my skills as a researcher.

To the many people I have met and worked with within Interactive Computing research area at the University of Illinois: Emily Hastings, Sneha Krishna Kumaran, Helen Wauck, Kristen Vaccaro, Patrick Crain, Dennis Wang, Cameron Merrill, Wayne Wu, Grace Yen, India Owens, Gina Do, Ziang Xiao, Silas Hsu, Joon Park, Tiffany Li, Vinay Koshy, Rizky Wellyanto, Ken Cheng, Stephanie Lin, Charlotte Yoder, and Rick Barber; along with non-IC computer science friends: Angello Astorga, Justin Szaday, Daniel McKee, Jacob Laurel, Liia Butler, and Wing Lam; thank you all for the conversations, support, games, and laughs! My

friendship will always extend beyond our time in graduate school.

To computer science faculty and staff: Professor Carl Evans, Professor Brad Solomon, Dr. Thierry Ramais, Professor Brian Woodward, Professor Brian Bailey, and Viveka Kudaligma. Carl, Brad, and Thierry, thank you for the teaching experience at CS 225, which quickly became one of my favorite experiences during graduate school. I have always loved teaching, and doing it at a larger scale reinvigorated me to continue pursuing mentoring beyond my academic responsibilities. Professor Woodward (and Ziang Xiao), thank you for the opportunity to fully leverage my engineering skills for a project that will benefit the educational development of the next generation of engineers. Brian and Viveka, thank you for your support during rough academic times, and for being the logistic powerhouse behind graduate computer science.

To the undergraduate students I mentored during my time: Harsh Deep, Jacqueline Chen, Jaewook Lee, and Drshika Asher; thank you for putting your faith in me as your mentor, and I hope you all leave the University of Illinois with a hearty experience of research. I am always a Discord server away, and I hope to see you there until we all retire. I am very proud of all your successes, and am honored to be a small part of your educational career.

To the Sloan University Center of Exemplary Mentoring at Illinois and Dr. Ellen Wang Althaus; thank you for providing a space where a first-generation doctoral student could find solidarity in a journey so unknown to me. Thank you for the funding, the guidance, the resources, the bonding, and the space provided to us Sloan scholars.

To even more people who kept my mind fresh by filling my life with dancing: Yoel Cortes-Peña, Saowaluck Khaophuan, Mishel Melendez, Bernardo Burbano, Katherine Bokenkamp, Brando Miranda, Sierra Schreiber, Xuejin Zhang, Alessandro Contento, Katherine Jo, Elizabeth Villegas, Accalia Boyle, Sylvia Kim, Lyan Padilla Velez, Kassie Miner, Pamela Rose, Krystal Montesdeoca, Amalia Marifatul, Santiago Aurelio, Jennifer Oesterling, Zhiyuan Lin, Daniela Gomez, Kat Kolumban, Rony Die, Lana Šteković, Yuri Sohn, Beatriz Perez, Nate Craig, Lorraine Lopez, Juan Perez, and Gene Santiago; thank you for all the fun evenings dancing, from Cowboy Monkey to Soma to Neil St. Blues. Salsa has been in my heart since childhood, and I am very grateful for all the time we have spent together listening to the clave and dancing the night away.

To my longstanding friends from Northwestern University: Michael Horst, Katrina Werner, Edgar Vazquez, Federico Paredes Garza, Daniel San Gabino, Carrie Willis, and Asher Rieck; thank you for your continued friendship and support as I continued to a graduate program. From both visiting you to having you visit, thank you for the connection to everything outside my academic world.

To my undergraduate research advisors at Northwestern University: Professor Corey Brady and Professor Michael Horn; Mike, thank you for sparking my interest in human-computer interaction one winter morning in 2015 – which culminated in this doctorate degree 7 years later. It was great to see you during your campus visit, and for the words of encouragement as I settled into being a graduate researcher. Corey, thank you for the research experience at the Center for Connected Learning, and for your guidance and support at every step during my application to graduate school.

To my mentors and collaborators at the U.S. Army Research Laboratory: Dr. Derrik Asher, Dr. Erin Zaroukian, Rolando Fernandez, Mark Mittrick, Michael Garber-Barron, Dr. Sean Barton, Justine Caylor, Jefferson Hoye, and Dr. Timothy Hanratty; thank you for making my stay at Aberdeen one of the most enlightening and hard-working experiences during my doctoral journey. On the western counterpart: Dr. James Schaffer, Johnathan Mell, Sara Rojas, and Parth Patel; thank you for providing such a great stay in Playa Vista and a well-needed escape from the midwestern weather.

To my mentors and friends during my internship at Meta: Daniel Gruner, Mary McCuistion, Amy Chong, Max Curran, Rachel Cultice, the entire Messenger team, and everyone who I interacted with during my internship; thank you for a great first industry experience and providing the perspective on how research is conducted in a non-academic setting.

To my collaborators at the EPIC group at Microsoft Research: Dr. Eyal Ofek, Dr. Payod Panda, and Raahul Natarrajan; thank you for the opportunity to contribute impactful work in the virtual reality space. This was a needed refresher from my research and had lots of joy from working in an engineering role.

To my friends and loved ones south of the Equator, in Peru: Lyon's Montoya Cuadrao, Magno Suarez Vives, Alejandro Gomez-Sanchez, Shadia Paz Majluf, Romina Bustamante Ventura, Katherine Ramos Lapa, and Javier Tovar Jaeger, among others and their families;

even at 6000 kilometers, you have always been a phone call away to provide support in my busiest days. With my doctorate complete, I look forward to planning more trips to Peru and spending more time with the friends I grew up with.

To Ximin Piao, for your support and motivation throughout this journey. Thank you for sharing your time even during our busiest days, supporting me through the tough times and celebrating with me during the good times.

And finally, to the never-ending support from my family: my mother Alyna, my father Juan, my brothers Juan Francisco and Juan Manuel, and Tobie and Trinity (may she rest in peace); thank you for taking care of me, indulging me, and pushing me to complete the final stretch of my education. Mami, thank you for always having my back in anything I needed, especially when I did not make time for you in my busiest days. Viejito, thank you for always imparting wisdom from the university of life as I navigated my adulthood, and for always sharing a laugh no matter the situation. Juancito, we both shared the start of our journey here in Urbana as you began your surgical residency and I began my doctoral studies. Residency sure kept you busy, but I cherished the days off when we would catch up on our medical dramas or play Resident Evil for hours on end. Manny, thanks for always lending an ear to chat or some time to spend, even during your busy time at Army JAG. To my pets, Tobie and Trinity (who are unable to read but are acknowledged), thank you for the unconditional love you give every day I return home, and all the *mlems* you give when asking for a treat.

To all of you, thank you.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The advancement of artificial intelligence and machine learning is pushing beyond traditional automation, enabling a greater level of intelligence that includes the ability to perceive information about the world and determine the appropriate action to achieve a desired outcome through their own computing procedures. [1]. Automation is utilized in many areas of human life – a few examples from the wide variety of applications include supporting our decision-making processes [2], recommending entertainment content [3], planning our commuting routes [4], reducing aviation crews for safety [5], and supporting warfighters in high-risk environments [6]. The benefits of automation and their perceived usefulness have resulted in the quick adoption of these systems in our day to day lives [7].

Throughout the advancement of the research, automation has gained a wide variety of synonyms and alternative nomenclature, such as the terms *autonomy*, *agents*, and *artificial intelligence* (AI). Notwithstanding the differing definitions, we focus on AI agents that are "designed to accomplish a specific set of largely deterministic steps... in order to achieve one of an envisaged and limited set of pre-defined outcomes." [8] As simplistic as this definition appears to be, these systems range from standard assembly line machinery to deployed machine learning models that aim to predict an outcome from patterns learned from training. These agents can function heuristically ("if this happens, I should do this") or learn new rules to adapt their behavior. Some agents exist as a physical entity to complete a task (e.g., a robotic arm lifting heavy equipment where a human would be incapable), while other automation have been virtualized to operate within computerized ecosystems (e.g., shopping recommendations on Amazon, or movie suggestions on Netflix). Furthermore, these are the systems we interact with on a daily basis, as we have yet to synthesize true computational autonomy (i.e., "systems that learn... and permanently change their functional capacities" [8]). The ubiquitous nature of these agents have resulted in humans and agents collaborating together to complete a task, often defined as Human-AI teams.

With this, however, there are many unknowns in how human-AI teams complete these tasks. Algorithmic advances have contributed to building automation that is perceived as correct, useful, and trustworthy. However, these technological advances are often accompa-

nied by drawbacks [9]. Human factors research has built the cast that automated systems can be used inappropriately through misuse and disuse [10] – often due to humans perceiving automation to be "too good" or "too bad" – resulting in humans being lulled into a sense of complacency where they believe their involvement is unnecessary to effectively complete a task, or often flat out ignoring the automation [11]. When tasks involve a certain level of risk, human disengagement can lead to critical or fatal outcomes.

It is important for humans interacting with agents to understand its purpose and capabilities in order to set proper expectations to what the system can handle versus what the human can handle [12]. To make sure this "calibration" is done correctly, one needs to adjust their expectation for the machine. This process is similar to evaluating a new employee or team member, where you may have some uncertainty about their skills despite having looked over their resume quickly [13]. Research shows that humans are not too adept at properly calibrating their trust towards agents; an agent that is perceived capable or high performing leads to instances of human over-trust, where it results in disengagement from the task [14]. Researchers have been able to identify models that take into account user characteristics, task representations, and agent features, and is able to theoretically predict scenarios where over-trust and complacent behavior could occur [14]. In this model, highly reliable systems are a strong predictor of over-trusting behavior; this effect is consistently found in recent human-AI interaction studies (e.g., [15, 16, 17, 18]), and even in deployed systems used by millions of users today (e.g., interpretability tools [19, 20] and semi-autonomous vehicles [21, 22]). This over-trusting behavior can in turn affect the development of proper decision-making, situation awareness, and knowledge – tenets of the day-to-day skill sets we use to interact with other humans or automated systems.

Human-AI teams operate in a wide variety of domains where its tasks are highly dependent on its requirements, constraints, and stakeholders. To support these tasks, agents are able to take roles with different level of involvement in the task. For instance, agents can take the form of recommender systems, where the system serves to enhance the decision-making processes of the human – yet, the human retains the final say in the outcome. Alternatively, agents can have a higher level of autonomy where they operate mostly without human intervention. These cases go beyond supervisory control, where the agents hold an active

stake in the task and completion would be very challenging (or even impossible) without their involvement. The conglomerate of human-technology research has explored a gamut of domains to investigate agent design such as performance [23], transparency [24], explanations [25, 26], control settings [27], level of automation [28], anthropomorphism [29, 30], failures [31, 32], presence [33, 34], among others. The amount of different scenarios that could be studied given a certain set of agent features becomes a very large space, thus we spend most of our time investigating a generalizable subset of tasks. However, gaps still remain on how the effect of these features vary per distinct task domains. For instance, one could envision that including anthropomorphism in a movie recommender system (e.g., Netflix) can evoke a distinct feeling or perception than anthropomorphism in a military drone. The type of task or domain often has high interaction with the observed effects in research, thus design elements and outcomes should not be observed in a vacuum [35].

This dissertation investigates how varying agent reliability at different levels affects human performance and trust in different domains using game-based tasks – to ideate how reliability can be framed as a requirement for effective human-AI teaming in a variety of distinct tasks. This work focuses on three distinct domains that could use expansion in the human-AI teaming literature with respect to reliability. We also investigate how the benefits of sub-optimality can vary depending on the scenario presented to the human-AI team. Research in AI development focuses on optimizing algorithm accuracy and speed, yet a threshold of minimum returns needs to be established to avoid endlessly pursuing the perfect agent. When deploying AI systems, the environment they operate it is seldom perfect and often subject to large amounts of noise (both in magnitude and variance). Leading to interaction failures, this then necessitates revisiting how to fit the human into the algorithm instead of the opposite, thus breaking a fundamental tenet of human-centered design. This central thrust aims to show the alternative to having perfect automated systems at all times, and "Good Enough" agents may serve for equal or better efficacy for both technical development and team composition in Human-AI teams. With this work, we open three distinct avenues for future work centered in investigating how to leverage unconventional (i.e., sub-optimal) agent reliability to guide AI teammates to be perceived as more effective.

In Chapter 3, we investigate how the presence of explanations and control can lead to

the emergence of over-trust affecting domain knowledge learning in an influence task with recommender systems (i.e., a decision support agent). This over-trust was mitigated by reducing the reliability of the agent, yet reducing reliability without explanations or control leads to worsened learning. From these findings, Chapter 4 presents a study with simulated physical agents in a continuous pursuit task and examine how a slight reduction in agent reliability is imperceptible to the human (through situation awareness and trust) but results in increased team performance due to task interactions. Additionally, we present a structural equation model accounting for a variety of effects studied in a vacuum in the human factors literature, and integrate human characteristics, agent features, and outcomes through mediation analysis. In Chapter 5, we examine a physical task and aim to investigate how agent embodiment and reliability affects human performance and trust in an uncertain decision-making task. With this work, we demonstrate how small changes in reliability can have a beneficial effect in human-AI teams in a wide variety of domains, and entertain the idea that for integrating humans and AI together, "Good Enough" agents could be better teammates than perfect agents.

1.1 DISSERTATION INQUIRIES

As mentioned, this dissertation investigates the effect of manipulating reliability – specifically, lowering it – on human trust calibration and performance in a variety of different task domains. Of additional interest is to observe how human individual differences and distinct task requirements affect how trust is formed and maintained, and how trust guides decision-making and outcomes. This leads to the following dissertation inquiries that guide the approach in this research:

1. How does lowering agent reliability affect task performance and human trust across different domains?

2. What is the relationship between human individual differences, agent reliability, and technology use?

3. How do different task requirements (e.g., task type, agent physicality, task urgency) interact with the human-AI interaction dynamic?

1.2 CONTRIBUTIONS

This research presents 4 major contributions that aim to inform the design of AI systems for improved calibrated trust, and thus, outcomes. All 3 studies have uncovered new and reinforced existing relationships between constructs that expand our knowledge on how humans trust automation in the presence of different reliabilities. The contributions are:

1. **Empirical results that demonstrate positive outcomes of reducing agent reliability in a variety of different contexts.** Throughout the dissertation studies, we find that slightly reducing reliability presents positive outcomes in learning (Chapter 3), knowledge retention (Chapter 3), situation awareness (Chapter 4), agent performance (Chapter 4), and decision-making (Chapter 5). While this is a surprising result, we should be cautious about its generalization towards the development of systems and is *not* definitive proof that lowered reliability causes positive outcomes. The integration and application of lowered reliability should be considered on a case-by-case basis – with specific consideration to the type of task and the risk tolerance defined by users. Instead, the effects found in this dissertation support the argument that reliability is not the only end-goal in system design, moreover, further experimentation and iteration is required to ascertain the benefits that imperfect reliability could bring to human-AI interaction.

2. **A holistic mediation model for human-AI interaction that connects reliability, individual differences, performance, and trust.** There exists a wide variety of research that have shown the relationships between all these variables in pairwise or co-variate comparisons, yet, the research lacks holistic understanding of the tradeoffs between these variables. A holistic understanding allows us to connect the wide variety of tested variables in the past decades of research. We address this gap by making extensive use of structural equation modeling (including path analy-

sis, mediation modeling, and growth modeling) to demonstrate how these variables, often studied in a vacuum, relate to each other from inception to outcome. Chapter 3 makes use of growth modeling to show the rate of change in knowledge (akin to velocity in a physics-based example), while Chapter 4 and 5 use pathway models to relate the reliability and performance through individual differences, task features, and human perception. Overall, we find that reducing reliability is not the direct cause for improved outcomes, but often it interacts with individual differences and levels of trust, which only strengthens the idea that these variables should be studied in a more holistic manner to gain deeper insight on how the human cognition works with respect to AI interactions, rather than studying these constructs in a vacuum.

3. **Further development of game-based abstract scenarios to model tasks within a given domain with appropriate ecological validity.** All studies in this dissertation make use of an interactive, goal-based scenario to incentivize participants to interact and make use of the automation. The Diner's Dilemma (Chapter 3), Predator-Prey game (Chapter 4), and the *Warehouse* game (Chapter 5) were designed with intrinsic motivation and competition in mind. Research participants sought the best way to maximize their outcomes by using the automation in distinct ways they believed was best (yet not necessarily the most optimal outcome, as evidence by over-trust in these and other game-based scenarios in the literature). This extends the idea pointing to using abstract games to capture research constructs, where they may model and be as valid as the real application of said automation [36]. Within these studies, we found replications and validations of the theories in over-trust and automation bias and complacency, which are researched in real-world applications, contributing to overall ecological validity of these studies [37]. All games are explained in detail for implementation and replication purposes, and the source code for *Warehouse* game is open-sourced and easily available online.

4. **A more thorough examination of the defining characteristics that underlie the effective generalization of research on reliability.** Of the surveyed literature, we find that attempts at exploring the effects of reliability on performance varied

depending on task and agent characteristics. We find variation on how reliability is perceived depending on the type of task (Chapter 3, 4, and 5), the representation and embodiment of the agent (Chapter 5), how the agent is invoked (Chapter 3), and the presence of time pressure during the task (Chapter 4 and 5). Although we find some central and replicated themes from previous reliability research (e.g., users perform well with highly reliable automation, users need longitudinal exposure for proper calibration), more specific metrics – such as trust, decision-making, situation awareness, knowledge, among others – are affected by domain and agent features, making the generalizability of reliability studies difficult towards scenarios outside any current study. In order to establish what aspects of a study can be generalized, it is crucial to identify which features have consistent outcomes across the research landscape.

In Chapter 2, we present a comprehensive background and review of definitions and prior research covering the interplay between performance and trust, uses – and misuses – of automated systems, the different representations an AI system can have, and the different domains and scenarios Human Factors and Human-Computer Interaction research have covered. The core studies introduced are presented in Chapters 3, 4, and 5. Finally, we conclude with a cross-study analysis and broader implications the results of this research brings forth in Chapter 6.

CHAPTER 2: BACKGROUND AND RELATED WORK

In order to establish the theory and research landscape behind the idea of agent reliability, we cover relevant background work from human factors, human-computer interaction, human-robot interaction, and game theory. We review 4 main concepts in this line of research. The following section gives an overview on how human factors has established human performance and its interplay with trust, and how automation begins to play a role in improving said performance. Next, we review a seminal concept in human factors research: which has a direct impact on the performance and trust exhibited by humans. Subsequently, we discuss different agent representations and how features can affect human perception. We then conclude with a comparative analysis of the different types of tasks automation can operate in and how research has addressed tasks in distinct domains.

2.1 RELIABILITY, TRUST, AND RELATED FACTORS

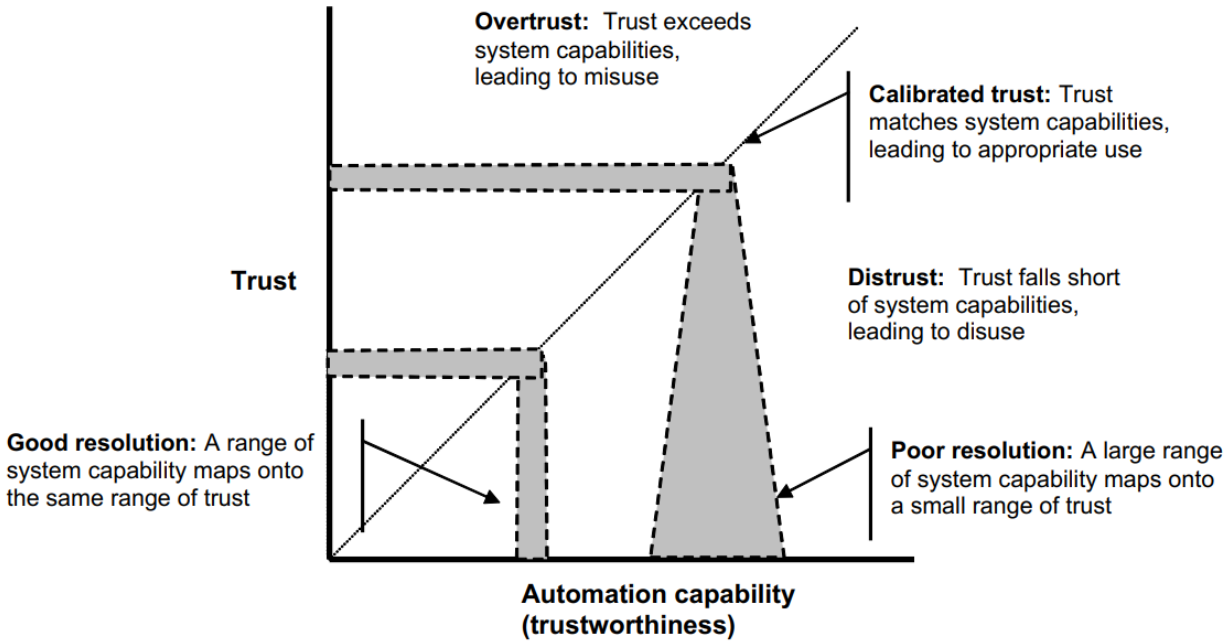


Figure 2.1: The relationship between calibration, automation capability, and resolution of trust in automated systems, adapted from Lee and See [38].

Research has identified that trust (and its associated attributions: predictability, depend-

ability, and faith) affects relationships between romantic partners [39], hierarchical roles [40], organizational entities [41, 42], commitments [43], and now, technology and automation [38, 44]. According to Lee and See, trust serves to accommodate the cognitive complexity that appears with uncertainty, as our society moves towards more complicated and highly structured organizations and technology [38]. With the introduction of automated systems into our daily lives, we follow the established societal paradigm: just as we estimate the amount of trust we should allocate to another person, we also estimate (i.e., calibrate) the amount of trust given to an automated system [45]. The amount of trust given should be proportional with the capability and reliability of the system. In the perfect scenario, a user is able to calibrate their trust according to the capability of the automated system, resulting in a one-to-one relationship; the user trusts the system enough to take advantage of its benefits, yet, they are aware of its limitations [12]. However, in adverse scenarios (e.g., a novice worker interacting with a system in a domain they have no experience with), the amount of trust allocated to the system does not match its potential, leading to either under-trust or over-trust. The consequences of under-trust are not too severe, as it leads to inefficiency due to reduced use and trust of automated systems, resulting in the user's increased cognitive effort when the automation's employment would have resulted in equal performance [46]. Increased cognitive effort may stem from either the user opting to complete the task without the use of automation, or if the automation is required, the user will suffer a higher attentional load due to monitoring and vigilance at excessive levels relative to the system's capability, which further varies with the level of involvement the system has in the task [28]. For instance, a driver who uses a route planning application (e.g., Google Maps) encounters a detour in their usual commute, which their application does not reflect. The driver begins to drive unassisted, as he no longer trusts the directions their application provide. The driver then incurs a cognitive load by recognizing and learning alternate routes to their destination as the usage of the application reduces due to lack of trust.

On the other hand, over-trust leads to potentially severe negative consequences if the system is less reliable than the trust it is warranted [10, 45]. Such issues can be succinctly summarized by "out-of-the-loop" behavior, as the user begins to lose perception of the system state and associated information due to loss of attention and sub-optimal monitoring of the

system. Thus, due to the inevitable unreliability (i.e., agents will never be perfect and are subject to error, even if in a singular case), users are sometimes required to re-enter the control loop, often unexpectedly, which results in an ineffective assumption of control. Such cases are often the result of the opaque nature of systems, where small errors are often subtly compensated by the automation, with no acknowledgement from the user. It is only until a dramatic failure occurs that the user is suddenly plunged into rectifying the issue, often with high cost to cognition, often dubbed the "automation surprise" or "return-to-manual-control-deficit" [47, 48, 49].

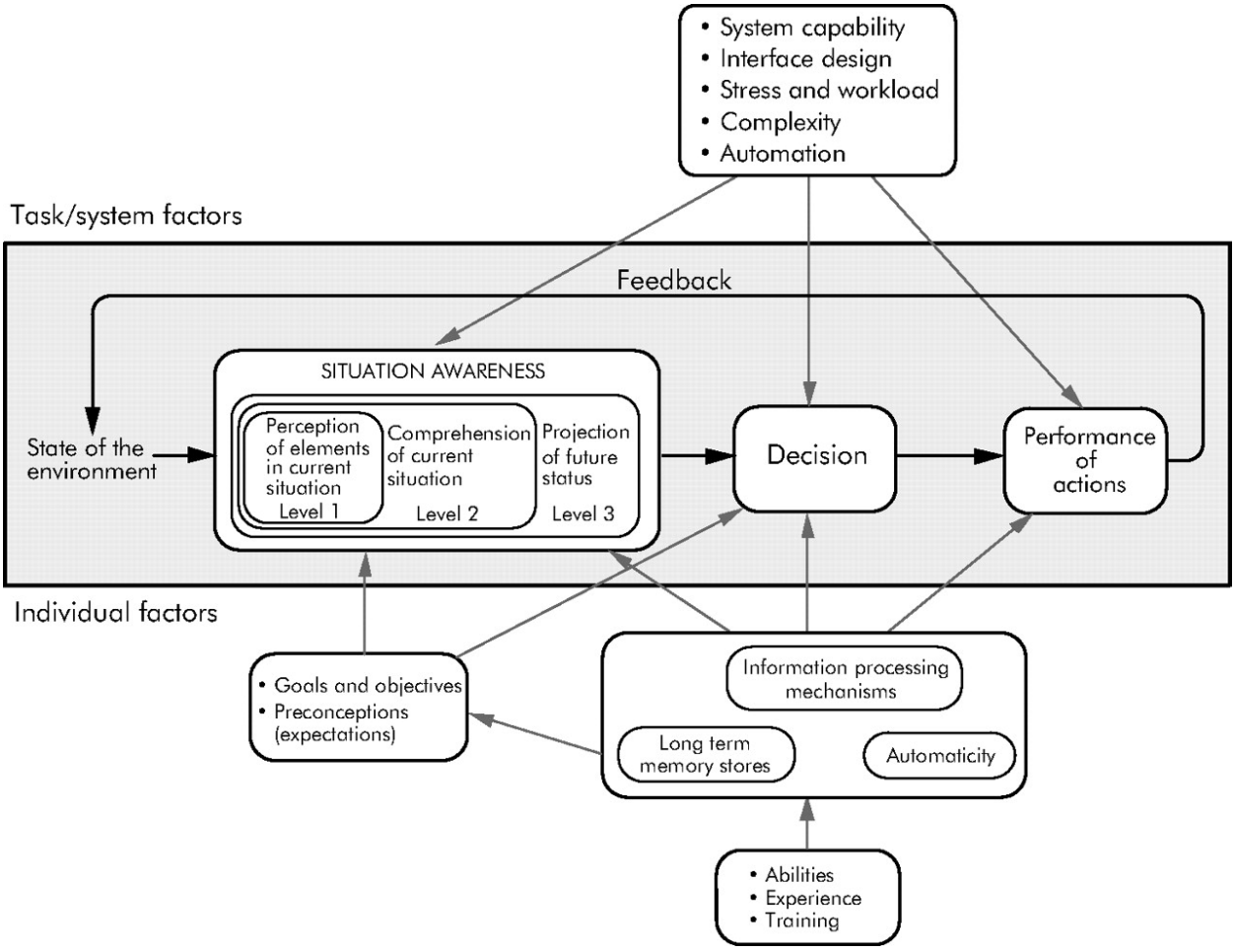


Figure 2.2: Endsley's situation awareness model, adapted from Endsley [50].

Situation awareness (sometimes referred to as situational awareness) is an important element in describing a human's effectiveness in a dynamic task. Situation awareness is defined as the perception of environmental elements and events with respect to time or space, the

comprehension of their meaning, and the projection of their future status. A human's situation awareness is constantly affected by personal, environmental, and task-related factors, and ultimately models the human's decision-making process and downstream consequences [51]. Prior research has shown that situation awareness is affected by increasing the level of automation and reliability, where it is most affected when the system has full control over the task, with no human intervention [52]. Situation awareness results in being very context-specific, relating to the current environment it is being employed in (i.e., prior knowledge or training in a different domain does not transfer over to a novel domain). When situation awareness is incomplete (i.e., it fails to properly project due to insufficient information at the lower levels), performance usually suffers. As mentioned, detriment of performance occurs in a high-risk scenario can lead to catastrophic consequences (e.g., the manufacturing and operation of Boeing 737 MAXs [53]). Situation awareness largely operates on a cyclic model that updates the operator's perception and understanding of the world as it interacts with it (see Figure 2.2). Chen et al. extended this framework towards understanding a particular intelligent agent or entity of automation [54]. In this case, automation can be designed such that a user can employ the situation awareness model to better understand how the automation is operating (such as understanding the automation's state of perception, comprehension, and projection – and reconcile this information with the user's own mental model), as opposed to Endsley's "global" situation awareness model, where the user perceives the entire environment (which may or may not include interactions with automation). Ultimately, situation awareness plays a role in trust calibration, as loss in situation awareness may lead to over-reliance on the system as the human attempts to reconcile their knowledge of the environment.

The development of trust and the resulting outcomes are additionally affected by prior experiences and personal factors, earmarked in the literature as "individual differences." Research notes that the level of trust calibration can be defined by a combination of these factors. Prinzel et al. found that perceived workload and self-efficacy play a significant role in trust calibration. Specifically, they noted that perceived workload can influence trust by either increasing or decreasing it, depending on the task demands and cognitive load of the operator [55]. Similarly, Sheridan et al. noted that self-efficacy, or the belief in one's

own ability to accomplish a task, can impact trust in automation [56]. They found that higher self-efficacy was associated with more positive attitudes towards automation, while lower self-efficacy was linked to greater reliance on automation. Research has also found that personality traits can impact trust in automation. Lyons et al. found that individuals with higher levels of openness, agreeableness, and conscientiousness were more likely to trust automation, while those with higher levels of neuroticism were less likely to do so [57]. Additionally, Esterwood and colleagues found that individuals' levels of trust in automation were affected by their perceived risk and familiarity with the technology, as well as their beliefs about the level of control they had over the automation [58]. Sanchez et al. found that age can also play a role in trust calibration, with younger individuals tending to exhibit greater trust in automation than older individuals [59]. However, they noted that this effect is likely to be complex, and may be influenced by a range of other factors such as experience with technology and situational factors.

A meta-analysis by Schaefer et al. has identified these individual differences to often drive mid-task outcomes such as trust propensity, attentional control, satisfaction, role interdependence, and perceived risk [35]. Ultimately, these individual differences affect how humans strategize their use of automation (e.g., the amount of attentional control one decides to use affects the level of trust given to the automation [60]), leading to distinct behavioral outcomes even when variables of interest are controlled for. As mentioned in the introduction (Chapter 1), this dissertation makes use of mediation modeling in order to estimate which individual differences affect other individual differences or behavioral outcomes. We aim to model a significant proportion of the individual differences that are discussed in the core studies of this dissertation. However, it may not be sufficient for a full understanding of how these differences influence trust calibration. In the core studies, each individual difference that is explored is elucidated further in the context of the relevant task. We defer those explanations until future chapters.

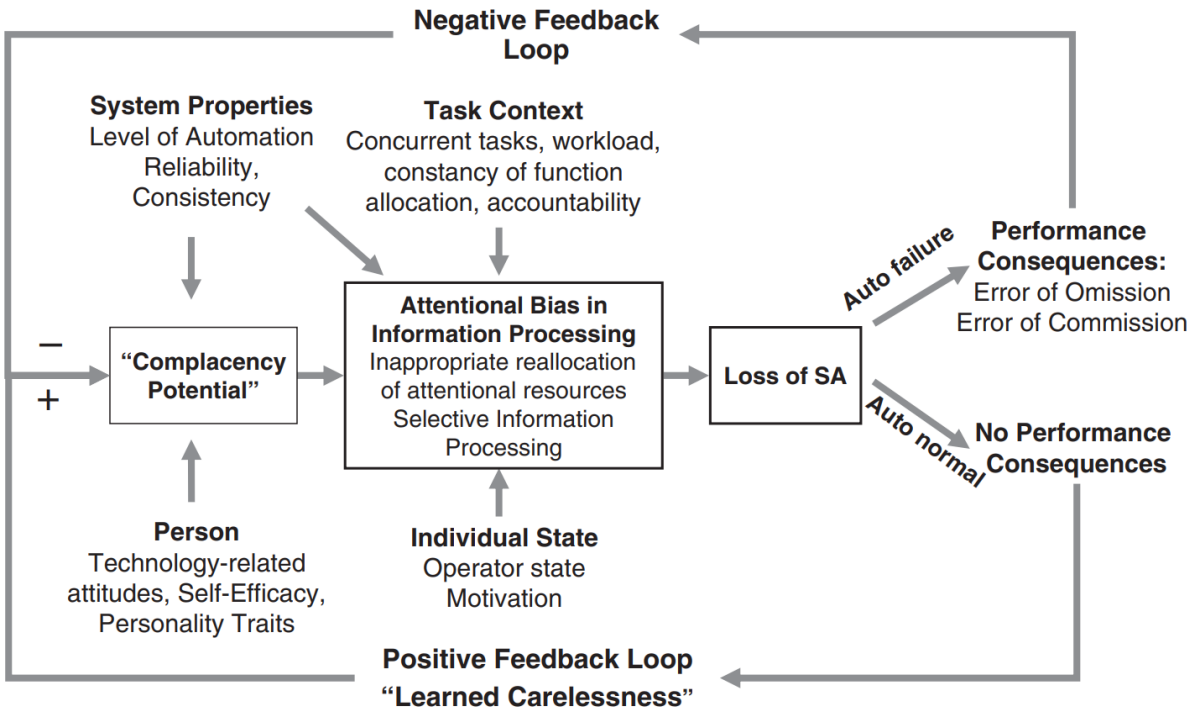


Figure 2.3: An integrated model of complacency and automation bias, adapted from Parasuraman and Manzey [14].

2.2 IMPROPER USE OF AUTOMATION

Technology always has the possibility of failure. Unexpected events, improper use, and system faults all plague even the most robust systems deployed across the globe. Researchers and engineers aim to build fail-proof systems that will minimize being affected by errors, and to recover as quickly as possible if it succumbs to failure. However, Parasuraman and Riley note that several studies show that automation failures do not prevent users from using the same automation again in the future, even after repeated and catastrophic failures in simulated scenarios [10, 61].

Complacency has been variably defined by the research literature since its inception, but some of the first definitions include being "a psychological state characterized by a low index of suspicion" [62], and a "self-satisfaction which may result in non-vigilance based on an unjustified assumption of satisfactory system state" [63]. These definitions initially stemmed from the aviation research community to describe sub-optimal performance of aircraft crews, which has been long implicated as a possible contributing factor in aviation

accidents [61, 62, 64]. Pilots were strongly recommended to "trust their instruments", as safety inside the cockpit increased with automated instruments. Unfortunately, this led pilots trained in this paradigm to highly believe in the reliability of their instruments and over-trust, resulting in biased and complacent behavior [65]. As the potential to automate directives and tasks came with the advancement of technology, complacency research became more prominent as it began to describe unwanted behavior in interactions between humans and systems [61], additionally appended with the prefix "automation-induced". There is a distinction to note between *automation bias* and *automation-induced complacency*: automation bias refers to erroneously trusting the automation's decision more than our own judgment, regardless of experience; whereas complacency refers to sub-optimal monitoring of the automation's decisions and actions [10]. As the definition states, the assumption of a satisfactory system state due to low suspicion leads to hampered performance, resulting in potential tragic outcomes in high-risk scenarios [10]. However, as automation evolves and moves beyond automated systems, the prefix "automation-induced" no longer describes many modern systems, as complacency remains prevalent in semi-autonomous and fully autonomous systems [66].

Biased behavior caused by over-trust was initially described as specific responses to binary cues: either compliance or reliance [67]. Compliance refers to the behavior of responding to a cue, whereas reliance describes avoiding taking action whenever there are no cues. Wiczorek et al. established that these behaviors are loosely related with respect to responding to a cue [68]. With the inclusion of cues, participants tend to comply much more often with the automated system if it is highly reliable. If we frame these responses in terms of trust miscalibration, compliance would relate to automation bias (i.e., "I will defer my judgment to the automation."), whereas reliance would refer to complacency (i.e., "Nothing bad seems to be going on, so it's fine."). The case changes if automation is not fully reliable: being reliant or compliant with sub-par automation robustness can result in being compliant with erroneous cues (automation bias) or reliant with the erroneous lack of cues (complacency). Thus, errors due to automation bias are more salient due to their direct relation with a cue, whereas errors due to complacency can often be subtle because of the lack of cues.

Research generally agrees about the causes of automation complacency, and occurs when:

(a) a user is monitoring an automated system, (b) the frequency of monitoring is below a normal standard, and (c) the sub-optimal monitoring leads to performance degradation [69]. When the automation is in a negative state, failure to respond to the task leads to a sub-optimal attempt at restoring the correct system state, whether by omission (i.e., lack of action towards a given cue) or commission (i.e., erroneous action with respect to an inaccurate cue). Parasuraman and Manzey posit that a factor of complacent behavior stems from the concept of "premature cognitive commitment", which states that a repeated experience engenders "autopilot"-esque behavior, leading to the reduced attention and situation awareness required to detect automation failures [14]. This serves as the main segue into a complacent feedback loop (see Figure 2.3, where regardless of the performance of the automated system, the user either produces an error or learns to exhibit complacent behavior, feeding back into the interaction with the automated system. Due to the repeated nature of the interaction, it becomes a challenge to break the inactive behavior demonstrated by the user, with potential severe consequences if left unchecked.

The seminal work discussing improper trust calibration and inappropriate use of automation was brought forth by Parasuraman [10]. This opened several lines of research where the aim is to investigate what system designs, human characteristics, and environmental scenarios cultivate a situation where trust is unable to be calibrated properly. Bagheri and Jamieson investigated the effect of different levels of reliability over time on failure detection [70] and how transparency serves to mitigate its resulting over-trust and performance loss. By providing the user with a description of the reliability of the automation, users were able to adjust their attention allocation strategy to prevent performance loss due to complacent behavior [71]. Using transparency has become one of the main interventions to increase user acceptance and trust, even in robot-operated domains [24]. Associated techniques, such as control settings over the automation [36] and uncertainty quantification [72], have been found to affect decision-making processes by increasing user cooperation and situation awareness with new information availability.

Trust building is dynamic but independent of individual differences [73] and correlates with the level of reliability the automation presents [18]. However, trust can be fragile and is easily broken by negative experiences [74]. Often, a first trust violation results in a dramatic

loss of trust and performance, as the user is plunged to rectify any incorrect outcomes (i.e., "automation surprise"), but any repeated violations result in a less pronounced effect. Wickens and Xu dubbed this phenomena as the "first failure effect" [23], and serves to show how calibrating trust is a longitudinal process of both positive and negative interactions. Following this, de Visser et al. propose a framework to maintain and repair – if needed – trust between humans and automation by using a varied amount of justifications, emulating as a human would do (e.g., apologizing, explaining, or recognizing the error) [75]. This framework has also been extended towards physical automation (i.e., robots), with similar mitigation strategies that emulate a human justifying their errors [76].

Development of a measurable construct to quantify the potential of users to be complacent came forth by Singh et al., where an individual's propensity to engage in sub-optimal monitoring behavior was modeled by four factors: confidence, reliance, trust, and safety [77]. As workload became a predictive factor in Parasuraman and Manzey's model (Figure 2.3), Merritt revised the complacency potential scale across two factors: workload alleviation and monitoring. This allowed for a more accurate prediction of an operator's complacent behavior with respect to their task domain [69]. These scales are used in the core studies, as complacency potential becomes an important individual difference between users.

In the literature, the experimental interventions are often shaped as intelligent agents with controlled features (e.g., agent explanations for results, control settings, confidence intervals). The research is limited in intervention approaches to prevent over-trust and bolster proper trust calibration. Satehi et al. used accountability in interactive control agents as a deterrent to over-trust; in a microworld hospital task, users who perceived to be accountable for their performance took deliberate, but slower, decisions to optimize their contribution [78]. Bahner et al. implemented training protocols as a mitigator to over-trust; by presenting rare automation failures and training participants to handle such cases, users were able to reduce, but not eliminate complacent behavior (with commission errors still prevalent) [79]. We contribute to available interventions by exploring how reduced reliability can allow an increased amount of error signals to set proper expectations of the system.

2.3 CATEGORIZING AGENT TYPES AND TASK SCENARIOS

Since the potential amount of tasks where automation can operate is uncountable, it becomes challenging to establish the true effect of reliability given all the different ways agents can be represented and what kinds of tasks they can fulfill. Research in human-robot interaction indicates that different tasks influence how humans perceive their agent teammates [80]. A meta-analysis by Esterwood et al. determines that 3 task categories exist for human-robot interactions: influence tasks, physical manipulation tasks, and social interaction tasks [58]. Although these tasks focus on human-robot interaction, they can also broadly apply to non-robotic agents. This dissertation will largely focus on influence and physical manipulation tasks, as these tasks have clearly defined goals where performance can be easily measured, in contrast to social interaction tasks that may lack a clearly defined goal and are studied more in the context of social perception and responses (e.g., chatbots).

The representation of the agent has long varied within the history of human-AI interactions. According to Li, representation of an agent depends on embodiment and presence. Embodiment refers to where the agent has a representation akin to mechanical construction (humanoid or otherwise – a "body") in contrast to a static virtual character. Presence refers the exposure of the agent to the human – if the agent was physically present with the human in the real world, the presence would be physical. If the agent was conveyed through a computer screen, the presence would be digital [81]. Beyond designing for robustness, agents may be augmented with anthropomorphic features in order to facilitate social interactions (whether they have a goal or not), having agents employ human-like features like humanoid bodies, voice, and socially acceptable traits (e.g., politeness or apologies) [82, 83]. Research has found that participants present stronger emotions when the agent is embodied, increasing the amount of trust and personal disclosure to the system ([84]) – which can serve as a double-edged sword in social interactions ([85]). However, there exist cases in other research threads (e.g., simulated risk [86], speech perception [87]) that have found no effect between embodiment and the way humans interact with agents. A meta-analysis by Johnson et al. state that research on reliability with embodied agents (and human-robot interaction in general) requires further depth [88]. It remains paramount to consider all different types of

agents when investigating the design of these systems – especially when the same effects are also found in physical agents (e.g., the first failure effect is also found with physical robots, outside decision-making tasks [89]).

A task where the goal of the agent is to influence (or support) the decision-making process of the user is denoted as an influence task. In these types of tasks, humans are tasked with making a decision under a certain amount of uncertainty, where they are required to search for relevant information, internalize results, and corroborate with experience to select the optimal outcome. The information acquired comes from a variety of different sources, requiring higher cognitive focus to process and derive insights to make a decision [90]. Now able to process large amounts of information with advancing hardware and algorithms, the agent's recommendations have become difficult to comprehend and predict. System designers have responded by increasing transparency (via explanations) and customizability (via control options) [91, 92, 93], and although these features have been shown to increase user adoption and agent trustworthiness [94, 95], it may further lead to states of over-trust due to high perceived reliability. Agents in influence tasks are not required to be embodied, as they only need to process and relay the recommendation back to the user, which can be done as simply as outputting text in a command line.

In other scenarios, physical manipulation tasks are what we usually envision when thinking about robotic interactions – for instance, assembly lines, navigation and exploration, physical assistance, robotic surgery, among others. These interactions usually require a physical representation in order to complete its goals. Investigating interactions between humans and agents in physical manipulation tasks becomes more challenging as it requires the design and construction of an agent with the capability to manipulate, such as robots and drones, which incurs high-research costs. As an alternative, these interactions are investigated in simulation – including the use of Virtual Reality thanks to its low cost of prototyping [96]. However, VR studies focus on task performance (from both the human and the agent) – with little emphasis on cognitive and psychological factors that affect the interaction (e.g., [97]). As agents become more efficient in handling uncertain and unconstrained environments, robots operating in tandem with humans will become a common occurrence, requiring further research on these domains.

Table 2.1: A random sample of identified human-AI interaction studies with operation domains common in human factors research (includes relevant constructs for this dissertation, including, but not limited to: reliability, trust, automation complacency and bias, and transparency).

Domain	**Task**	**Related Publications**
Process Control	AutoCAMS (1.0/2.0)	[79, 98, 99, 100]
Aviation	Multi Attribute Task Battery (I/II)	[11, 31, 101, 102, 103, 104]
	Flight Simulations	[64, 105, 106]
	Autopilot Monitoring	[107, 108]
	Air Traffic Control	[109, 110]
Vehicular	Driving Simulations	[111, 112]
Security	X-Ray Screening	[113, 114, 115, 116, 117]
Healthcare	Medical Diagnosis	[118, 119]
	Surgical Procedures	[120]
Military	Command and Control	[22, 121, 122]
	UAV Control	[123, 124, 125]
	Convoy Security	[126, 127]

Human factors research has addressed and contributed to our understanding of trust calibration through individual experiments on the different facets that comprise human cognition and automation interface design. To consider, however, is the variety of scenarios that an AI system may be deployed in, and how research in human factors and human-computer interaction as addressed the effect of the scenario itself on the interaction between the human and the agent. Hopko, Mehta, and McDonald stipulate that the effects between reliability and trust vary widely per scenario [128].

In the past, human factors research has relied on carefully designed and validated tasks to study human performance and workload in a laboratory setting. A couple of example frameworks that have been extensively used for research are AutoCAMS [129] and the Multi Attribute Task Battery [130]. These tasks facilitate the research of humans supervising and controlling automated processes – supervisory control as to establish a definition. Consequently, the research community would continue to focus on domains where the operator

is largely subject to time-critical, high-risk scenarios where results would be catastrophic if the operator's performance were to suffer or fail. Domains such as process control, aviation, vehicular control, national security, healthcare, and military usually comprise stressful and high-stakes scenarios, and complacency has been largely studied under these simulated tasks, albeit with the guarantee of safety (i.e., whether the participant or operators fail during the task carries no consequence) – a sample of such studies can be found in Table 2.1. A systematic literature review from healthcare research [131] identified a large amount of publications within aviation and medicine, with small focus on low-risk fields such as general human-computer interaction. Granted, the term "general human-computer interaction" is a vague and wide-covering term, with a massive amount of potential scenarios that improperly calibrated trust can affect, yet we constantly see the effect of sub-optimal cognitive processing and build software solutions and tools in order to measure and combat against adverse behavior.

New research and development on programming languages, graphical engines, and user interfaces have allowed for a renewed facility to program graphical and immersive tasks that are tailored to model a specific scenario. This presents a divergence in the scenarios that are researched and how automation reliability affects outcomes in domains that could be conceptually distinct. On one hand, there is extensive research on a small subset of tasks that model a specific set of human-AI interactions, but may have limited ecological validity [37] due to the specific domain and task conditions required to evoke the effects found (e.g., environmental interactions, agent features, feedback timing). On the other hand, newly-designed simulations and tasks present a small portion of the effects found in research, requiring a longitudinal comparison or meta-analysis on how the scenario affects the outcomes. This dissertation does not include a meta-analysis, but offers a comparison on how reliability affects the outcome between the scenarios of the core studies presented (Chapter 6).

CHAPTER 3: VIRTUAL AGENTS IN DECISION SUPPORT SYSTEMS AND KNOWLEDGE COMPLACENCY

This chapter centers on studying decision support systems as agents in Human-AI teams operating in an influence task. As reviewed in Chapter 2, human-AI interaction and human factors often study what agent features directly affect human performance. However, little is researched about the factors that mediate this loss in performance, and whether improperly calibrated trust could be a culprit in this dynamic. To this goal, we designed two studies that aim to measure insight of decision-making when interacting with a decision support system. One task mimics an activity that would be done in routine interactions with information retrieval systems, while the other task follows a more abstract, game-theoretic framework.

We discuss the design, results, and implications of a study which manipulated the presence of explanations, control settings, and the reliability of two recommender systems: a collaborative filtering algorithm and a forecasting algorithm. In this study, users interacted with two recommender systems: the Movie Miner, and the Diner's Dilemma game. The results support the following:

- User insights about domain knowledge are reduced whenever they perceive the agent as reliable (i.e., explainable, controllable, and correct agents lead to over-trust – improper trust calibration).
- This over-trust is mitigated by an agent that is imperfect – domain learning was improved. However, benefits from imperfections are interactions; by itself, imperfections also hinder learning.

3.1 STUDY OVERVIEW

The main research questions surrounding this study are:

RQ3.1: How do recommender system design affect domain knowledge acquisition?

RQ3.2: How do different levels of reliability interact with recommender system design?

To this end, we extended Schaffer's work on cognitive modeling in recommender systems [36, 132] to denote the presence of improper trust calibration and over-trusting behavior in users interacting with decision support systems. Decision support systems are used rather ubiquitously, helping users to navigate from point A to point B, recommending movies and music, assisting in military operations and decision-making, and even high performance computing. When these systems are used, the user benefits from the knowledge ingrained in the system [27], but would users benefit from long-term usage given that a reliable system suggests potential over-trust [32]? When users fail to exercise their judgment and knowledge, it may lead to the atrophying of these skills [120]. As decision support systems are widely deployed in systems used by the standard consumer, it provides us general insight into how over-trust affects the human-AI interaction dynamic in these systems.

Past research on virtual monolithic and multi-agent systems indicate that all automated agents can be described by the presence of three features: explanations, control settings, and reliability[1] (ECR). Explanations and control settings allow for improved transparency of the agent, leading to improved cognitive insight and maintained situation awareness, thus avoiding averse behavior [36, 66, 132]. Explanations and control settings are system features, and as such, can be manipulated by developers and designers in an optimal way to the needs of the domain and situation. Reliability pertains to the amount of error in the system; the output of the agent is a function of the robustness of the system. Although we expect automated systems to be generally reliable [66], occasionally there can be no guarantees about the quality of the input that the system receives. Unreliable systems result in frustration and non-adoption from the user. In this experimental setup, we manipulate ECR in order to investigate how transparent systems engender the presence of improper trust calibration and affect user domain knowledge.

By manipulating the agents' ECR in both tasks, we establish the following four hypotheses that guide the approach in this study:

> H_1: Error in (i.e., inaccurate) decision support systems (DSS) lead to increased domain knowledge over time.

[1] Changed from *error*, to emphasize the established nomenclature in this dissertation.

H_2: Explanations from DSS prevent domain knowledge from decreasing over time.

H_3: Control over DSS leads to increased domain knowledge over time.

H_4: Control over DSS along with explanations leads to increased domain knowledge over time.

The rationale behind the hypotheses are grounded on intuition of the human cognition. For H_1, we expect that an unreliable agent would compel the participant to further understand the domain as to make decisions that allows them to complete the task successfully. For H_2, we expect explanations to allow a better understanding of the agent, under the idea that allowing a user to grasp the rationale of the automation would provide increased insight into how it uses domain knowledge to justify its decision. For H_3, allowing control over the agent allows the user to experiment with different parameters and develop insight on how data flows through the system under different conditions, leading to increased knowledge on the domain. Lastly, for H_4, the interaction between allowing control and providing explanations to the user serves as transparency, a common method for designers to allow predictability, trustworthiness and adoptability of their systems [94], thus we expect positive interaction to increase their domain knowledge after use.

In many cases, the "effectiveness" of decision support systems is defined by the goal the user has. According to Schaffer et al. [133], decision support systems can be applied to task domains that are either subjective or objective. A subjective task domain has a criterion of success that aligns more with the personal satisfaction of the user, such as recommending a relevant item for a user purchasing items online. An objective task domain covers scenarios where success and failure is a clear dichotomy; completion of the goal was either achieved or missed, and is verifiable by any third party. As this extends the coverage of our applied domains, we defined two abstract tasks to study agent reliability: the Movie Recommendation (MR) study and the Diner's Dilemma (DD) study.

3.2 EXPERIMENTAL DESIGN

3.2.1 Movie Recommendation Study

Background

Personalized content is now a staple of systems that provide consumer content. Popular feed services (e.g, YouTube, Facebook, Twitter) hinge on providing dynamic recommended content to the user based on their profile, interests, or preferences. A recommender system is a type of decision support system where it uses any sort of algorithmic process in order to produce a list of items of interest. In the previous examples, such algorithms can be seen curating the content feed in Facebook (posts and pictures from friends) or YouTube (recommending content based on previously watched videos). With recommended content shown in a static interface, such as a scroll down list for feed-based content (e.g., Twitter, Facebook), or in a grid for selective content (e.g., Netflix movie recommendations, Amazon products), many users are largely unaware of the workings of an algorithm behind the scenes [134].

Curating the personalized content is a form of automation, since the process occurs largely without the user's explicit instructions or intervention. For instance, Twitter automatically curates recent and popular posts from account the user follows, without prompting the user to select specific topics of interest from that profile. On YouTube, watching or liking videos automatically builds a profile upon which the algorithm suggests future content. Other recommendation or curation services follow the same paradigm: to automatically build and adjust the user's profile in order to provide a seamless experience through use and navigation of their service.

The Movie Miner is a simulated movie recommendation tool that provides movie suggestions as the user builds their profile by rating movies. By interacting with the tool, users form insights and intuitive knowledge of the movie metadata domain (e.g., what movies are popular, what movies have the highest rating). By introducing the automated process of recommendation, the user's perception and cognition is now part of the human-automation interaction dynamic: it is subject to benefits and detriments of automation. We aim to

study effects of recommender system features on over-trust, as an initial attempt to judge how does reliability interact with commonly used features, as well as observing how does the recommendation domain affect these findings.

System Design

Figure 3.1: A screenshot of the Movie Miner, a movie discovery interface, which was used in the Movie Recommendation study. The browsing tool is shown on the left (blue) and the recommendation tool is shown on the right (brown). Users could search, rank, and filter on all movies in the dataset using the browser, and the recommendations on the right interactively updated as rating data was provided (center, yellow). In the task, users were asked to find a set of interesting movies to watch (center, green) using whichever tool they most preferred.

We reused the design of the system as outlined by Schaffer [27]. Schaffer states that the Movie Miner was implemented with two goals in mind: a) to make it familiar to modern web users as much as possible, and b) to make it similar to deployed recommender systems. Novel design choices were minimized to prevent deviations of results to the current practice of recommender systems.

The Movie Recommendation study was designed to simulate users selecting a movie to watch using available information tools online. Participants were presented with the Movie Miner (Figure 3.1), a movie search and recommendation tool akin to IMDb [135] or Movielens [136]. On the left (blue) side, participants could search, rank, and filter the movie dataset with the provided basic features. The right (brown) side provided the participants with a ranked list of recommendations, interactively updated with a collaborative filtering algorithm as the participant provided ratings. The Movie Miner was powered by the Movielens 20M dataset [137], which contains updated movie references and ratings, and has been validated through wide use in industry and academic research. 4 million ratings were randomly sampled from the 20 million total ratings to prevent computation delays from the Movie Miner.

According to Schaffer [27], a traditional collaborative filtering approach was selected to generate movie recommendations from the dataset. Collaborative filtering is a technique known and well understood within recommender systems research [138], where recommendations are generated through user-user similarity, which yielded movies from other users' profiles if they had similar ratings on titles common between the users. The similarity function (i.e., the similarity between two users u and v) was defined by the Pearson correlation over the user profile of ratings, specified as:

$$sim(u,v) = \frac{\hat{u} \cdot \hat{v}}{||\hat{u}|| \cdot ||\hat{v}||} = \frac{\sum_i \hat{r_{ui}} \hat{r_{vi}}}{\sqrt{\sum_i \hat{r_{ui}}^2} \sqrt{\sum_i \hat{r_{vi}}^2}} \tag{3.1}$$

r_{ui} is the rating given by user u to movie i, where $i \in I_u \cap I_v$ (I_u is the set of movies rated by user u, and i must have been rated by user u and v). $\hat{r_u}i$ is the normalized rating $r_{ui} - \mu_u$, where μ_u is the mean rating given to movies by user u. Herlocker damping [139] was applied to $sim(u,v)$ in order to prevent high similarity between users through mutual rating of popular movies. Then, a predicted rating for user u, calculated as r_{ui} when $i \notin I_u$, is specified by:

$$r_{ui} = b(\mu_u + k \sum_{v \in U} sim(u,v) \hat{r_{vi}}) \tag{3.2}$$

k is a normalizing scalar imposed on all users where $k = \frac{1}{\sum_{v \in U} |sim(u,v)|}$. Another normalization was applied by b, defined as $b = \max_{i \notin I_u} r_{ui}$, where it allows the predicted ratings to spread over the full rating range (0.5 stars to 5 stars) rather than a user's predicted maximum rating.

Interface Design

Following the design of Schaffer [27], the Movie Miner was divided into 2 main interactables: the browser and the recommender, as seen in Figure 3.1.

In the browser, the users could search, rank, and filter movies according to their needs. This interaction paradigm models the most standard features found in movie recommendation systems (e.g., Netflix). Search is based on text matching, and returns all movies that match a particular keyword. Users could rank and sort the movies by individual metadata (e.g., title, rating, release date, genre). Finally, users could create multiple range filters that would match a certain metadata (e.g., users could filter all movies with a rating of at least 2 stars).

The recommender features were available according to the independent variables (discussed in the next section). The recommender would present a list of suggested movies based on the filtering in the browser, along with the user's recommendations. This list was sorted by the predicted rating from the collaborative filtering algorithm, and was not able to be rearranged under any condition. However, similar to the browser, users could filter the recommended movies and relay to the recommender that they were "Not interested" in a suggested movie, permanently hiding it from the recommendation list.

Experimental Design

Independent Variables. Two levels of each ECR feature in the Movie Miner were manipulated, as per Schaffer [27]. All manipulations were done in between-subjects in this experiment. Thus, we had 8 manipulations in total, including the baseline condition with explanations and control absent, with high reliability. The independent variable manipulations for each feature were as follows:

Explanation

- Absent: Movie recommendations were provided without explanation.
- Present: The recommender explained with a short sentence how ratings were calculated: "Movie Miner matches you with other people who share your tastes to predict your rating." Additionally, the recommender would display the list of items in the participant's profile that most affected the recommendation.

Control

- Partial: Allowed participants to manipulate their profile by adding, deleting, or re-rating to receive recommender feedback.
- Full: In addition to Partial, participants could define custom filters to narrow recommendations, and remove individual movies (i.e., "Not interested" button) from the recommender feedback.

Reliability

- High: The collaborative filtering algorithm with Herlocker damping was used.
- Low: A vector of noise was generated (up to 2 stars of difference) and added to r_{ui}, perturbing the algorithm. This resulted in a reordered list of recommendations.

Dependent Variable. The response variable – knowledge – was assessed through the knowledge test. The knowledge test was designed to gauge the participant's knowledge of movie metadata (e.g., genres, ratings, popularity). High scores would be indicative of a good internalization of the movie metadata domain. IMDb records were used as the ground truth for the questions. The questions administered in the knowledge test are outlined in Table 3.1.

The dependent variable was analyzed with a Structural Equation Model (SEM) to study multiple correlates of change. Raykov [140, 141] developed an SEM suited for change through time, which allows us to not make any assumptions about the distribution of the dependent variable or the homogeneity of variance between samples. We did not make any assumptions

Table 3.1: Knowledge test questions administered in the Movie Recommendation study. Questions were evaluated on a dichotomy (correct or incorrect).

Code	Item	Responses
ins1	Online, which genre has the highest current average audience rating?	Multiple choice
ins2	Online, which of these genres tends to be the most common among the movies with the highest average audience rating?	Multiple choice
ins3	Online, which of these genres has the highest current popularity?	Multiple choice
ins4	Generally, which of these genres has the highest current popularity?	Multiple choice
ins5	Online, which of these decades has the highest current average audience rating?	Multiple choice
ins6	How many movies haven an average audience rating greater than 9/10?	Multiple choice
ins7	Popular movies tend to have an average rating that is _____.	Multiple choice: lower/average/higher
ins8	Movies with an average rating of 9/10 or higher tend to have _____ votes.	Multiple choice: fewer/average/more

about the participants' prior knowledge; the participant could have been very knowledgeable in films, or have never watched a movie in their life. Additionally, we did not make any assumptions about their ability to learn about the domain in a short period of time. Thus, an SEM serves appropriate to study predictors of longitudinal change. We refer to this SEM from this point as a Raykov change model. The established Raykov change model is visualized in Figure 3.2.

Procedure. Participants were recruited on Amazon Mechanical Turk[2] – an online crowdsourcing platform which allows researchers to collect large amounts of participants for experiment and data collection. Participants completed the study through 4 phases: pre-study, priming phase, watchlist phase, and post-study. The procedure is identical as done by Schaffer [27].

The pre-study and post-study contained the knowledge test. Participants responded to

[2]https://www.mturk.com/

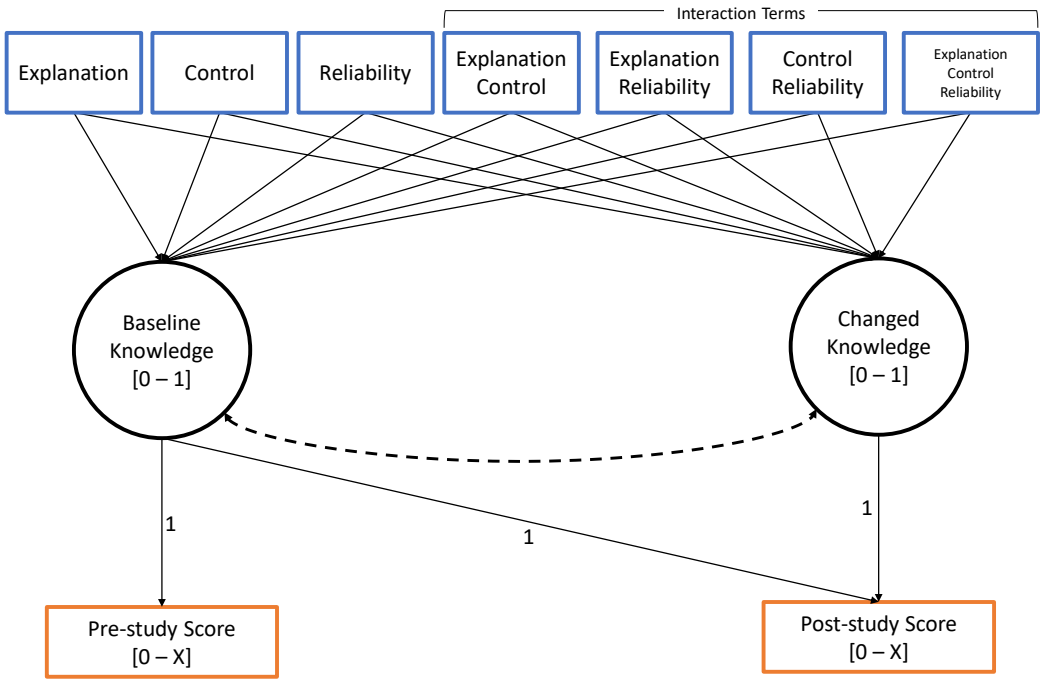


Figure 3.2: The Raykov change model specified to test the hypotheses in both the Movie Recommender study and the Diner's Dilemma study. X defines the number of questions in the knowledge test for each study.

the same test before and after the task.

The participants then began the task with the priming phase, where they were asked to rate at least 10 movies that they believed would best represent their preferences, using only the blue Movie Database list. In many cases, participants rated more than the required 10. The priming allowed the Movie Miner to calculate the proper recommendations according to the participant's profile.

In the watchlist phase, participants were allowed to use the brown Recommended For You in addition to the blue Movie Database list. Participants were told to use whichever tool they would like in order to find 5 to 7 new movies to watch, requiring them to at least spend 12 minutes using the Movie Miner. The post-study was conducted after the participants finished interacting with the interface. The experiment flow is visualized in Figure 3.3.

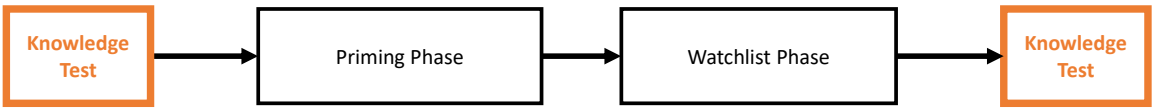


Figure 3.3: Experiment flow of the Movie Recommendation study.

3.2.2 Diner's Dilemma Study

Background

We selected the iterated Prisoner's Dilemma, a known game-theoretic task which has been long used to study cooperative behavior between conflicting entities, to investigate how complacency affects domain knowledge in an abstract adversarial game.

For the uninitiated, the Prisoner's Dilemma demonstrates and analyzes why two fully rational individuals might not cooperate, even if it might be in their best interest to do so. Many past studies have used the cooperation/defection paradigm to study economics, politics, sociology, biology, and computer science [142, 143, 144]. A summary of the game is as follows: Two prisoners in separate rooms without the ability to communicate are given a bargain by their prosecutors in order to lessen their sentence. Prisoners can either stay silent (i.e., cooperate) or testify that the other committed the crime (i.e., defect). If both prisoners stay silent, they serve 1 year in prison. If both prisoners betray the other, they both serve 2 years in prison. If one of the prisoners stays silent but the other betrays, the silent prisoner will serve 3 years and the betrayer will go free. The *dilemma* here is pursuing one's own self-interest results in a worse outcome than if both players had cooperated. Thus, the Nash equilibrium (i.e., the best possible strategy without considering the strategy of the other player) resolves to always defect, as the opportunity to go free outweighs going to jail for 2 years, which is much better than the possibility of staying silent and being locked for 3 years. However, this equilibrium does not hold if there are multiple rounds of the game, where memory of past actions affects the decisions a player may take in the future. This repeated version is referred to as the iterated Prisoner's Dilemma.

In a lighthearted version of the iterated Prisoner's Dilemma, Teng [145] developed the

Diner's Dilemma, originally to study the relationship between trust and situation awareness with different variations of interface design. We re-purpose this abstract task to study how do players, who may be familiar with Prisoner's Dilemma or their variations, exercise their knowledge in the presence of a decision support system with varying features of ECR.

The Diner's Dilemma is an abstract way to study discrete decision-making and cooperative behavior. At every round in the game, the user must make a single unit of judgment in other to maximize their reward while traversing cooperative dynamics with other diners by determining their strategies. Albeit an interesting concept by itself, as the Prisoner's Dilemma can be applied to almost any situation that intends to balance cooperation and competition, the focus is to study how knowledge (which shapes judgment and decision-making) is affected when the decision is supported by automation. Demonstrating how does improper trust calibration and knowledge loss occurs in a discrete judgment scenario is a comparison point against other similar domains where knowledge can affect decision-making and overall performance.

System Design

We re-purposed Schaffer's Diner's Dilemma web game [27] for this study. Participants visit a restaurant with two of their friends and must choose between an inexpensive dish with has low nutritional value (a hotdog), or an expensive dish with high nutritional value (a lobster). The diners all agree to split the bill equally no matter what is ordered. As the game progresses, participants must form strategies of cooperation or defection to maximize the *dining points* they score in each round. The *dining points* of a diner are defined as the nutritional value of the food divided by the price paid by that diner. Thus, the game is centered around evaluating the benefits and costs of selecting the appropriate strategy against the strategy and dishes the other diners select.

The participant plays Diner's Dilemma with two other simulated co-diners. The co-diners were driven by a heuristic rule set, mainly centered in playing variations of Tit-for-Tat. Tit-for-Tat is a rudimentary strategy where the co-diner selects the same dish the participant did the previous round. Strategies occasionally varied from Tit-for-Tat through two behav-

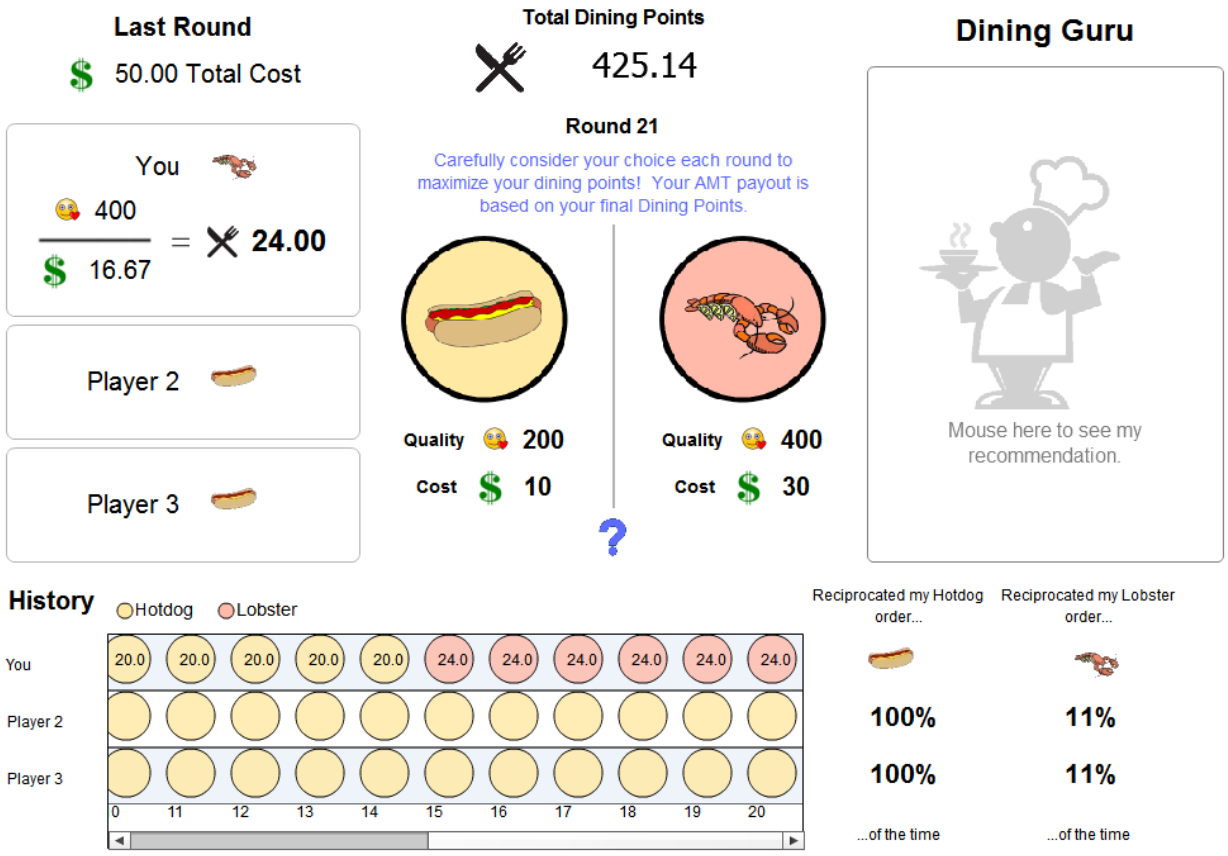


Figure 3.4: Interface for the Diner's Dilemma study, showing the core game information (center, left), the history tool (bottom), and the Dining Guru (right side). Participants could seek recommendations by mousing over the Dining Guru or alternatively use the history panel to devise strategies.

iors: rate of forgiveness and rate of betrayal. When a co-diner forgives, they will respond to the previous expensive order with a cheap order. Conversely, when a co-diner betrays, they will respond to a previous cheap order with an expensive order. Co-diners responded independently to the participant's selections, in order to make the game more understandable. Co-diner strategies had no bearing on analysis.

While playing, participants were provided with a visualization of choices made and points earned by the group in previous rounds. Participants also had the choice of receiving recommendations from the *Dining Guru* (Figure 3.4), which analyzed the behavior of the diners and suggested a selection based on which item was expected to result in the most number of points. Unlike the Movie Recommendation study, users could access recommendations on

demand by mousing over the *Dining Guru* panel. Participants were trained in the interface panel and the *Dining Guru* before the game started.

Experimental Design

Independent Variables. Two levels of explanation, two levels of control, and three levels of reliability in the *Dining Guru* were manipulated, as per Schaffer [27]. All manipulations were done in between-subjects in this experiment. Thus, we had 12 manipulations in total, including the baseline condition with no explanations, no control settings, and high reliability (i.e., the *Dining Guru* was a highly accurate black box). The modified components are visualized in Figure 3.5 and examples of the interfaces per condition can be seen in Figure 3.6. The independent variable manipulations for each dimension were as follows:

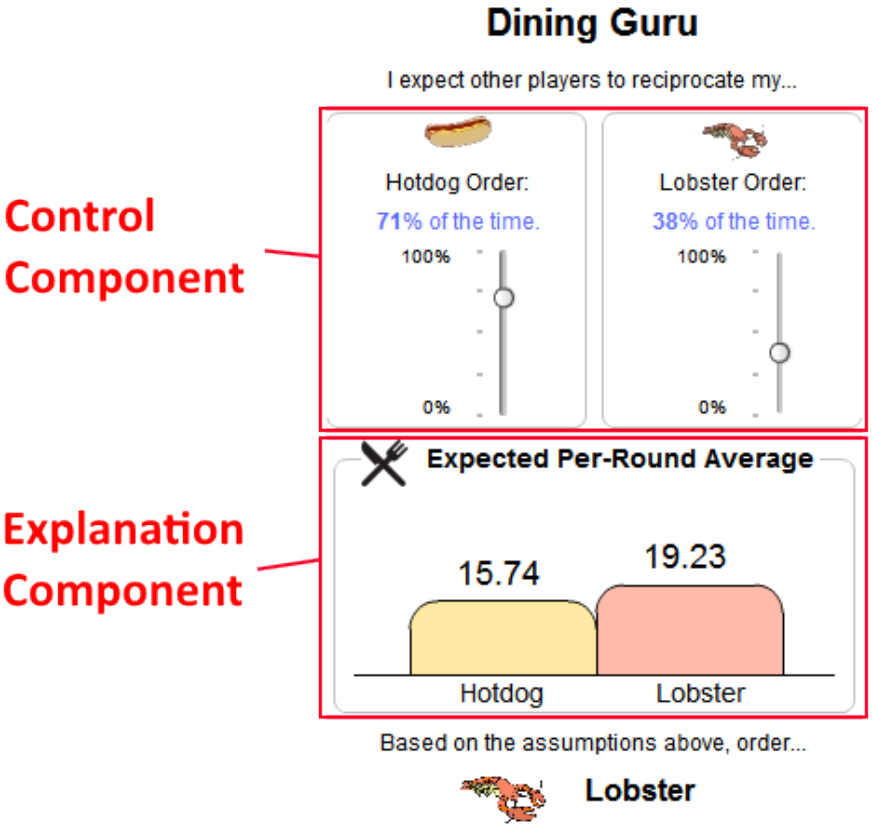


Figure 3.5: *Dining Guru* components varied based on treatment. The control component and explanation component are outlined. In the baseline treatment, the *Dining Guru* would only recommend the item (bottom).

Explanation

- Absent: The recommender would only show the recommended dish to select.
- Present: The recommender would display the expected dining points average for that round, taking into account either previous decisions made by the co-diners or the participant defined expected reciprocity, depending on the Control condition.

Control

- None: Participants were not able to control expected reciprocity (i.e., selecting the same dish as the participant) from other co-diners in order to explore possibilities. If explanations were present, the explanation was based on the co-diners' previous decisions.
- Full: Participants could set a percentage of reciprocity in order to change the recommendation given by the *Dining Guru*. If explanations were present, the explanation was based on the percentage input per dish.

Reliability

- High: The *Dining Guru* would provide correct and accurate recommendations based on the reciprocity rates for each dish. Participants would perform extremely well following recommendations.
- Mid: The reciprocity rates for each dish was modified randomly from the truth up to $\pm$ 25% every round. This resulted in the *Dining Guru* to occasionally switch recommendations between rounds. If the recommendations were followed, participants still performed relatively well.
- Low: The reciprocity rates for each dish was modified randomly from the truth up to $\pm$ 50% every round. This resulted in the *Dining Guru* to behave erratically, with recommendations hampering performance.

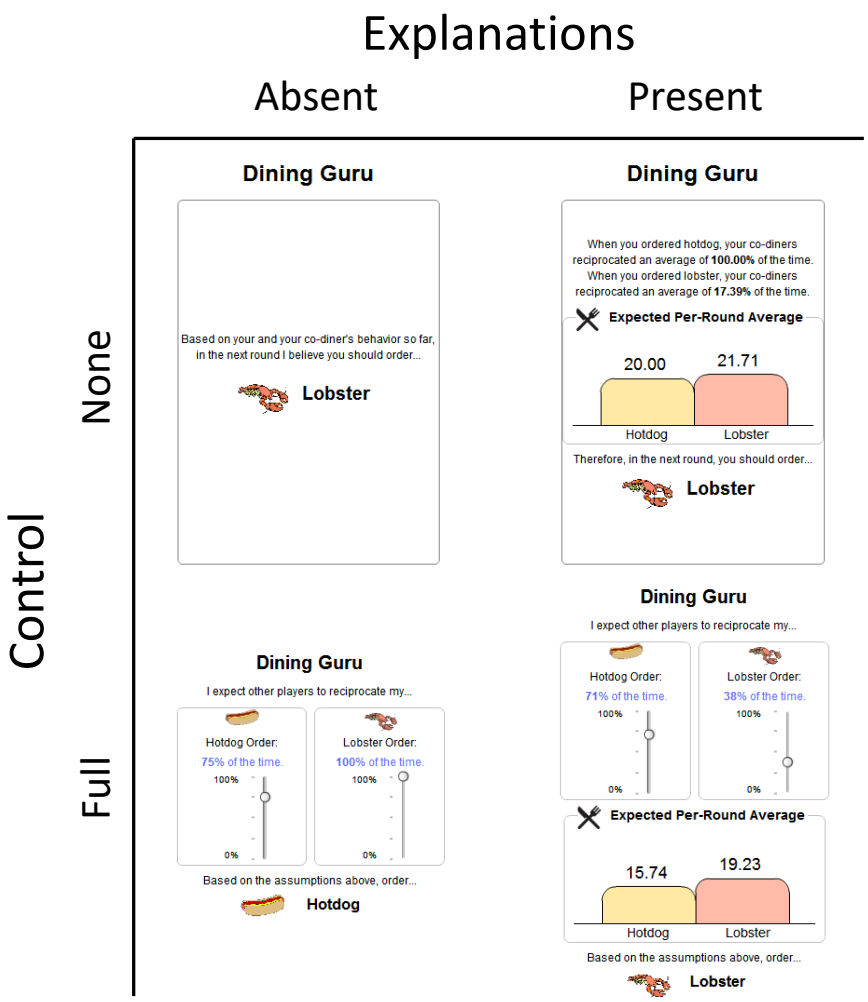


Figure 3.6: Screenshots of the *Dining Guru* with all Explanation and Control variations.

Dependent Variable. The response variable – knowledge – was assessed through the knowledge test. The knowledge test was designed to gauge the participant's familiarity and knowledge of Diner's Dilemma and the rules of the game. High scores would be indicative of a good internalization of the game's rules. The questions administered in the knowledge test are outlined in Table 3.2. As with the Movie Recommendation study, a Raykov change model was also used to investigate change in knowledge over time (visualized in Figure 3.2).

Procedure. Participants were recruited on Amazon Mechanical Turk[3]. Participants completed the study in 4 phases: training phase, pre-study, game phase, and post-study. The

[3]https://www.mturk.com/

Table 3.2: Knowledge test questions administered in the Diner's Dilemma study. Questions were evaluated on a dichotomy (correct or incorrect).

Code	Item	Responses
ins1	How much does a Hotdog cost?	Slider (0 to 50, increment: 1)
ins2	How much does a Lobster cost?	Slider (0 to 50, increment: 1)
ins3	What is the quality of a Hotdog?	Slider (0 to 500, increment: 10)
ins4	What is the quality of a Lobster?	Slider (0 to 500, increment: 10)
ins5	In a one-round Diner's Dilemma game (one one restaurant visit), you get the least amount of dining points when...	Multiple choice
ins6	In a one-round Diner's Dilemma game (one one restaurant visit), you get the most amount of dining points when...	Multiple choice
ins7	Which situation gets you more points?	Multiple choice
ins8	Which situation gets you more points?	Multiple choice
ins9	Suppose you know for sure that your co-diners reciprocate your Hotdog order 100% of the time and reciprocate your Lobster order 100% of the time. Which should you order for the rest of the game?	Hotdog or Lobster
ins10	Suppose you know for sure that your co-diners reciprocate your Hotdog order 0% of the time and reciprocate your Lobster order 100% of the time. Which should you order for the rest of the game?	Hotdog or Lobster
ins11	Suppose you know for sure that your co-diners reciprocate your Hotdog order 50% of the time and reciprocate your Lobster order 50% of the time. Which should you order for the rest of the game?	Hotdog or Lobster

procedure is identical as proposed by Schaffer [27] and the previous Movie Recommendation study.

The pre-study and post-study contained the knowledge test. Participants responded to the same test before and after the task.

The participants began the task with the training phase, where they were introduced to the rules of the game and the utility of the *Dining Guru*. Participants learned how to access and operate the *Dining Guru*, and were advised that the *Dining Guru* was not guaranteed to provide the optimal decision at every round, and following said advice was up to the participant. After participants familiarized themselves with the game and the tool, the first knowledge test was administered in the pre-test.

In the game phase, participants played 3 games of Diner's Dilemma with two simulated co-diners, each varying in strategy per game. Participants completed the games at their own

pace, and the final knowledge test was administered during the post-study. The experiment flow is visualized in Figure 3.7.

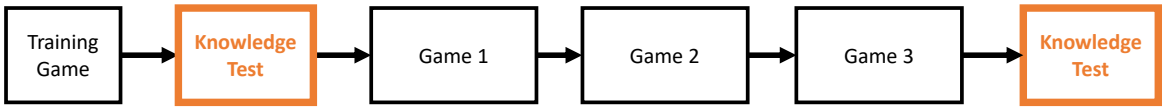


Figure 3.7: Experiment flow of the Diner's Dilemma study.

3.3 RESULTS

A total of 1055 participants completed both studies, apt for extensive inferential statistics and structural equation modeling. A summary of demographic information is outlined in Table 3.3.

3.3.1 Movie Recommendation Study

526 participants were recruited via Amazon Mechanical Turk. Participants were between 18 and 71 years of age (μ = 35 years, σ = 11 years, 45% male) and spent between 25 and 60 minutes interacting with the Movie Miner. All participants were compensated 10 USD for completing the experiment.

A McNemar's Chi-squared test with continuity correction revealed that knowledge test performance significantly differed after the intervention ($\chi^2(1, 4208) = 10.263$, p = 0.0013, $\phi = 0.05$, odds ratio is 10.3).

The Raykov change model reveals that explanation, control, and reliability all caused incorrect knowledge to be formed. Users in the "control only" condition scored less than half a standard deviation lower in the final knowledge test (solution = -0.527). This was somewhat mitigated with the presence of explanations (solution = -0.276) and low reliability (solution = -0.347). "low reliability only" led to incorrect knowledge (solution = -0.391). Explanations, control, and reliability have significant interactions which led to incorrect knowledge (solution = -0.309).

Table 3.3: Resulting demographics for the Movie Recommendation and Diner's Dilemma studies. Categories with 0 participants in all conditions were not included. Highest education demographic information was not collected for the Movie Recommendation study.

	Movie	**Diner**
Sample Size (n)	526	529
Age		
18 - 24	69	72
25 - 34	234	263
35 - 44	121	126
45 - 54	63	45
55 - 64	31	21
65+	8	2
Gender		
Male	234	284
Female	292	243
Non-Binary	0	2
Highest Education Completed		
High School	-	51
2-year College	-	172
4-year College	-	235
Graduate	-	65
Terminal	-	6

H_1 (inaccurate DSS lead to increased knowledge over time) was rejected, with the model predicting a knowledge decrease between pre-test and post-test when the DSS was error-prone (B = -0.391, $p < 0.01$). H_2 (explanations from DSS prevents knowledge from decreasing over time) was failed to be rejected (although it presents marginal significance), with the model predicting a knowledge decrease between pre-test and post-test when explanations were present (B = -0.298, p = 0.088). H_3 (control over DSS leads to increased knowledge over time) was rejected, with the model predicting a knowledge decrease between pre-test and post-test when control over the DSS was allowed (B = -0.527, $p < 0.01$). H_4 (control over DSS along with explanations leads to increased knowledge over time) was rejected, with the model predicting a knowledge decrease between pre-test and post-test when both explanations and control were present (solution = -0.276, $p < 0.05$).

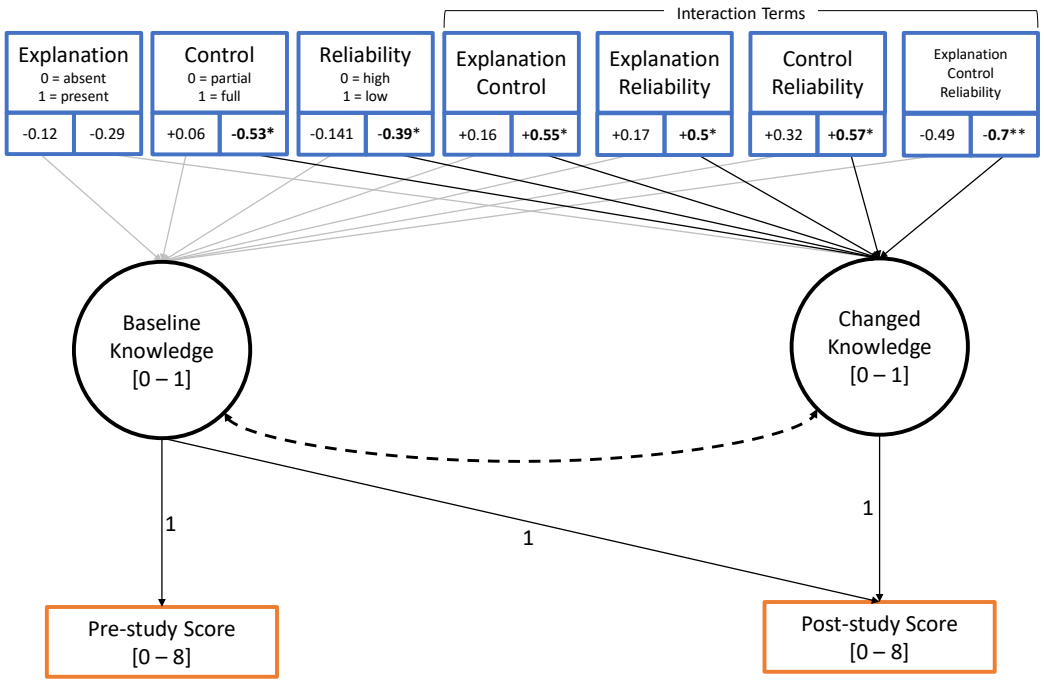


Figure 3.8: The fitted Raykov change model for the Movie Recommendation study. Non-significant regressions have been grayed out.

3.3.2 Diner's Dilemma Study

529 participants were recruited via Amazon Mechanical Turk. Participants were between 18 and 70 years of age (μ = 34 years, σ = 10 years, 54% male) and spent between 30 and 50 minutes playing Diner's Dilemma. All participants were compensated 10 USD for completing the experiment.

A McNemar's Chi-squared test with continuity correction revealed that knowledge test performance significantly differed after treatment ($\chi^2(1, 5819) = 37.081$, $p < 0.001$, $\phi = 0.08$, odds ratio is 11.3).

The Raykov change model reveals that explanation, control, and reliability all caused incorrect knowledge to be formed. Lowest scores in the knowledge test originate from the "control only" (solution = -0.563), "explanation, control, and low reliability" (solution = -0.576), and "explanation and reliability" (solution = -0.614) conditions.

H_1 (inaccurate DSS lead to increased knowledge over time) was rejected, with the model predicting a knowledge decrease between pre-test and post-test when the DSS was error-

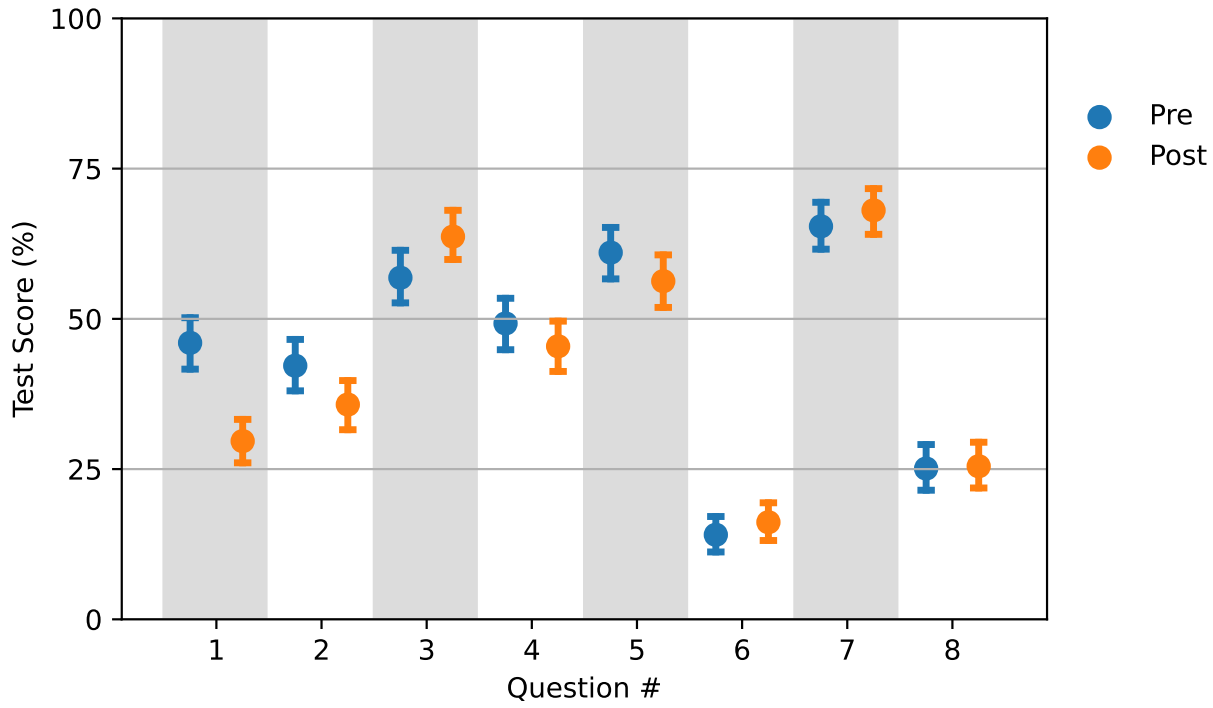


Figure 3.9: Mean correct answers for each knowledge question in the Movie study across time. Error bars indicate 95% confidence intervals.

prone (B = -0.477, $p < 0.01$). H_2 (explanations from DSS prevents knowledge from decreasing over time) was rejected, with the model predicting a knowledge decrease between pre-test and post-test when explanations were present (B = -0.464, $p < 0.01$). H_3 (control over DSS leads to increased knowledge over time) was rejected, with the model predicting a knowledge decrease between pre-test and post-test when control over the DSS was allowed (B = -0.563, $p < 0.01$). H_4 (control over DSS along with explanations leads to increased knowledge over time) was rejected, with the model predicting a knowledge decrease between pre-test and post-test when both explanations and control were present (solution = -0.459, $p < 0.05$).

3.4 DISCUSSION

We found that the data supported rejecting all of the initial null hypotheses. However, we found a very interesting effect: the opaque, non-customizable systems were the best for domain knowledge learning and retention. Knowledge loss was observed in both studies, but only under specific conditions of explanation, control, and reliability. In the Movie

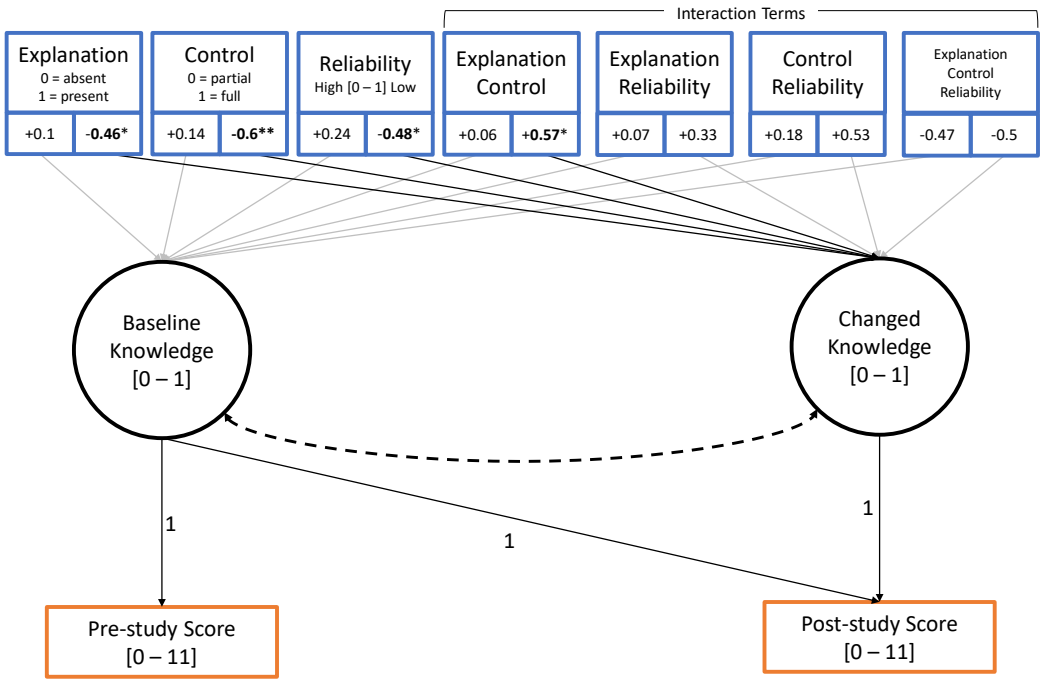


Figure 3.10: The fitted Raykov change model for the Diner's Dilemma study. Non-significant regressions have been grayed out.

Recommendation study, there was an overall drop in knowledge score from 45.0% correct to 42.5% correct on average, regardless of experiment treatment. In the Diner's Dilemma study, there was an overall increase from 73.5% to 77.3%. To better conceptualize the differences between treatments, we calculated the solutions to the multiple regression model – shown in Table 3.4 and visualized in Figure 3.12. These predictions imply that more interaction with the agent via control settings leads to decreased knowledge. For instance, in the Movie Recommendation study, our data showed that participants were much more likely to use the agent if control settings were available. Increased interaction with the recommendation agent could have lead to a skewed perception of what was in the movie database – for instance, a horror-movie aficionado may only see horror movies in the recommender, perhaps subtly convincing him that horror movies are the most highly rated genre (thus influencing the scores for knowledge questions 1 and 2). Figure 3.12 also supports the notion that trust is not being calibrated properly and over-trust may be a serious issue. For instance, in the Diner's Dilemma study, the low reliability recommender would exhibit "flip-flopping" behavior, where it changed its answer almost every round, making the sense-making process

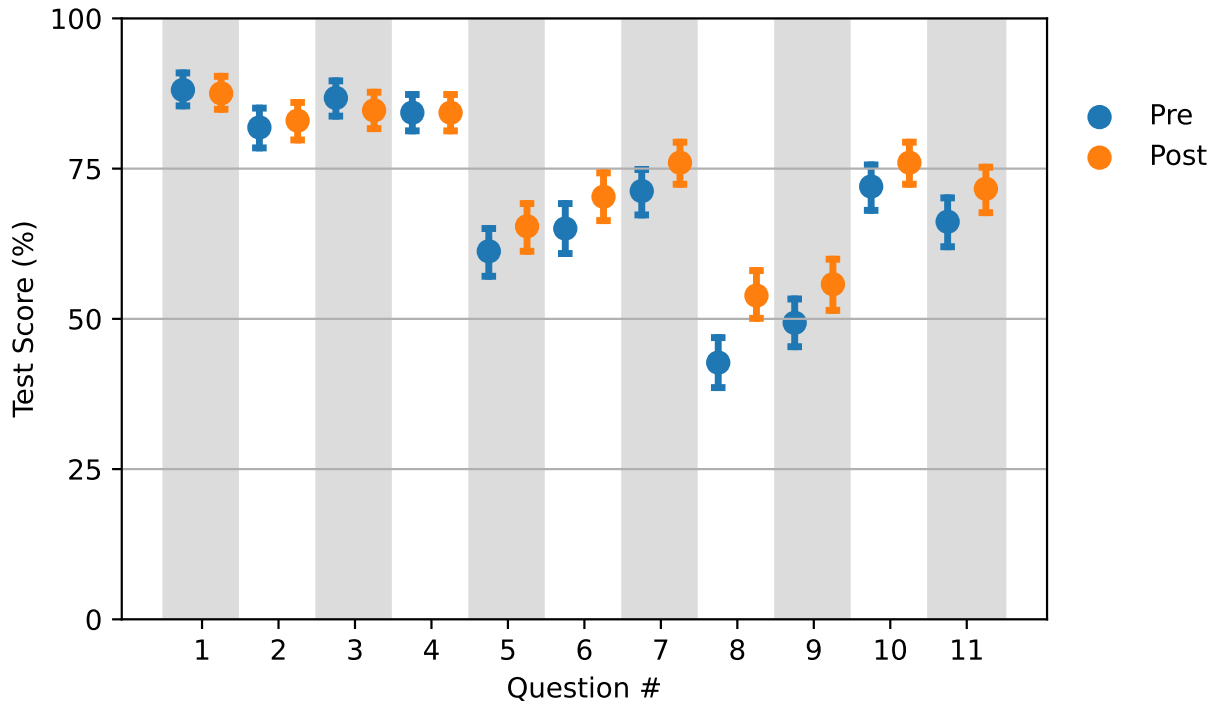


Figure 3.11: Mean correct answers for each knowledge question in the Diner study across time. Error bars indicate 95% confidence intervals.

difficult. The explanations and control features that were provided only exacerbated the issue, as the estimates for those conditions are even lower.

An interesting difference done in these tasks in contrast to the original Diner's Dilemma study conducted by Schaffer et al. [36] was in the presentation of the agent: the original study framed the system as a tool while the present studies framed the system as an agent. Upon comparison, we found that the explanations and control treatment caused decreased learning in the present Diner's Dilemma study. In fact, the best treatment for learning in the Diner's Dilemma study was the no explanations, no control settings, high reliability treatment – a treatment which had a fairly static interface (the *Dining Guru* would "lock-in" to the best answer early on and remain there). In contrast, [36] saw equivalent levels of learning in all treatments. This suggests that it may be a combination of the agent framing, combined with features that make an agent appear competent (explanations and control settings) that trigger knowledge loss.

We note that of the two studies described here, two types of invocation strategies were

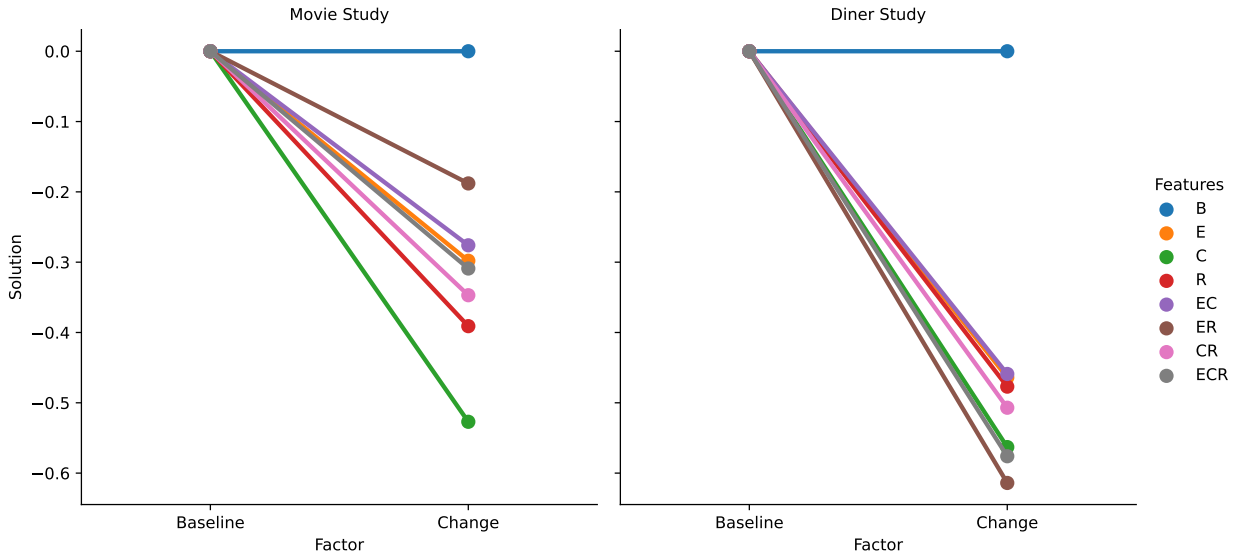


Figure 3.12: Plotted Raykov solutions for both the Movie and Diner study. The Baseline factor (blue – no explanations, no control, high reliability) was rotated to have a slope of 0, and all other factors are rotated accordingly. We can observe that the presence of all other factors worsened potential learning.

attempted: always on (Movie Miner) and on-demand (*Dining Guru*), however, other types of invocation methods and explanation strategies have been studied in expert systems research [146]. While we initially believed prompting the user for input to get them involved in the process (*Dining Guru*) or allowing the user to customize the agent's output (Movie Miner) would mitigate any bias, it has become clear that more complex interaction strategies may be required. Providing recommendations adaptively, rather than making them available all the time, may be part of the solution, but further research as well as novel methods for interaction may be required.

The effect sizes found in this study tended to be small, at most half of a knowledge unit (i.e., a question). Even if the effect sizes were large, does it really matter if humans are making correct decisions regardless of knowledge loss? We provide an argument to answer each of these two questions. First, small effect sizes can imply high cost depending on the domain. For instance, in movie recommendation, one user forming a misconception about a movie database might not be particularly devastating, however, if the user is a military command and control operative and the domain is high-risk, small policy changes can save hundreds of lives. On the other hand, having agents with lower reliabilities in high-risk domains can in-

Table 3.4: Solutions to the Raykov Change factor for each study. Bolded solutions present a larger knowledge loss than other factors.

	Movie Study	**Diner Study**
Baseline	0.0 (at mean)	0.0 (at mean)
Explanation	-0.298	-0.464
Control	**-0.527**	**-0.563**
Reliability	-0.391	-0.477
Explanation + Control	-0.276	-0.459
Explanation + Reliability	-0.188	**-0.614**
Control + Reliability	-0.347	**-0.507**
Explanation + Control + Reliability	-0.309	**-0.576**

troduce unnecessary complications. Striking a balance between maintaining agent reliability and human performance is an ongoing research challenge. Second, computational systems are not perfect and are (currently [147]) not legally liable for poor decisions, meaning for the foreseeable future humans will always be part of the decision loop. If agents make errors or fail electronically, humans will suddenly become responsible for any task they might have been automating, even if the agent is performing at a high reliability. Thus, maybe the ideal agent is not the one that attempts to complete everything perfectly (as it may fail), but the one that brings out the best of its operator, whether by helping them remain attentive or practice their domain knowledge a little longer.

3.5 SUMMARY

We conducted two user studies: the Movie Recommendation study (n = 526) and the Diner's Dilemma study (n = 529). In the Movie Recommendation study, participants interacted with an agent that provided automatic movie recommendations. In the Diner's Dilemma study, participants played multiple rounds of a variation of the iterated Prisoner's Dilemma with an agent that provided them a recommendation for their next choice. Each agent's ECR profile was manipulated to determine what agent features affect learning of domain knowledge. Analysis of this manipulation revealed that the user's knowledge about the domain can be reduced whenever they perceive an agent to be capable, whether it is through a combination of transparency, controllability, or reliability, echoing past research

on improper trust calibration in human factors. Although task dependencies and interactions exist, we observe from the Raykov change model that lowering reliability can incur learning benefits when the agent is capable as to hamper the effects of over-trust. Findings suggest that for learning, a fully transparent and controllable agent (the gold standard) may do harm to our knowledge, and that an agent with imperfections or lack of features could be a viable alternative for decision support systems.

3.5.1 Afterword

The two studies discussed have been conducted and the results published in a conference paper at the IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA) in 2019 [148]. This work received a best paper award. Other deliverables include a dataset including the results of the knowledge tests, and the knowledge tests themselves for the Movie Recommendation study and the Diner's Dilemma study. This study was done in collaboration with the University of California, Santa Barbara. This chapter has been extended from the original publication with permission from IEEE.

Table 3.5: Fitted Raykov change model coefficients for Movie Recommendation and Diner's Dilemma studies. The baseline factor estimates were predictably non-significant (treatment was independent of user knowledge). Bold p-values indicate significant regressions. * = significant at $\alpha = 0.05$; ** = significant at $\alpha = 0.01$; *** = significant at $\alpha = 0.001$

	Movie					**Diner**				
	B	**Std. Error**	**z-value**	**p(> | z |)**	**Sig.**	**B**	**Std. Error**	**z-value**	**p(> | z |)**	**Sig.**
Baseline Factor										
Explanation (E)	-0.120	0.175	-0.688	0.492		0.099	0.190	0.519	0.603	
Control (C)	0.062	0.169	0.368	0.713		0.138	0.197	0.703	0.482	
Reliability (R)	-0.141	0.171	-0.829	0.407		0.236	0.213	1.105	0.269	
EC	0.160	0.246	0.651	0.515		0.059	0.273	0.217	0.828	
ER	0.167	0.248	0.672	0.502		0.066	0.296	0.222	0.824	
CR	0.321	0.243	1.320	0.187		0.182	0.302	0.602	0.547	
ECR	-0.489	0.349	-1.400	0.162		-0.473	0.424	-1.114	0.265	
Change Factor										
Explanation (E)	-0.298	0.175	-1.707	0.088		-0.464	0.190	-2.438	**0.015**	*
Control (C)	-0.527	0.169	-3.128	**0.002**	**	-0.563	0.197	-2.858	**0.004**	**
Reliability (R)	-0.391	0.171	-2.293	**0.022**	*	-0.477	0.213	-2.234	**0.025**	*
EC	0.549	0.246	2.231	**0.026**	*	0.568	0.273	2.081	**0.037**	*
ER	0.501	0.248	2.017	**0.044**	*	0.327	0.296	1.103	0.270	
CR	0.571	0.243	2.349	**0.019**	*	0.533	0.302	1.761	0.078	
ECR	-0.714	0.349	-2.043	**0.041**	*	-0.500	0.424	-1.178	0.239	

CHAPTER 4: SIMULATED PHYSICAL AGENTS IN MEDIATING PERFORMANCE IN HETEROGENEOUS HUMAN-AI TEAMS

This chapter focuses on investigating Human-AI team dynamics in a simulated continuous task. Following the knowledge complacency study, we note that most of the Human-AI interaction research focuses on tasks that are meant to influence, as in recommend, suggest, or support a decision. However, the other side of the gamut have agents operating in physical environments, such as automation in assembly lines, unmanned military vehicles, robot-assisted surgery, among others. It becomes challenging to authentically observe any interactions between humans and agents in these domains, and for the case of varying reliability, interaction models are even harder to validate. For this, we integrate agent reliability and human individual differences in a simulated pursuit task to inform how trust, performance, and task factors interplay in a holistic model.

We discuss the design, results, and implications of a study which manipulated the reliability of agents completing a team-based continuous task alongside the human. The results support the following:

- Lowering agent reliability increases team performance without any cost to trust or situation awareness. However, it does not indicate that it increases human performance or is applicable to every task. "Good Enough" agents are situational.
- Our resulting structural equation model shows applicable similarities to the integrated model of bias and complacency formulated by Parasuraman and Manzey [14], with a key difference: trust positively predicted situation awareness – a direct contradiction to the automation bias literature.

4.1 STUDY OVERVIEW

This chapter we approach a distinct domain of human-AI teaming, where instead of the user having supervisory control of the agent in a task, the agent holds an active stake in the task by constructing a scenario unable to be completed by only the human. Work investigating different facets in human-AI interaction and human-agent teams (HATs) often

make use of experiments that operate in discrete domains: at a specific point in the task, the experiment prompts the user to make a choice to affect the outcome, often with no time pressure or accounting for other variables in the environment – variables which may significantly affect performance [149]. However, the reality is that our world is continuous, and many advances in AI take advantage of continuity in order to provide more precise operation. For instance, drone technology for aerial domains (e.g., first responders, military, aviation) often acts beyond discrete decision-making; determining whether they are operating correctly or not and what decision to make at the moment depends highly on a variety of different factors (often taking the form of continuous random variables, such as position, height, velocity, and goal), rather than a dichotomous summary of its actions. The agent's reliability can often shift from intended behavior to erratic movements, relying on internal system components to account for and correct the ill behavior. In these cases, we must ensure that established cognitive models of trust generalize towards operation in continuous spaces. As far as we are aware, no prior research has attempted to investigate human performance in HATs and perceptions of AI in continuous environments, and thus we present our main contributions.

In this study, we present an experiment that demonstrates the interplay of performance and perceptions of AI systems by varying agent reliability in a HAT operating in a continuous environment. Agents are assigned to complete a continuous pursuit task (dubbed the *Predator-Prey game* – PPG) by actively collaborating with a human. We measure performance and subjective perceptions of the agents to inform the discussion of trust and performance models in continuous tasks. This experiment explores a between-subjects design where we note changes in performance through perceptual factors and human predispositions. We aim to answer the following research questions:

> RQ4.1: How is human performance affected by varying reliability in collaborating agents in a continuous pursuit task?
>
> RQ4.2: How do human individual differences mediate perception of agents, situation awareness, and performance in continuous domains?

4.2 BACKGROUND

Teams are often defined as organizations that employ dynamic and adaptive behavior between individuals in order to achieve a common goal [150]. For many scenarios, good team performance can be the difference between success or failure of the goal. With advances in computational technologies, the use of artificial intelligence and machine learning has allowed us to spring automated agents forth to autonomy: non-living entities which have the capability to be intelligent and make their own decisions. Because of this, the fundamental structure of teaming has changed – teams can now be comprised of a combination of human members and artificial agents. With automation and autonomy becoming increasingly ubiquitous in the 21st century, human-agent teams already exist in a wide variety of domains (in research and practice), such as embodied agents for military operations [151], partially-autonomous driving [152], content recommendations [36], and algorithmic decision-making systems prominent in data analytics [92, 146], with more examples reviewed in Chapter 2. This results in a growing need to study the cooperative dynamic in HATs to allow effective and intuitive interaction for the growing usage of automation towards autonomy.

Thanks to advanced mathematical techniques in machine learning, complex and unpredictable operating environments are effectively actionable by agents. The capabilities of AI systems can often match or surpass human performance in specific tasks that utilize an agent's inherent advantages – such as speedy processing of the operating environment's data for decision-making. In other cases, the human's ability to visually perceive and adapt to the environment with high levels of judgment serve to complement the agent's processing speed [153]. Upon integration into HATs, general intuition established that the human and AI would collaborate effectively, resulting in increased performance across the board. Far from it, however, is that upon interaction with a capable AI, humans often end up over-trusting the system resulting in penalties to both individual and team performance [148, 154]. However, in these domains, the AI system is often treated as an assistant, providing decision aid to users in order to complete a task, with the human retaining the role of the primary decision-maker. However, if AI research is geared towards eventually reaching autonomy, these systems must be able to act independently to complete their task with limited to no

intervention [66]. Human-agent teaming research often addresses situations where agents hold equal or higher responsibility than human operators, but few focus on the effect of varying agent reliability (e.g., [127, 155]). The level of performance brought by these agents might significantly differ from or exceed human performance, resulting in potential over-trusting behavior, in part because the agent's performance is less likely to degrade over time [156]. In combination with a simulated scenario, we investigate what factors affect proper trust calibration, and later bring comparisons on how reliability affects human-AI interaction in this chapter's context.

4.3 SYSTEM DESIGN

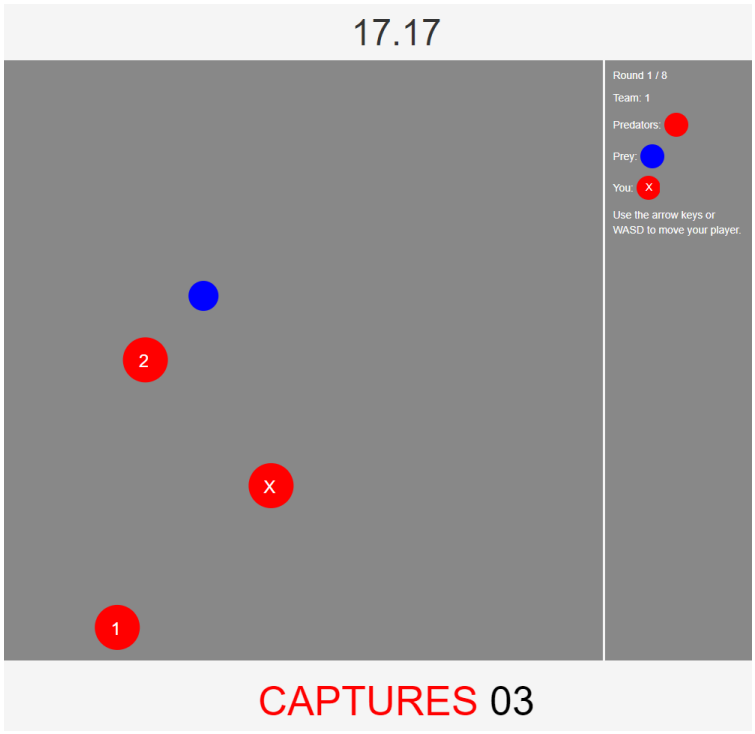


Figure 4.1: Screenshot of the Predator-Prey game. Predators (red dots) are tasked to collectively work to capture the Prey (blue dot). The participant always assumes the role of the predator marked with an *X*. Predator teammates are labeled with either *1* or *2*. The round time remaining and number of team captures can be seen at the top and bottom, respectively. Instructional information is displayed in the sidebar.

In this section, we describe the implementation of our experimental task. We had participants play the Predator-Prey game, where they are tasked with teaming up with two au-

tomated agents (predators) to "capture" (i.e., collide with) a third automated agent (prey), which is constantly evading in a continuous space. This serves to help our understanding of cooperative real-world pursuit tasks and strategy formation, as opposed to a simplified, discrete space. A screenshot of the PPG is shown in Figure 4.1.

Game theoretic-scenarios are often employed to research the performance of algorithms that drive AI behavior. Heuristic algorithms and reinforcement learning policies are often benchmarked in strategy-based games [157, 158, 159, 160, 161, 162]), with modern human-AI interaction and collaboration paradigms using game-based tasks as well [163, 164, 165, 166, 167]). However, a large amount of work focuses on scenarios where participants encounter a discrete choice to make (but not to be confused with a discrete choice experiment). For instance, a turn-based game with or against an AI may allow infinite time to respond or interact with the agent, allowing user impressions not only to be formed given characteristics of the agent, but also through time variation. Interactions such as these can be generalized with real-life tasks associated with low-risk or with no time pressure, such as movie or product recommendations. However, previous research has yet to address cases with time-critical constraints beyond supervisory control. Scenarios like these are of high interest for domains that require sophisticated embodied agents that cooperate to achieve a goal (e.g., robotics, elder care, drone control), often with little to no time to allow judgment or intervention from the human. Often, the technical advancement of autonomy supersedes human-centered uses and concerns when deploying these systems, and many have called to focus on human-centered issues when developing independent systems [8, 66]).

4.3.1 Simulation Environment

The following description of the simulation environment maps 1 SI unit to 1 virtual unit (e.g., 1 meter = 1 unit of distance in the environment). The task environment is comprised of a closed square arena of 2 m of width and 2 m of height. The players in the PPG move their circular avatar on a physics-based system by applying a force to their agent. In order to give predators and prey an equal chance to succeed, as well as to encourage the emergence of coordination among predators, the predators were made slower than the prey.

The predators had a maximum speed of 1 m/s and accelerated at a maximum rate of 3 m/s^2. The prey had a maximum speed of 1.3 m/s and accelerated at a maximum rate of 4 m/s^2. The mass of all players was set at 1 kg. The diameters of the players were 0.15 m and 0.1 m for predators and prey, respectively. Upon capture, the capturing predator and prey would knock each other back at an impulse force of 1 $\frac{m}{kg*s}$ until losing all momentum (by reducing the absolute velocity at the rate of -0.25 m/s^2). Additionally, the prey is granted 0.5 seconds of invincibility after being captured, such that subsequent captures are not counted if multiple predators capture the prey at the same time. This task was designed according to a reinforcement learning testbed presented by OpenAI [168], which has been of recent interest for investigating human-agent teams in embodied collaborative contexts [169]. A summary of this description is found on Table 4.1.

Table 4.1: Game features of the PPG. Units should be transferred to the equivalent SI unit in the virtual space.

Feature	**Specification**
Play Area	2 x 2 m (4 m^2 area)
Predator	Size: 0.15 m diameter Weight: 1 kg Acceleration: 3 m/s^2 Max Velocity: 1 m/s
Prey	Size: 0.1 m diameter Weight: 1 kg Acceleration: 4 m/s^2 Max Velocity: 1.3 m/s
Mechanics	Impulse on Capture: 1 $\frac{m}{kg*s}$ Invulnerability after Capture: 0.5 *seconds*

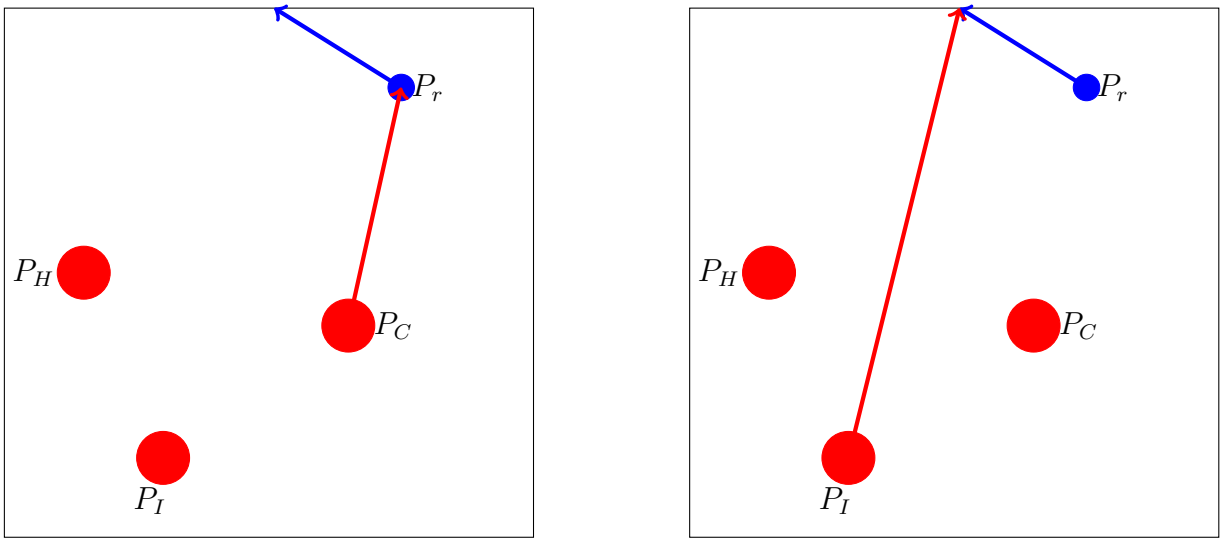
The proposed task in the PPG presents a specific case of interaction within an HAT: the agents are fully autonomous, have no line of communication with the human, and operate in a fully continuous space. Example real-world scenarios that this task would effectively model are interaction with fully-autonomous unmanned aerial vehicles (UAVs) [170]), human-swarm interaction [171]), or team-dynamics with drones as part of the Internet of Battlefield Things (IoBT) [172]); these are situations where supervisory control is either

not desired nor feasible to attain. There has been longstanding interest in using continuous pursuit models to investigate emergent collaborative behavior between humans and agents [35, 169, 173]). Although prior work has focused on establishing the parameter space [174]) and algorithmically identifying cooperative behavior from agents [175, 176]), the question of the viability of these agents in operation with respect to models in human-AI interaction remains unaddressed. Ultimately, this study aims to shed some light on long-established models in human-AI interaction and HAT in a novel context, which will allow engineers and designers incorporating AI-based systems in these domains to be aware of their benefits and shortcomings.

4.3.2 Modeling Agent Behavior

The agents investigated in this study can be defined according to the Levels of Automation (LOA) continuum established by Sheridan and Verplank [177]). The PPG implements agents at the highest level (level 10), granting them complete independence from human intervention. According to the definition of autonomy, these agents are fully autonomous and effectively operational only in the continuous pursuit domain. Thus, this shifts away from supervisory control and moves towards equal collaboration, which some prior work has addressed (e.g., [178, 179]). Kessler et al. note that as the role of automation grows from passive heuristics (e.g., recommendations, monitoring) to near-autonomous systems that guide their own behavior, the trust processes we draw to trust another human being might also be transferred to machine systems, as we perceive them to be competent and self-sufficient in the context of the task [13].

A heuristic rule set powered the agents' decision-making process in the PPG. Their behavior was governed by a procedure involving multiple geometric calculations per second to determine a target position toward which to move. For our experiment, agents calculated a new target position once every 0.25 seconds (4 Hz) – we refer each calculation as a *step*. Each predator was assigned a distinct strategy (either *chaser* or *interceptor* behavior), and the prey was assigned an evasive strategy. The predator chaser strategy attempts to close the distance to the prey at every step; i.e., the target position is always the position of the



(a) Chaser strategy.

(b) Interceptor strategy.

Figure 4.2: A single frame of the agent predators' heuristic calculations. The agents are labeled P_C for the chaser and P_I for the interceptor.

prey. This is a rudimentary form of pursuit with an easy-to-establish a mental model for its behavior. The agent calculates a difference vector from the target position to their current position, and it moves in the direction of the vector. The predator interceptor strategy accounts for the prey's position and velocity in order to intercept it at some point in the near future, as it is found to be an optimal strategy for pursuers in differential games [180]. The prey strategy calculates a set of candidate points on the edge of the play area to move towards based on the positions of the predators. The candidate points are generated by creating a triangle using the predator agents as vertices, followed by drawing an line that crosses the midpoint of each side of the triangle, that intersects with the bounds of the arena – this results in the farthest midpoints between every pair of predators, as the prey will attempt to maximize distance between itself and all predators. This results in 6 candidate points: the selected point is then the farthest from the predators based on the squared sum of the Euclidean distance from the predators to the a point. A visualization of the prey's heuristic can be seen in Figure 4.3.

In perfect scenarios, agents have the potential to respond perfectly to the environment, but it is often far from reality. For instance, robotics research establishes that frequent errors often plague robots, even after much investment and research to make them reliable [181].

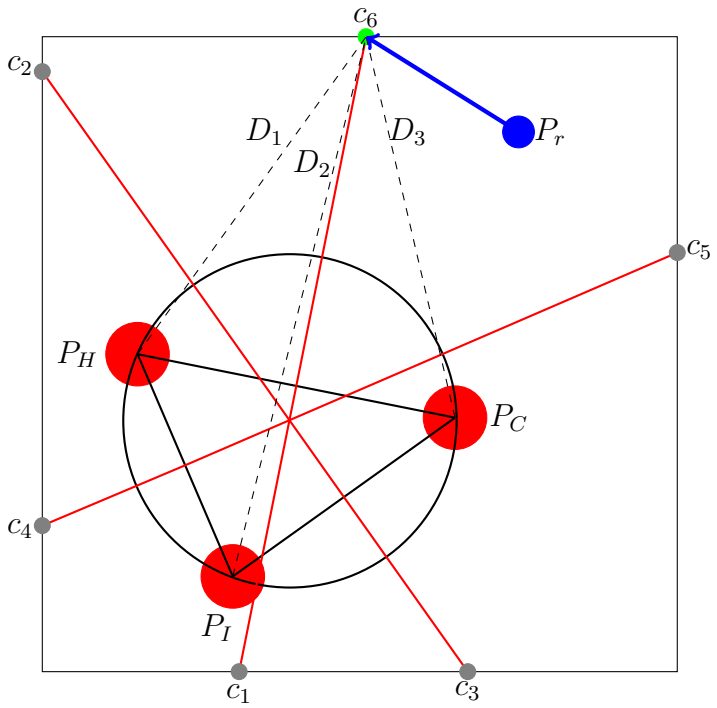


Figure 4.3: A single frame of the prey's heuristic calculation. At this step, c_i are the candidate points, and c_6 has been selected as the target point by the largest sum of D_i.

Likewise for computational systems, limitations in information or processing power can lead to unexplained and sudden errors [133]. Because errors in an AI system can manifest in different magnitudes and frequencies [181], we investigate a particular occurrence of error that hampers (but does not eliminate) the AI system's ability to complete their objectives.

For a continuous pursuit-evasion game, the optimal solutions are often formulated to require a pursuer to reach a certain target point in the action space under a certain time constraint (e.g., before a target escapes), after which the conditions have changed and require recalculation of said target point [180, 182]. By introducing error that completely hampers the functionality of the AI system (e.g., by making it non-functional or non-responsive), we trivialize the proposed research questions, as prior work has widely addressed the relationship between faulty systems and subjective attitudes [10, 61, 183]. Instead, we opted to add error that introduces erratic behavior without preventing the agent from reaching their calculated target points. Such behavior is reminiscent of "juking" movement seen in gridiron football, albeit used from an offensive perspective.

We control this behavior by using a Markov Decision Process (MDP) to have agents behave according to two states: Clean and Noisy. Our independent variable – reliability – is the set of transition probabilities that govern which state the agent remains in most of the time. In the Clean state, the agent moves towards the calculated target position without any perturbation following its appointed strategy. In the Noisy state, the agent's target position is perturbed by adding a random vector with a magnitude of 0.8. The random vector is recalculated every time a new target position is generated, such that when the agent is in the Noisy state, it is constantly thrown off-course yet still following the trajectory of the target position. The magnitude of 0.8 was selected to contain the perturbed target position within the arena. The transition probabilities were selected by using Monte Carlo simulations to calculate the amount of time an agent would remain in a given state.

4.4 EXPERIMENTAL DESIGN

4.4.1 Independent Variable

The independent variable of interest for these experiments is the reliability of the predator teammates. We define the agent's reliability as its potential capability to retrieve information from the environment and perform as designed. For instance, an agent with 80% reliability is akin to a real-world robot with sensors that function correctly 80% of the time. Prior research relating automation bias and complacency with human trust in automation indicates that reliability of the automation is a strong predictor of whether a human remains cognitively focused on the task at hand [10, 127]. As discussed, the reliability is manipulated by changing the transition probability of the MDP such that we have distinct frequencies of transitioning and remaining in a specific state, resulting in distinct overall behavior. A visualization of each condition's MDP is shown in Figure 4.4. Thus, we define two different levels of reliability manipulated in this study:

Reliability

- High: Remains in the Clean state for a large amount of time (self-transition: 98%), and jumps into the Noisy state at a probability of 2%.

– Low: Remains in the Clean state for some amount of time (self-transition: 70%), but jumps into the Noisy state at a higher probability of 30%.

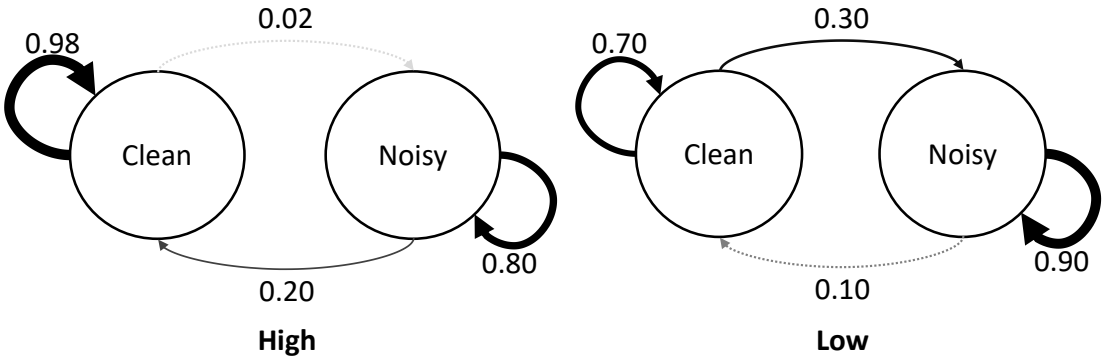


Figure 4.4: MDP transition probabilities for the High condition (left), and Low condition (right). The thickness and darkness of the arrow represents the probability of transition.

4.4.2 Dependent Variables

We measure performance, situation awareness, and subjective predispositions and attitudes towards automated agents through survey interventions. Our selected metrics to record from our task closely match prior research done in the human-agent teaming domain, as consolidated in a literature and meta-research review by O'Neill et al. [184].

Performance

Performance is measured by tallying the amount of captures by the participant and captures by the team. As the prey is allowed an intangibility period after a capture to prevent repeat captures, the tally does not include captures while the prey is intangible. Performance is the main dependent variable analogous to success in real-life scenarios. The resulting performance of an agent will be a function of its reliability, the environment, and its interactions with other agents. All positional, input, and agent decision data is recorded for replay and re-simulation, resulting in multiple time series apt for analysis.

Situation Awareness

We employed a Situation Awareness Global Assessment Technique (SAGAT), often defined as the gold standard in measuring situation awareness [51, 185]. A SAGAT consists of interrupting the current task, disabling all interfaces and hiding all relevant artifacts that would provide knowledge about the scenario, and asking the participant to recreate the situation from memory. For the PPG, during a specified round at a random time, the game freezes and all players are removed from the play area. Then, the participant is tasked to recreate the position and direction of all players by dragging icons into the play area. This SAGAT measures perception of data and the scenario, often referred to as Level 1 SA [50, 51]. After participants complete the queries, they are allowed to resume the remainder of that round, with that data omitted when analyzing performance. An overview of the situation awareness probe can be seen in Figure 4.5.

To measure SA, we establish 2 equations to quantify the amount of awareness from the participant's response (R) to the truth (T), i.e., the state of the game when the task froze. Positional SA (SA_p) is defined as the Euclidean distance from the participant's positional response to the truth, divided by 2.83 (which is the maximum distance achievable in a 2 m by 2 m area), complemented with 1:

$$SA_p(R,T) = 1 - \frac{\sqrt{(T_x - R_x)^2 + (T_y - R_y)^2}}{2.83} \tag{4.1}$$

Directional SA (SA_d) is the cosine similarity from the participant's vector response to the truth, normalized to a unit value:

$$SA_d(R,T) = \frac{\left(\frac{R \cdot T}{\|R\|\|T\|}\right) + 1}{2} \tag{4.2}$$

Thus, as both measures approach 1, situation awareness is maximized. Both SA_p and SA_d are averaged to calculate a final value for situation awareness:

$$SA = \frac{SA_p + SA_d}{2} \tag{4.3}$$

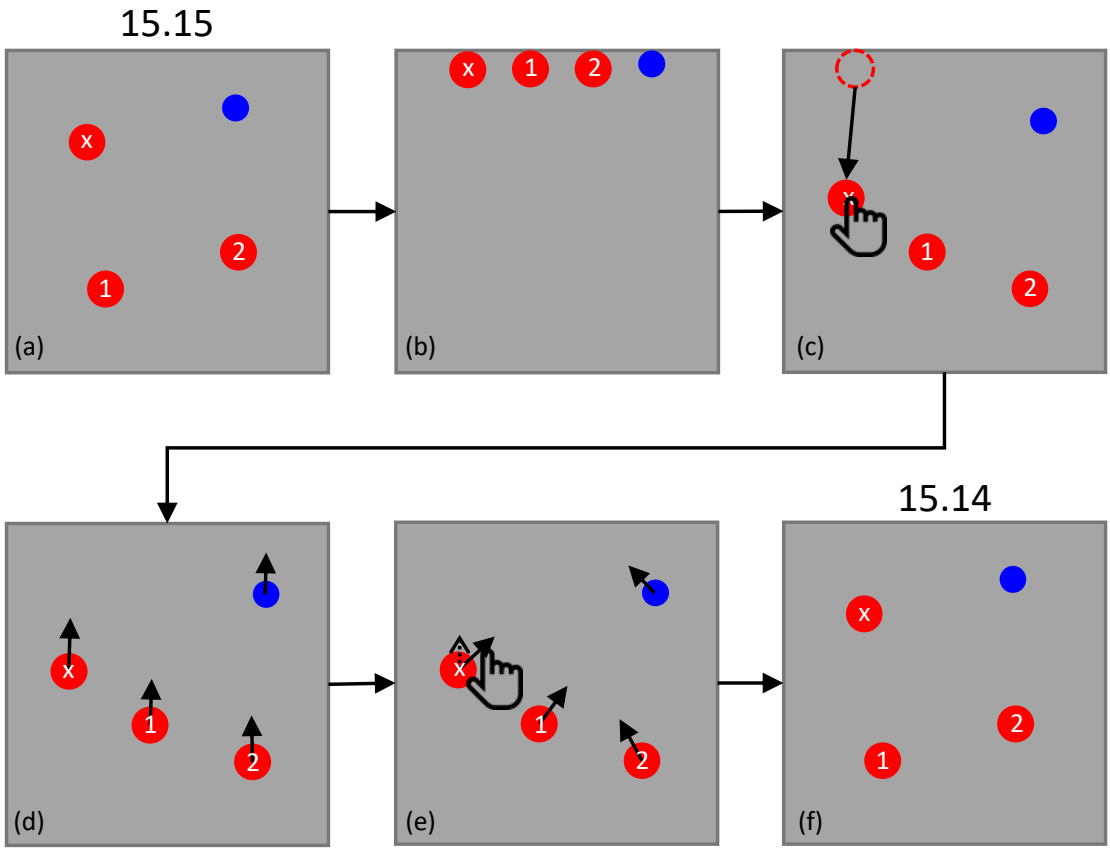


Figure 4.5: Situation awareness freeze probe done mid-round. (a) Participants play the game until a random time in the trial. (b) At the start of the probe, all players are moved at the top, and participants are prompted to (c) move the players to their last known position (as seen in (a)). (d) Arrows then appear in the position of all players, and participants are prompted to (e) drag the arrows in the directions toward which they were moving. (f) After completing the probe, participants are allowed to resume the round.

Survey Intervention

We used subsets of multiple validated surveys in order to measure relevant individual differences and human perception to model attitudes of trust on the agents. These surveys were abridged by selecting items with the highest factor loadings per survey factor to reduce the time needed to complete the experiment, as we had strict time limitations due to the nature of crowdsourcing payouts. We divide our surveys into two categories by utility: predisposition instruments and perception instruments. The selected questions and factor loadings are outlined in Table 4.2.

Predisposition Instruments. The predisposition instruments are comprised of the Automation-induced Complacency Potential (AICP) scale and the adapted Propensity to Trust Technology (aPTT) scale.

The AICP scale measures a participant's tendency towards sub-optimal monitoring patterns through 2 factors: workload alleviation and frequency of monitoring [69]. The highest factor loadings for the AICP was included in its manuscript, resulting in 2 questions.

The aPTT scale measures a participant's tendency to trust technology, modified to include language explicitly referring to "automated agents" rather than technology [186]. According to Jessup et al., using the specific language of "automated agents" allows the measure to predict behavioral trust apt for the PPG. The aPTT did not include factor loadings, thus we selected items that were unique and non-congruent in nature (e.g., "Automated agents are reliable" and "I rely on automated agents" are conceptually similar, so only one of them was included). Both of these scales measure a participant's bias to trust or distrust technology, which serves useful for predicting potential issues with trust calibration. We expect both scales to positively correlate the participants' propensity to trust automation and their perceived trust of the automated teammates.

Perception Instruments. The perception instruments are comprised of the Intrinsic Motivation Inventory (IMI), Trust in Automated Systems scale (TAS), and the NASA Task Load Index (NASA-TLX).

The IMI focuses on task evaluation, measuring interest and enjoyment, perceived competence, effort and importance, and pressure and tension [187]. Designed to be reworded to contextualize the inventory for the task, questions were modified to make them relevant to the PPG (e.g., "I felt pretty skilled when cooperating with this team" instead of "I felt pretty skilled at this task"). The IMI was given to measure changes in motivation, as has been known to accurately predict performance and engagement in game-based tasks [192]. The IMI factor loadings were given by a confirmatory factor analysis [188], resulting in 4 questions.

The TAS scale measures how trustworthy the participants perceived the system that they just interacted with – in our case, the automated teammates [189] – to be. The TAS

factor loadings were given by a confirmatory factor analysis [190], resulting in 2 questions. The questions in this scale were rephrased with the automated teammates as the object of reference. We expect the observed trust measured by this scale to be related to the previous aPTT.

The NASA-TLX is a prevalent and strongly validated tool to measure perceived workload throughout a task [191]. Since complacency might relate to the amount of workload a participant perceives, we expect to find a correlation between reliability and perceived workload. Additionally, any variance demonstrated from predicting complacent behavior using reliability could be clarified using workload.

Additionally, a demographic survey was employed to collect participant information, including their age, gender, race/ethnicity, education, and experience with video games (on hours per week). Game experience was collected to model any performance variance due to familiarity with game-based tasks.

4.4.3 Latent Growth Modeling and Structural Equation Modeling

In order to answer our research questions, we turn to latent variable modeling as a method to investigate hidden relationships between measures and extract valuable information from the variance often ignored in traditional inferential statistics approaches such as ANOVAs [193]. In addition to finding any effects given by varying reliability, we aim to discover any changes in participants' perception towards the agents and situation awareness, as valuable information is found when noting changes as the participants interact with the agents [32].

Latent growth modeling establishes two factors that capture differences between groups and over time, often referred to as the *Intercept* factor and the *Slope* factor. For clarity, we refer to group differences as the *Base* factor and growth differences as the *Change* factor. The Base factor determines whether there is a difference between observed measures in groups, whereas the Change factor determines if there is a difference in the trajectory of the observations over time between groups. This nomenclature is similar to the Raykov change model presented in Chapter 3. Therefore, two groups may begin at the same quantification of a given observation (e.g., trust), and then diverge over time. This allows us to determine

not only if but how much attitudes and perceptions of the agents change through the trials.

Furthermore, we use structural equation modeling (SEM) to determine mediations through variables. Most effects found in HAT models cannot be observed in a vacuum, as there is a wide variety of interdependence between human characteristics and individual differences that interplay and affect performance and SA in any given task. Guided by literature, we hypothesize an initial SEM and iterate the model by modifying pathways and gauging the strength of the relationships between variables, aiming to find a pathway from reliability to performance. The initial SEM consisted of the following associations:

- Reliability drives the performance of the automated agents. Thus it will directly affect team performance (2 automated agents and 1 human operator) and will correlate with individual performance.

- Reliability affects the perceived trust by the operator, as trust and reliability are known to be correlated in prior research [45]. As reliability is the independent variable that can be controlled, it directs towards trust.

- Propensity to trust technology (measured by aPTT) predicts the estimated trust by the human to an agent, thus driving the demonstrated trust measured by the TAS [186].

- Complacency potential (measured by AICP) also drives the perceived trust by the human, as over-trusting automation usually leads to the presence of complacent behavior [10, 14, 61]. Thus, if there is varying complacency potential, we expect it to drive perceived trust.

- Propensity to trust technology and complacency potential are correlated through trust (i.e., over-trust) [14, 45].

- Perceived trust affects motivation, under the rationale that trusting automation may lead to less incentive from the operator to perform (i.e., the value of their contribution is lessened by the capability of the automation, "The agent is doing well, I think I'll sit back and allow it to do it's job") [194].

- The operator's motivation and complacency potential drive their individual performance in the PPG, as combinations of these factors lead operators to retain certain mental states during the task (e.g., absentminded vs. disengaged).

4.4.4 Procedure

Participants were recruited through Amazon Mechanical Turk[4]. Participants selected the associated Human Intelligence Task (HIT) from AMT and signed an initial consent form. Participants were then redirected to VolunteerScience[5] to complete the PPG. Participants first filled the demographic survey, read instructions on how to play the PPG, and then completed 2 practice rounds (no recorded data) followed by 20 recorded rounds of 30 seconds each. Both the predisposition and perception instruments were administered upon completion of rounds 2, 10, and 18. The situation awareness probe freeze was administered at a random interval between 10 and 20 seconds into rounds 4, 10, and 16. Participants were allowed 4 rounds (+2 practice rounds) to interact with the predator agents before the first situation awareness probe freeze, as any changes in situation awareness would not be reflected immediately upon interaction, but requires a certain amount of time for complacency to impact the participant's cognitive state. Reliability was treated between-subjects: participants completed the entire PPG with one level of reliability, where both predator teammates were set to the same level. The experiment flow is visualized in Figure 4.6.

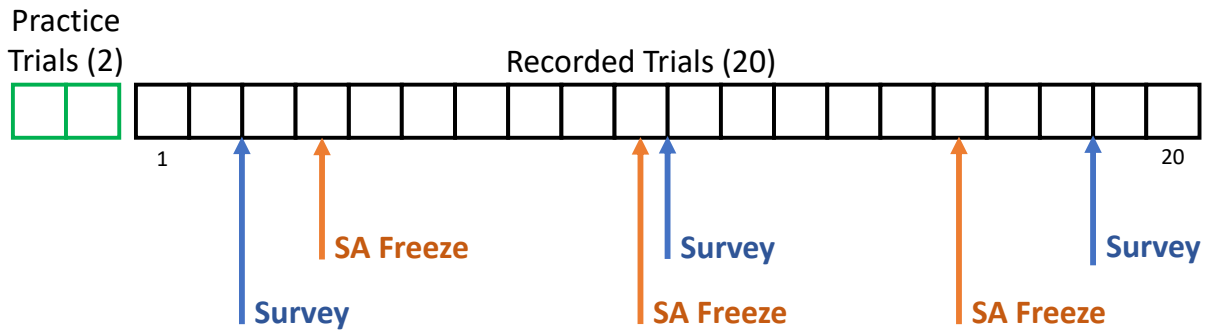


Figure 4.6: Experiment flow of the PPG experiment.

[4]https://www.mturk.com/
[5]https://volunteerscience.com/

4.5 RESULTS

4.5.1 Demographics

104 participants completed the PPG. Demographics are reported in Table 4.3. In summary, 60% of participants were male, 46% of participants were in the age range of 25 to 34, 65% finished a 4-year college education, and 35% played videogames for 2 to 4 hours a week. All participants were compensated 10 USD for completing the experiment. The collected data was checked for satisficing and completion, resulting in no records dropped.

4.5.2 Inferential Statistics, Behavior, and Growth Modeling

Performance

We compare the performance of both the individual participant and the whole team across the two reliability conditions. Since the number of captures is count data (i.e., non-negative integers with a low tendency of centrality), we conduct non-parametric statistical inferences from participants playing the PPG. The score of every trial is treated as a single data point. Plotted scores across trials and condition are visualized in Figure 4.7.

We conducted a Mann-Whitney U test to determine significant differences between conditions in individual performance and team performance. For individual performance, the test revealed no difference between the High and Low conditions (μ: 1.45 vs. 1.38, M: 1 vs. 1, U = 555210, n = 2080, p = 0.27). For team performance, the test revealed that there was a statistically significant difference between the High and Low conditions (μ: 3.5 vs. 3.87, M: 3 vs. 4, U = 485150.5, n = 2080, $p < 0.001$).

We compare the performance of each predator (including the participant) within the team per condition. In the High reliability condition, a Kruskal-Wallis test and follow-up Dunn's test indicate that each predator's performance was significantly distinct ($\chi^2(2, 3119) = 82.93$, $p < 0.001$; all pairwise comparisons: $p < 0.05$). The highest performing member was the participant, followed by the Interceptor, with the Chaser trailing last. In the Low reliability condition, a Kruskal-Wallis test found no significant differences in the performance of each

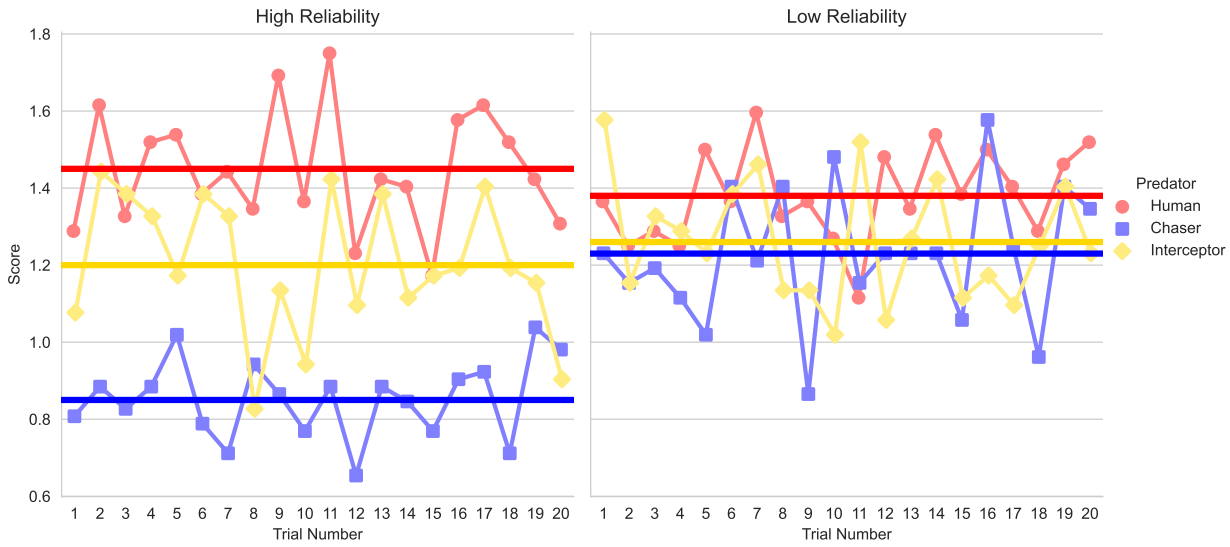


Figure 4.7: Plotted scores for every member of the predator team across trials. Each point in the x-axis is the mean of all scores for a given trial. The horizontal lines denote the average for the predator according to the legend. Comparing the High to Low conditions, the Chaser predator substantially increased its performance (0.85 → 1.23 captures; blue). Non-significant effects were found in the Human predator's performance (1.45 → 1.38 captures; red) and the Interceptor predator's performance (1.2 → 1.26 captures; yellow).

member of the team ($\chi^2(2, 3119) = 0.15$, p = 0.92).

To ascertain the contribution of each agent predator, we conduct Mann-Whitney U tests across reliability conditions for each predator. The Chaser had a significant increase in its performance in the Low condition (μ: 0.85 vs. 1.26, M: 1 vs. 1, U = 436695, n = 2080, p $<$ 0.001), increasing its contribution by 0.41 captures per trial. In contrast, the Interceptor had a close but non-significant increase in its performance in the Low condition (μ: 1.20 vs. 1.26, M: 1 vs. 1, U = 515826.5, n = 2080, p $<$ 0.001), increasing its contribution by 0.06 captures per trial.

Movement Behavior

We visualize the positional data of all players through heatmaps in Figure 4.8. This denotes how frequently a player was located in a given area of the arena. From here we draw several important comparisons. First, we see the static strategy the High reliability Chaser exhibits: the prey's best strategy to escape was to circle around the area away from

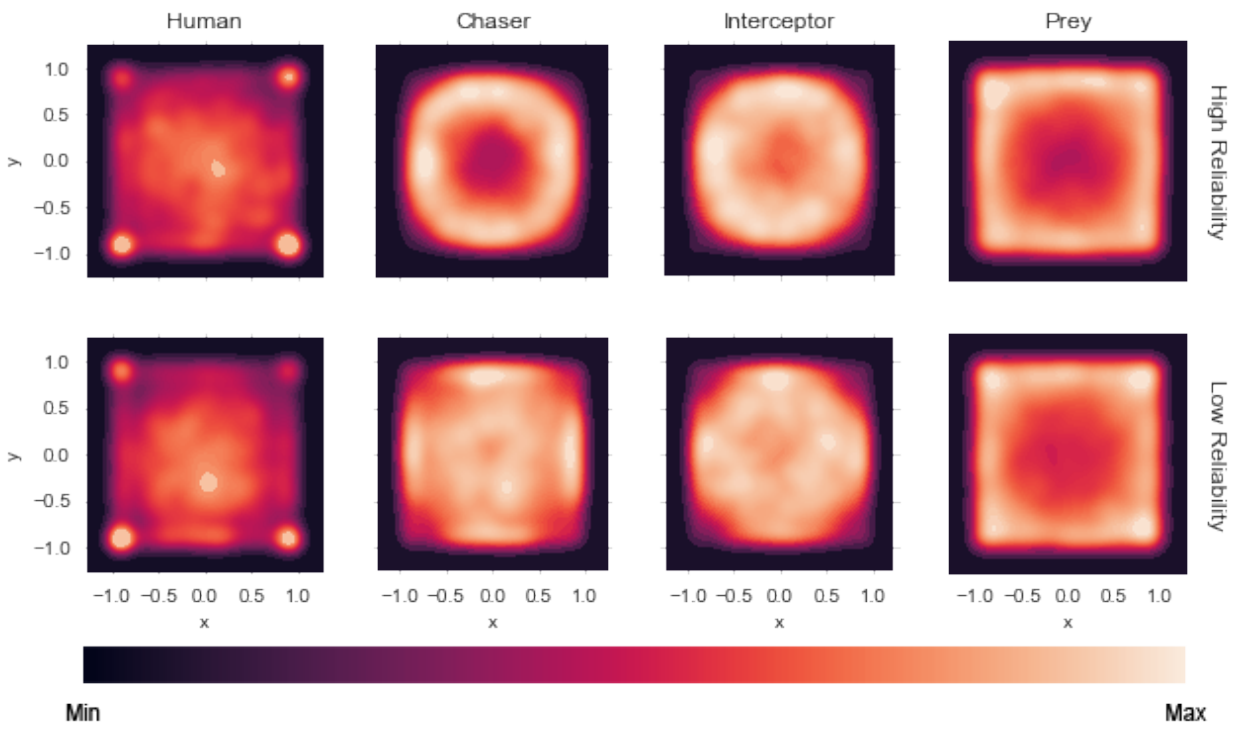


Figure 4.8: Position heatmaps for all players (n = 104) and conditions in the PPG across all trials. Positions were recorded at 4 Hz, resulting in nearly 1,000,000 data points. The top row covers the High condition, the bottom row covers the Low condition. The gradient represents the frequency of positions.

the predators, and the Chaser followed through by using the same circling as defined by its strategy. It is only when noise is added to the Chaser where it chases all across the arena, similar to an Interceptor (both in High and Low reliability). Interestingly, the human has a high incidence of remaining in the corners of the area, potentially demonstrating a "wait and attack" strategy, where it allows the other agents to draw the prey to the corner of the arena where the human awaits for a quick capture. Second, one argument to be brought forth is that the more coverage an agent has, the higher number of captures it can potentially achieve, as exemplified by the Chaser and Interceptor. The number of Chaser and Interceptor captures increased in the Low condition, which aligns with the increased coverage shown in their respective heatmaps. Similarly, coverage subtly decreases for the human in the Low condition (areas beyond radius 0.5) along with its performance (although this was found to be non-significant). Finally, the distinct coverages from the High and Low

reliabilities led to the prey to adopt a distinct strategy, as the Prey could mostly avoid the predators by staying adjacent to the edges of the arena, but once coverage was increased, the prey adjusted by considering new routes closer to the center of the arena.

Survey Interventions and Situation Awareness Probe

As discussed, survey responses and the SA probe were analyzed using latent growth curve modeling [140, 193] in order to shed light not only on subjective differences between the High and Low conditions but on how these differences change over time as participants interact with the automated agents. Therefore, we analyze each survey question with a growth curve model, along with aggregating the questions with their respective instrument. A summary of the growth curve model results can be found in Table 4.4.

For the predisposition instruments, we had expected and found no change between conditions or through time (i.e., non-significant Base and Change factors) since complacency potential and propensity to trust technology is a predisposed attitude rather than being affected by the intervention. For the perception instruments, there was no significant change per condition or through time, demonstrated by the non-significant p-values in both the Base and Change factors. Responses over time are plotted in Figure 4.10 and Figure 4.11.

For all positional and directional situation awareness in the growth curve model, there were no significant differences between conditions or through time for all PPG players, except for prey directionality in the Base factor ($p < 0.05$). Summary statistics of the probe results along with the p-values of the growth curve model are found in Table 4.5, and responses over time are plotted in Figure 4.12.

4.5.3 Structural Equation Model Fit

Prior research has established numerous complex relationships between system reliability, human performance, and individual differences that are difficult to investigate with inferential techniques alone. To gain a clearer understanding of these relationships, we developed a structural equation model (SEM) to examine factors that affect performance and situation awareness in continuous pursuit tasks. The final SEM is visually represented in Figure 4.9,

which outlines relationships between exogenous and endogenous variables and includes fit and regression coefficients. The SEM was constructed using R 3.5.2 with lavaan 0.6-7 [195].

To aggregate individual and team performance for SEM analysis, we used the median of a participant's trials. We also aggregated survey items through parceling to account for reverse scoring and common factors, which can improve estimates and model fit by reducing measurement error through aggregation [196]. To control for the issue of multiple comparisons (as every model fit is a new hypothesis), we employed the False Discovery Rate (FDR) correction, which is recommended for exploratory SEM analysis [197]. Specifically, we controlled the FDR through the Benjamini-Hochberg procedure [198] with Q set at 0.15, a standard choice that balances the need for statistical power with the need to control the false positive rate. In total, we tested 46 models for the PPG and effects.

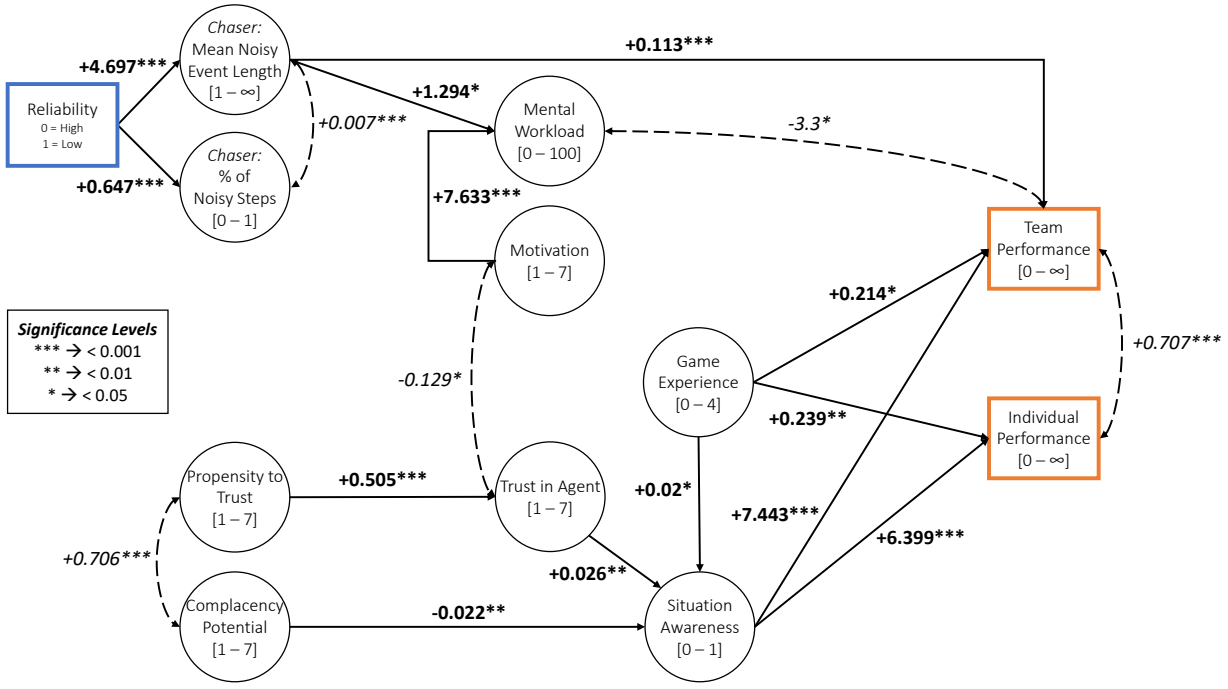


Figure 4.9: Fitted SEM for the PPG experiment. Model fit: $N = 104$ with 36 free parameters. $RMSEA = 0.037$ ($CI = [0, 0.074]$), $TLI = 0.991$, $CFI = 0.994$, over null baseline model, $\chi^2(78) = 1363.238$. Solid lines indicate significant regressions, and dashed lines indicate correlations. Numbers represent the fitted coefficient along with its level of significance.

4.6 DISCUSSION

We situate our findings by first discussing the results from inferential statistics and growth curve modeling, and we use those to guide our exploratory SEM analysis. Next, we discuss the implications of these effects in the context of continuous cooperative multi-agent systems, and finally, we answer our research questions and highlight future research challenges.

4.6.1 Statistics and Growth Curve Modeling

We begin by noting the performance of the predator team during the PPG. We find that individual human performance is slightly higher when the other team members are not impacted by noise ($\mu_H = 1.45$, $\mu_L = 1.38$). Such behavior could indicate the human's ability to predict future positions of their teammates and optimize their possibilities to capture the prey. Albeit the mean difference between the individual performance of the two conditions is 0.07, the calculated effect size is 0.91 – i.e., if we generate all ordered pairs between participants in the High condition and participants in the Low condition, in 91% of the pairs, the High condition participant outperforms the Low condition participant.

More than individual performance, however, we observe that the Chaser agent had far superior performance in the Low condition compared to the High condition, in contrast to the reduced effects from the human and the Interceptor agent. We hypothesize that by adding a random perturbation vector to the target point of the Chaser agent during its Noisy state, its actions were subtly altered to match the Interceptor behavior (as the Interceptor strategy is theoretically equal to the Chaser, with a consistent target position offset accounting for the prey's position and velocity). Additionally, the perturbation could have led the predators to adopt unpredictable behavior that became difficult for the prey to account for. By having 2 interceptors in a team – with interceptor behavior being a very viable strategy in a continuous pursuit game [199, 200, 201] – we note the difference in team performance across conditions. This hypothesis aligns with the position heatmaps visualized in Figure 4.8; to reiterate, a Noisy Chaser had more coverage across the arena, and had a similar coverage to an Interceptor agent, overall forcing the prey to cover new routes. Further analysis can be conducted through histogram analysis, as previous work has shown that the

distribution of the distance between predators and prey is the most characteristic feature of agent strategy (according to a Principal Component Analysis), and is a strong indicator of Chaser vs. Interceptor behavior within teams [202]. It is interesting to note how unintended noise can bring benefit to the agent, as it begins exploring alternative strategies that may help it achieve its goal at a more unorthodox pace, much akin to research that suggests that a perfect, transparent agent may not be the perfect teammate [19, 203, 204]. How well this generalizes to real-world scenarios is highly up to debate, as the perturbations introduced in the Noisy state, albeit random, were controlled. The complexity of real-world environments can make it difficult to predict the outcomes of introducing unexpected perturbations (e.g., drones navigating during a windy day versus a sandstorm that completely blocks visibility). We should be cognizant of potential over-trusting that can occur if agents indeed receive unintended benefits from a lowered reliability.

We opted to take measurements for the survey instruments at the early, middle, and late interactions with the agents to determine any changes through time, as interactions and impressions with automated agents are often conditioned by the worst recalled interaction [38, 44]. Thus, if any negative interactions occurred with the Low condition agents, it would rapidly reflect on the items. For the predisposition instruments, we found no significant change through time, as we expected that beliefs already held before interaction with the agents would not change with a short amount of interaction (it took participants around 15 to 20 minutes to complete the experiment, with 10 minutes directly interacting with the agents). In a similar vein, the perception instruments show that participants did not perceive a change in motivation or trust across conditions or through time. We take note of the most significant factors to inform the construction of our exploratory SEM.

The SA growth curve model demonstrates consistent SA across all probes, conditions, and time, except for the awareness of prey direction. This indicates that participants had adjusted to an information processing strategy that was consistent across 20 trials, regardless of agent reliability. This may suggest that participants did not perceive the Low reliability agents to be any different from the High reliability agents or require a different strategy to keep track of their state (such as vigilantly monitoring a vehemently unreliable agent). The exception present in the prey direction probe where the Base factor was significant indicates

a distinct level of initial situation awareness by completion of the first probe (High SA_d: 0.594 vs. Low SA_d: 0.469, $p < 0.05$). This may be explained by an initial adjustment of expectations with respect to behavior from the predators in the Low condition: the perceived erraticness of the participants' teammates could have led to higher vigilance towards them while reducing awareness of the prey's direction. At future probes, the situation awareness of Low condition participants increased linearly, indicating that they had adjusted to such behavior, whereas the High condition participants saw a linear reduction, possibly indicating over-trust and, consequently, loss of situation awareness. However, the change factor indicates non-significance, even though the change is opposite per condition (most likely due to the small magnitude of the change slope: -0.031 vs. 0.042). Future research may address specific points of interests and objects of focus that lead to varying situation awareness.

4.6.2 Structural Equation Model Effects

We use our SEM to outline relationships between factors and answer our research questions. As this exploratory model aims to inform future research, we encourage readers to not take the described model as a definitive description of the human cognitive process with respect to reliability and performance, but to consider the mediated effects when formulating future research.

An initial point of discussion is how well each MDP executed to produce distinct reliability behavior between the two agents. After all, if at every step the agent would have alternated between the Clean and Noisy states, behaviorally, the agent would have continued its original trajectory after correcting for the incorrect state with non-perceptible error. Hence, we hypothesized that a longer stay in the Noisy state would produce erratic behavior perceived by a human. Thus, we considered both the percent of Noisy steps (i.e., how much time the agent was in the Noisy state), and the average length of a Noisy event (i.e., how long the agent remained in the Noisy state before transitioning out to the Clean state). We found that for both the Chaser and Interceptor agents, the High condition had 10% of total Noisy steps, with a mean Noisy length of 4 steps (1 second), whereas the Low condition had 75% of total Noisy steps, with a mean Noisy length of 8 steps (2 seconds). As the surveys and

probes indicate, the slight difference in the length of the Noisy event might be imperceptible to humans but results in distinct behavior from the agents – behavior which leads to altered performance. The SEM shows that the MDP was able to manipulate the agents' behavior ($p < 0.001$), with the caveat that since the Interceptor behavior did not influence any further variables downstream, it was removed from the model. As expected, the percent of Noisy steps and the average Noisy length are correlated.

As to control a portion of the performance variance given participants' experience with game-based entertainment (since our task is a game-style task), we categorized participants based on their video gaming experience. This allows us to use their experience with such entertainment as a proxy to engagement, as we expect participants to remain interested throughout the task due to familiarity. Engagement with the PPG led to higher individual and team performance ($p < 0.05$). This strengthens other factors in the model to explain the remaining variance.

The length of the Noisy state and the participant's current motivation affected the mental workload they perceived ($p < 0.05$). The erratic behavior directly influenced the team's performance positively ($p < 0.001$), as a direct effect from the increase in Chaser performance as discussed earlier. However, a negative correlation existed between a participant's mental workload and the performance of the team ($p < 0.05$). Mental workload may have been increased by the nature of the task, as the time pressure to capture the prey along with erratic behavior of the predator teammates may have affected coordination [149, 205].

The predisposition instruments accurately predicted participants' attitudes on the reliability of the AI, validating the instruments' utility in continuous tasks. A participant's propensity to trust accurately predicted their amount of trust in the agents after the PPG ($p < 0.001$). Complacency potential predicted their level of situation awareness ($p < 0.01$), validating over-trust and automation complacency paradigms discussed in the literature [14, 79]. Both of the predisposition instruments strongly correlated with each other, of almost 1 point on each scale ($p < 0.001$).

Trust and situation awareness were strong mediators between individual characteristics and observed performance. Complacency potential may describe an initial state of situation awareness (lowering SA more as the predisposition to complacency is higher) but is simul-

taneously increased by trust ($p < 0.01$) and engagement ($p < 0.05$). Interestingly, this is an opposite effect than what is usually found in SA-based studies and metrics, where often an increase in trust hampers awareness [52, 66] due to improper trust calibration [38]. We are unable to determine whether the increase in situation awareness due to trust is driven by trust calibration since the model did not find a pathway from the agents' reliability to perceived trust. Another explanation is given by the nature of the task: the participant's active role in the PPG may have influenced them to exercise higher awareness due to their dependency on the other agents to capture the prey. That is, if the participant was alone with the prey, it would become incredibly challenging for them to capture a prey due to the asymmetry of their capabilities. Much prior research often places participants hierarchically superior to automated or decision support systems, where little time pressure is found for participants to make a decision, often deferring their judgment to an AI. Further investigation about the benefits of the nature of the task is warranted, as trust, engagement, and complacency potential only describe a small amount of the effect ($R^2 = 0.142$), albeit very significantly. A possible explanation is that video gaming experience allows for increased multitasking performance, which allows for more effective situation awareness [206]. On the other hand, we found that motivation and trust in the agents are negatively correlated ($p < 0.05$), echoing an effect similar to social loafing in working groups (i.e., the less one is motivated, the more they will trust the AI out of inactivity) [207, 208].

Ultimately, performance was largely mediated by situation awareness ($p < 0.001$) and engagement, explaining almost half of the variance in both individual and team performance ($R^2 = 0.428, 0.411$). Breaking down the variance contributors, we find that for individual performance, the variance was explained by game experience (26%) and situation awareness (74%). Similarly, for team performance, the variance was explained by game experience (18%), how long agents were in a Noisy state (18%), and situation awareness (64%). The behavior of the agents influenced participants' perceived workload, but performance in the task was predicted by situation awareness. A correlative bridge exists between motivation and trust, yet further research is required to solidify these relationships with respect to continuous contexts. In light of this model, we return to our research questions.

4.6.3 Answering Research Questions

RQ4.1: How is human performance affected by varying reliability in collaborating agents in a continuous pursuit task?

According to the resulting SEM, human performance was not directly affected by agent reliability, but reliability was mediated by predispositional and perceptual factors that ultimately influenced resulting performance. A caveat does exist with team performance, as slight noise added to the Chaser agent resulted in increased performance due to its hypothesized similarity to an Interceptor, largely due to the nature of the task facilitating the use of optimal strategies. This resulted in a more capable teammate and a higher share of captures for the agents, all while being minimally perceptible to the human operator.

RQ4.2: How do human individual differences interplay with performance, situation awareness, and perception of agents in continuous domains?

Constructs such as complacency, engagement, motivation, and social loafing were observed in Human-Agent Teams in game-theoretic continuous tasks. We observed that multiple individual characteristics (such as propensity to trust and complacency potential) affected performance, mediated by trust and situation awareness. In continuous tasks where the reliability of a system may not be defined through a number or probability, the system's behavior can provide a strong signal for the operator to make a judgment on whether the agent is behaving correctly. Depending on the situation or scenario, agents may be subject to a varied amount of noise (either internal or external), which may result in inaccurate perceptions of the agents' efficacy. This work has shown that a system may be highly erroneous due to system failure or environmental constraints yet can be perceived as effective as an agent operating correctly.

4.6.4 Limitations and Future Work

This study is not without its limitations. Determining trust during a continuous task is a challenging prospect that research should focus on addressing, as we develop embodied AI agents that are active operators of a task holding a certain amount of responsibility and, consequently, risk. We were limited in measuring trust at discrete points, thus losing

granular information on what events can cause trust to change during the task. Additionally, this particular task (continuous pursuit) ends up being abstract and specific for a certain kind of utility, and thus factors discussed here may change depending on the domain, task, or situation. As this work is meant to be exploratory, research can focus on solidifying and replicating certain effects in the context of continuous tasks, as the growth of our AI-based systems now often operate in these domains. For instance, the perceptible difference of AI performance is a common effect found when controlling for reliability [36, 148], yet this threshold is not well defined. Furthermore, this study only considers the highest level of automation, where the agent presents no explicit cues or explanations on their behavior. It becomes important to understand how can autonomous agent teammates can present their system state and processing to users, with some recent proposals on how to achieve this [209].

4.7 SUMMARY

We performed a study demonstrating how lowering reliability of an agent in a continuous pursuit task (n = 104) can affect how people perceive, trust, and perform in human-agent teams. We additionally emphasize the use of mediation models to connect cognitive factors often studied in a vacuum, and present a holistic view on how agent performance can ultimately affect individual performance, similar to prior models established for automation bias. Albeit this work uses an abstract task to demonstrate the studied effects, many operative environments may have similar structures or goals as the ones presented in the PPG. In the future, replacing the task with well-defined goals will allow us to verify the effectiveness of modeling cognition in HATs through mediation analysis. Within the bounds of this domain, we see that reducing reliability increased team performance with minimal cost to trust or situation awareness, suggesting that a "Good Enough" agent may provide better cohesion than a perfect agent. However, simulation behind a screen will not be enough to model true physical interactions between humans and agents.

4.7.1 Afterword

The study discussed has been conducted and the results published as an article in the Journal of Cognitive Engineering and Decision Making, under a special issue in Human-AI Teaming in 2022 [210]. Other deliverables include a dataset including the observed behaviors of the participants and agents, along with survey responses. All software developed belongs to the U.S. Army Combat Capabilities Development Command Army Research Laboratory, but may be replicated for any future studies.

This research was sponsored by the U.S. Army Combat Capabilities Development Command Army Research Laboratory. The views and conclusions contained in this dissertation are those of myself and of close collaborators, and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. This chapter has been adapted from the journal publication with permission from SAGE Publishing.

Table 4.2: Survey interventions for the Predator-Prey Game. Items marked with (R) are reverse scored. The factor loadings are given by the related literature; n/a indicates no factor loadings were found, thus multiple items were used.

Code	Item	Factor	Loading
Predisposition Instruments			
	Automation-induced Complacency Potential [69]		
aicp1	If life were busy, I would let an automated system handle some tasks for me.	Alleviation	0.78
aicp2	It's not usually necessary to pay much attention to automation when it is running.	Monitoring	0.67
	Adapted Propensity to Trust Technology [186]		
aptt1	Generally, I trust automated agents.	Trust Propensity	n/a
aptt2	Automated agents help me solve many problems.		n/a
aptt3	I don't trust the information I get from automated agents. (R)		n/a
aptt4	Automated agents are reliable.		n/a
Perception Instruments			
	Intrinsic Motivation Inventory [187, 188]		
imi1	I enjoyed playing with this team very much.	Interest	0.80
imi2	I think I am pretty good at playing with this team.	Competence	0.97
imi3	I tried very hard while playing with this team.	Effort	0.85
imi4	I felt pressured while playing with this team.	Tension	0.72
	Trust in Automated Systems [189, 190]		
tas1	My team is dependable.	Trust	0.88
tas2	I am wary of my team. (R)	Distrust	0.87
	NASA Task Load Index [191]		
tlx-me	How mentally demanding was playing with this team?	Mental Workload	n/a
tlx-ph	How physically demanding was playing with this team?	Physical Workload	n/a
tlx-te	How hurried or rushed was the pacing of this team?	Tension	n/a
tlx-pe	How successful were you in capturing the prey with this team?	Competence	n/a
tlx-ef	How hard did you have to work to capture the prey with this team?	Effort	n/a
tlx-fr	How discouraged, stressed, and annoyed were you while playing with this team?	Tension	n/a

Table 4.3: Resulting demographics for the PPG experiment. Categories with 0 participants in all conditions were not included.

	High	**Low**
Sample Size (n)	52	52
Age		
18 - 24	2	4
25 - 34	25	24
35 - 44	16	13
45 - 54	6	6
55 - 64	3	4
65+	0	1
Gender		
Male	32	30
Female	20	22
Race/Ethnicity		
White	47	45
African American	2	2
Native American	2	2
Asian	0	2
Hispanic/Latino	1	1
Highest Education Completed		
High School	8	6
2-year College	3	3
4-year College	34	33
Graduate/Professional	10	7
Hours/Week Playing Videogames		
< 1	0	4
1 - 2	14	15
2 - 4	20	16
4 - 7	11	11
> 7	7	6

Table 4.4: Survey results and growth curve modeling summary. The bolded represents the parceled aggregation of the scale.

	High M (MAD)	**Low M (MAD)**	**Base p(> | z |)**	**Change p(> | z |)**
Predisposition instruments				
aicp1	6 (0.99)	5 (1.03)	0.34	0.58
aicp2	5 (1.38)	5 (1.39)	0.61	0.73
aicp	**5.5 (1.05)**	**5 (1.09)**	**0.40**	**0.55**
aptt1	5 (0.87)	5 (0.97)	0.50	0.67
aptt2	5 (0.96)	5 (1.14)	0.83	0.48
aptt3	5 (1.44)	5 (1.33)	0.90	0.87
aptt4	5.5 (0.85)	5 (1)	0.87	0.22
aptt	**5 (0.66)**	**5 (0.7)**	**0.97**	**0.43**
Perception instruments				
imi1	6 (0.91)	6 (0.88)	0.37	0.77
imi2	5 (1.04)	5 (1.13)	0.85	0.60
imi3	6 (0.71)	6 (0.99)	0.17	0.06
imi4	5 (1.2)	5 (1.28)	0.82	0.85
imi	**5.5 (0.62)**	**5.5 (0.66)**	**0.85**	**0.73**
tas1	5 (0.91)	5 (1.04)	0.27	0.40
tas2	5 (1.32)	5 (1.18)	0.23	0.33
tas	**4.5 (0.72)**	**4 (0.76)**	**0.09**	**0.21**
tlx-me	70 (13.71)	75 (12.56)	0.50	0.09
tlx-ph	65 (18.25)	70 (22.7)	0.73	0.54
tlx-te	70 (12.8)	75 (12.95)	0.67	0.73
tlx-pe	70 (15.64)	70 (16.7)	0.85	0.99
tlx-ef	75 (11.15)	75 (10.7)	0.65	0.47
tlx-fr	65 (22.52)	65 (22.31)	0.88	0.24

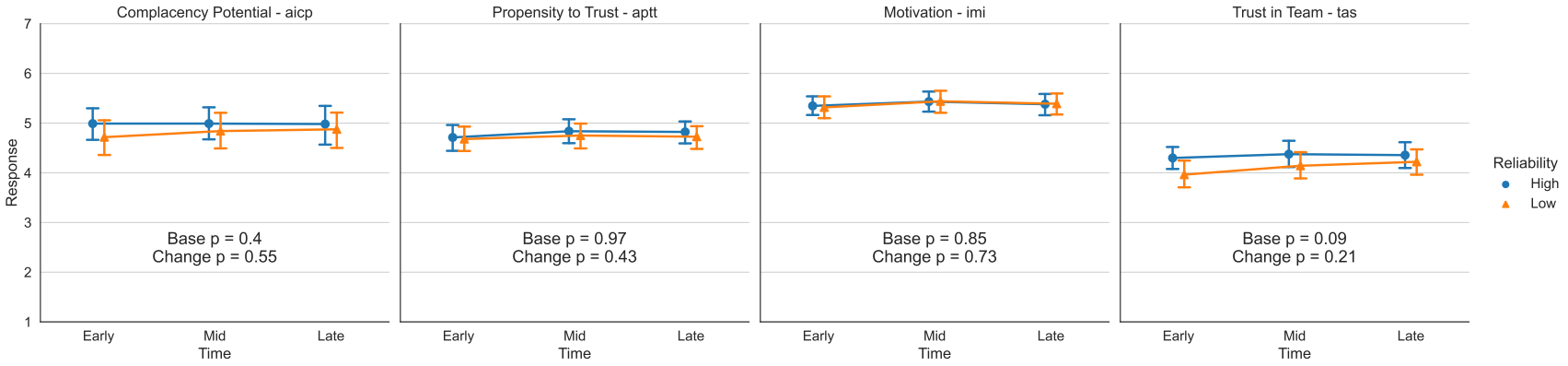


Figure 4.10: Plotted Likert response average with error bars over time for parceled survey interventions (except NASA-TLX responses). The p-value of the growth curve model factors (Base and Change) are annotated below. 1 = Lowest, 7 = Highest.

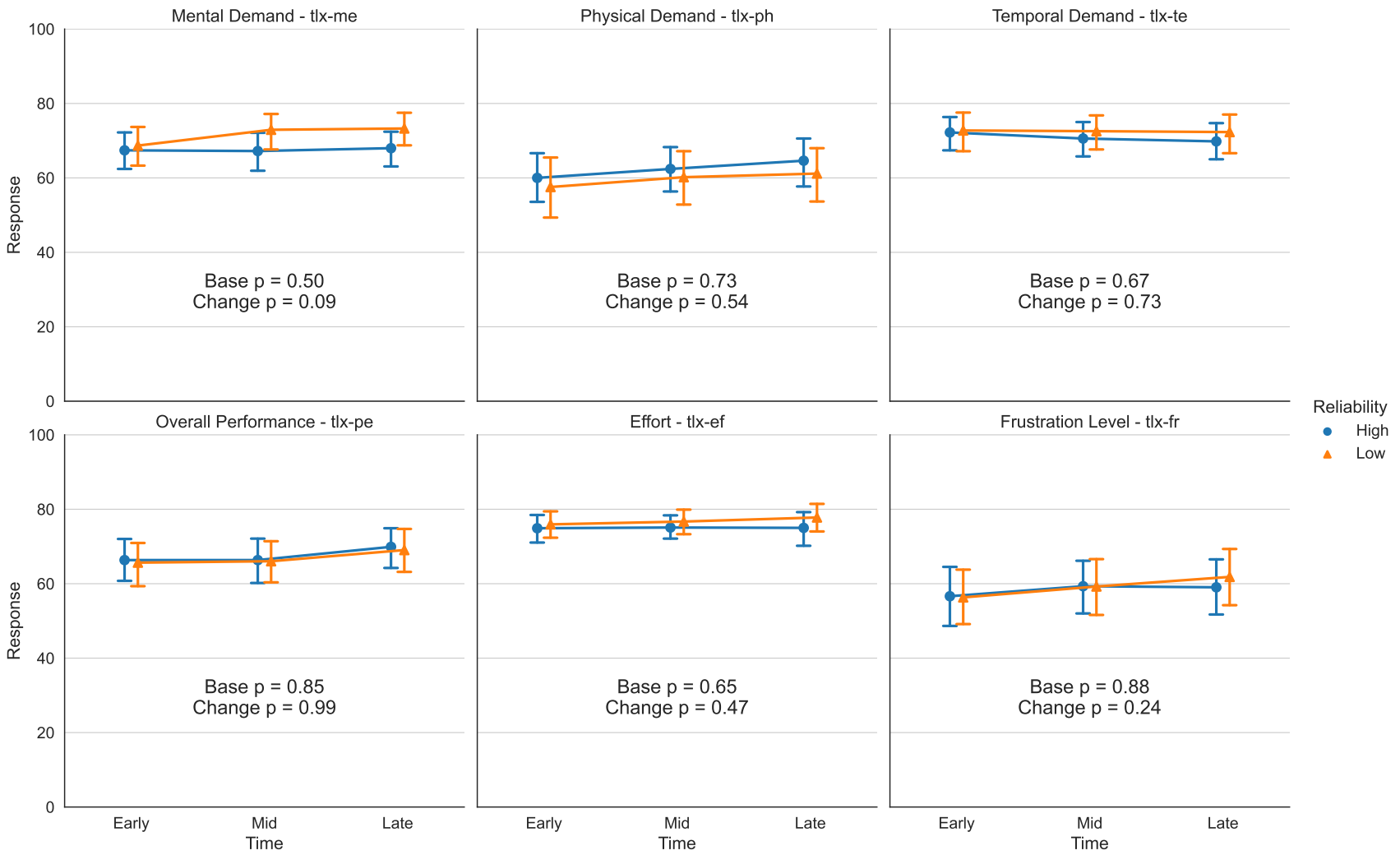


Figure 4.11: Plotted response average with error bars over time for NASA-TLX items. The p-value of the growth curve model factors (Base and Change) are annotated.

Table 4.5: Situation Awareness probe results and growth curve modeling summary. Values were rounded to 2 significant figures. H = High condition, L = Low condition. Bold indicates significant differences.

Probe	**Cond.**	**N**	**Mean**	**StDev**	**Min**	**Max**	**Skewness**	**Kurtosis**	**Base p(> |z|)**	**Change p(> |z|)**
Position										
Human	**H**	156	0.68	0.18	0.17	0.98	-0.52	-0.55	0.81	0.41
	L	156	0.71	0.18	0.20	0.98	-0.59	-0.34		
Chaser	**H**	156	0.68	0.16	0.27	0.98	-0.26	-0.55	0.84	0.89
	L	156	0.68	0.16	0.24	0.98	-0.26	-0.63		
Interceptor	**H**	156	0.70	0.14	0.32	0.97	-0.24	-0.64	0.33	0.23
	L	156	0.70	0.16	0.33	1.00	-0.37	-0.54		
Prey	**H**	156	0.69	0.19	0.17	0.99	-0.60	-0.42	0.54	0.2
	L	156	0.70	0.19	0.13	0.99	-0.64	-0.26		
Direction										
Human	**H**	156	0.59	0.33	0.00	1.00	-0.38	-1.13	0.3	0.73
	L	156	0.54	0.34	0.00	1.00	-0.13	-1.37		
Chaser	**H**	156	0.58	0.36	0.00	1.00	-0.39	-1.43	0.47	0.29
	L	156	0.58	0.35	0.00	1.00	-0.37	-1.31		
Interceptor	**H**	156	0.56	0.36	0.00	1.00	-0.23	-1.52	0.69	0.35
	L	156	0.54	0.33	0.00	1.00	-0.14	-1.32		
Prey	**H**	156	0.56	0.35	0.00	1.00	-0.25	-1.45	**0.049***	0.13
	L	156	0.51	0.34	0.00	1.00	-0.05	-1.43		

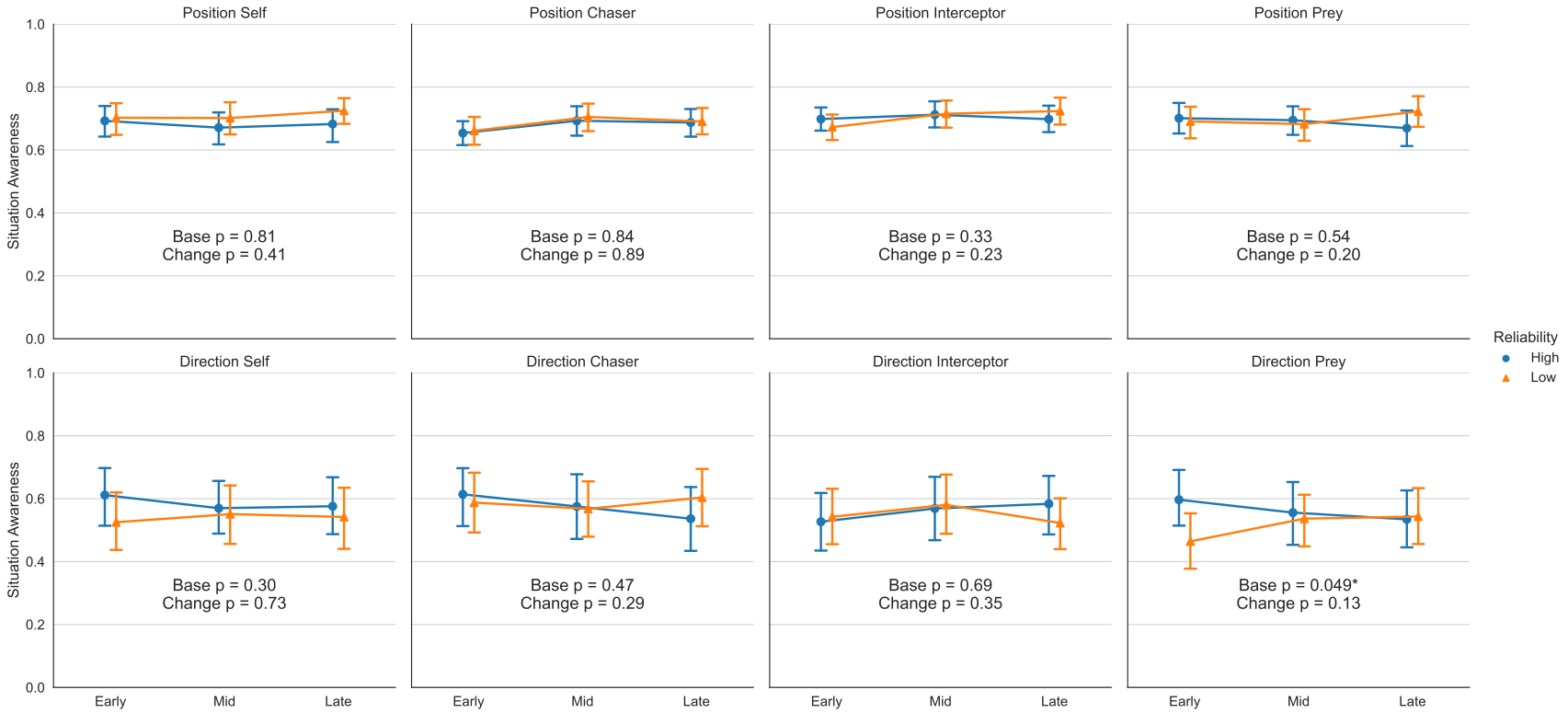


Figure 4.12: Plotted situation awareness response average with error bars over time for the situation awareness probe. The p-value of the growth curve model factors (Base and Change) are annotated.

CHAPTER 5: EMBODIED AGENTS IN HUMAN-ROBOT COLLABORATION ON DECISION-MAKING

This chapter now jumps from the virtual interactions we observed in Chapters 3 and 4 to a physical domain to investigate how does reliability affect trust calibration in the presence of embodied interactions. The other half of the human-AI interaction domain lies within agents that have a physical entity, often represented in the form of robots. Prior research has demonstrated that robots engender different responses than virtual agents [34, 211], often positive [81] due to social embodiment [212]. Agents can have differing representations (intangible vs. tangible) or the interaction itself can be distinct (non-embodied vs. embodied). Researching robotic applications in the real world often require interactions beyond making a selection in a computer screen.

We hypothesize that embodied interactions and agent tangibility will be distinct from usual interaction paradigms used in human-AI research, where a participant is tasked to complete an objective where the agent "lives" behind the computer monitor. To this end, we designed a physical task to be completed under 2 different interaction paradigms: tangible and non-tangible. We then control agent reliability and investigate how people discern different agent reliabilities within an embodied setting. This chapter then completes the scope of this dissertation: investigating imperfect agents in distinct domains where an AI can operate in – virtual, simulated, and embodied.

We discuss the design, results, and implications of a study which manipulated the reliability and representation of a robot completing a supported decision-making task alongside a human. The results support the following:

- Varying reliability has significant effects on decision-making and the adherence to automation, replicating prior research found on automation reliance and compliance.
- Embodied agents polarize the levels of trust between the human and the system.

5.1 STUDY OVERVIEW

The studies presented in Chapters 3 and 4 rely on a research paradigm that requires participants to interact with the agents through a screen (i.e., the agents are intangible). These tasks, however, do not reflect all the possible AI systems that humanity may encounter as computation progresses. Outside of the virtual realm and into our physical world, agents own a tangible embodied representation and complete tasks that require interaction with other physical entities, whether human or agent. Robots are often used in item assembly, packing and packaging, earth and space exploration, medical surgery, reconnaissance, mass production, and safety. Human-robot interaction research investigates the interaction dynamics between humans and robots (i.e., agents with a physical representation) and have long focused on trust and how humans react to different robot capabilities, features, and representations [16, 58, 76, 81, 213, 214, 215]. However, there has been a limited number of studies focusing on measuring trust through manipulation of robot representation – whether they are tangible or non-tangible – and much less through reliability. Setting up experiments and interactions with physical robots requires special considerations for budgeting and safety. Virtual reality (VR) has gained popularity in the recent years for its capacity to simulate and invoke embodied physical interactions in a safe, high-fidelity environment – an approach that has found high success in research methodology [216, 217]. In this vein, we are able to simulate and study the physicality of these interactions in a reliable manner by additionally using motion tracking to capture the full-bodied interaction. The key gap here is a missing comparison on how trust models hold between virtual and physical domains, and how embodiment overall affects trust calibration and decision-making. We investigate the following research questions:

RQ5.1: How do embodied interactions affect performance and trust in a collaborative decision-making task?

RQ5.2: How is capability perception of different agent reliabilities affected by embodied interactions?

5.2 BACKGROUND

The findings presented in this dissertation suggest that the effectiveness of different system features, such as reliability and transparency, varies depending on the context in which they are used. In the realm of human-robot interactions, the concept of embodied agents is crucial. An embodied agent is a physical entity that is controlled by an agent that is embedded within it, whether in the physical or virtual world. These physical entities are often referred to as "bodies," and they can have various appendages, such as heads or hands, which enable them to interact with their surroundings through the use of sensors and motors [218, 219]. It is worth noting that these "bodies" do not have to be human-like; they can be constructed with mechanoid or animal-like representations [220].

Embodied agents have a significant impact on users compared to non-embodied agents [221], influencing users' trust [222], empathy [223], and attention [224]. Humans tend to generally prefer humanoid and anthropomorphized agents, viewing them as more trustworthy and conscientious [152, 220, 225], resulting in improved trust dynamics, such as trust repair [29]. System designers often choose to embody their agents for user interaction in the form of avatars, such as chatbots or recommender systems.

However, there is conflicting research on whether embodiment positively affects the formation of trust in human-agent teams. For example, Herse et al. found no effect of embodiment in a simulated high-risk task, even when presented with urgency [86]. Mollahosseini et al. argued that embodiment helped participants understand facial expressions from a robot but was only effective for a limited set of emotions (e.g., anger), and could not capture other major emotions or contextual speech patterns [87]. These inconsistencies may be due to distinct research approaches and task parameters, indicating that the success of embodiment depends highly on the context in which it is deployed.

An embodied agent can vary by its tangibility [133]. Tangible agents have a physical body or shell that can be interacted with, such as robots, drones, or the physical speaker device of a virtual assistant (e.g., an Amazon Echo). Intangible agents, on the other hand, exist only in virtual space, and may have the same visual features as their tangible counterparts. The strong influence of social cues, cultural norms, differing expectations of the system,

(a) PC version (Screen condition). (b) VR version (VR condition).

Figure 5.1: Two screenshots of the Warehouse game for this study. (a) presents the game as a standard computer game interactable with a mouse and keyboard. (b) presents a sideview of the game scene. Participants are able to grab and interact with objects in the virtual environment.

and levels of acceptance of anthropomorphic representations [226] are unique to tangible embodied agents, and can affect interactions with them. For instance, humans have been shown to have different psychophysiological responses (i.e., stress, anxiety, happiness) when dealing with tangible robot swarms compared to an intangible simulation [227]. Human-Agent Teams are particularly sensitive to the effect of tangibility, as there is a higher level of individual interaction between humans and agents. While humans tend to be more polite and trusting of tangible embodied agents, other research suggests that the effects of trust may not last longitudinally [228]. Intangible agents, which are less expensive and more prevalent, continue to be studied and offer potential advantages for certain applications.

To bring the breadth of scenarios discussed in this dissertation to a close, a signal detection theory (SDT) task will be devised in the presence of an embodied robot. The robot's tangibility will be emulated with the use of virtual reality, and the study will aim to observe how reliability in decision-making tasks interacts with the embodiment of an agent. The goal is to inform whether the effects of reliability found in previous studies can replicate towards physical domains, and to verify whether using virtual reality to emulate embodied interactions can be used as a viable approach to human-robot interaction studies.

5.3 SYSTEM DESIGN

For this study, we use SDT to design a task that models uncertain decision-making, akin to prior studies conducted in human-AI teaming models and human factors [68, 114, 229]. When a task is uncertain, users often use tools of support and re-assurance to ensure they are making the correct decision, especially during high-risk scenarios [229]. We designed *Warehouse*, an interactive SDT-based decision-making game where participants must identify true cases (i.e., true positives and true negatives) and avoid false cases (i.e., false positives and false negatives). The premise presented to the participants is that they are quality assurance workers at a warehouse that ships polygonal trinkets (see Figure 5.2) and need to ensure that a given package contains a specific trinket according to an order. Given a package and an order, they must choose either send the package out to the customer or reject the package back to the warehouse. To determine the contents of the package, participants must use a scanner that displays the trinket inside; however, the scanner is impacted by a high amount of visual noise, requiring focused visual acuity to determine the contained trinket. The human-AI collaboration component is provided by a robotic teammate that is able to determine the package's contents and provide a recommendation on whether the package should be sent or rejected.

To investigate both the effects of agent reliability and embodiment for joint decision-making, the interactive features in *Warehouse* can be modified. For reliability, the robot teammate can be controlled with different levels of reliability, changing the quality of its recommendations. For embodiment, *Warehouse* was built as a PC game executable and a virtual reality experience for standalone VR headsets. By leveraging the virtual reality experience (i.e., stereoscopic vision, spatial audio, and embodied interactions), we bring participants into a more immersive version of the task, where we hypothesize embodiment may change behavior and trust patterns under the same task paradigm.

5.3.1 Reward Structure and Instrumentation

In this scenario, we denote the positive case to be the positive ideal outcome for the customer (i.e., there are no issues with the package and it can be sent without hesitation).

This is contrary to other SDT-based studies, where the positive case indicates an alarm or an issue (e.g., the package contains the wrong trinket). Nevertheless, defining the positive case leads to a symmetric negative case, leaving the analysis of outcomes unchanged. The participant may choose to rely on their own visual recognition abilities or use the robot's recommendation to make a decision. Furthermore, time pressure is introduced to place an ongoing urgency, much like real-world workers must fulfill quotas. At every order, participants can earn points if they complete an order correctly (true cases – true positives and false positives) or lose points if they fail to do so (false cases – false positives and false negatives). Initially, the participant can receive a maximum reward or a minimal penalty, respectively. As time progresses, the reward and penalty are scaled accordingly to de-incentivize participants from delaying their decision. Let $f(t)$ be the reward penalty applied after time t. At the start of the order, the participant will have a reward of $+X$ for fulfilling the order, and a penalty of $-Y$ for making a mistake. As time progresses, the reward will reduce to $+X - f(t)$ for a correctly fulfilled order, while the penalty increases to $-Y - f(t)$. We expect that participants will use and adhere to the robot's recommendation in their search to be speedy and accurate workers [230], although they may still choose to rely on their own visual acuity for their decisions. This reward structure then translates into a bonus for their given compensation; prior research has shown this to be apt at motivating participants to make sound and appropriate decisions [26, 204]. Table 5.1 outlines the distribution of rewards and penalties per trial during the Warehouse game.

Table 5.1: Reward matrix for the Warehouse game.

		Recommendation			
		Send		Reject	
Package	Match	TP		FN	
		$t = 0$	$t = 10$	$t = 0$	$t = 10$
		+5	+1	0	-2
	Mismatch	FP		TN	
		$t = 0$	$t = 10$	$t = 0$	$t = 10$
		-1	-5	+2	0

We implemented the scanner with a high amount of visual noise to introduce the feeling of uncertainty in the participants, and push them to adhere to the robot's recommendation. The implementation of the scanner is as follows: for a given trinket inside a package, we generate a random image using a combination of the trinket images (Figure 5.2a-e). The generated image is synthesized using the following procedure: we define a percentage X, where $X\%$ of pixels will be taken from the package's trinket (the signal), and $(100 - X)\%$ of pixels will be taken from all other trinkets at random (the noise). For a completely unidentifiable image, X should be set at 20%; this results in an image where 20% of pixels are of each of the 5 trinkets (20% from the signal, 80% from the 4 noise), making it impossible to distinguish the signal, as every trinket image is equally represented in the output. To make the signal recognizable with enough visual acuity, we generated multiple images with a random percentage X ranging from 74% to 80% (for examples, see Figure 5.2f-j). This results in a very subtle flicker when viewing the images in a timed sequence (i.e., a video). To prevent easy recognition due to different color contrasts, 10% of pixels were grayscaled. Participants were instructed that the predominant and flickering trinket is the trinket contained in the package, and to make their decision accordingly. 3 videos were generated for every trinket, resulting in 15 different scans used during the task. During the task, the scan for a given trinket was randomly selected. After extensive testing, pilot participants were able to determine the trinket correctly 70% of the trials, which we believe is a good amount for participants to feel they could rely on their visual recognition skills, yet feel uncertain enough to follow the robot's recommendations.

5.3.2 Task Modeling and Process

To further immerse participants in the *Warehouse* scenario, participants first begin by watching a series of training videos, similar to an onboarding process one would find at a process control factory. In these videos, participants are instructed how to interact with the interface and fulfill orders. Participants are informed about the asymmetrical reward structure (i.e., incorrect sends are very costly, while incorrect rejects are minor setbacks) and the fact that the robots could make mistakes (as the over-trust intervention stipulated in

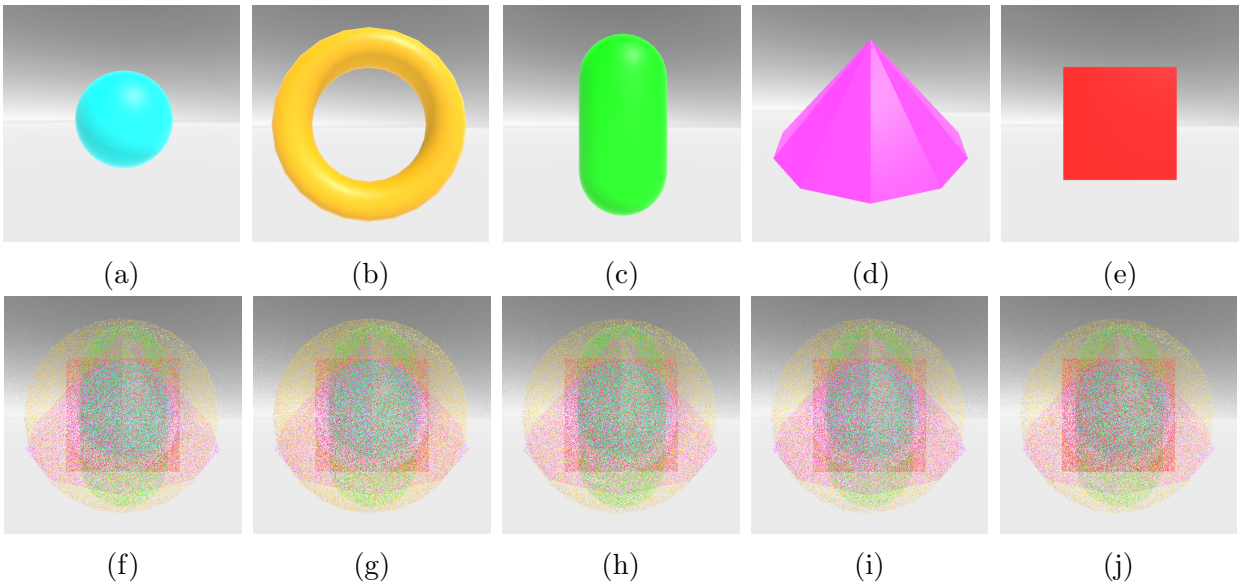


Figure 5.2: (a - e) The 5 trinkets shipped by the Warehouse. (f - j) A single frame of the perturbed scanner during the task for every trinket. Each trinket is matched vertically with its scan (i.e., the cyan sphere (a) is matched with its scan (f). A single frame is insufficient to represent the challenge and visual acuity required during the task, but it is included for reference. The scan is a video feed comprised by randomly generated pixels from the trinkets.

Chapter 2). The robots were framed as having uncalibrated sensors that detect the trinket inside the package, but in reality, all robot mistakes were controlled. Participants are then instructed they will complete 4 shifts (blocks), where they will interact with a different robot each. Each robot is painted a different color to emphasize this distinction.

As mentioned before, depending on condition, participants will either complete the task using a computer monitor with mouse input (Screen Representation condition), or in virtual reality using a Meta Quest 2 (VR Representation condition). The Meta Quest 2 includes handheld motion-tracked controllers, which allows hand-tracking and interactions with the virtual environment. This allows us to introduce embodied interactions in the VR experience when processing orders in-game. All participants are instructed in the procedure to process an arriving order. First, an order is received on the clipboard (a UI popup in the Screen condition or a virtual object in the VR condition); this order contains the trinket that must be sent, along with the colored tray that must be placed before the robot brings the package from storage. In the Screen condition, participants simply need to click the button

pertaining to the correct colored tray, but in the VR condition, participants must grab the trays at the left of their station (see Figure 5.1b), and place it in the inspection area. This brings physical cognitive load in which we hypothesize that the presence of this effort will lead participants to optimize their interactions (including their decision-making), compared to no physical cognitive load. Once the correct tray is placed, the robot places the package on the tray and awaits for the participant to make an initial decision (send or reject). Once a decision has been made, the robot states its recommendation in the form of an alarm or lack thereof. When the robot sounds its alarm, it is a direct cue that it detected the trinket inside the package does not match the order. Participants are then given the opportunity to make a final decision considering the robot's recommendation. This procedure is then repeated for all orders (visualized in Figure 5.3). In the VR condition, participants could grab and interact with all objects in the virtual environment, including the clipboard and the scanner.

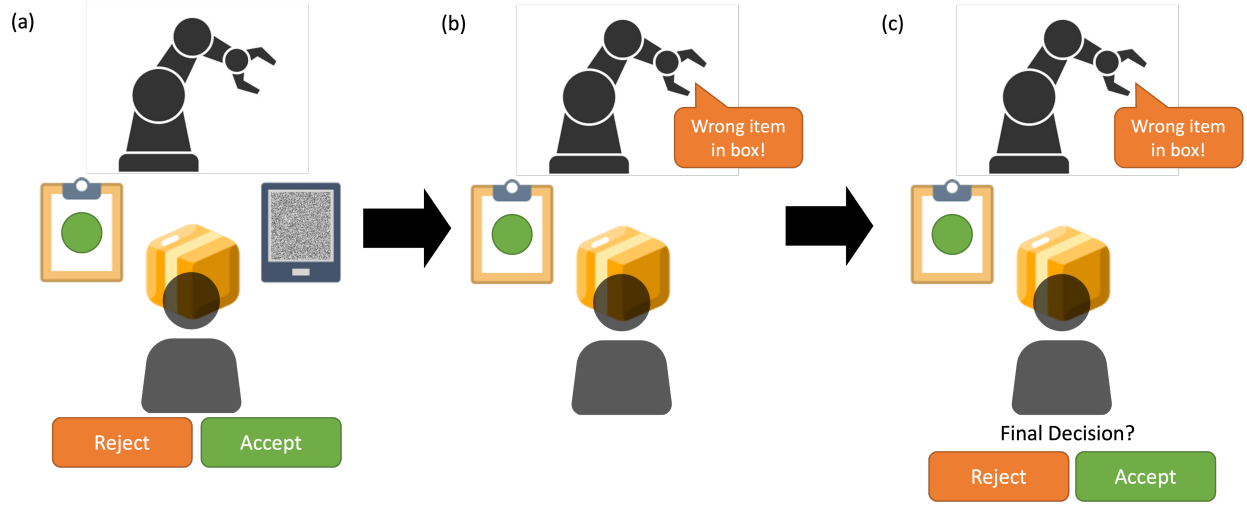


Figure 5.3: Decision-making procedure for a *Warehouse* trial. (a) Once the tray is placed, the robot places the package in front of the participant. The participant then makes an initial decision by checking the scanner to see if it matches with the correct order. (b) The robot will then sound an alarm if it has detected if the package contains the wrong item. The scanner is disabled at this point. (c) The participant is then prompted to make a final decision.

Careful consideration needs to be taken upon providing the participant feedback on their decision. If the participant decided to follow the robot's recommendation and the recom-

mendation was incorrect, trust towards the robot can rapidly deteriorate, and can prove very challenging to repair without proper intervention and care from the robot [75]. This presents the following concern: should participants be given immediate or intermittent feedback throughout the task? Immediate feedback may heavily affect the participants' behavior, as they may be able to discern the reliability of the robot through trial and error, and follow a heuristic to maximize their performance. To remediate this, we provided feedback halfway through every shift, in the form of the confusion matrix of their decisions. With this, we are able to determine how participants perceive the competence of the robotic agents, and how individual differences affect the amount of trust allocated. This type of feedback is often referred to as Knowledge of Result, which can help with calibration in difficult and cognitive-heavy tasks (such as visual acuity) [231, 232].

SDT-based decision-making tasks are often investigated in human factors, and generalize to a large amount of scenarios where a user is required to separate the signal from the noise under an unknown amount of uncertainty. However, the external validity of these types of tasks often suffer due to varying task (e.g., time pressure, correctness, reward structure and symmetry) and agent (e.g., transparency, uncertainty quantification, explanations) features per study. *Warehouse* models a time-sensitive task with asymmetric rewards, with no uncertainty quantification by the agent. In Section 5.6, we discuss this study's findings with respect to this specific variation of task and agent features, and consider how it could affect findings in the SDT-based task paradigm.

5.4 EXPERIMENTAL DESIGN

5.4.1 Independent Variables

As discussed in the background, we expect different responses from participants according to the Representation condition. Our running hypothesis is that invoking embodiment or embodied interactions will result in distinct trust responses according to the theory of social embodiment [212].

We control the reliability of the robot by setting the robot's Positive Predictive Value

Table 5.2: SDT parameters for the 4 agent reliabilities. PPV = positive predictive value; NPV = negative predictive value; C = cue behaviors, C $<$ 0 indicates a liberal policy (flags more often), C $>$ 0 indicates a conservative policy (flags less often); d' = distance between signal and noise (a.k.a., sensitivity index); Z_h = hit rate inverse CDF of the normal distribution; Z_{fa} = false alarm rate inverse CDF of the normal distribution; b = ratio of signal to noise at a random point.

Reliability	**PPV**	**NPV**	**C**	**d'**	$\mathbf{Z_h}$	$\mathbf{Z_{fa}}$	**b**	**No. Outcomes**	
								True	**False**
Perfect	1.00	1.00	0	∞	∞	0	∞	24	0
Ideal	0.91	0.91	0	2.77	1.38	-1.38	1	22	2
Good Enough	0.75	0.75	0	1.35	0.67	-0.67	1	18	6
No Info	0.50	0.50	0	0	0	0	1	12	12

(PPV) and Negative Predictive Value (NPV) according to condition. The selected parameters for every reliability is outlined in Table 5.2. For every order, the robot will determine whether the trinket contained in the package matches the current order, returning a binary response. The PPV and NPV indicate how useful the recommendation is when it is either positive or negative, respectively. Fixing the PPV and NPV to 1 presents a perfect robot: all positive recommendations denote the order matches the trinket, and all negative recommendations denote the order mismatches the trinket. As PPV and NPV are reduced, the robot starts miscategorizing boxes, presenting false positives and false negatives. Prior research has investigated the asymmetry of an agent's PPV and NPV and how does it affect trust-focused interactions [68, 233, 234], so we focus on symmetrical PPV and NPV values for this study. Each robot will be framed as distinct and painted a unique color to avoid carry-over effects. The presented independent variables are then as follows:

Representation

- Screen: Participants in the Screen condition complete the task behind a screen with an embodied robot akin to a telepresent interaction. Participants can retrieve information through interactions with the user interface, and execute actions all actions with a mouse and keyboard. The task requires no physical exertion.
- Virtual Reality (VR): Participants in the VR condition complete the task in

VR with the Meta Quest 2 headset, with interactions resembling a copresent dynamic. Participants grab the trays, use the scanner, and make decisions by using the motion-tracked controllers. These embodied interactions will guide distinct reactions to each agent representation.

Reliability

- Perfect: The robot never errs in categorizing a box. We predict the perfect agent will lead to the most trust, up to and potentially, over-trusting behavior (i.e., participants will forgo the scanner in favor of speedy decisions) [10, 14].
- Ideal: The robot makes 2 errors when categorizing boxes: 1 false positive and 1 false negative. A single mistake can vastly affect perception of the robot [100].
- Good Enough: The robot miscategorizes 6 of the boxes presented: 3 false positives and 3 false negatives. This follows the cutoff reliability stipulated by Wickens and Dixon where anything below would provide no benefit to the user [183].
- No Info: The robot miscategorizes half of the boxes presented. Therefore, the robot provides no useful information about categorization and is equivalent to guessing.

5.4.2 Dependent Variables

The dependent variables of interest are a variety of task and subjective measures to determine performance, trust, behavior, and perception of different kinds of agent Representation and Reliability.

Pre-Survey. The administered pre-survey contains demographic inquiries and measures of individual differences. For demographics, we measured age, gender, race and ethnicity, highest level of education completed, experience playing videogames (measured on hours per week), experience with virtual reality, and VR headset ownership. We expect familiarity effects due to experience with games or embodied experiences with VR reality to emerge in the data, and collecting this information allows us to control for variations.

For individual differences, we measured their propensity to trust technology [186] and their cognitive reflection using an alternate test by Thomson and Oppenheimer [235] than the usual test by Frederick [236]. We opted to use a different cognitive reflection test to ensure participants have not seen the questions before, given that participants in crowdsourcing participant pools (e.g., Amazon Mechanical Turk) are often exposed to cognitive reflection questions by researchers [235]. We do this for conservative measures, even if it is stipulated that the original cognitive reflection test by Frederick is robust to multiple exposures over a long period of time [237]. We expect participants' individual differences will serve as useful covariates while modeling their behavior and perceptions. Table 5.3 outlines the instruments used for the pre-survey.

Table 5.3: Pre-survey for the Warehouse game. Items marked with (R) are reverse scored. Loading factors were omitted because all items were used for both instruments (except for CRT-2, where one item was omitted due to confusing wording).

Code	**Item**	**Factor**
Cognitive Reflection Test-2 [235]		
crt1	Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?	Cognitive Reflection
crt2	How many cubic feet of dirt are there in a hole that is 3ft deep x 3ft wide x 3ft long?	
crt3	A farmer had 15 sheep and all but 8 died. How many are left?	
Propensity to Trust Technology [186]		
aptt1	Generally, I trust technology.	Trust Propensity
aptt2	I rely on technology.	
aptt3	Technology helps me solve many problems.	
aptt4	Automated agents are reliable.	
aptt5	I do not trust the information I get from technology. (R)	
aptt6	Technology is reliable.	

Task Behavior. A wide range of participants' behavior can be recorded and inferred from their interaction with the *Warehouse* game. The main behaviors of interest are the participant's time to complete an order (considering only the time to make an initial and final decision) and the decision made (package sent or rejected). Along with the sequence of orders, we are able to infer the following behavioral patterns:

- Correctness: Whether the final decision is correct (i.e., a true case, or decision accuracy),
- Reliance: Whether an order was sent when the robot did not sound its alarm,
- Compliance: Whether an order was rejected when the robot sound its alarm,
- Adherence: Whether the final decision matches the robot's recommendation,
- Switches: Whether the final decision matches the robot's recommendation and is distinct from the initial decision,
- Deferral: Whether participants did not use the scanner and completed the order in less than 2 seconds (i.e., the participant deferred their decision to the robot's recommendation).

We also aggregate the amount of deferred trials to obtain a measure of trust calibration. The number of trials that are deferred should be correlated with the reliability of the robot. If the robot is of high reliability, accepting the robot's recommendation and optimizing for speed is proper, calibrated use for this scenario. If the robot is of low reliability, using the scanner at a higher rate allows the user to gain insight of the robot's reliability, assigning a proper level of trust to it. Trust calibration is then compared against the true reliability of the robot to measure a calibration difference. A calibration difference of 0 indicates perfect trust calibration with respect to the robot's reliability. A positive value indicates over-trust (misuse) and a negative value indicates under-trust (disuse). For properly calibrated trust, on average, the proportion of trials that are deferred should match the reliability of the robot.

Usage of the scanner was measured as a binary outcome on how long was the scanner visible on the participant's display. In the Screen condition, the scanner could be displayed by clicking a button on the interface. In the VR condition, we only consider the time when the scanner's screen intersected with the foveal region of the VR display, which is approximately a 20° region at the center of the display [238]. Since the Meta Quest 2 does not have eye-tracking integration, we used the center of the display as fixed position for the

foveal region. This presents a limitation, since the participant could focus on the scanner using other regions of the display. The threshold of use was 2 seconds: if the participant stared at the scanner for more than 2 seconds, we consider it conscious use and record that the participant used the scanner for that order. This threshold was defined through pilot testing; the fastest conscientious (i.e., some pilot participants were instructed to use the scanner) order fulfillment was 4 seconds. Setting a threshold for half that time sets a hard limit to prevent casual glances at the scanner to be considered conscious use.

Survey Intervention. We used subsets of multiple validated surveys to measure subjective perceptions of how participants trust and cogitate during the task. These surveys were abridged by selecting items with the highest factor loadings per survey factor to reduce the time needed to complete the experiment, to fit within the time limit allocated to participants during crowdsourcing.

We measure task workload with the highly validated NASA Task Load Index (NASA-TLX) [191], as done previously in Chapter 4. As previously hypothesized, we expect that perceived workload has a relation to the amount of complacent behavior the participant exhibits during the task.

We use the Trust in Automated Systems (TAS) scale to measure trust and distrust in a more extensive manner than in Chapter 4. [189] stipulates that trust and distrust are distinct factors, and not necessarily opposite of each other, which has been confirmed through a factor analysis by [190]. We take the highest 2 loadings per factor, and include their associated questions in the survey.

We additionally measure the perception of agent competency through the Elements of Computer Credibility scale [239]. This scale is particularly aged, but the specific language and keywords used to evoke a response from the participant is necessary to measure a specific paradigm: competency and capability are often used as synonyms, but rather it should be noted that competency is an extension of capability. Capability is defined as having the initial capacity to perform a given task, while competence is the skill level demonstrated beyond being capable of performing the task. Thus, we expect the Reliability condition to predict the response in competency, but have a weaker effect on capability.

Additional questions were included to model constructs of interest, such as perception of self-competency, how much participants attributed their decision to the robot, and the strategy used to make a decision. Finally, a manipulation check was introduced to ensure participants' perception was being affected due to differing agent reliabilities.

The surveys are administered at the middle and end of every block. This allows us to model any changes given by partial and full interaction with the robots, and see how these are affected over time. The selected questions, timings, and factor loadings are outlined in Table 5.4.

Post-Survey. The post-survey measures the participant's ability to recognize and rank reliabilities along with the Igroup Presence Questionnaire [240], a validated survey measuring spatial presence, involvement, and realism in a virtual environment. This serves as the manipulation check for the Representation factor. We expect that presence is overall higher for participants who completed the game in the VR condition. Finally, a question verifying if a blue light filter was used to ensure behavioral outcomes in the task was not affected by the reduction of blue colors (e.g., the cyan and purple trinkets). To reduce the length of the post-survey, the Igroup Presence Questionnaire was also abridged using 2 questions with the highest factor loading according to the factor (with the exception of the Sense of Being There factor, which only had 1 question). The questions used in the post-survey are outlined in Table 5.5.

5.4.3 Procedure

Participants were recruited through Prolific[6], an online crowdsourcing platform similar to Amazon Mechanical Turk, as the COVID-19 pandemic presented challenges in recruiting in-person participants. Prolific was compared against other crowdsourcing platforms, and was found to contain higher quality data and with less instances of gaming the system [241]. There has been a body of upcoming research investigating the quality of remote virtual reality experiments (e.g., [96]), and this work indirectly contributes to the validation of this

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

paradigm. Participants selected the study (Screen or VR) in Prolific and signed an initial screening and consent form. Participants were screened for colorblindness, corrected vision, and hearing impairments (Prolific enforces these screens for data integrity); participants who failed the screening were not accepted into the study. Participants filled a demographic survey along with our pre-survey metrics, and were given instructions to install *Warehouse* on their device. Once the application is opened, they were instructed to watch a sequence of instructional videos, as outlined in Section 5.3.2. Once participants completed the videos, a quiz is administered to test their understanding of the goal as an attention check.

Participants then continue onto the working area. Participants complete 10 practice trials with no robot recommendations. This is done to acquaint participants with the game interactions and the visual acuity required to use the scanner. Participants then complete 4 blocks of 24 trials, each with a different robot reliability (to reiterate, each robot was colored differently and emphasized to have a different calibration setting). For every block, the orders in a trial were set in a predefined sequence, i.e., all participants receive the same sequence of orders, packages, and type of errors in a given Reliability condition. This is done to counteract the first-failure effect, as once the expectation of trust is broken, it gradually recovers with good performance [14]; this rate of recovery must be controlled for all participants at the risk of adding unnecessary noise to the interaction. *Warehouse* does not provide feedback after every trial, but provides it at the midpoint and end of every block, in the form of a confusion matrix of their decisions along with their decision speed. Participants completed the respective surveys after the midpoint and end feedback during a block.

Once participants completed the game, they completed a post-survey. Representation was treated between-subjects; 2 studies were posted in Prolific: a posting for a computer game and a posting for a VR game. This allowed us to target a specific population who owned a Meta Quest 2. Reliability was treated within-subjects with a randomized balanced Latin square design to avoid ordering and carry-over effects. The experiment flow is visualized in Figure 5.4

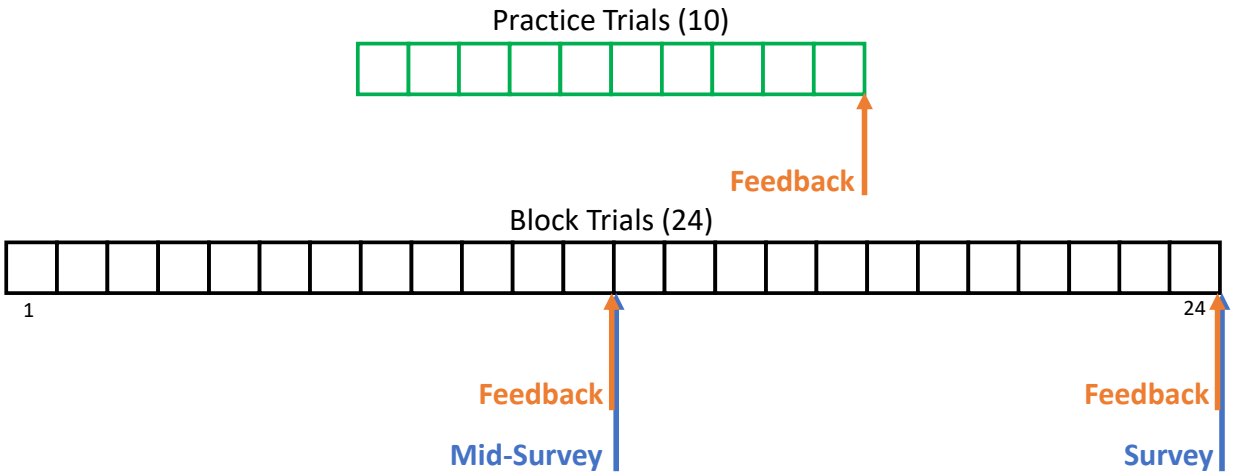


Figure 5.4: Experiment flow of the *Warehouse* game. Performance feedback is provided after the midpoint (trial 12) and end (trial 24) of each block, followed by their respective survey. The block trial is completed for every Reliability condition.

5.5 RESULTS

This study had a variety of quantitative variables. In this section, we summarize the most important findings, and provide specific details in their respective sections.

- As expected, agent reliability predicted overall performance. Scores during the *Warehouse* game decreased as agent reliability decreased. Analysis of participants' decision behaviors during the tasks showed a contrasting change in reliance and compliance for an agent that was erroneous at a single trial (Ideal) and an agent that was erroneous half of the trials (No Info).

- Trust behaviors, such as deferring one's decision to the robot, were generally higher with embodied agents and scaled according to the agent's reliability. Over time, deferred decisions increased as participants became more familiar with the robot in VR, compared to the Screen condition, which stayed low for the entire intervention.

- Embodiment allowed for better trust calibration when the agent was presented as imperfect. Participants gave appropriate levels of trust to each agent, rather than a static level of trust for all agents in the Screen condition. Notably, a single point of feedback can lead to significant changes in behavior, and allow for calibration of

expectations and performance to improve decision-making.

5.5.1 Demographics

120 participants completed the *Warehouse* game. Demographics are reported in Table 5.6. All participants were compensated 10 USD, along with a bonus of 0.02 USD for every point they earned in the game (outlined in Table 5.1). The average bonus payment was 3.46 USD. 1 participant did not complete the post-survey after completing the game, and thus was dropped for a total sample size of 119.

5.5.2 Inferential Statistics

The behavioral and survey data retrieved from the *Warehouse* game will be analyzed using a mixed ANOVA across 3 factors: Representation, Reliability, and Time. The Representation and Reliability factors are our independent variables of interest, while the Time factors refers to the aggregate of trials before and after the mid-block report. This factor is taken into consideration because the mid-block reports are the only point during the task where participants receive actionable feedback, and are able to adjust their behavior accordingly – as the end-block report does not allow for adjustment since a robot with a different reliability follows.

We recorded a total of 11424 trials. We then remove any trials that had a fulfillment time greater than 3 standard deviations from the mean (i.e., decisions that took an excessive amount of time), resulting in 172 trials dropped for a total of 11252 trials.

Task Outcomes

We begin by discussing the behavior the participant demonstrated during the task.

Performance. Score was the main outcome observed from the task, and as mentioned earlier, is a function of correct decision-making and timeliness. The ANOVA revealed the Reliability factor had a main effect in score ($F(3, 298) = 45.33$, $p < 0.001$), along with an interaction effect with Representation ($F(3, 329) = 27.06$, $p < 0.001$) Time ($F(3, 329) =$

27.06, $p < 0.001$). Further inspection reveals that a higher robot reliability is expectedly correlated with higher scores. The interaction with Time shows that after participants received the mid-block report with the No Info robot, their score dropped during the second half. Furthermore, the Ideal robot performed better under the VR Representation condition than the Screen Representation condition. The scores are visualized in Figure 5.5.

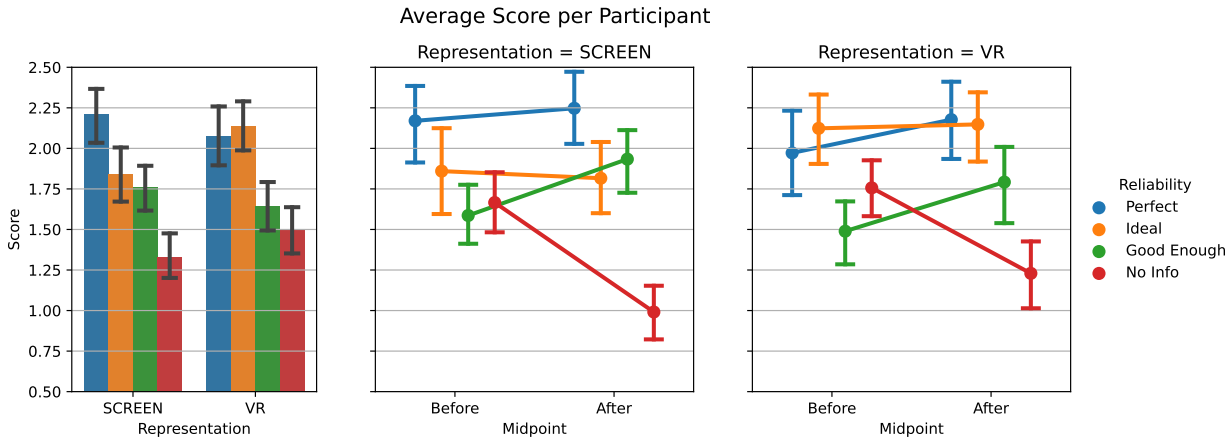


Figure 5.5: Average score per participant. The Representation factor is deconstructed.

We divide the participant's decision-making accuracy into the positive (sends) and negative (rejects) cases. As rewards were asymmetrical and the robot gave an alarm only when it believed a package was mismatched, we expect distinct performance in each case, as it has been shown that reliance and compliance are often not complementary [68]. For send accuracy, the ANOVA revealed a main effect in the Time factor ($F(1, 117) = 57.04$, $p < 0.001$), an interaction between Representation and Reliability ($F(3, 326) = 3.2$, $p < 0.05$), and an interaction between Reliability and Time ($F(3, 312) = 21.55$, $p < 0.001$). Interestingly, after receiving feedback, participants interacting with the No Info robot suffered a drop in correct sends, with all other conditions (with the exception of Perfect reliability) trending downwards as well. For reject accuracy, the ANOVA revealed only an interaction between Reliability and Time ($F(3, 351) = 10.09$, $p < 0.001$). Across time, reject accuracy trends upwards in the No Info condition. No changes are found in any other reliabilities. The send and reject accuracies are visualized in Figure 5.6 and 5.7, respectively.

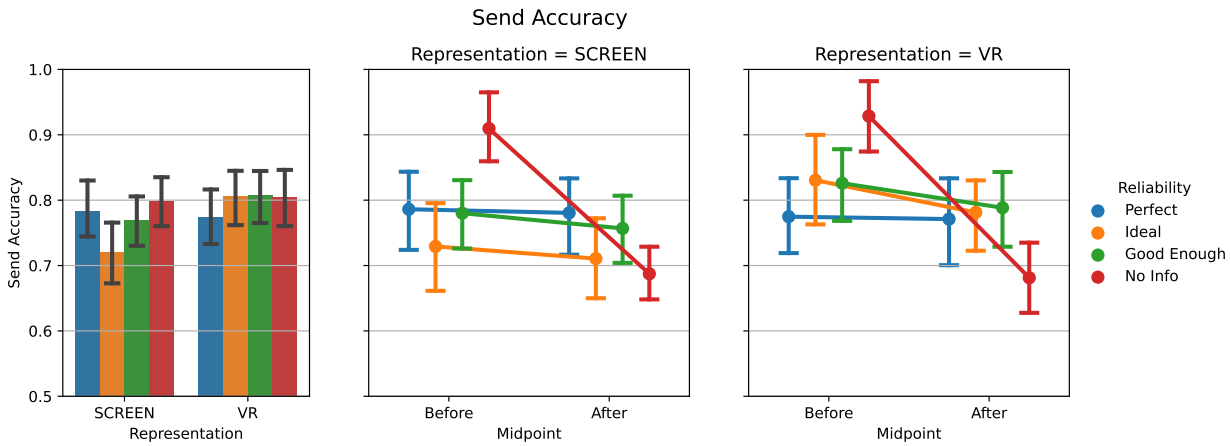


Figure 5.6: Send accuracy per participant. The Representation factor is deconstructed.

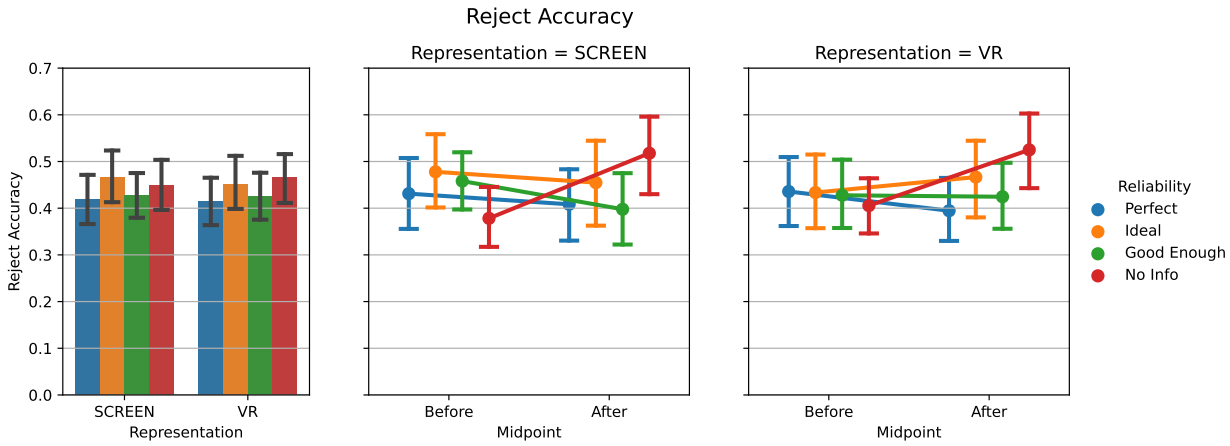


Figure 5.7: Reject accuracy per participant. The Representation factor is deconstructed.

Behavioral Trust. To establish patterns of behavioral trust, first, we focus on reliant and compliant behavior when interacting with the robots. Reliance is defined as the lack of a given action given the absence of an alarm (for this task, sending a package when no alarm is presented). In contrast, compliance refers to taking an action when an alarm is given (similarly, rejecting a package when the robot sounds its alarm). Both reliance and compliance have been found to relate to over-trusting and complacent behavior [242]. For reliance, the ANOVA revealed a main effect in the Reliability ($F(3, 313) = 11.12$, $p < 0.001$) and Time ($F(1, 117) = 5.59$, $p < 0.05$) factors, with an interaction between them ($F(3, 351) = 45.18$, $p < 0.001$). Opposite to compliance, after the mid-block feedback, reliance decreased in the No Info condition, where it increased in the Ideal condition. Other reliability

conditions remained unchanged. For compliance, the ANOVA revealed a main effect in the Reliability factor ($F(3, 351) = 17.47$, $p < 0.001$), with an interaction between Reliability and Time ($F(3, 351) = 69.03$, $p < 0.001$). Compliance was correlated with reliability. After the mid-block feedback, compliance greatly increased in the No Info condition, whereas it decreased in the Ideal condition. Other reliability conditions remained unchanged. The reliance and compliance rates are visualized in Figure 5.8 and 5.9, respectively.

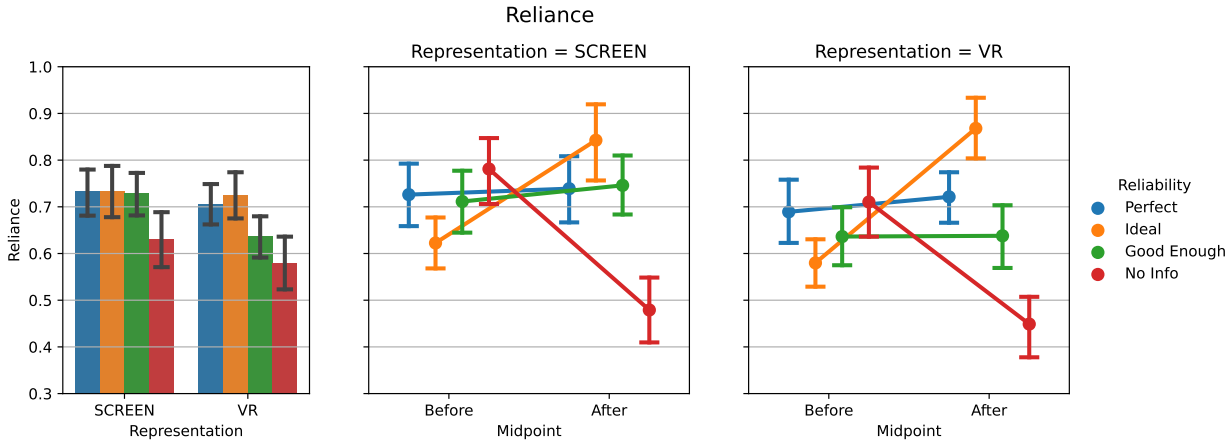


Figure 5.8: Amount of Reliance. The Representation factor is deconstructed.

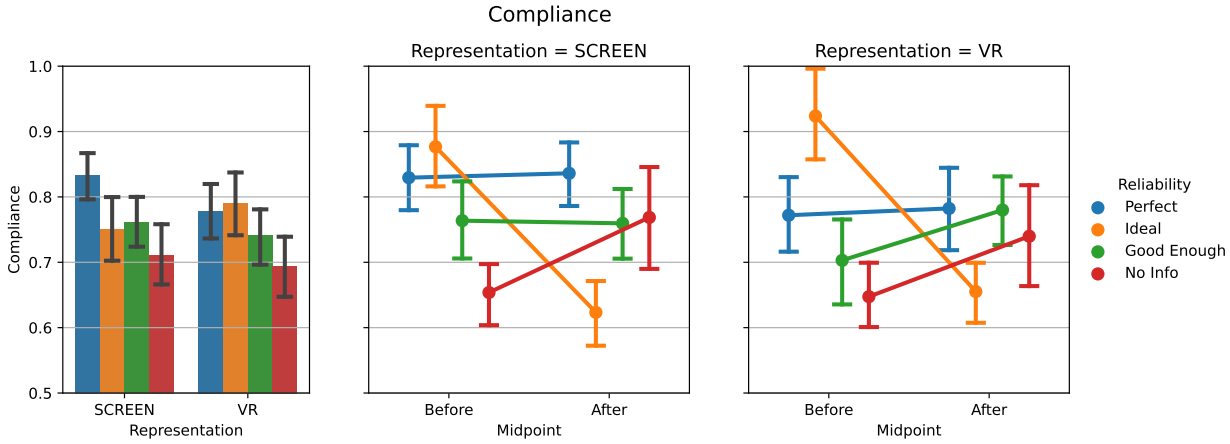


Figure 5.9: Amount of Compliance. The Representation factor is deconstructed.

Two decision-making metrics were used to further determine behavioral trust, as prompted by Zhang et al. [26]. First, the participant's switches quantifies the proportion of trials where the participant decided to switch their initial decision to match the robot's recommendation. A conscious switch to match the recommendation given by the robot is indicative of

trust. The ANOVA revealed an interaction effect between Reliability and Time ($F(3, 327) = 16.23$, $p < 0.001$). We observe that switching increases over time with higher reliabilities, but decreases in lower reliabilities. Second, we measured how much a participant adhered to the robot's recommendation. This metric covers additional cases from switches where the participant and the robot did not initially disagree. The ANOVA revealed main effects in the Reliability ($F(3, 330) = 24.11$, $p < 0.001$) factor, with an interaction between Reliability and Time ($F(3, 330) = 7.81$, $p < 0.01$). Adherence increases with reliability, and predictably, decreases in lower reliabilities after receiving the mid-block feedback. The switch and adherence rates are visualized in Figure 5.10 and 5.11, respectively.

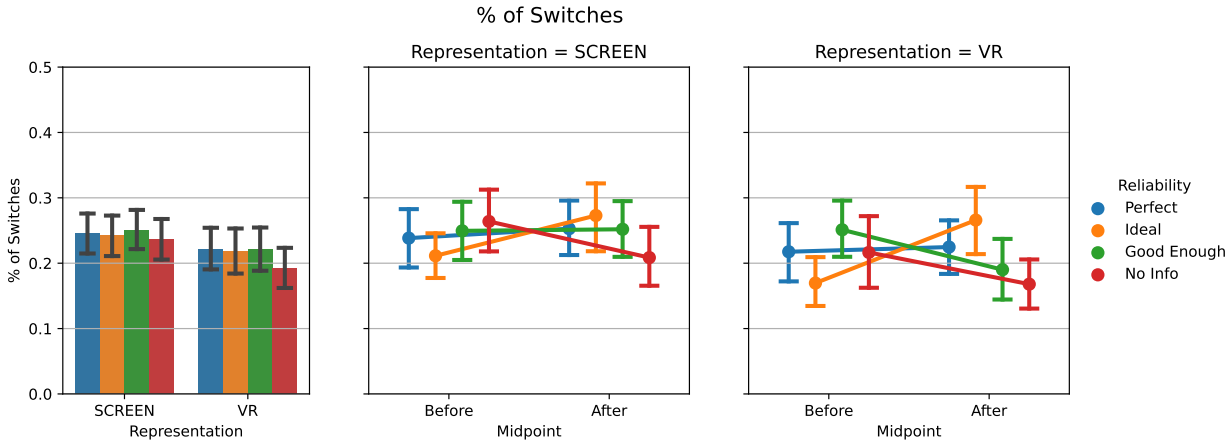


Figure 5.10: Proportion of Switches. The Representation factor is deconstructed.

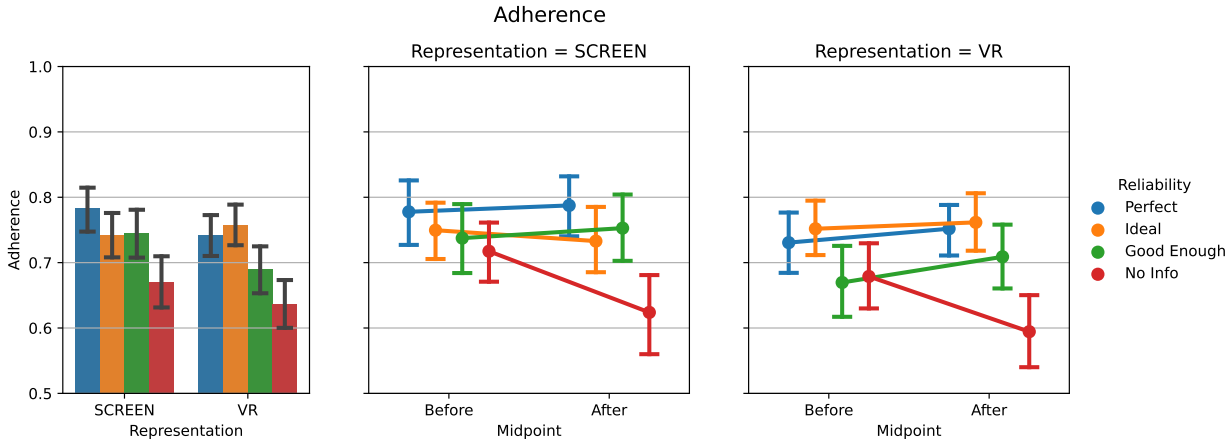


Figure 5.11: Proportion of Adherence. The Representation factor is deconstructed.

We inferred deferral from a combination of 2 explicit behaviors: for a given trial, if par-

ticipants used the scanner and fulfilled the order in a combined time less than 2 seconds. This behavior could be construed as the participant automatically adhering to the robot's recommendation. The ANOVA revealed only a main effect in the Time factor ($F(1, 25) = 18.04$, $p < 0.001$), along with a near-significant interaction with the Representation factor ($F(1, 25) = 3.43$, $p = 0.07$). It is interesting to note that instances of deferral was higher when robots are embodied in the VR condition within Representation. Additionally, this interacts with Time, as the amount of deferrals increased after the mid-block feedback, likely an effort from participants to increase their decision-making speed and rewards. If we plot the amount of deferred trials over time, we can see a trending increase that is higher in the VR Representation condition (see Figure 5.14b). The amount of deferrals is visualized in Figure 5.12.

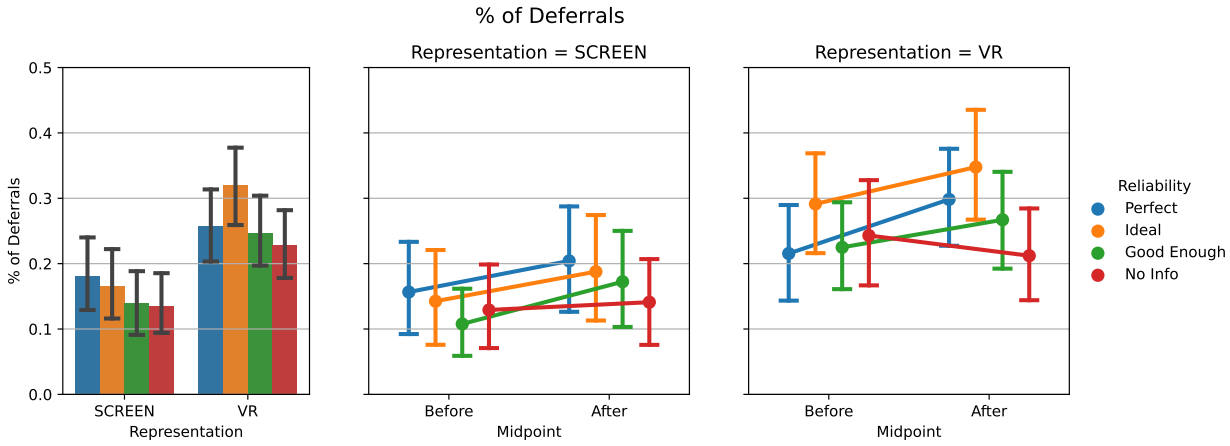


Figure 5.12: Proportion of Deferred Trials. The Representation factor is deconstructed.

Surveys

In this section, we discuss results from all surveys administered during the task (pre, in-game, and post-surveys). The items for each construct were parceled by taking the average of the related items. [243] stipulates that using a parametric test for Likert-type questions is appropriate and yields near-identical significance compared to non-parametric approaches (which are often used for ordinal data). All Likert scale questions were measured with 5-point scales.

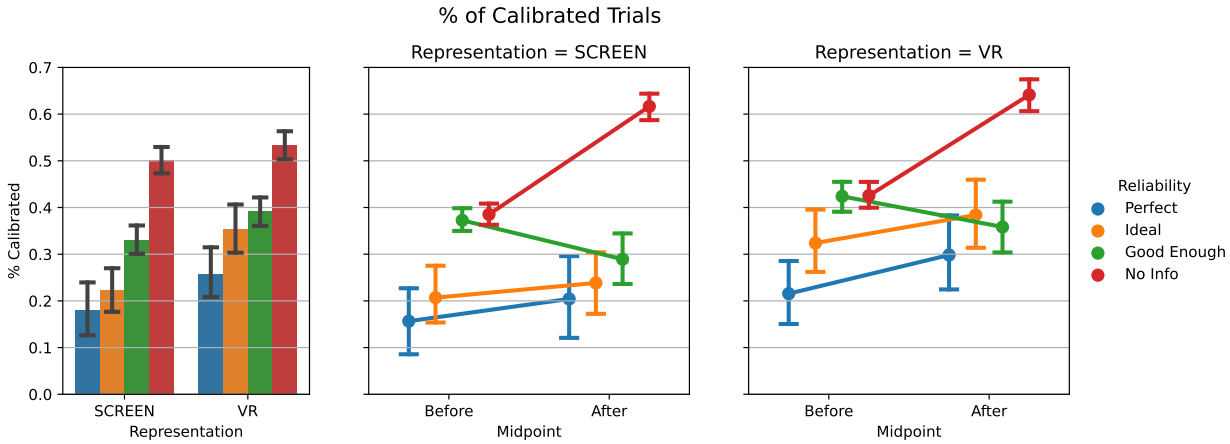
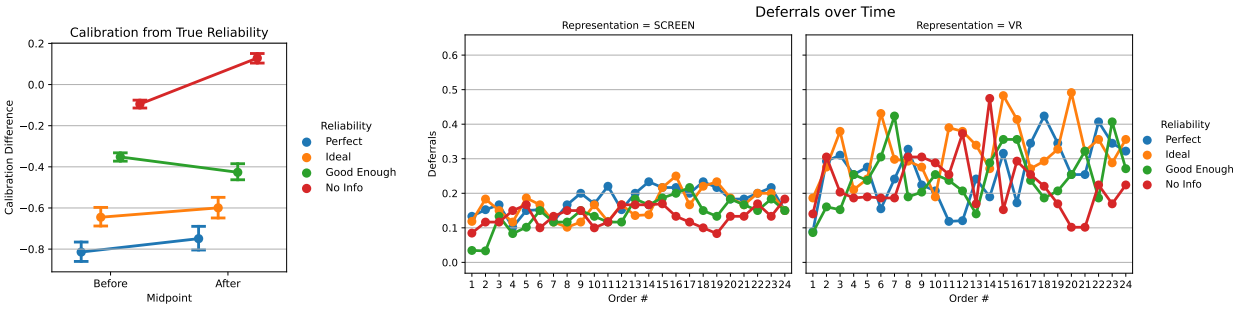


Figure 5.13: Trust calibration per condition. The Representation factor is deconstructed.



(a) (b)

Figure 5.14: (a) Calibration difference from the agents' true reliability. (b) Deferred trials over time. The VR condition had a greater variance in deferred trials.

Subjective Trust. For trust in the robot, the mixed ANOVA revealed a main effect in the Reliability ($F(3, 325) = 33.6$, $p < 0.001$) factor, along with an interaction with Time ($F(3, 351) = 2.96$, $p < 0.05$). Trust scaled appropriately with reliability, but only the No Info reliability led to participants adjusting their trust downwards even further after the mid-block report. Trust is visualized in Figure 5.15.

For distrust in the robot, the mixed ANOVA revealed a main effect in Reliability ($F(3, 325) = 21.48$, $p < 0.001$) and Time ($F(1, 117) = 6.21$, $p < 0.05$), with an interaction between them ($F(3, 351) = 4.16$, $p < 0.01$). Distrust scales inversely with reliability, and is more pronounced in the VR representation condition. After the mid-block feedback, distrust increased for the No Info reliability condition. Distrust is visualized in Figure 5.16.

For how competent an agent was perceived, the mixed ANOVA revealed a main effect

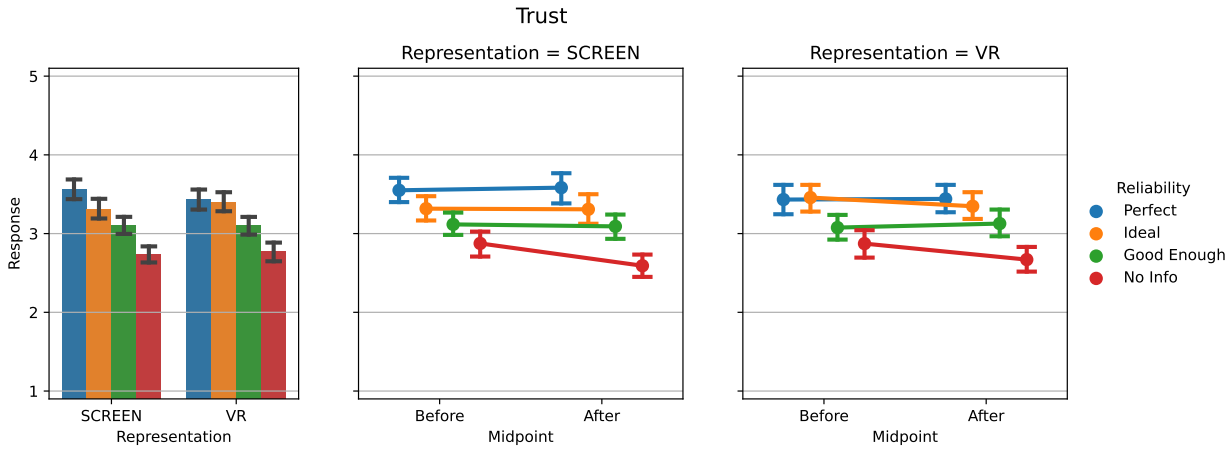


Figure 5.15: Trust responses. The Representation factor is deconstructed.

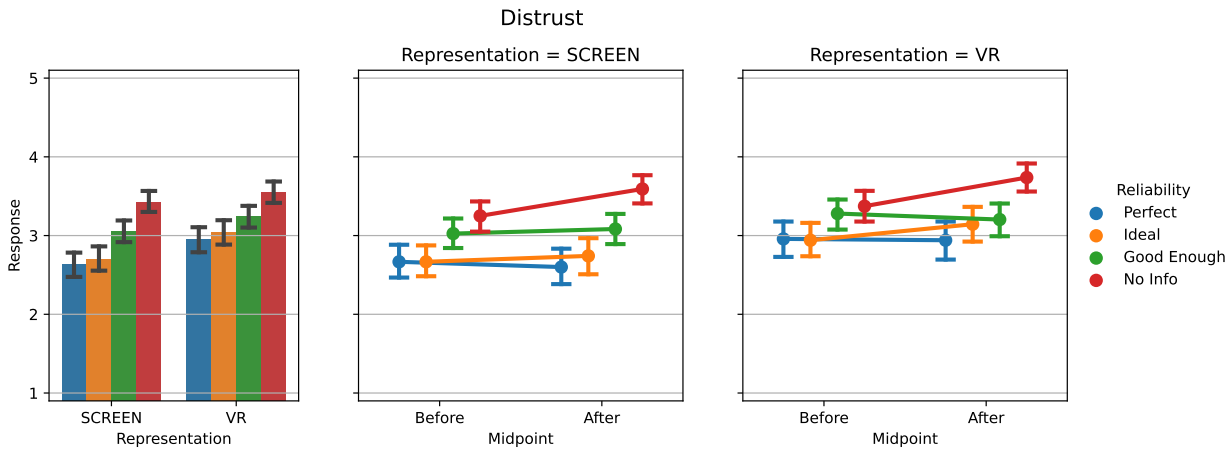


Figure 5.16: Distrust responses. The Representation factor is deconstructed.

in Reliability ($F(3, 351) = 30.65$, $p < 0.001$) and Time ($F(1, 117) = 4.91$, $p < 0.05$), with an interaction between them ($F(3, 351) = 2.7$, $p < 0.05$). Competency scales accordingly with reliability, and is reduced after the mid-block feedback only for the No Info reliability condition. Competence perception is visualized in Figure 5.17.

Strategy and Sensitivity. We define decision attribution as how much did a participant feel the robot influenced their decisions. For decision attribution, the mixed ANOVA revealed a main effect in Reliability ($F(3, 351) = 22.54$, $p < 0.001$) and Time ($F(1, 117) = 4.72$, $p < 0.05$). Expectedly, the amount of credit given to the robot for the decision made scales accordingly with reliability, with attribution decreasing after the mid-block feedback for the

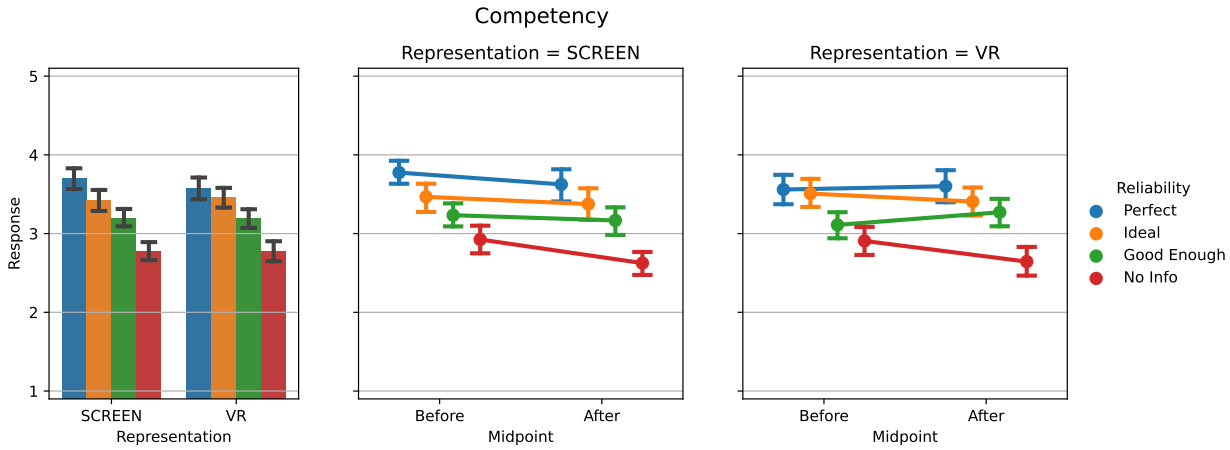


Figure 5.17: Competence perception responses. The Representation factor is deconstructed.

No Info reliability condition. Decision attribution is visualized in Figure 5.18.

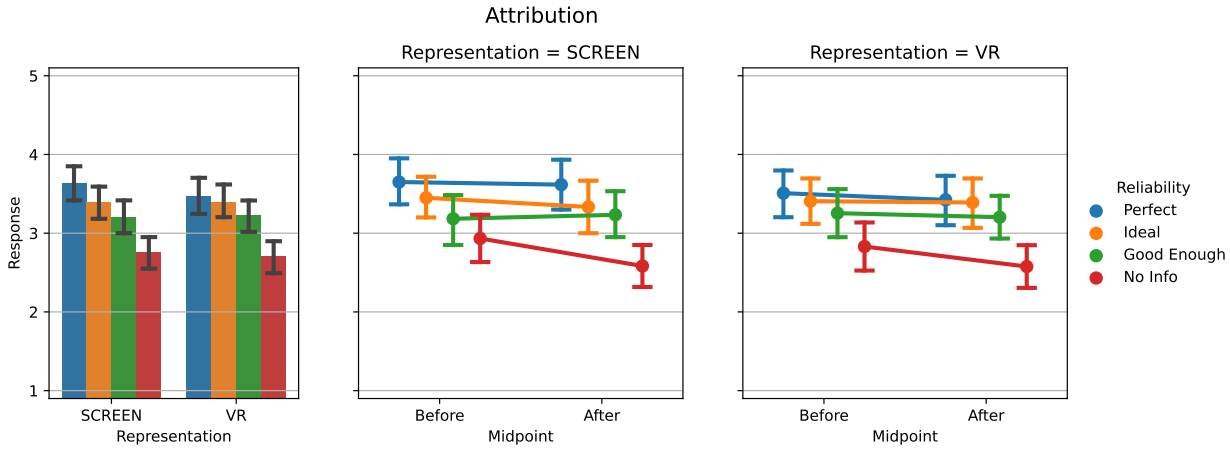


Figure 5.18: Attribution responses. The Representation factor is deconstructed.

We measured the participant's ability to correctly rank the robots based on their reliability – dubbed sensitivity. We score the accuracy of their ranks from 0 to 4 (based on correct rankings ascertained). A Mann-Whitney U-test found no significant differences between Representation conditions in sensitivity to reliability ($U = 1946.5$, $p = 0.33$). Furthermore, a Friedman test to determined that the mean ranking assigned to robot reliabilities were significantly different ($\chi^2(3) = 33.94$, $p < 0.001$). Interestingly, the Perfect and Ideal robots had a very similar mean rank ($\mu_P = 2.17$, $\mu_I = 2.18$), followed by the Good Enough robot ($\mu_{GE} = 2.68$), with the No Info robot trailing last ($\mu_{NI} = 2.95$). Nemenyi post-hocs found a threshold between the Perfect/Ideal conditions and Good Enough/No Info conditions (i.e.,

non-significant differences between Perfect and Ideal; and Good Enough and No Info).

Finally, we asked participants at the end of each block which signal they believed was the most informative for them to make a decision from 4 options: the robot's alarm (compliance), the robot's silence (reliance), the scanner (self-reliance), or none of the above. Visual confirmation indicates that the preferred signal was significantly different, but we also conducted chi-square tests of independence of the observed frequencies using contingency tables. We found no different preference for a given signal across reliabilities – but was near significance ($\chi^2(9) = 16.4$, $p = 0.06$). However, we found significant differences across robot representations ($\chi^2(3) = 21.7$, $p < 0.001$). We observe that in the Screen representation condition, there was high value placed on the robot's alarm, whereas in the VR condition, value shifts more towards the lack of alarm and the scanner (the alarm still remains as the most preferred). Information preference is visualized in Figure 5.19.

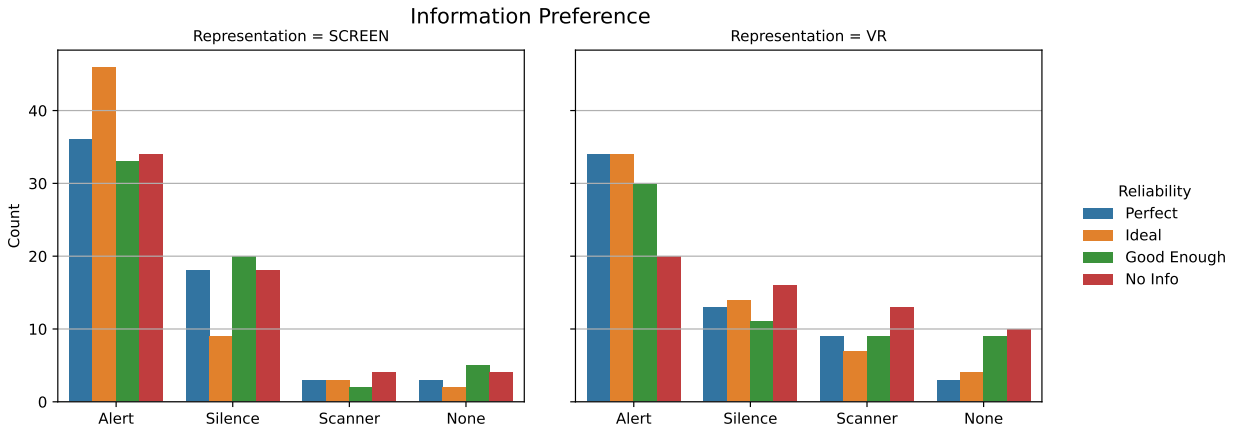


Figure 5.19: Information preference responses. The Representation factor is deconstructed.

Workload. Aggregate workload was measured by taking the average of all 6 individual NASA-TLX factors, which is often viable and reported [244] for complete perceived workload instead of the standard procedure indicated by [191]. The NASA-TLX was administered only at the end of a block, thus we analyze it with a two-way mixed ANOVA, without considering the Time factor introduced earlier. For this section, the ANOVA refers to this redefinition. The ANOVA showed a main effect in both the Representation ($F(1, 117) = 9.29$, $p < 0.01$) and Reliability ($F(3, 351) = 3.4$, $p < 0.05$) factors. Overall, workload was higher in the VR

representation condition regardless of reliability. We now present specific analysis from the findings of each individual NASA-TLX factor. Each workload factor is visualized in Figure 5.20.

For mental workload, the ANOVA found a near-significant main effect in the Reliability factor ($F(3, 351) = 2.57$, $p = 0.054$). Mental workload was overall trending higher in the VR Representation condition, yet, no significant effects were found. For physical workload, the ANOVA found a main effect in the Representation factor ($F(1, 117) = 20.16$, $p < 0.001$). For temporal workload, the ANOVA found a main effect in the Representation factor ($F(1, 117) = 10.81$, $p < 0.01$), along with an interaction with Reliability ($F(3, 351) = 3.56$, $p < 0.05$). For effort, the ANOVA revealed a main effect in Reliability ($F(3, 351) = 8.92$, $p < 0.001$). Expectedly, we see that the effort participants exert in the task scales with robot reliability, but with no perceptual difference across representations. For satisfaction, the ANOVA revealed a main effect in the Reliability factor ($F(3, 351) = 28.44$, $p < 0.001$), along with an interaction with Reliability ($F(3, 351) = 3.05$, $p < 0.05$). Expectedly, satisfaction in the task correlated with reliability. Although non-significant, it is interesting to note that satisfaction in VR had lower variance than the Screen representation condition. For frustration, the ANOVA showed a main effect in the Reliability factor ($F(3, 351) = 12.74$, $p < 0.001$). We see that frustration increased with the lowered reliability of the robots.

5.5.3 Structural Equation Model Fit

We once again employ a structural equation model (SEM) to analyze the relationships between various metrics collected in this study. The final SEM is presented in Figure 5.21, which displays the relationships between exogenous and endogenous variables, along with fit and regression coefficients. The SEM model was built using R 4.2.1 with lavaan 0.6-12 [195].

We focus on five outcome metrics: score, calibration difference, trust calibration, adherence, and switches. Score represents the participant's performance, as a function of correct decisions and time taken. Trust calibration captures the participants' level of trust behavior from the truth, where a value of 1 indicates perfect trust calibration. Calibration differ-

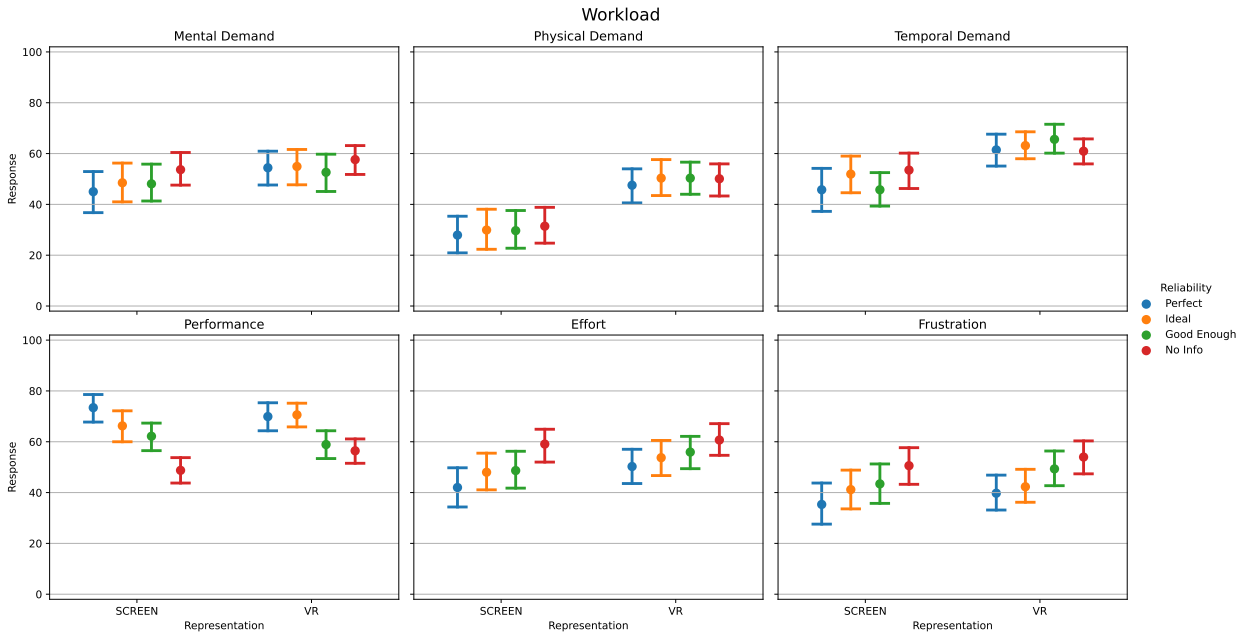


Figure 5.20: Workload responses across all 6 dimensions.

ence captures how much would participants over-trust or under-trust, with a positive value indicating over-trust and a negative value indicating under-trust. Adherence models the agreement between the participant's final decision and the robot's recommendation. Adherence also models reliance and compliance, and thus were removed from the model. Finally, switches models whether a participant switched their final decision to match the robot's recommendation (as explained in Section 5.4.2.

We follow a similar approach as detailed in Chapter 4: we aggregated all factors and outcomes by taking the mean of each participant's data points. Survey items were aggregated for analysis via parceling, which reduces measuring error through aggregation [196]. We also controlled the False Discovery Rate (FDR) to prevent the issue of multiple comparisons, which is recommended for exploratory SEM analysis [197]. To control the FDR, we used the Benjamini-Hochberg procedure [198] with Q set at 0.15, allowing us to determine which factors in the final SEM model are expected to be false positives. A total of 38 models were tested for the outcomes and their effects.

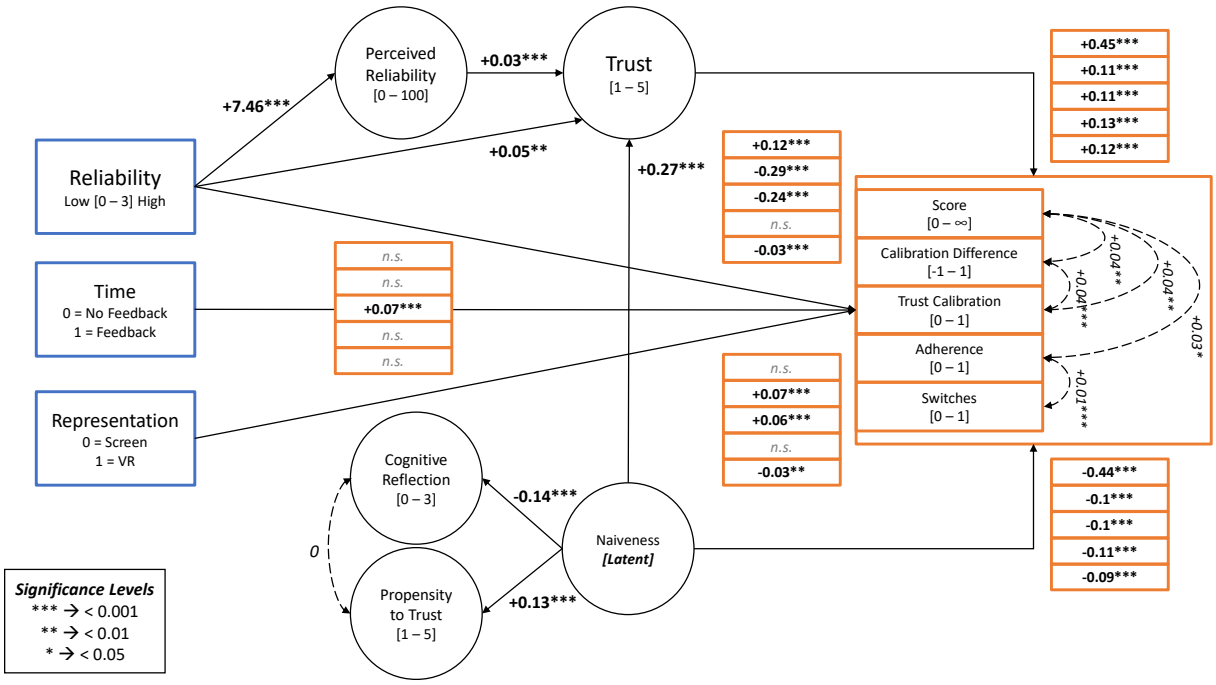


Figure 5.21: Fitted SEM for the *Warehouse* game. Model fit: N = 119 with 44 free parameters. $RMSEA = 0.073$ ($CI = [0.063, 0.084]$), $TLI = 0.977$, $CFI = 0.947$, over null baseline model, $\chi^2(28) = 171.79$. Solid lines indicate significant regressions, and dashed lines indicate correlations. Numbers represent the fitted coefficient along with its level of significance. The coefficients in boxes match the outcomes, from top to bottom respectively.

5.6 DISCUSSION

This study has demonstrated an extensive amount of findings in the relationship between performance, trust, embodiment, and reliability, which serve useful to consider how small changes in the design of AI can lead to distinct outcomes. We summarize the findings of the results in Table 5.7.

5.6.1 Reliability, Embodiment, and Decision-Making

This study largely aimed to discover the effect of embodied interactions and reliability on trust, both through task behaviors and subjective survey responses. Contrary to reliability and trust, the literature on embodiment and trust is not clear, with many studies having contradictory results on whether embodiment or physicality has a positive effect on trust, and determining whether a virtual agent or a physical agent is preferred by humans in tasks where

they can be interchanged [222, 245]. Additionally, prior work has extensively researched the effect of automation reliability on performance, and we find multiple replications and comparisons across different representations. As mentioned, prior studies are often done simulating a given embodied context – which may dilute the constructs measured, as these contexts bring additional cognitive load that affects how reliability is perceived and how humans decide to make decisions.

Score and accuracy (the main metrics for performance) were found to correlate with reliability, similar to various previous studies (e.g., [88, 229, 246]). This was expected, and we additionally expected to see not much difference across Representation conditions, as completion of the task itself requires the same rational processes regardless of whether participants click a button with a mouse or with their hands in the virtual environment. Regardless of robot interaction, the task was designed in such a way that participants would have an innate intuition on what trinket was shown in the scanner, up to certain level of personal confidence (around 70%). This "gut feeling" is not expected to change across Representations, and is considered one of the factors that drives decision-making before considering an agent's recommendation [236].

We observe interesting effects within the development of reliance and compliance. Prior research has noted that compliance and reliance are two distinct types of cognitive states associated with the interaction of automated systems, and are expected to decrease with imperfect automation – false alarms (i.e., matching packages recommended to be rejected) reduces compliance and misses (i.e., mismatching packages recommended to be sent) reduce reliance [68, 242]. Thus, we would expect reliance and compliance to correlate with reliability, either increasing or decreasing through time according to the robot's performance. In *Warehouse*, however, we observed contradictory patterns between reliance and compliance for the Ideal (91% reliability) and No Info (50% reliability) conditions: reliance increased (Ideal) and decreased (No Info) accordingly, demonstrating the capability of the human to calibrate towards the "all is good" signal, but for compliance we see that it decreases (Ideal) and increases (No Info) accordingly, opposite to what was expected. Two interesting points are inferred from this: 1) a single false case (either a false alarm or a miss) is impactful enough to change their behavior, with a false alarm causing a dramatic drop in compliance

– opposite to what Kaplan stipulates where systems that miss are judged more harshly than systems that give false alarms [247] and, 2) reliance and compliance were largely stable for the Perfect (100% reliability) and Good Enough (75% reliability) conditions, hinting at the potential loss in compliance due to the "first failure effect" [248], as the Ideal condition presented only 1 false alarm and 1 miss. This amount of failure may have been insufficient for the human to fully calibrate their trust – as they may have been expecting more failures. In combination with time pressure, participants interacting with the Ideal robot may have opted to rely more on it to make decisions [249].

We measured behavioral trust through the number of decision switches, adherence, and deferred orders the participants did per block. These behaviors are analyzed with the consideration that research on trust metrics indicate that subjective (i.e., self-reported) trust is not a strongly reliable indicator for trust behavior [250], and has been utilized in prior studies [25, 26, 251]. Switches were not affected by embodiment, but instead were affected by the reliability of the robot across time. This is expected behavior, as humans will opt to trust highly reliable robots more over time than robots that are not reliable, consistent with longitudinal studies of trust and acquaintance [252]. The level of adherence (i.e., whether the human's final decision matches the robot's recommendation, regardless of changes) was found to expectedly correlate with reliability. Finally, we quantified the amount of trials where participants deferred their decision to the robot (i.e., participants did not use the scanner to make their decision and immediately followed the robot's recommendation for their final decision). The interaction between agent representation and time had a near-significant effect ($p < 0.08$), where it was higher in VR, but varied across time. We attribute this to the time pressure introduced in the task: to maximize points, participants likely sped up their decisions to maximize the rewards per trial, and since perceived workload was overall higher in VR – participants may have began to prioritize the robot's feedback based on their perception of the robot's reliability, matching prior findings where time pressure increases dependence on automated systems [230]. One interesting theme to note is whether this perception is conscious or unconscious. Revisiting the Predator-Prey game in Chapter 4, the reliability difference between the tested agents was very subtle, and in cases participants were not able to make distinct trust judgments between the high and low reliabilities (as

seen in the survey results). However, behaviorally, participants were able to adapt to different reliabilities, indicating that this process may be more subconscious than what we often believe. Overall, this lead to better trust calibration, as deferrals in capable systems (i.e., Perfect, Ideal) increased whereas it decreased in incapable systems (i.e., No Info); compared to the constant level of deferrals for the Screen condition.

Subjective trust is still measured as a complement to the behavioral metrics discussed. Embodiment mainly affected the level of distrust towards the robot and workload. A running hypothesis posed for this study was that the VR representation would lead to polarized (i.e., more extreme) trust effects. Although trust was not found to be affected by representation, distrust was, echoing a sentiment found by [189] during the development of the Trust in Automated Systems scale: trust and distrust are not necessarily opposites. Distrust was likely affected in a more extreme manner than trust due to the "first failure effect" [248]: as prior research has found, trust is challenging to build and is affected by a wide set of initial parameters and characteristics [253], but is easily destroyed by any sort of trust violation (e.g., recommending an incorrect outcome) [254]. It is likely that the negative sentiment is overall higher in an embodied representation, as it mimics interactions with physical entities, such as another human. Workload was expected to be higher in VR than in the Screen condition, as the embodied interactions would incur higher effort from participants in order to complete the Warehouse game, even if the time pressure was present in both representation conditions. Potential interpretations for this include workload leading to a higher level of deferred trials, matching the literature on the relationship between workload and complacent behavior [255], and active workload could have contributed to investment during the task, taking away from the boredom of pointing and clicking in the Screen condition.

During the *Warehouse* game, participants can use various signals to rationalize their decision. Each signal brings certain value to participants, thus we ascertain which signal is most reliable during SDT-based task according to representation condition. In the Screen condition, participants mostly valued the alarm given by the robot, much higher than the lack of an alarm (a measure of reliance) or the scanner itself. During embodied interactions, however, even though participants again valued the alarm much higher than other signals, the lack of an alarm and the scanner both gained prominence as having the highest value.

A way to rationalize the change in preference could be the presence brought by the task itself, as sitting on a computer screen can easily bring a static pattern of signals that merely has to be followed, while in VR, using one's whole body to complete a task (along with the presence of spatial audio cues and an embodied robot) brings higher immersion, prompting participants to consider other signals in the environment. This matches a finding by Kulms and Kopp [228], where embodiment can facilitate the acceptance of certain signals given by the robot, before it stabilizes over time. An interesting point of future research would be to investigate how signal preferences and value change over time, and how this could potentially change the decision taken by the participant (i.e., someone who values the alarm may have a higher reject accuracy and be more prone to comply).

5.6.2 Transparency and the Timeliness of Feedback

All of the effects measured in the task outcomes or survey responses are affected by measuring across different points in time. This reinforces an important point in automation reliability research: when and how often is feedback presented affects how humans calibrate their expectations of the system. Prior work has investigated the gamut on feedback frequency: from immediate and transparent feedback, to sporadic feedback, to delayed feedback, up to no feedback at all [26, 232, 256]. The general consensus of the effect of feedback is that providing it in a timely manner (e.g., immediately after a decision) allows for immediate calibration of expectations, leading to immediate changes in behavior and decision-making depending on whether the feedback is positive or negative.

However, not providing feedback does not indicate that participants are unable to calibrate. As shown in this study, the humans' own capability at recognizing the correctness of a task also serves as a strong signal to calibrate, depending on whether the task is solvable without the help of automation. Based on the manipulation check, participants were overestimating the accuracy of the No Info robot, and underestimating all other robots with higher reliabilities. Two possible reasons could be inferred from this finding: 1) the robots were framed as imperfect, and may have primed participants to rely on their own capabilities at first, and 2) because the scanner was designed to have a 70% success rate, it is likely that

participants were forming some incorrect perception of the correctness of the scanner, which affected their perception on the robot's accuracy. It was not until they received feedback at the mid-block report that participants were able to calibrate, significantly affecting their behavior. The implication of this finding is that it is important to design AI systems such that they exhibit a strong signal of their capabilities, as human behavior – and consequently, the outcome – changes depending on their perception of the system. Feedback systems are not the only approach; Barg-Walkow and Rogers demonstrated that an explicit statement of the agent's capabilities affects perceptions and overall performance [257].

A transparent system may indeed help users calibrate to a system by clearly defining its capabilities [15], and by combining it with proper training, better trust calibration may be fostered. For this study, we combined a delayed-feedback system with training, leading to a certain amount of under-trust for agents above the reliability cutoff threshold (70%). In the VR condition, however, we can see the level of over-trusted trials increase with time (Figure 5.14b), indicating that it may be easier to calibrate upwards with an embodied agent.

5.6.3 Mediation Analysis

The fitted SEM reveals an interesting finding within a participant's individual differences: the features describing individual differences can be combined to a latent variable representing "naiveness." This latent "naiveness" positively predicts a participant's propensity to trust, and negatively predicts their cognitive reflection. Cognitive reflection have been used to describe a quantitative measure of the mode of thought used: intuition or reasoning – mostly popularized by Kahneman and Tversky [258]. Cognitive reflection often requires thoughtful, but slow reasoning, thus people who rely on automatic thinking tend to score lower on cognitive reflection tests [236], and in a way, making them more "naive." This latent "naiveness" positively regresses onto trust, which is an intuitive explanation for this latent factor (i.e., the more "naive" one is, more trust tends to occur). Overall, "naiveness" negatively affected the task outcomes, impacting score, trust calibration, and frequency of adherence and switches.

We record 2 self-reported measures related to agent competency: perceived reliability and

trust. The SEM notes that how reliable an agent was perceived to be is affected by the agent's reliability and whether the participant has received feedback about the agent. Expectedly, the higher an agent's reliability, the higher participants will perceive its reliability. Perceived reliability, however, was down by the presence of feedback, possibly indicating it as a calibration mechanism to prevent potential over-trust in the agents. Sequentially, perceived reliability, the agent's reliability, and the "naiveness" latent factor all affected self-reported trust. Considering the coefficients of the model, a very reliable agent can receive a high amount of trust (i.e., a perceived reliability of 90% would bring trust into an already high and appropriately calibrated level), but the presence of a "naive" user may tilt the scale excessively into over-trust (B = 0.27, $p < 0.001$). We confirm this by verifying the coefficients of the "naiveness" factor: with increasing "naiveness", absolute trust calibration is worsened (B = 0.1, $p < 0.001$). "Naiveness" mediates into self-reported trust, bringing trust up (as mentioned), but regularizing by having an under-trusting effect in trust calibration (B = -0.1, $p < 0.001$).

For the overall outcomes; trust and the agent's reliability increased the resulting score, while a participant's "naiveness" reduced it. Since the agents were framed as imperfect, participants often found themselves under-trusting the agents, when they were often more capable at recognizing the trinket than what participants thought they were. Allocating higher trust would result in increased performance, yet, improper trust calibration (in the form of "naiveness") would then decrease overall performance. We then finally observe how does agent representation affects the outcomes, and if embodiment brings a positive effect to collaboration metrics. An embodied agent brings effects into 3 outcomes: it tilts the calibration difference positive (B = 0.07, $p < 0.001$) along with improving trust calibration (B = 0.06, $p < 0.001$), and reduces the number of switches done (B = -0.03, $p < 0.01$). The variance of the outcome variables were described well, with trust calibration and calibration difference almost reaching an R^2 of 0.7, and other behavioral variables at 0.4. The exploratory nature of structural equation models often prevents high levels of correlation to be found, as more unmeasured variables can be into play.

5.7 SUMMARY

We conducted a study to investigate the effect of reliability and embodiment on performance and trust (n = 119), to fill the research gap on how dynamics in the human-agent teaming paradigm holds within a human-robot collaboration context, which is one of the representative goals in the development of AI agents. In this chapter, we designed a signal detection theory-based task with either an embodied or non-embodied decision support system in the form of a robot. We controlled robot reliability and type of representation to observe how it affected a wide range of behaviors and subjective perceptions. Analysis of the results reveals 2 major insights: 1) when an agent is presented as imperfect, embodiment improves trust calibration, as users defer their decisions to the agent at greater degree. Conversely, in the Screen condition, where agents were perceived as imperfect regardless of embodiment, users demonstrated a consistent, low level of deferral. Embodiment allowed for a higher level of behavioral trust and improved calibration, resulting in more accurate and appropriate deferrals; and 2) humans are able to reasonably calibrate their expectations over AI systems with little to no feedback or transparency, with a single point of feedback being enough to cause significant changes in behavior. Future work should focus on comparing the results from a similar decision-making task (such as the one presented in this study) with a physical robot, to validate how decision-making is processed in the real world.

5.7.1 Afterword

At the writing of this dissertation, the studies discussed have been conducted and the results are currently being compiled in a manuscript aimed for a human-computer interaction, human-robot interaction, or human factors conference. Other deliverables include a dataset including the observed task behaviors of the participants, and the pre and post-survey results.

Table 5.4: In-game survey administered in the Warehouse game, after the midpoint and endpoint of every block. The factor loadings are given by the related literature; n/a indicates no factor loadings were found, thus multiple items were used. M = question administered at the midpoint of the block, E = question administered at the end of the block.

Code	Session	Item	Factor	Loading
NASA Task Load Index [191]				
tlx-m	E	How mentally demanding was this shift with this robot companion?	Workload	n/a
tlx-p	E	How physically demanding was this shift with this robot companion?		n/a
tlx-t	E	How hurried or rushed was the pacing of this shift with this robot companion?		n/a
tlx-d	E	How successful were you in making the correct decisions with this robot companion?		n/a
tlx-e	E	How hard did you have to work to make the correct decisions with this robot companion?		n/a
tlx-f	E	How discouraged, stressed, and annoyed were you while completing this shift with this robot companion?		n/a
Trust in Automated Systems [189, 190]				
tru1	ME	This robot companion is dependable.	Trust	0.88
tru2	ME	This robot companion has integrity.		0.86
dist1	ME	I am wary of this robot companion.	Distrust	0.87
dist2	ME	I am suspicious of the robot companion's recommendations.		0.80
Elements of Computer Credibility [239]				
comp1	ME	This robot companion is competent.	Competency	n/a
comp2	ME	This robot companion is capable.		n/a
Additional Questions				
manip	ME	How accurate do you think this robot companion is?	*Manipulation Check*	n/a
attrib	ME	How much did this robot companion help you in making decisions?	Attribution	n/a
strategy	E	Which source of information did you think was the most useful during this shift?	Strategy	n/a

Table 5.5: Post-survey for the Warehouse game.

Code	Item	Factor	Loading
Igroup Presence Questionnaire [240]			
bt	In the Warehouse, I had a sense of "being there."	Sense of Being There	1
sp1	I did not feel present in the Warehouse. (R)	Spatial Presence	0.79
sp2	I felt present in the Warehouse.		0.74
inv1	I was not aware of my real environment.	Involvement	0.85
inv2	I still paid attention to the real environment. (R)		0.78
real1	How real did the Warehouse seem to you?	Experienced Realism	0.77
real2	The Warehouse seemed more realistic than the real world.		0.73
Additional Questions			
rank	Rank the 4 robot companions that you worked with today, from worst to best.	Ability Perception	n/a
bluelight	Did you have a blue light filter enabled during the game?	*Screening*	n/a

Table 5.6: Resulting demographics for the Warehouse experiment. Categories with 0 participants in all conditions were not included. Race/Ethnicity categories add up to more than the sample size due to some participants being multi-racial.

	Screen	**VR**
Sample Size (n)	60	59
Age		
18 - 24	16	15
25 - 34	17	33
35 - 44	15	11
45 - 54	8	0
55 - 64	3	0
65+	1	0
Gender		
Male	33	43
Female	25	15
Non-Binary	2	1
Race/Ethnicity		
White	45	31
African American	2	5
Asian	10	20
Hispanic/Latino	5	6
Native American	3	0
Middle Eastern	0	1
Highest Education/Degree Completed		
Some High School	2	2
High School	24	17
2-year College	6	6
4-year College	18	23
Graduate	8	10
Terminal/Professional	2	1
Hours/Week Playing Videogames		
< 1	8	7
1 - 2	6	2
2 - 4	6	1
4 - 7	12	10
7 - 12	11	9
12 - 20	8	15
> 20	9	15
Owns a VR Headset	11	48
Prior Experience with VR	32	59

Table 5.7: Effects found in the Warehouse study. Generalized Eta squares (η^2_G) are reported along the significance level. * = significant at $\alpha = 0.05$; ** = significant at $\alpha = 0.01$; *** = significant at $\alpha = 0.001$; ! = near significant ($p < 0.1$). - = effect not tested.

Metric	**Main Effect (η^2_G)**			**Interactions**	**Findings**
	Rep.	**Rel.**	**Time**		
Task					
Score		*** (0.11)		Rep./Rel. ** Rel./Time ***	-Score corr. Reliability -No Info trends downwards after feedback -Ideal better in embodied conditions
Send Accuracy			*** (0.03)	Rep./Rel. * Rel./Time ***	-No Info trends downwards after feedback -Others trend downwards slightly -Perfect remains constant
Reject Accuracy				Rel./Time ***	-No Info trends upwards after feedback
Reliance		*** (0.03)	* (0.004)	Rel./Time ***	-No Info trends upwards after feedback -Ideal trends downwards after feedback
Compliance		*** (0.03)		Rel./Time ***	-Compliance corr. Reliability -No Info trends downwards after feedback -Ideal trends upwards after feedback
Switches				Rel./Time ***	-P/I trends upwards after feedback -GE/NI trends downwards after feedback
Adherence		*** (0.05)		Rel./Time ***	-Adherence corr. Reliability -No Info trends downwards after feedback
Deferral			*** (0.02)	Rep./Time (!)	-Higher in VR (n.s.) -Trends upwards after feedback
Calibrated Decisions	** (0.03)	*** (0.2)	*** (0.02)	Rel./Time ***	-Higher in VR -Reduced calibration in Good Enough
Surveys					
Trust		*** (0.11)		Rel./Time ***	-Trust corr. Reliability -No Info trends downwards after feedback
Distrust	!	*** (0.06)	* (0.003)	Rel./Time ***	-Distrust inv. corr. Reliability -Higher in VR (n.s.) -No Info trends upwards after feedback
Competency		*** (0.1)	* (0.003)	Rel./Time ***	-Competency corr. Reliability -No Info trends downwards after feedback
Decision Attribution		*** (0.06)	* (0.002)		-Attribution corr. Reliability -No Info trends downwards after feedback
Sensitivity		-	-	-	-No difference between Screen and VR -Threshold between P/I and GE/NI
Information Preference	***	!	-	-	-Alert valued highest in Screen -Scanner and No Info increases in VR
Workload	** (0.06)	* (0.006)	-		-Higher in VR
Mental		!	-		-Highest in No Info (n.s.)
Physical	*** (0.12)		-		-Higher in VR
Temporal	** (0.06)		-	Rep./Rel. **	-Higher in VR
Effort		*** (0.03)	-		-Effort inv. corr. Reliability
Satisfaction		*** (0.11)	-	Rep./Rel. *	-Satisfaction in VR was more centralized
Frustration		*** (0.04)	-		-Frustration inv. corr. Reliability

CHAPTER 6: CONCLUSION

This chapter presents a discussion of all the findings presented in the core studies, in order to answer the dissertation inquiries posed in the introduction (Chapter 1). We then turn our attention to new avenues of research this dissertation proposes, and how the implications of this research affects the design of AI systems for human-AI integration.

6.1 CROSS-STUDY ANALYSIS

We focused on investigating the effects of varied reliability in a wide variety of domains. We limited the scope of the domains to be as general as possible in order to maximize the external validity of our studies, hence we studied on a virtual space, a simulated space, and a physical space. The focused tasks present a slight overlap, as we investigated 2 iterations of a collaborative decision-making task (the Diner's Dilemma study and the *Warehouse* game). Regardless, the presented tasks are distinct enough for the effects to be considered varied and informative.

We aimed to explore independent variables that were exclusive to a specific researched domain (e.g., explanations for decision support systems), but kept one independent variable in common across domains: reliability. In that vein, we focused on a wide variety of dependent variables (e.g., knowledge, team performance, adherence) that would help us inform how reliability affects outcomes while keeping trust and human performance paramount across domains as well. Table 6.1 presents an overview of the measurements and treatments of the core studies.

Table 6.1: A comparison of interventions and effects observed in this dissertation across the core studies.

	Study 1	**Study 2**	**Study 3**
	Movie Miner Diner's Dilemma	**Predator-Prey Game**	**Warehouse**
Motivation	Investigate XAI features in the presence of imperfect agents	Mediating individual differences between imperfect reliability and performance	Exploring imperfect agents in embodied scenarios
Domain	Decision Support Systems (virtual agents)	Simulation Systems (physical agents)	Human-Robot Collaboration (embodied agents)
Task	Preference selection	Multi-agent real-time pursuit	Collaborative decision-making
IVs	Presence of explanations Degree of control **Reliability level**	**Reliability level**	Degree of embodied interaction **Reliability level**
DVs	Knowledge	Performance Trust	Performance Trust
Results	XAI worsens calibration, hampers potential learning; imperfections help	Imperfections improved agent/team performance, no differences in perception	Small imperfections affect decision-making; embodiment affects trust judgments, workload
Implications	Forming proper domain knowledge helps in decision-making; perfect agents/XAI may make this challenging	Imperfections can cause interactions with the domain, inadvertently changing performance; transparency needed	HAI is distinct across embodiment (consider teleoperation), and additionally affects reliability perception

6.1.1 Research Formulation

These studies stem from a curiosity about how features of decision support systems affected human performance. A long line of research beginning from recommender systems (e.g., [27, 36, 146]) has investigated the benefits and drawbacks of assisted decision-making, encountering a large set of individual characteristics (e.g., trust propensity, personality traits, domain expertise) affects how decision-making and performance is formed. Research often focuses on how improper trust calibration can affect overall user performance, but an interesting avenue was to investigate how covariates to performance are affected. We selected knowledge due to its relevance with the role of human operators, whose domain expertise can likely affect outcomes beyond the regular novice users, especially when trust calibration is at place. That is, one could expect a domain expert to be more aware of the system's limitations, and overall prevent biased or complacent behavior from occurring. Increasing knowledge when interacting with the system may be a viable resource for helping the users understand the capabilities and limitations of the system by allowing domain knowledge to serve as the ground truth, and any deviations from it can be met with a skeptic eye by the user. Overall, this allows the user to exercise their judgment at a higher frequency and preventing inappropriate reliance.

We were interested in deviated from the usual discrete decision-making paradigm that most human-AI interaction research opts for (e.g., [26, 114, 154]), so we jumped to a physical domain for the next study. We investigate how would these interactions hold between humans and robots, in a scenario that would require continuous actions and constant decision-making. Such a scenario exists within the U.S. Army Research Laboratory's efforts to integrate autonomous systems with soldiers[7], where the capabilities of humans and agents are clearly heterogeneous, and the constant, active pressure in the battlefield may affect how soldiers perceive and think of their agent teammates, much more than low-risk or relaxed domains where a joint human-AI decision can be taken with slow consideration. The Predator-Prey game (Chapter 4) approaches this with a simulation of one of the common scenarios in this domain: pursuit of a high value target. Although the scenario remains ab-

[7]https://www.arl.army.mil/business/collaborative-alliances/current-cras/iobt-cra/

stract, a high fidelity environment can be replicated for further validation, including adding a third dimension into the simulation to map the real world even closer. In an ideal scenario, immersive technologies (such as virtual reality) could serve as a conduit to reach even higher validation, as it is unrealistic to conduct these types of studies in the field. This observation led the development of the final core study.

The final study aimed to investigate the effects of embodiment in a task common to decision support and recommender systems: signal detection. The rationale behind this study was to fill the gap on robotic-type interactions, where embodiment plays a larger role in the cognitive processing of the user. This addresses the question posed in the previous paragraph and extends to interactions introduced in Chapter 3: there is interest garnered from designers and researchers to develop social robots that conference and assist users in their day-to-day decision-making. The same machine learning models deployed in recommender systems would serve as an assisting robot if the model was given a physical representation (or a "body", as discussed in Chapter 5). This brings interesting implications for not only comparisons between embodied and non-embodied interactions, but tele-operations of remote systems. For instance, a surgical robotic system at a remote location may have features that automate or assist certain procedures or maneuvers, where a surgeon may opt to hand off control to the automation for simpler procedures (e.g., a measured, longitudinal cut across the abdomen). If the robot was operated behind a computer screen, the reduction of workload and increase in boredom might prompt an operator to become demotivated and relegate its decisions to the automation. When embodied, however, the spatial perception may prompt a higher level of involvement from the operator. Overall, it is interesting to note how embodiment can play a role in altering the amount of workload perceived by the user, which in turn affects their decision-making.

Together, these studies explore a representative subset of task domains where AI systems can operate in, and up to a successful degree. There still remain a wide variety of domains that yet interact with the main factors discussed in this dissertation, but by drawing these longitudinal connections, we begin observing overall patterns on how trust calibration and reliability interplay with one another, in the presence of varied scenarios. We discuss some of these common points in the following section.

6.1.2 Dissertation Themes

The Impact of Imperfect Reliability

In the standard use-case scenario of deployed technology, the automation is already performing at a very high level of reliability (and in commercial applications, they can even be perceived as perfect agents by users). This can lead to unreasonable expectations, opening a segue into over-trust. By keeping expectations artificially low (through low reliability), we can improve performance in unforeseen situations as human facet of the system is more alert to failures.

We emphasize the manipulation of reliability in this dissertation due to it being often defined by the performance of the model rather than a feature to design. Machine learning engineers, data scientists, and model designers spend several hundreds of working hours optimizing the performance of a model, where they can hit a point of minimal returns. Does a potential 1% improvement justify 100 additional work hours? The threshold of what "ideal" performance should be is so not well defined, that often domain experts use their judgment to select an arbitrary threshold, and then adjust in the future if a higher level of performance or quality is needed. But in reality, could there be a threshold where a reduced reliability may actually incur benefits to how humans process information and decision-make? We posed in the introduction (Chapter 1) whether a "Good Enough" reliability could engender better trust calibration and performance in human-AI integration.

The Movie Recommendation and Diner's Dilemma study (Chapter 3) demonstrate evidence of a negative effect caused by common explainable AI (XAI) features in decision-making scenarios, caused by increased interactions with the decision support system that led to the emergence of over-trust and the incorrect insights to be formed. Novice users may be especially susceptible to this, as they tend to use decision aids in a feed-forward manner (i.e., seek a recommendation from the aid before acting) [92]. Transparency with an imperfect agent may help users to compare their judgment against the explanations given and rationalize a good recommendation against a bad recommendation, allowing trust to be better distributed throughout the interaction. Hoffman et al. stipulate that a proper understanding of the capabilities of the decision support systems allows for better trust calibration

[259], and in this case, repeated exposure to both high and low quality recommendations allow for a better mental model to be formed. In this vein, future work may focus on developing training regimes that focus on both correct behavior and incorrect behavior, and allow users to draw comparisons between behaviors and prime a mental model based on "what's correct," in contrast to approaches that focus on manual training [255] or addressing rare, catastrophic scenarios [79]. While these studies did not measure any form of mental model processing, the behavioral constructs recorded provide a reasonable approximation of how the mental model influenced their decision-making.

The Predator-Prey game (Chapter 4) demonstrates how the environment and task parameters can interact with the perceived reliability of a system, potentially changing how users calibrate their trust. The agent with the imperfect *chaser* strategy had better performance than a perfect chaser, which led to increased team performance. Initially, we inferred that the lower reliability of this agent would invoke a lower level of trust, but as its performance increased (again, due to lower reliability), the contribution of the human teammate decreased, as the *chaser* agent took a more active role in the task. Additionally, there was no perceptual difference between the perfect agent and the imperfect agent. The implications of these findings are that designers should be aware of how the environment itself can bring adverse reactions to the altered behavior of the agent. Hypothetically, if the imperfect agent would interact with the user for an extended period of time (e.g., multiple scenarios, multiple missions, multiple days), one could expect that the higher perceived reliability could lead to instances of over-trust and performance loss, even if the true reliability is much lower. If the agent were to fail while the user is in this state (which is likely to happen since the agent is already imperfect), a higher penalty would be incurred. If the environmental factors would change (e.g., changing the strategy, the size of the play area, the number of agents), it is likely we would see reliability affecting performance and behavior in distinct ways. Future research could consider a taxonomy of task features that interacts with agent features, and begin interconnecting how features affect performance in different experiments. We expand the concept of perceived reliability and how it affects outcomes in Chapter 5.

Finally, in *Warehouse*, we observe how a small amount of imperfections can lead to a drastic change in reliability perception. The "ideal" agent (set at 91% recommendation accuracy)

led to a large change in reliance and compliance (increase and decrease, respectively), an effect not observed in any of the other reliabilities (100% or 75%). The reliability threshold for what is considered an "ideal" or "good" agent is varies per study, so it is interesting to note why – for this particular scenario – the presence of 1 error of each type led to this divergence in behavior, which is not observed in the case where there are no errors or multiple errors exist (the 50% case is ignored, as it is equivalent to random guessing). The intuition here is that the presence of a single error may have not been enough for users to calibrate to negative outcomes, leading to inappropriate amounts of reliance and compliance. The presence of a single false alarm was enough to bring compliance down by 30% during the trials. It is likely that with more errors, the user learns to calibrate their trust better across time, and have a better expectation of what the system is capable of doing. One mistake can be considered a fluke, and in cases it may serve useful to be salient about the capabilities of the agent (given that no other XAI features exist that could relay this information as well). This study could then be expanded with the addition of XAI features that allow transparency, and observe how users decide to use the automation, or even if it is used at all for lower reliabilities. In such cases, would the presence of XAI features discourage the use of automation that is not of a certain reliability threshold?

With the results demonstrated in this dissertation, a relevant application is in the development of training sessions for human-AI collaborations that optimize future human behavior by using "Good Enough" agents. By emphasizing the role of these agents as training tools, we can avoid potential arguments about whether people would like or dislike working with subpar agents directly. In this way, the focus shifts from designing an agent that meets certain performance criteria (without over-investing resources) to designing a training environment that fosters optimal human-AI interaction, and to bring those interactions forward when the time comes to complete a real task. Additionally, framing "Good Enough"-ness this way allows for greater flexibility in terms of how much "Good Enough" should be, as it is ultimately intended to improve the overall effectiveness of human-AI collaborations, which is already largely shown to be very domain dependent.

Meta-Research: Virtuality and Physicality

This dissertation demonstrated how different agent types and reliability levels can affect perceptions of decision support systems versus robots. We demonstrated in the *Warehouse* game (Chapter 5) that embodiment can affect users' perceived workload and self-reported trust, indicating that physicality can affect rationalization of one's feelings more than actual behavior (as we saw that adherence was more prevalent with a non-embodied agent). This same effect was found in the Movie Recommendation and Diner's Dilemma studies (Chapter 3), where the non-embodied agents did affect behavioral outcomes more than self-reported measures. There could be a process that interaction with an embodied agent or an embodied interaction triggers that drives unconscious behavior more than self-reports. Along with research that stipulates that self-reported measures may not be the most reliable for capturing a particular construct [26, 250], this elevates the importance of capturing both self-reported and behavioral metrics in research, as depending on the type of agent representation (and could extend to other factors, such as task features), responses may be overall distinct.

A minor contribution of the *Warehouse* study (Chapter 5) relevant to the rest of the dissertation is to investigate the viability of using simulation proxies to what the real world represents. For instance, human-robot interaction research often does make use of real, physical robots that can cost a substantial amount of a research group's budget. These robots need not be large or industrial in size, but small robots with highly-precise sensors and functionality can also incur a large cost. As discussed in Chapter 5, the proliferation of virtual reality and ease of development has allowed research to be conducted in a fully simulated setting, as to capture the feeling of the real world as close as possible [81, 97, 217]. Could a virtual reality experience then emulate human-robot interaction findings in the real world? A future avenue of research can aim to answer this question through a design where a group completes a task in virtual reality versus a group that completes it in the real world. If the findings and effects are then similar, a major cost reduction could potentially be achieved by focusing on virtual reality research within human-robot interaction, while using physical robots to validate the findings encountered during research. Chapter 5 is a starting point to this approach, as to compare a task that can be completed behind a computer screen can

be vastly different if it was completed using virtual embodied interactions. These embodied interactions can then be applied to the real world, and observe how much variation in trust, performance, and self-reported perceptions there is.

Individual Differences

Thanks to the mediation models demonstrated in this dissertation, we are able to appreciate how individual human characteristics fit within the entire human-AI interaction dynamic. In the Predator-Prey game (Chapter 4), we observe how trust propensity, complacency potential, and intrinsic motivation affect the level of workload, trust, and situation awareness. All the predisposition variables had a meaningful and interpretable relationship with the behavioral variables recorded during the task. This also serves to solidify the existence of these effects in continuous simulated scenarios, often distinct from the standard human-computer interaction and human factors research where users are placed in supervisory control of the automation and make a decision at discrete points in time (e.g., [26, 36, 57]).

We do not shy away from supervisory control, however, as the interaction within the *Warehouse* game (Chapter 5) does model supervisory control, as the robot only provides a recommendation, but the user has the authoritative say on what the final decision should be. We hypothesized that individual difference would drive trust and performance during the *Warehouse* game as well, and indeed corroborated this finding through a "naiveness" factor. This could be described using Tversky and Kahneman's model of judgment and decision-making, where a decision is made using one of two cognitive systems: an emotional system and a logical system. The emotional system makes use of the "gut feeling" sensation, often without consideration of slower rationalization of environmental information [258]. Without actually considering what the system is capable of doing, naive decisions follow their predisposed beliefs, which can be an inappropriate level of allocated trust or making a decision without logical thought. The level of "naiveness" here drove not only user trust towards the automation, but also behavioral trust metrics, such as calibration and adherence. The more "naive" a user was, the less capable they were at calibrating trust correctly. In retrospect and relation to prior research, experiments often take individual factors (e.g.,

trust propensity, complacency potential) and use them as co-variates to predict trust and performance outcomes, creating multiple implications across studies that become hard to reconcile what is relevant versus what is not. A possibility to entertain could be using mediation modeling to aggregate multiple individual differences into one latent factor that could more easily describe the variance in trust and outcomes, as found in this study.

6.2 FUTURE WORK

This dissertation has scratched the surface of the wide amount of available tasks and agent features that could be employed in a given scenario, and we have observed how the effects of reliability vary between these. Some relevant approaches to extend this dissertation include the varying the timeliness of system feedback to the user, the presence of system confidence and how it affects reliability perception (similar to [26] and [260]), real-time trust measurements in pursuit scenarios (similar to [32]), or investigating other scenarios in general to contribute to the growing body of reliability research across different domains. The integration between human and AI systems will require a much deeper understanding between domains, and the construction of taxonomies (such as Esterwood et al.'s task taxonomy [58] and Li's agent taxonomy [81]) will serve to connect domain experts from distinct fields into how to best optimize the research so it may generalize at a greater scale.

The AI systems built in the core studies are referred to as agents, entities, and teammates. There is current vivid discussion on whether these AI systems should even *be* considered as having human-like teammate qualities, and that instead these systems should be considered as "supertools", as the primary metaphor to guide their design and functionality [261]. The findings of this research are all contextualized under the fact that the systems were presented as teammates, and it can indeed be very different if the systems were framed as tools [262]. However, an important addition that this dissertation makes into that discussion is that it is likely that, without any priming, individual differences could define the metaphor humans use when interacting with these systems (in addition to other environmental and societal factors, such as social media and journalism inflating the capabilities of AI systems [263]). Without any prior education on these systems, it is likely users will use the metaphor they see

fit defined by prior experiences and individual differences. A worthwhile endeavor would be to investigate how individual differences then define the interaction metaphor that humans choose to use (a "teammate" or a "tool") to inform how engineers should design for the users – a reverse approach to Matthews et al. [264] –, as to match their mental model as close as possible.

6.3 SUMMARY

This dissertation has investigated how the effect of lowered reliability affects the human's ability to perform, trust, and decision-make across different domains. We identified and manipulated several system and user factors that predict performance and trust (both behavioral and self-reported) in order to contribute to the human-AI interaction dynamic and emphasize that much of the factors encountered in this research are so interconnected, it becomes challenging to interpret smaller findings in a vacuum. Additionally, every experimental task is so specific, any variation in task or user factors may change how reliability affects the human-AI interaction, resulting in different outcomes. We fully employed inferential statistics and mediation methodologies across all studies and expect that findings should generalize well into scenarios within each domain.

In the introduction (Chapter 1) of this dissertation, we presented the following inquiries:

1. How does lowering agent reliability affect task performance and human trust across different domains?

2. What is the relationship between human individual differences, agent reliability, and technology use?

3. How do different task requirements (e.g., task type, agent physicality, task urgency) interact with the human-AI interaction dynamic?

We present an answer to each of these inquiries, as follows:

1. We entered this dissertation with the expectation that agent reliability would predict trust and performance, as widely indicated by prior human factors research (e.g., [17,

22, 246]). Although this relationship generally holds true, when interacting with other factors (such as agent features or individual differences), a low reliability *can* bring forth positive effects; evidence demonstrated in this dissertation includes: over-trust and knowledge loss is prevented when an agent is transparent (Chapter 3), environmental interactions that lead to improved agent performance at no perceptual cost (Chapter 4), and improved trust calibration under no system transparency (Chapter 5).

2. We reveal through mediation modeling that individual differences have a large influence between reliability, trust, and performance. Chapter 4 demonstrates how trust propensity, complacency potential, intrinsic motivation, and prior experience mediates reliability and performance. Chapter 5 shows evidence of a "naiveness" factor – a latent variable stemming from predispositions or positive impressions from technology – which increases trust but negatively affects task outcomes (e.g., performance). Although individual differences are recognized by research to be influential in the development of trust [253], viewing these effects through mediation analysis gives us a holistic view on how different constructs relate to each other, often not done in human factors studies (i.e., individual differences are often relegated to co-variates against the dependent variable, without describing correlations with other relevant individual differences).

3. The domains explored in this dissertation follow a subset of the taxonomy defined by Esterwood et al.: influence tasks and physical manipulation tasks [58]. We discussed the findings with respect to its task domain, where we can infer the task generalizes to other associated contexts: Chapter 3 describes decision support systems – often found in domains such as medicine or finance; Chapter 4 covers simulated human-agent teams – the simulation itself serves as a validation bridge to bring virtual agents into the physical world, such as approaches in the Internet of Battlefield Things [169, 173]; Chapter 5 investigates human-robot interactions, a representative goal for embodied AI interactions. Each of these domains have distinct task requirements, and when designing agents to complement and fulfill these (e.g., transparency in decision support systems), it may affect how users calibrate their trust in the presence of differently

reliable systems, as automation is not expected to be perfect.

Based on these studies, there are several recommendations that could be made to system designers to enhance human-AI integration. First, we can design agents to push artificially low expectations to improve performance in unforeseen situations, as reduced reliability may benefit how humans process information and make decisions. Granted, this is not a universal solution, and the context in which the agent is deployed should be heavily considered (e.g., a surgical robot in trauma medicine versus a collaborative AI in a card game). Next, transparency with an imperfect agent may help users to compare their judgment against the explanations given and rationalize a high-quality recommendation against a low-quality one, allowing trust to be better distributed throughout the interaction. Designers should also be aware of how the environment itself can bring adverse reactions to the altered behavior of the agent, and that the perceived reliability of an agent can change with even small amounts of imperfections, leading to distinct mental models and trust habits. Therefore, for high-risk interactions, designers should focus on developing training regimes that allow users to draw comparisons between correct and incorrect behavior, and prime a mental model based on "what's correct." Conversely, for low-risk, agents should be designed to be informatively purposeful and explanatory (e.g., agents should state clearly when they are having issues providing a high-quality recommendation and why, and whether it is transient or persistent), as to help users build an understanding of the system's capabilities through repeated interactions.

Fundamentally, we must understand that the development and maintenance of calibrated trust is the key for proper use of AI systems that will complement and extend our capabilities in ways that we could have not imagined several decades ago. Technical advances keep pushing the limit of automated and AI systems, with no signs of stopping contributions in machine learning, computer vision, automated processing, recommender systems, and a wide amount of other technical fields. But at the end of the day, do the impact of these advances matter if the human is destined to remain "out-of-the-loop"? If these systems cannot be trusted properly by a human user to be used in an appropriate manner, then would there be a purpose to building these automated systems? The resounding answer to

these questions is, "No." Humans *must* remain a piece of the puzzle, and this dissertation is a step – alongside the entire human-AI interaction research community – to find the piece's fit.

REFERENCES

[1] E. A. Feigenbaum, "Artificial intelligence research," *IEEE Transactions on Information Theory*, vol. 9, no. 4, pp. 248–253, 1963.

[2] D. Manzey, J. Reichenbach, and L. Onnasch, "Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience," *Journal of Cognitive Engineering and Decision Making*, vol. 6, no. 1, pp. 57–87, 2012.

[3] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, and B. Livingston, "The YouTube video recommendation system," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 293–296.

[4] L. Santos, J. Coutinho-Rodrigues, and C. H. Antunes, "A web spatial decision support system for vehicle routing using Google Maps," *Decision Support Systems*, vol. 51, no. 1, pp. 1–9, 2011.

[5] N. Ho, W. Johnson, K. Panesar, K. Wakeland, G. Sadler, N. Wilson, B. Nguyen, J. Lachter, and S. Brandt, "Application of human-autonomy teaming to an advanced ground station for reduced crew operations," in *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*. IEEE, 9 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8102124/ pp. 1–4.

[6] J. Y. C. Chen and M. J. Barnes, "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, 2 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6697830/

[7] P. Lai, "THE LITERATURE REVIEW OF TECHNOLOGY ADOPTION MODELS AND THEORIES FOR THE NOVELTY TECHNOLOGY," *Journal of Information Systems and Technology Management*, vol. 14, no. 1, 4 2017. [Online]. Available: http://www.jistem.tecsi.org/index.php/jistem/article/view/10.4301%252FS1807-17752017000100002/643

[8] P. A. Hancock, "Imposing limits on autonomous systems," *Ergonomics*, vol. 60, no. 2, pp. 284–291, 2 2017. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00140139.2016.1190035

[9] N. Postman, *Technopoly: The Surrender of Culture to Technology*. Vintage, 1992.

[10] R. Parasuraman, "Humans and Automation: Use, Misuse, Disuse, Abuse," Tech. Rep. 2, 1997.

[11] U. Metzger and R. Parasuraman, "Automation-Related "Complacency": Theory, Empirical Data, and Design Implications," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 45, no. 4, pp. 463–467, 10 2001. [Online]. Available: http://journals.sagepub.com/doi/10.1177/154193120104500442

[12] J. C. F. de Winter and D. Dodou, "Why the Fitts list has persisted throughout the history of function allocation," *Cognition, Technology & Work*, vol. 16, no. 1, pp. 1–11, 2 2014. [Online]. Available: http://link.springer.com/10.1007/s10111-011-0188-1

[13] T. Kessler, K. Stowers, J. Brill, and P. Hancock, "Comparisons of Human-Human Trust with Other Forms of Human-Technology Trust," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, no. 1, pp. 1303–1307, 9 2017. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1541931213601808

[14] R. Parasuraman and D. H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 52, no. 3, pp. 381–410, 6 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720810376055

[15] N. Du, Q. Zhang, and X. J. Yang, "Evaluating effects of automation reliability and reliability information on trust, dependence and dual-task performance," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 174–174, 9 2018.

[16] S. Y. Chien, M. Lewis, K. Sycara, J. S. Liu, and A. Kumru, "The effect of culture on trust in automation: Reliability and workload," *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 4, 11 2018.

[17] C. Wang, C. Zhang, and X. J. Yang, "Automation reliability and trust: A Bayesian inference approach," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 202–206, 9 2018.

[18] R. W. Wohleber, G. L. Calhoun, G. J. Funke, H. Ruff, C.-Y. P. Chiu, J. Lin, and G. Matthews, "The Impact of Automation Reliability and Operator Fatigue on Performance and Reliance," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1, pp. 211–215, 9 2016.

[19] S. Stumpf, "Explanations Considered Harmful? User Interactions with Machine Learning Systems," *ACM SIGCHI Workshop on Human-Centered Machine Learning*, pp. 1511–1524, 2016.

[20] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning," pp. 1–14, 2020.

[21] C. Ward, M. Raue, C. Lee, L. D'Ambrosio, and J. F. Coughlin, "Acceptance of Automated Driving Across Generations: The Role of Risk and Benefit Perception, Knowledge, and Trust," 2017, pp. 254–266. [Online]. Available: https://link.springer.com/10.1007/978-3-319-58071-5_20

[22] L. Rodriguez Rodriguez, C. Bustamante Orellana, J. Landfair, C. Magaldino, M. Demir, P. G. Amazeen, J. S. Metcalfe, L. Huang, and Y. Kang, "Dynamics of Trust in Automation and Interactive Decision Making during Driving Simulation Tasks," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 786–790, 9 2021.

[23] C. Wickens and X. Xu, "How does automation reliability influence workload?" *University of Illinois at Urbana-Champaign, Aviation Human Factors Division, Tech. Rep*, 2002.

[24] M. Barnes, L. R. Elliott, J. L. Wright, and A. Scharine, "Human-Robot Interaction Design Research: From Teleoperations to Human-Agent Teaming Human-Robot Interaction Design Research: From Teleoperations to Human-Agent Teaming Human-Agent Teaming Agent Transparency View project Sustained C4ISR operations View project," Tech. Rep., 2019. [Online]. Available: https://www.researchgate.net/publication/335589582

[25] V. Lai and C. Tan, "On Human Predictions with Explanations and Predictions of Machine Learning Models," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, 1 2019, pp. 29–38.

[26] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, 1 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3351095.3372852 pp. 295–305.

[27] J. Schaffer, "A Quantitative Investigation into the Design Trade-offs in Decision Support Systems," Ph.D. dissertation, University of California, Santa Barbara, 2016. [Online]. Available: https://escholarship.org/uc/item/2x55z1dz

[28] "Attentional Control and Performance Across Increasing Degrees of Unreliable Automation," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 64, no. 1, pp. 1238–1241, 12 2020.

[29] E. J. de Visser, S. S. Monfort, R. McKendrick, M. A. B. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman, "Almost human: Anthropomorphism increases trust resilience in cognitive agents," *Journal of Experimental Psychology: Applied*, vol. 22, no. 3, pp. 331–349, 9 2016. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/xap0000092

[30] P. Kulms and S. Kopp, "More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human–agent cooperation," *ACM International Conference Proceeding Series*, pp. 31–42, 2019.

[31] R. Molloy and R. Parasuraman, "Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 38, no. 2, pp. 311–322, 6 1996. [Online]. Available: http://journals.sagepub.com/doi/10.1177/001872089606380211

[32] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 3 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6483596/ pp. 251–258.

[33] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, "The effect of presence on human-robot interaction," *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, pp. 701–706, 2008.

[34] K. M. Lee, "Presence, Explicated," *Communication Theory*, vol. 14, no. 1, pp. 27–50, 2 2004. [Online]. Available: https://academic.oup.com/ct/article/14/1/27-50/4110793

[35] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems," *Human Factors*, vol. 58, no. 3, pp. 377–400, 2016.

[36] J. Schaffer, J. O'Donovan, L. Marusich, M. Yu, C. Gonzalez, and T. Höllerer, "A study of dynamic information display and decision-making in abstract trust games," *International Journal of Human-Computer Studies*, vol. 113, pp. 1–14, 5 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1071581918300028

[37] J.-M. Hoc, "Towards ecological validity of research in cognitive ergonomics," *Theoretical Issues in Ergonomics Science*, vol. 2, no. 3, pp. 278–288, 1 2001. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14639220110104970

[38] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 1 2004. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50_30392

[39] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships." *Journal of Personality and Social Psychology*, vol. 49, no. 1, pp. 95–112, 1985. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.49.1.95

[40] H. H. Tan and C. S. Tan, "Toward the differentiation of trust in supervisor and trust in organization." *Genetic, social, and general psychology monographs*, vol. 126, no. 2, pp. 241–260, 5 2000.

[41] R. M. Morgan and S. D. Hunt, "The Commitment-Trust Theory of Relationship Marketing," *Journal of Marketing*, vol. 58, no. 3, p. 20, 7 1994. [Online]. Available: https://www.jstor.org/stable/1252308?origin=crossref

[42] G. Müller, "Secure communication Trust in technology or trust with technology?" *Interdisciplinary Science Reviews*, vol. 21, no. 4, pp. 336–347, 12 1996. [Online]. Available: http://www.tandfonline.com/doi/full/10.1179/isr.1996.21.4.336

[43] R. C. Nyhan, "Changing the Paradigm," *The American Review of Public Administration*, vol. 30, no. 1, pp. 87–109, 3 2000. [Online]. Available: http://journals.sagepub.com/doi/10.1177/02750740022064560

[44] B. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 3 1996. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00140139608964474

[45] J. Lee, C. Wickens, Y. Liu, and L. Boyle, *Designing for People: An Introduction to Human Factors Engineering*, 3rd ed. CreateSpace Independent Publishing Platform, 2017.

[46] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe, "The perceived utility of human and automated aids in a visual detection task," *Human Factors*, vol. 44, no. 1, pp. 79–94, 2002.

[47] A. Kirlik, K. Ackerman, B. Seefeldt, E. Xargay, K. Riddle, D. Talleur, R. Carbonari, L. Sha, and N. Hovakimyan, *Visualizing automation in aviation interfaces*, 2019, no. September.

[48] C. Billings and D. Woods, "Concerns about adaptive automation in aviation systems," in *Human performance in automated systems: Current research and trends*, M. Mouloua and R. Parasuraman, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1994, pp. 24–29.

[49] G. A. Hadley, L. J. Prinzel, F. G. Freeman, and P. J. Mikulka, "Behavioral, subjective and psychophysiological correlates of various schedules of short-cycle automation," in *Automation technology and human performance*, M. W. Scerbo and M. Mouloua, Eds. NJ: Erlbaum: Mahwah, 1999, pp. 139–143.

[50] M. R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," in *Human Error in Aviation*. Routledge, 7 1995, vol. 37, no. 1, pp. 217–249. [Online]. Available: https://www.taylorfrancis.com/books/9781351563475/chapters/10.4324/9781315092898-13

[51] M. R. Endsley, "Design and Evaluation for Situation Awareness Enhancement," *Proceedings of the Human Factors Society Annual Meeting*, vol. 32, no. 2, pp. 97–101, 10 1988. [Online]. Available: http://journals.sagepub.com/doi/10.1177/154193128803200221

[52] M. R. Endsley and E. O. Kiris, "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 1995.

[53] D. Campbell, "Redline: The many human errors that brought down the Boeing 737 Max," 2019. [Online]. Available: https://www.theverge.com/2019/5/2/18518176/boeing-737-max-crash-problems-human-error-mcas-faa

[54] J. Y. C. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. J. Barnes, "Situation Awareness–Based Agent Transparency," *US Army Research Laboratory*, no. April, pp. 1–29, 2014.

[55] L. J. Prinzel, H. Devries, F. G. Freeman, and P. Mikulka, "Examination of Automation-Induced Complacency and Individual Difference Variates," Tech. Rep., 2001. [Online]. Available: https://ntrs.nasa.gov/search.jsp?R=20020021642

[56] T. B. Sheridan, "Individual differences in attributes of trust in automation: Measurement and application to system design," *Frontiers in Psychology*, vol. 10, no. MAY, pp. 1–7, 2019.

[57] J. B. Lyons and S. Y. Guznov, "Individual differences in human–machine trust: A multi-study look at the perfect automation schema," *Theoretical Issues in Ergonomics Science*, vol. 20, no. 4, pp. 440–458, 2019.

[58] C. Esterwood, K. Essenmacher, and H. Yang, "A meta-analysis of human personality and robot acceptance in human-robot interaction," *Conference on Human Factors in Computing Systems - Proceedings*, 2021.

[59] J. Sanchez, W. A. Rogers, A. D. Fisk, and E. Rovira, "Understanding reliance on automation: effects of error type, error distribution, age and experience," *Theoretical Issues in Ergonomics Science*, vol. 15, no. 2, pp. 134–160, 3 2014.

[60] C. Textor and R. Pak, "Paying Attention to Trust: Exploring the Relationship Between Attention Control and Trust in Automation," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 817–821, 9 2021.

[61] R. Parasuraman, R. Molloy, and I. L. Singh, "Performance Consequences of Automation-Induced 'Complacency'," *The International Journal of Aviation Psychology*, vol. 3, no. 1, pp. 1–23, 1 1993. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327108ijap0301_1

[62] E. L. Wiener, "Complacency: Is the term useful for air safety," in *Proceedings of the 26th Corporate Aviation Safety Seminar*, vol. 117, 1981, pp. 116–125.

[63] C. E. Billings, J. K. Lauber, H. Funkhouser, E. G. Lyman, and E. M. Huff, "NASA AVIATION SAFETY REPORTING SYSTEM QUARTERLY REPORT NUMBER 76-1," NASA, Tech. Rep., 1976. [Online]. Available: https://ntrs.nasa.gov/search.jsp?R=19760026757

[64] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *The International Journal of Aviation Psychology*, vol. 8, no. 1, pp. 47–63, 1 1998. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327108ijap0801_3

[65] J. Ferraro, L. Clark, N. Christy, and M. Mouloua, "Effects of Automation Reliability and Trust on System Monitoring Performance in Simulated Flight Tasks," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 1232–1236, 9 2018.

[66] M. R. Endsley, "From Here to Autonomy: Lessons Learned from Human-Automation Research," *Human Factors*, vol. 59, no. 1, pp. 5–27, 2017.

[67] J. Meyer, "Conceptual issues in the study of dynamic hazard warnings," *Human Factors*, vol. 46, no. 2, pp. 196–204, 2004.

[68] R. Wiczorek, J. Meyer, and T. Guenzler, "On the Relation Between Reliance and Compliance in an Aided Visual Scanning Task," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1, pp. 253–257, 9 2012. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1071181312561060

[69] S. M. Merritt, A. Ako-Brew, W. J. Bryant, A. Staley, M. McKenna, A. Leone, and L. Shirase, "Automation-Induced Complacency Potential: Development and Validation of a New Scale," *Frontiers in Psychology*, vol. 10, 2 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2019.00225/full

[70] N. Bagheri and G. A. Jamieson, "Considering Subjective Trust and Monitoring Behavior in Assessing," Tech. Rep. 1993, 2004.

[71] N. Bagheri and G. Jamieson, "The impact of context-related reliability on automation failure detection and scanning behaviour," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, vol. 1. IEEE, 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1398299/ pp. 212–217.

[72] K. I. Fletcher, M. L. Bartlett, S. J. Cockshell, and J. S. McCarley, "Visualizing Probability of Detection to Aid Sonar Operator Performance," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, no. 1, pp. 302–306, 9 2017. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1541931213601556

[73] J. Lee, G. Abe, K. Sato, and M. Itoh, "Developing human-machine trust: Impacts of prior instruction and automation failure on driver trust in partially automated vehicles," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 81, pp. 384–395, 8 2021.

[74] J. Reichenbach, L. Onnasch, and D. Manzey, "Misuse of automation: The impact of system experience on complacency and automation bias in interaction with automated aids," in *Proceedings of the Human Factors and Ergonomics Society*, vol. 1, 2010, pp. 374–378.

[75] E. J. de Visser, R. Pak, and T. H. Shaw, "From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018. [Online]. Available: https://doi.org/10.1080/00140139.2018.1457725

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

[77] I. L. Singh, R. Molloy, and R. Parasuraman, "Automation-Induced "Complacency": Development of the Complacency-Potential Rating Scale," *The International Journal of Aviation Psychology*, vol. 3, no. 2, pp. 111–122, 1993. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84963396144&doi=10.1207%2Fs15327108ijap0302_2&partnerID=40&md5=924a70c4ea12495ddffe3c5ec392e12d

[78] P. Satehi, E. K. Chiou, and A. Wilkins, "Human-agent interactions: Does accountability matter in interactive control automation?" in *62nd Human Factors and Ergonomics Society Annual Meeting, HFES 2018*, vol. 3. Department of Human Systems Engineering, Arizona State University, Mesa, AZ, United States: Human Factors and Ergonomics Society Inc., 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85072750094&partnerID=40&md5=d82f329b40ce74aba38d7836ed17d345 pp. 1643–1647.

[79] J. E. Bahner, A.-D. Hüper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *International Journal of Human-Computer Studies*, vol. 66, no. 9, pp. 688–699, 9 2008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1071581908000724

[80] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2015-March, pp. 141–148, 2015.

[81] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 5 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S107158191500004X

[82] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 177–190, 3 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0921889002003743

[83] J. Cassell, "Embodied Conversational Agents Representation and Intelligence in User Interfaces," *American Association for Artificial Intelligence*, pp. 67–84, 2001.

[84] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, 8 2014.

[85] J. Gratch, "The promise and peril of interactive embodied agents for studying non-verbal communication: a machine learning perspective," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 378, no. 1875, 4 2023.

[86] S. Herse, J. Vitale, M. Tonkin, D. Ebrahimian, S. Ojha, B. Johnston, W. Judge, and M. A. Williams, "Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System," *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, no. August, pp. 7–14, 2018.

[87] A. Mollahosseini, H. Abdollahi, T. D. Sweeny, R. Cole, and M. H. Mahoor, "Role of embodiment and presence in human perception of robots' facial cues," *International Journal of Human-Computer Studies*, vol. 116, pp. 25–39, 8 2018.

[88] R. C. Johnson, K. N. Saboe, M. S. Prewett, M. D. Coovert, and L. R. Elliott, "Autonomy and automation reliability in human-robot interaction: A qualitative review," in *Proceedings of the Human Factors and Ergonomics Society*, vol. 3. Human Factors an Ergonomics Society Inc., 2009, pp. 1398–1402.

[89] F. Legler, D. Langer, F. Dittrich, and A. C. Bullinger, "I don't care what the robot does! Trust in automation when working with a heavy-load robot," in *Human Factors and Ergonomics Society*, 2019.

[90] J. Hollan, E. Hutchins, and D. Kirsh, "Distributed cognition," *ACM Transactions on Computer-Human Interaction*, vol. 7, no. 2, pp. 174–196, 6 2000. [Online]. Available: https://dl.acm.org/doi/10.1145/353485.353487

[91] N. Tintarev and J. Masthoff, "A Survey of Explanations in Recommender Systems," in *2007 IEEE 23rd International Conference on Data Engineering Workshop*. IEEE, 4 2007. [Online]. Available: http://ieeexplore.ieee.org/document/4401070/ pp. 801–810.

[92] Arnold, Clark, Collier, Leech, and Sutton, "The Differential Use and Effect of Knowledge-Based System Explanations in Novice and Expert Judgment Decisions," *MIS Quarterly*, vol. 30, no. 1, p. 79, 2006. [Online]. Available: https://www.jstor.org/stable/10.2307/25148718

[93] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa, "Inspectability and control in social recommenders," in *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*. New York, New York, USA: ACM Press, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2365952.2365966 p. 43.

[94] J. O'Donovan and B. Smyth, "Trust in recommender systems," in *Proceedings of the 10th international conference on Intelligent user interfaces - IUI '05*. New York, New York, USA: ACM Press, 2005. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1040830.1040870 p. 167.

[95] M. S. Luster and B. J. Pitts, "Trust in Automation: The Effects of System Certainty on Decision-Making," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 32–36, 9 2021.

[96] V. Villani, B. Capelli, and L. Sabattini, "Use of Virtual Reality for the Evaluation of Human-Robot Interaction Systems in Complex Scenarios," *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 422–427, 2018.

[97] O. Liu, D. Rakita, B. Mutlu, and M. Gleicher, "Understanding human-robot interaction in virtual reality," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, vol. 1, no. 8. IEEE, 8 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8172387/ pp. 751–757.

[98] S. Röttger, K. Bali, and D. Manzey, "Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task," *Ergonomics*, vol. 52, no. 5, pp. 512–523, 2009.

[99] J. Reichenbach, L. Onnasch, and D. Manzey, "Human performance consequences of automated decision aids in states of sleep loss," *Human Factors*, 2011.

[100] C. D. Wickens, B. A. Clegg, A. Z. Vieane, and A. L. Sebok, "Complacency and Automation Bias in the Use of Imperfect Automation," *Human Factors*, vol. 57, no. 5, pp. 728–739, 2015.

[101] U. Metzger, J. A. Duley, R. Abbas, and R. Parasuraman, "Effects of Variable-Priority Training on Automation-Related Complacency: Performance and Eye Movements," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, no. 11, pp. 346–349, 7 2000. [Online]. Available: http://journals.sagepub.com/doi/10.1177/154193120004401104

[102] N. Bagheri and G. A. Jamieson, "A Sampling Model to Ascertain Automation-Induced Complacency in Multi-Task Environments," in *Human Error, Safety and Systems Development*. Boston: Kluwer Academic Publishers, 2004, pp. 131–145. [Online]. Available: http://link.springer.com/10.1007/1-4020-8153-7_9

[103] I. L. Singh, R. Molloy, and R. Parasuraman, "Automation-induced monitoring inefficiency: Role of display location," *International Journal of Human Computer Studies*, 1997.

[104] L. J. I. Prinzel, "The Relationship of Self-Efficacy and Complacency in Pilot-Automation Interaction," *Nasa/Tm-2002-211925*, no. September, 2002.

[105] N. R. Bailey and M. W. Scerbo, "Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust," *Theoretical Issues in Ergonomics Science*, vol. 8, no. 4, pp. 321–348, 2007.

[106] S. M. Casner, R. W. Geven, M. P. Recker, and J. W. Schooler, "The Retention of Manual Flying Skills in the Automated Cockpit," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 56, no. 8, pp. 1506–1516, 12 2014. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720814535628

[107] J. Gouraud, A. Delorme, and B. Berberian, "Autopilot, mind wandering, and the out of the loop performance problem," *Frontiers in Neuroscience*, vol. 11, no. OCT, 2017. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85030675307&doi=10.3389%2Ffnins.2017.00541&partnerID=40&md5=e8fa19045e0e225eaf36327d021759ce

[108] J. Gouraud, A. Delorme, and B. Berberian, "Out of the Loop, in Your Bubble: Mind Wandering Is Independent From Automation Reliability, but Influences Task Engagement," *Frontiers in Human Neuroscience*, vol. 12, 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054886832&doi=10.3389%2Ffnhum.2018.00383&partnerID=40&md5=b685ff63b5c5d277229e19c3a84193a2

[109] F. Trapsilawati, X. Qu, C. D. Wickens, and C.-H. Chen, "Human factors assessment of conflict resolution aid reliability and time pressure in future air traffic control," *Ergonomics*, vol. 58, no. 6, pp. 897–908, 6 2015.

[110] J. Langan-Fox, M. J. Sankey, and J. M. Canty, "Human Factors Measurement for Future Air Traffic Control Systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 51, no. 5, pp. 595–637, 10 2009.

[111] E. Tivesten, T. W. Victor, P. Gustavsson, J. Johansson, and M. Ljung Aust, "Out-of-the-loop crash prediction: the automation expectation mismatch (AEM) algorithm," *IET Intelligent Transport Systems*, vol. 13, no. 8, pp. 1231–1240, 8 2019. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-its.2018.5555

[112] F. Shahini, J. Park, and M. Zahabi, "Effects of unreliable automation and takeover time budget on young drivers' mental workload," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 1082–1086, 9 2021.

[113] S. M. Merritt, H. Heimbaugh, J. Lachapell, and D. Lee, "I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Human Factors*, vol. 55, no. 3, pp. 520–534, 2013.

[114] J. Davis, A. Atchley, H. Smitherman, H. Simon, and N. Tenhundfeld, "Measuring Automation Bias and Complacency in an X-Ray Screening Task," in *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 4 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9106670/ pp. 1–5.

[115] S. M. Merritt and D. R. Ilgen, "Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 2, pp. 194–210, 4 2008. [Online]. Available: http://journals.sagepub.com/doi/10.1518/001872008X288574

[116] K. Korbelak, J. Dressel, D. Tweedie, W. Wilson, S. Erchov, and B. Hilburn, "Teaming with Technology at the TSA: An Examination of Trust in Automation's Influence on Human Performance in Operational Environments," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 656–660, 9 2018.

[117] P. Madhavan and D. A. Wiegmann, "Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 49, no. 5, pp. 773–785, 10 2007.

[118] A. Acharya, A. Howes, C. Baber, and T. Marshall, "Automation Reliability and Decision Strategy: A Sequential Decision Making Model for Automation Interaction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 144–148, 9 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1541931218621033

[119] M. Sujan, D. Furniss, K. Grundy, H. Grundy, D. Nelson, M. Elliott, S. White, I. Habli, and N. Reynolds, "Human factors challenges for the safe use of artificial intelligence in patient care," *BMJ Health & Care Informatics*, vol. 26, no. 1, p. e100081, 11 2019. [Online]. Available: https://informatics.bmj.com/lookup/doi/10.1136/bmjhci-2019-100081

[120] D. Manzey, S. Röttger, J. E. Bahner-Heyne, D. Schulze-Kissing, A. Dietz, J. Meixensberger, and G. Strauss, "Image-guided navigation: the surgeon's perspective on performance consequences and human factors issues," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 5, no. 3, pp. 297–308, 9 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/rcs.261

[121] E. Rovira, K. McGarry, and R. Parasuraman, "Effects of imperfect automation on decision making in a simulated command and control task," *Human Factors*, vol. 49, no. 1, pp. 76–87, 2007.

[122] R. Parasuraman, E. de Visser, M.-K. Lin, and P. M. Greenwood, "Dopamine Beta Hydroxylase Genotype Identifies Individuals Less Susceptible to Bias in Computer-Assisted Decision Making," *PLoS ONE*, vol. 7, no. 6, p. e39675, 6 2012. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0039675

[123] M. Cummings, L. Huang, H. Zhu, D. Finkelstein, and R. Wei, "The Impact of Increasing Autonomy on Training Requirements in a UAV Supervisory Control Task," *Journal of Cognitive Engineering and Decision Making*, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071448317&doi=10.1177%2F1555343419868917&partnerID=40&md5=05a95c5eb2058a688049d9ed096bd670

[124] J. Y. Chen and M. J. Barnes, "Supervisory control of multiple robots: Effects of imperfect automation and individual differences," *Human Factors*, vol. 54, no. 2, pp. 157–174, 2012.

[125] J. Lin, G. Matthews, R. Wohleber, C.-Y. P. Chiu, G. Calhoun, G. Funke, and H. Ruff, "Automation Reliability and Other Contextual Factors in Multi-UAV Operator Selection," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1, pp. 846–850, 9 2016.

[126] J. L. Wright, "Transparency in Human-agent Teaming and its Effect on Automation-induced Complacency," *Procedia Manufacturing*, vol. 3, pp. 968–973, 1 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235197891500150X?via%3Dihub#aep-article-footnote-id1

[127] J. L. Wright, J. Y. Chen, and S. G. Lakhmani, "Agent Transparency and Reliability in Human-Robot Interaction: The Influence on User Confidence and Perceived Reliability," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 3, pp. 254–263, 2020.

[128] S. K. Hopko, R. K. Mehta, and A. D. McDonald, "Trust in Automation: Comparison of Automobile, Robot, Medical, and Cyber Aid Technologies," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 462–466, 9 2021.

[129] D. Manzey, M. Bleil, J. E. Bahner-Heyne, A. Klostermann, L. Onnasch, J. Reichenbach, and S. Röttger, "AutoCAMS 2.0," 2008.

[130] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock, "The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide," 2011.

[131] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, 2012.

[132] J. Schaffer, J. O'Donovan, and T. Höllerer, "Easy to please: Separating user experience from choice satisfaction," *UMAP 2018 - Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pp. 177–185, 2018.

[133] J. Schaffer, J. Humann, J. O'Donovan, and T. Höllerer, "Quantitative Modeling of Dynamic Human-Agent Cognition," *Contemporary Research*, pp. 137–186, 2020.

[134] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig, "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2015-April, pp. 153–162, 2015.

[135] "Internet Movie Database." [Online]. Available: imdb.com

[136] "Movielens." [Online]. Available: movielens.org

[137] F. M. Harper and J. A. Konstan, "The MovieLens Datasets," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, pp. 1–19, 1 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2827872

[138] B. N. Miller, "GroupLens: An Open Architecture for Collaborative Filtering," *Proceedings of CSCW94*, pp. 175–186, 1995.

[139] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*. New York, New York, USA: ACM Press, 1999. [Online]. Available: http://portal.acm.org/citation.cfm?doid=312624.312682 pp. 230–237.

[140] T. Raykov, "Studying Correlates and Predictors of Longitudinal Change Using Structural Equation Modeling," *Applied Psychological Measurement*, vol. 18, no. 1, pp. 63–77, 1994.

[141] T. Raykov and G. A. Marcoulides, "A First Course in Structural Equation Modeling," 2000.

[142] J. Andreoni and J. H. Miller, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence," *The Economic Journal*, vol. 103, no. 418, p. 570, 5 1993. [Online]. Available: https://academic.oup.com/ej/article/103/418/570-585/5157331

[143] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson, "Rational cooperation in the finitely repeated prisoners' dilemma," *Journal of Economic Theory*, vol. 27, no. 2, pp. 245–252, 1982.

[144] R. Axelrod, *The Evolution of Cooperation: Revised Edition*. Basic Books, 2009. [Online]. Available: https://books.google.com/books?id=GxRo5hZtxkEC

[145] Y. Teng, R. Jones, L. Marusich, J. O'Donovan, C. Gonzalez, and T. Höllerer, "Trust and situation awareness in a 3-player diner's dilemma game," *2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2013*, pp. 9–15, 2013.

[146] V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton, "Explanation provision and use in an intelligent decision aid," *Intelligent Systems in Accounting, Finance & Management*, vol. 12, no. 1, pp. 5–27, 1 2004. [Online]. Available: http://doi.wiley.com/10.1002/isaf.222

[147] T. Claburn, "Europe mulls treating robots legally as people ... but with kill switches." [Online]. Available: https://www.theregister.co.uk/2017/01/13/eu_treat_robots_as_people/

[148] S. S. Rodriguez, J. O'Donovan, J. A. Schaffer, and T. Hollerer, "Knowledge Complacency and Decision Support Systems," in *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 4 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8724175/ pp. 43–51.

[149] E. E. Entin and D. Serfaty, "Adaptive Team Coordination," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 41, no. 2, pp. 312–325, 6 1999. [Online]. Available: http://journals.sagepub.com/doi/10.1518/001872099779591196

[150] H. Arrow, J. McGrath, and J. Berdahl, *Small Groups as Complex Systems: Formation, Coordination, Development, and Adaptation*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2000. [Online]. Available: http://sk.sagepub.com/books/small-groups-as-complex-systems

[151] A. Kott, "Intelligent Autonomous Agents are Key to Cyber Defense of the Future Army Networks," *The Cyber Defense Review*, vol. 3, no. 3, 2018. [Online]. Available: https://cyberdefensereview.army.mil/Portals/6/Documents/CDRJournalArticles/Fall2018/KOTT_CDR_V3N3.pdf?ver=2018-12-18-101632-597

[152] C. Lawson-Guidigbe, N. Louveton, K. Amokrane-Ferka, B. LeBlanc, and J. M. Andre, "Impact of Visual Embodiment on Trust for a Self-driving Car Virtual Agent: A Survey Study and Design Recommendations," *Communications in Computer and Information Science*, vol. 1226 CCIS, pp. 382–389, 2020.

[153] J. E. H. Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, "Human- versus Artificial Intelligence," *Frontiers in Artificial Intelligence*, vol. 4, 3 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2021.622364/full

[154] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, p. 19, 2019. [Online]. Available: www.aaai.org

[155] N. J. McNeese, M. Demir, E. K. Chiou, and N. J. Cooke, "Trust and Team Performance in Human–Autonomy Teaming," *International Journal of Electronic Commerce*, vol. 25, no. 1, pp. 51–72, 1 2021. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10864415.2021.1846854

[156] R. Langner and S. B. Eickhoff, "Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention," *Psychological Bulletin*, vol. 139, no. 4, pp. 870–900, 7 2013. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030694

[157] M. Campbell, A. Hoane, and F.-h. Hsu, "Deep Blue," *Artificial Intelligence*, vol. 134, no. 1-2, pp. 57–83, 1 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0004370201001291

[158] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, I. Dunning, S. Mourad, H. Larochelle, M. G. Bellemare, and M. Bowling, "The Hanabi Challenge: A New Frontier for AI Research," 2 2019. [Online]. Available: http://arxiv.org/abs/1902.00506http://dx.doi.org/10.1016/j.artint.2019.103216

[159] E. Gibney, "Google AI algorithm masters ancient game of Go," *Nature*, vol. 529, no. 7587, pp. 445–446, 1 2016. [Online]. Available: http://www.nature.com/articles/529445a

[160] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 12 2018. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aar6404

[161] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, J. Quan, S. Gaffney, S. Petersen, K. Simonyan, T. Schaul, H. van Hasselt, D. Silver, T. Lillicrap, K. Calderone, P. Keet, A. Brunasso, D. Lawrence, A. Ekermo, J. Repp, and R. Tsing, "StarCraft II: A New Challenge for Reinforcement Learning," 8 2017. [Online]. Available: http://arxiv.org/abs/1708.04782

[162] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castañeda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, N. Sonnerat, T. Green, L. Deason, J. Z. Leibo, D. Silver, D. Hassabis, K. Kavukcuoglu, and T. Graepel, "Human-level performance in 3D multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 5 2019. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aau6249

[163] C. Liang, J. Proft, E. Andersen, and R. A. Knepper, "Implicit Communication of Actionable Information in Human-AI teams," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 5 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3290605.3300325 pp. 1–13.

[164] K. I. Gero, Z. Ashktorab, C. Dugan, Q. Pan, J. Johnson, W. Geyer, M. Ruiz, S. Miller, D. R. Millen, M. Campbell, S. Kumaravel, and W. Zhang, "Mental Models of AI Agents in a Cooperative Game Setting," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 4 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376316 pp. 1–12.

[165] S. Havrylov and I. Titov, "Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols," 5 2017. [Online]. Available: http://arxiv.org/abs/1705.11192

[166] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-Agent Cooperation and the Emergence of (Natural) Language," 12 2016. [Online]. Available: http://arxiv.org/abs/1612.07182

[167] Z. Ashktorab, C. Dugan, J. Johnson, Q. Pan, W. Zhang, S. Kumaravel, and M. Campbell, "Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction," *CHI 2021*, 2021.

[168] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," 6 2017. [Online]. Available: http://arxiv.org/abs/1706.02275

[169] S. Barton and D. Asher, "Reinforcement learning framework for collaborative agents interacting with soldiers in dynamic military contexts," *SPIE*, vol. 1065303, no. April 2018, p. 2, 2018.

[170] N. J. McNeese, M. Demir, N. J. Cooke, and C. Myers, "Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming," *Human Factors*, vol. 60, no. 2, pp. 262–273, 2018.

[171] A. Kolling, P. Walker, N. Chakraborty, K. Sycara, and M. Lewis, "Human Interaction With Robot Swarms: A Survey," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 9–26, 2 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7299280/

[172] A. H. DeCostanza, A. R. Marathe, A. Bohannon, A. W. Evans, E. T. Palazzolo, J. S. Metcalfe, and K. McDowell, "Enhancing Human–Agent Teaming with Individualized, Adaptive Technologies: A Discussion of Critical Scientific Questions," *IEEE Brain*, no. June, p. 38, 2018. [Online]. Available: https://apps.dtic.mil/docs/citations/AD1051552

[173] D. Asher, E. Zaroukian, and S. Barton, "Adapting the Predator-Prey Game Theoretic Environment to Army Tactical Edge Scenarios with Computational Multiagent Systems," 7 2018. [Online]. Available: http://arxiv.org/abs/1807.05806

[174] T. H. Chung, G. A. Hollinger, and V. Isler, "Search and pursuit-evasion in mobile robotics A survey," *Autonomous Robots*, vol. 31, no. 4, pp. 299–316, 2011.

[175] E. Zaroukian, S. Rodriguez, S. Barton, J. Schaffer, B. Perelman, N. Waytowich, B. Hoffman, and D. Asher, "Algorithmically identifying strategies in multi-agent game-theoretic environments," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, T. Pham, Ed., vol. 11006. SPIE, 5 2019. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11006/2518609/Algorithmically-identifying-strategies-in-multi-agent-game-theoretic-environments/10.1117/12.2518609.full p. 38.

[176] D. Asher, M. Garber-Barron, S. Rodriguez, E. Zaroukian, and N. Waytowich, "Multi-Agent Coordination Profiles through State Space Perturbations," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 12 2019. [Online]. Available: https://ieeexplore.ieee.org/document/9070901/ pp. 249–252.

[177] T. B. Sheridan and W. L. Verplank, "Human and Computer Control of Undersea Teleoperators," MASSACHUSETTS INST OF TECH CAMBRIDGE MAN-MACHINE SYSTEMS LAB, Tech. Rep., 1978.

[178] P. J. Hinds, T. L. Roberts, and H. Jones, "Whose job is it anyway? A study of human-robot interaction in a collaborative task," 2004.

[179] M. Q. Azhar and E. I. Sklar, "A study measuring the impact of shared decision making in a human-robot team," *International Journal of Robotics Research*, vol. 36, no. 5-7, pp. 461–482, 2017.

[180] W. Lin, Z. Qu, and M. A. Simaan, "A Design of Entrapment Strategies for the Distributed Pursuit-Evasion Game," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 9334–9339, 1 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1474667016451116

[181] S. Honig and T. Oron-Gilad, "Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development," *Frontiers in Psychology*, vol. 9, 6 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2018.00861/full

[182] I. E. Weintraub, M. Pachter, and E. Garcia, "An Introduction to Pursuit-evasion Differential Games," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2003.05013

[183] C. D. Wickens and S. R. Dixon, "The benefits of imperfect diagnostic automation: A synthesis of the literature," *Theoretical Issues in Ergonomics Science*, vol. 8, no. 3, pp. 201–212, 2007.

[184] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature," *Human Factors*, no. October, 2020.

[185] M. R. Endsley, "Situation awareness global assessment technique (SAGAT)," in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. IEEE, 1988. [Online]. Available: http://ieeexplore.ieee.org/document/195097/

[186] S. A. Jessup, T. R. Schneider, G. M. Alarcon, T. J. Ryan, and A. Capiola, "The Measurement of the Propensity to Trust Automation," Wright State University, Dayton, OH, United States, pp. 476–489, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069691988&doi=10.1007%2F978-3-030-21565-1_32&partnerID=40&md5=2a3d0294463242010bc1107ddc5f5be5

[187] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American Psychologist*, vol. 55, no. 1, pp. 68–78, 2000. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.55.1.68

[188] E. McAuley, T. Duncan, and V. V. Tammen, "Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis," *Research Quarterly for Exercise and Sport*, vol. 60, no. 1, pp. 48–58, 3 1989. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02701367.1989.10607413

[189] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 3 2000. [Online]. Available: https://doi.org/10.1207/S15327566IJCE0401_04

[190] R. D. Spain, E. A. Bustamante, and J. P. Bliss, "Towards an Empirically Developed Scale for System Trust: Take Two," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, no. 19, pp. 1335–1339, 9 2008. [Online]. Available: http://journals.sagepub.com/doi/10.1177/154193120805201907

[191] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," 1988, pp. 139–183. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0166411508623869

[192] R. M. Ryan, C. S. Rigby, and A. Przybylski, "The Motivational Pull of Video Games: A Self-Determination Theory Approach," *Motivation and Emotion*, vol. 30, no. 4, pp. 344–360, 12 2006. [Online]. Available: http://link.springer.com/10.1007/s11031-006-9051-8

[193] T. E. Duncan and S. C. Duncan, "The ABC's of LGM: An Introductory Guide to Latent Variable Growth Curve Modeling," *Social and Personality Psychology Compass*, vol. 3, no. 6, pp. 979–991, 12 2009. [Online]. Available: http://doi.wiley.com/10.1111/j.1751-9004.2009.00224.x

[194] J. A. Shepperd, "Productivity loss in performance groups: A motivation analysis," *Psychological Bulletin*, vol. 113, no. 1, pp. 67–81, 1993.

[195] Y. Rosseel, "lavaan : An R Package for Structural Equation Modeling," *Journal of Statistical Software*, vol. 48, no. 2, 2012. [Online]. Available: http://www.jstatsoft.org/v48/i02/

[196] M. Matsunaga, *Item Parceling in Structural Equation Modeling: A Primer*, 2008, vol. 2, no. 4.

[197] R. A. Cribbie, "Multiplicity Control in Structural Equation Modeling," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 14, no. 1, pp. 98–112, 1 2007. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10705510709336738

[198] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 3 1995. [Online]. Available: http://www.jstor.org/stable/2346101

[199] Y. Ho, A. Bryson, and S. Baron, "Differential games and optimal pursuit-evasion strategies," *IEEE Transactions on Automatic Control*, vol. 10, no. 4, pp. 385–389, 10 1965. [Online]. Available: http://ieeexplore.ieee.org/document/1098197/

[200] M. Foley and W. Schmitendorf, "A class of differential games with two pursuers versus one evader," *IEEE Transactions on Automatic Control*, vol. 19, no. 3, pp. 239–243, 6 1974. [Online]. Available: http://ieeexplore.ieee.org/document/1100561/

[201] R. Fernandez, E. Zaroukian, J. Humann, B. Perelman, M. Dorothy, S. Rodriguez, and D. Asher, "Emergent Heterogeneous Strategies from Homogeneous Capabilities in Multi-Agent Systems," *International Conference on Artificial Intelligence*, 2021.

[202] B. Howell, E. Zaroukian, D. Asher, and L. Parker, "Identification of Emergent Collaborative Behaviors in Multi-Agent Systems," Knoxville, TN, 2021.

[203] G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld, "Optimizing AI for Teamwork," 2020. [Online]. Available: http://arxiv.org/abs/2004.13102

[204] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, "Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 2429–2437, 2019.

[205] X. Fan, M. McNeese, and J. Yen, "NDM-Based Cognitive Agents for Supporting Decision-Making Teams," *Human-Computer Interaction*, vol. 25, no. 3, pp. 195–234, 7 2010. [Online]. Available: http://www.informaworld.com/openurl?genre=article&doi=10.1080/07370020903586720&magic=crossref%7C%7CD404A21C5BB053405B1A640AFFD44AE3

[206] J. Y. C. Chen, M. J. Barnes, S. A. Quinn, and W. Plew, "Effectiveness of RoboLeader for Dynamic Re-Tasking in an Urban Environment," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, no. 1, pp. 1501–1505, 9 2011. [Online]. Available: http://pro.sagepub.com/lookup/doi/10.1177/1071181311551312

[207] S. J. Karau and K. D. Williams, "Social Loafing: A Meta-Analytic Review and Theoretical Integration," *Journal of Personality and Social Psychology*, vol. 65, no. 4, pp. 681–706, 1993.

[208] B. Latané, K. Williams, and S. Harkins, "Many hands make light the work: The causes and consequences of social loafing," *Small Groups: Key Readings*, vol. 37, no. 6, pp. 297–308, 2006.

[209] T. Sakai and T. Nagai, "Explainable autonomous robots: a survey and perspective," *Advanced Robotics*, vol. 36, no. 5-6, pp. 219–238, 3 2022.

[210] S. S. Rodriguez, E. Zaroukian, J. Hoye, and D. E. Asher, "Mediating Agent Reliability with Human Trust, Situation Awareness, and Performance in Autonomously-Collaborative Human-Agent Teams," *Journal of Cognitive Engineering and Decision Making*, vol. 17, no. 1, pp. 3–25, 3 2023.

[211] J. Kennedy, P. Baxter, and T. Belpaeme, "Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 293–308, 4 2015. [Online]. Available: http://link.springer.com/10.1007/s12369-014-0277-4

[212] L. W. Barsalou, P. M. Niedenthal, A. K. Barbey, and J. A. Ruppert, "Social Embodiment," 2003, pp. 43–92. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0079742103010119

[213] R. van den Brule, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, and P. Haselager, "Do Robot Performance and Behavioral Style affect Human Trust?" *International Journal of Social Robotics*, vol. 6, no. 4, pp. 519–531, 11 2014. [Online]. Available: http://link.springer.com/10.1007/s12369-014-0231-5

[214] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 33–42, 2020.

[215] L. Tian, P. Carreno-Medrano, A. Allen, S. Sumartojo, M. Mintrom, E. Coronado Zuniga, G. Venture, E. Croft, and D. Kulic, "Redesigning Human-Robot Interaction in Response to Robot Failures: A Participatory Design Methodology," *Conference on Human Factors in Computing Systems - Proceedings*, 2021.

[216] S. Lee-Cultura and M. Giannakos, "Embodied Interaction and Spatial Skills: A Systematic Review of Empirical Studies," *Interacting with Computers*, vol. 32, no. 4, pp. 331–366, 2020.

[217] D. Saffo, C. Yildirim, S. Di Bartolomeo, and C. Dunne, "Crowdsourcing virtual reality experiments using VRChat," in *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 4 2020.

[218] M. Tonkin, J. Vitale, S. Ojha, J. Clark, S. Pfeiffer, W. Judge, X. Wang, and M.-A. Williams, "Embodiment, Privacy and Social Robots: May I Remember You?" 2017, pp. 506–515. [Online]. Available: http://link.springer.com/10.1007/978-3-319-70022-9_50

[219] T. Ziemke, "Disentangling Notions of Embodiment," 2001.

[220] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, "Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams," in *Lecture Notes in Computer Science*, 2018, vol. 10809, pp. 56–69. [Online]. Available: http://link.springer.com/10.1007/978-3-319-78978-1_5

[221] M. Hertzum, H. H. Andersen, V. Andersen, and C. B. Hansen, "Trust in information sources: Seeking information from people, documents, and virtual agents," *Interacting with Computers*, vol. 14, no. 5, pp. 575–599, 2002.

[222] I. Rae and L. Takayama, "In-body Experiences : Embodiment , Control , and Trust in Robot-Mediated Communication," pp. 1921–1930, 2013.

[223] S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young, "Poor Thing! Would You Feel Sorry for a Simulated Robot?: A comparison of empathy toward a physical and a simulated robot," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2015-March, pp. 125–132, 2015.

[224] J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Matarić, "The role of physical embodiment in human-robot interaction," *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 117–122, 2006.

[225] M. L. Walters, K. L. Koay, D. S. Syrdal, K. Dautenhahn, and R. Te Boekhorst, "Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials," *Adaptive and Emergent Behaviour and Complex Systems - Proceedings of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2009*, pp. 136–143, 2009.

[226] C. Breazeal, "Social interactions in HRI: The robot view," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 34, no. 2, pp. 181–186, 2004.

[227] G. Podevijn, R. O'Grady, N. Mathews, A. Gilles, C. Fantini-Hauwel, and M. Dorigo, "Investigating the effect of increasing robot group sizes on the human psychophysiological state in the context of human–swarm interaction," *Swarm Intelligence*, vol. 10, no. 3, pp. 193–210, 2016.

[228] P. Kulms and S. Kopp, "The effect of embodiment and competence on trust and cooperation in human–agent interaction," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10011 LNAI, pp. 75–84, 2016.

[229] J. M. Ross, J. L. Szalma, P. A. Hancock, J. S. Barnett, and G. Taylor, "The Effect of Automation Reliability on User Automation Trust and Reliance in a Search-and-Rescue Scenario," 2008.

[230] S. Rice, J. Hughes, J. S. McCarley, and D. Keller, "Automation Dependency and Performance Gains under Time Pressure," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, no. 19, pp. 1326–1329, 9 2008.

[231] A. W. Salmoni, R. A. Schmidt, and C. B. Walter, "Knowledge of results and motor learning: A review and critical reappraisal." *Psychological Bulletin*, vol. 95, no. 3, pp. 355–386, 1984.

[232] E. Brunsen, I. Murph, A. C. McLaughlin, and R. B. Wagner, "The Influence of Feedback Types on the Use of Automation During Learning," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 143–147, 9 2021.

[233] G. Doisy and J. Meyer, "Responses to warnings and the effect of notifications in a simulated robot-control task," *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp. 1594–1598, 2013.

[234] G. Kedar, J. Meyer, and Y. Bereby-Meyer, "Responses to alerts and subjective reports: Evidence for partial dissociation between processes," *Proceedings of the Human Factors and Ergonomics Society*, pp. 144–148, 2013.

[235] K. S. Thomson and D. M. Oppenheimer, "Investigating an alternate form of the cognitive reflection test," *Judgment and Decision Making*, vol. 11, no. 1, pp. 99–113, 2016.

[236] S. Frederick, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, vol. 19, no. 4, pp. 25–42, 11 2005. [Online]. Available: https://pubs.aeaweb.org/doi/10.1257/089533005775196732

[237] M. Bialek and G. Pennycook, "The cognitive reflection test is robust to multiple exposures," *Behavior Research Methods*, vol. 50, no. 5, pp. 1953–1959, 10 2018. [Online]. Available: http://link.springer.com/10.3758/s13428-017-0963-x

[238] S. Jabbireddy, X. Sun, X. Meng, and A. Varshney, "Foveated Rendering: Motivation, Taxonomy, and Research Directions," 5 2022. [Online]. Available: http://arxiv.org/abs/2205.04529

[239] B. J. Fogg and H. Tseng, "The elements of computer credibility," in *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*. New York, New York, USA: ACM Press, 1999. [Online]. Available: http://portal.acm.org/citation.cfm?doid=302979.303001 pp. 80–87.

[240] T. W. Schubert, "The sense of presence in virtual environments:," *Zeitschrift für Medienpsychologie*, vol. 15, no. 2, pp. 69–71, 4 2003. [Online]. Available: https://doi.org/10.1026//1617-6383.15.2.69

[241] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, vol. 70, pp. 153–163, 5 2017.

[242] S. R. Dixon and C. D. Wickens, "Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 48, no. 3, pp. 474–486, 9 2006. [Online]. Available: http://journals.sagepub.com/doi/10.1518/001872006778606822

[243] J. De Winter and D. Dodou, "Five-Point Likert Items: t Test Versus Mann-Whitney-Wilcoxon," *Practical Assessment, Research, and Evaluation*, vol. 15, no. 11, 2010. [Online]. Available: https://www.researchgate.net/publication/266212127

[244] S. G. Hart, "Nasa-Task Load Index (NASA-TLX); 20 Years Later," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 9, pp. 904–908, 10 2006. [Online]. Available: http://journals.sagepub.com/doi/10.1177/154193120605000909

[245] S. Reig, J. Forlizzi, and A. Steinfeld, "Leveraging Robot Embodiment to Facilitate Trust and Smoothness," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2019-March, pp. 742–744, 2019.

[246] B. Oakley, M. Mouloua, and P. Hancock, "Effects Of Automation Reliability On Human Monitoring Performance," in *Human Factors and Ergonomics Society*, 2003.

[247] A. Kaplan, "Trust in Imperfect Automation," 2019, pp. 47–53. [Online]. Available: http://link.springer.com/10.1007/978-3-319-96071-5_5

[248] A. Sethumadhavan, "Effects of first automation failure on situation awareness and performance in an air traffic control task," in *Proceedings of the Human Factors and Ergonomics Society*, 2011, pp. 350–354.

[249] T. Rieger, E. Roesler, and D. Manzey, "The Imperfect Automation Schema - Evidence for Increased Trust in Human Support Agents," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 1020–1020, 9 2021.

[250] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your AI: expertise and explanations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, 3 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3301275.3302308 pp. 240–251.

[251] M. Yin, J. W. Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12, 2019.

[252] A. Van Maris, H. Lehmann, L. Natale, and B. Grzyb, "The influence of a robot's embodiment on trust: A longitudinal study," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 313–314, 2017.

[253] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[254] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, "Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions," *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 4, pp. 1–30, 2018.

[255] A. Singh, T. Tiwari, and I. Singh, "Effects of Automation Reliability and Training on Automation-Induced Complacency and Perceived Mental Workload," *Journal of the Indian Academy of Applied Psychology*, vol. 35, no. Special, pp. 9–22, 2009.

[256] Y. Seong and A. M. Bisantz, "The impact of cognitive feedback on judgment performance and trust with decision aids," *International Journal of Industrial Ergonomics*, vol. 38, no. 7-8, pp. 608–625, 2008.

[257] L. H. Barg-Walkow and W. A. Rogers, "The Effect of Incorrect Reliability Information on Expectations, Perceptions, and Use of Automation," *Human Factors*, vol. 58, no. 2, pp. 242–260, 3 2016.

[258] D. Kahneman, *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011. [Online]. Available: https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7

[259] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," no. December, 2018. [Online]. Available: http://arxiv.org/abs/1812.04608

[260] J. Duncan-Reid and J. S. McCarley, "Strategy Use in Automation-Aided Decision Making," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 96–100, 9 2021.

[261] B. Shneiderman, *Human-Centered AI*. Oxford University Press, 2022. [Online]. Available: https://books.google.com/books?id=YS9VEAAAQBAJ

[262] F. Cabitza, A. Campagner, and C. Simone, "The need to move away from agential-AI: Empirical investigations, useful concepts and open issues," *International Journal of Human-Computer Studies*, vol. 155, p. 102696, 11 2021.

[263] J. S. Brennen, P. N. Howard, and R. K. Nielsen, "What to expect when you're expecting robots: Futures, expectations, and pseudo-artificial general intelligence in UK news," *Journalism*, vol. 23, no. 1, pp. 22–38, 1 2022.

[264] G. Matthews, J. Lin, A. R. Panganiban, and M. D. Long, "Individual Differences in Trust in Autonomous Robots: Implications for Transparency," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 3, pp. 234–244, 2020.