Factors influencing Trust, Reliance, Performance and Cognitive Workload in Human-Agent Collaboration

Sylvain Daronnat
Computer and Information Sciences
University of Strathclyde

A thesis submitted for the degree of
Doctor of Philosophy

Glasgow 2021

Declaration of Authenticity and Author's Rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date: 22/09/2022

Acknowledgements

I would like to thank my supervisors Leif Azzopardi and Martin Halvey for supporting me during this PhD.

I would also like to thank Mateusz, Olivia, Zuzana, Vasilis and Amine for their continuous support during this journey.

And of course, thanks to my parents and family for having supported me in my choices, no matter how far away I would go.

Pour Theo.

Contents

List of Figures

List of Tables

Abbreviations

FPS Frame Per Second

GUI Graphical User Interface

HAC Human-Agent Collaboration

HAI Human-Agent Interaction

HAT Human-Agent Teaming

HMI Human-Machine Interface

NASA-TLX NASA Task Load Index

SA Situational Awareness

SAGAT Situation Global Assessment Technique

SART Situation Rating Technique

SPAM Situation Present Assessment Method

UAV Unmanned Aerial Vehicle

Glossary

Agent Automated system capable of interacting with a user.

Aiming Agent Automated agent whose sole purpose is to provide assistance with aiming.

Cognitive Workload The reported mental effort associated with accomplishing a task.

Collaborative Agent Automated agent designed to assist user(s) in conducting a task.

Game Theory A branch of mathematics and statistics dealing with the maximisation of gains and minimisation of losses in competitive or collaborative situations involving constraints and uncertainty.

Goal Oriented Task Tasks in which performance is measured according to pre-defined goals.

On-boarding The process of introducing a system, its functionalities and modalities of interaction to users..

Robot Physical entity capable of interacting with other human or systems.

Situational Awareness Users' perception of their environment of interaction with respect to time, space, understanding of the situation, and ability to predict future outcomes.

Spawning In this thesis, "spawning" describes the apparition of new missiles from the top of the screen.

User Human operators interacting with a system.

Viewport The area rendered by the computer that can be visualised by the user..

Virtual Agent Synthetic non-physical entity processing information and capable of interacting with other human or artificial systems.

Visual Agent Automated agent whose sole purpose is to provide assistance with the display of visualisations.

Abstract

Increasingly, automated systems are being incorporated in collaborative environments where they are used to alleviate the cognitive load of human operators while increasing task performance. Automated agents are present in a variety of domains, from safety critical environments to leisure-oriented activities, and more and more, they are being considered as a virtual teammate rather than simple decision-aid tools.

Trust is a key factor that will determine how much a human operator is willing to take into account or rely on the help provided by an automated agent. Past research on trust in automation highlights key elements that will influence its development, such as how the automated agent is perceived, how reliable the agent appears to be and how transparent its actions are. However, most related work make use of turn-based tasks where trust is measured post-hoc, which does not entirely capture the evolving aspect of trust.

This thesis presents the development and use of a real-time collaborative game where human operators can choose the extent to which they rely on the help of automated agents displaying different behaviours and various levels of performance. We used different levels of task difficulty as well as survey instruments and the logging of task-specific behavioural information to elicit and measure variables that are important to understand the human-agent relationship such as trust, reliance, task performance, cognitive load or situational awareness.

We ran four user-studies using this apparatus. The first study tested the effects of different levels of agent reliability and predictability on the human-agent relationship while the second study experimented with different types of agent errors. The third study tested the impact of different types of environmental uncertainty on the human-agent relationship while the fourth and final study measured the benefits of different kinds of visualisation-based decision-aid systems.

Overall, this work sheds lights on under-investigated issues in Human-Agent Collaboration scenarios by providing insights on factors that are most likely to harm the human-agent relationship and underline how the behaviour of agents as well as the context of interaction can drastically alter a person's attitude toward an automated agent.

Part I

Introduction, Background and Methodology

Chapter 1

Introduction

1.1 Motivation

The idea of human operators and automated agents working together to solve problems has been theorised and studied since the very early days of computer science [111]. Thanks to technological advances in computer sciences and a better understanding of user behaviours, we are starting to shift from perceiving automation as simple decision-aid tools to virtual teammates actively collaborating with users. Human-agent collaborative scenarios now involve tasks where duties and responsibilities are shared between agents and human operators [16,97]. To fulfil the potential of human-agent interaction, not only are automated agents required to be reliable and trustworthy, human operators also have to be willing to rely on their decisions and trust the agents they are interacting with. A significant amount of work has examined how trust in agents is affected over the course of both human(s)-human(s) [197] and human(s)-agent(s) [103] interactions. Less work, however, has gone into the study of trust during real-time collaboration, and how specific types of agent error and visual attributes influence human-agent teaming.

This thesis seeks to better our understanding of the properties in either **(a) the agents** or **(b) the context of interaction** that are influencing users' propensity to trust and rely on automated agents. In particular, we are focusing on how the way agents make errors and uncertainty or added information in the environment of interaction affects the human-agent relationship. To answer our research questions, we designed a collaborative aiming-game in which human operators have to cooperate with agents in tasks of various difficulty and visual uncertainty. We chose an aiming task to craft a real-time human-agent collaboration scenario, as opposed to past HAI work which mostly made use of turn-based tasks. We used both validated survey instruments and behavioural information to measure, infer and study changes in the human-agent relationship. The ecological validity and real-world relevance of this task are discussed in Section 8.7.

1.2 Context

In this thesis, we focus on human-agent interaction in a real-time goal-oriented task where users and agents have explicit goals and responsibilities. The framework we developed in our work, which is described in Chapter 3, allows for the study of the human-agent relationship via validated pre- and post-hoc survey instruments, as well as task-specific behavioural metrics logged during the study.

While the focus of our work is on trust, we also study other important variables related to the human-agent relationship such as: **reliance**, **task performance** (studied via behavioural metrics), **cognitive load** (as reported by users) and **situational awareness** (when relevant). Every study was conducted using the same framework. The first two (about agent predictability in Chapter 4 and errors in Chapter 5) took place in a controlled lab-environment while the last two (about visual uncertainty in Chapter 6 and visual aid in Chapter 7) were conducted remotely.

1.3 Research Questions

This thesis focuses on how (a) **the behaviour of automated agents** and (b) **the environment of interaction** impact the human-agent relationship in a real-time, goal-oriented scenario. More specifically, we seek to answer the following sub research questions:

RQ1 How do changes in agent predictability (how easy it is to guess its next actions) and reliability (how good the agent is at the task) impact the human-agent relationship?

RQ2 How do different types of agent errors defined from previous related work such as slips, mistakes and lapses affect the human-agent relationship?

RQ3 How do different types of environmental conditions (static or moving), which impair vision and induce uncertainty, affect the human-agent relationship?

RQ4 How do different types of visual help (designed to elicit different levels of situational awareness) influence the human-agent relationship?

1.4 Contribution

The main contributions of this thesis are based on empirical findings gathered throughout four studies using an interactive human-agent collaborative framework. We show how agents behaviours, visual changes in environment and added transparency about agents' actions impact the human-agent relationship. More specifically, this work:

- Details changes in user behaviour during a collaborative human-agent interactive task using quantitative data captured during the interaction.

- Provides insights on the different ways users perceive collaborative agents from a qualitative point of view in terms of trust in the system, cognitive load and situational awareness.

- Investigates the impact of different types of error-prone agents, as well as the implications for the design of future, more trustworthy systems (Chapters 4 and 5).

- Investigates how adverse visual conditions in the environment of interaction can impact upon the human-agent relationship (Chapter 6).

- Details the process that unfolds when decision-making in human-agent collaboration is supported through different levels of visual help (Chapter 7).

- Provides an empirically tested framework that allows for the manipulation of variables important for most HCI work (task difficulty, agent performance, behaviour and transparency) and that can be used to get insights on a wide range of issues relevant to real-time human-agent collaboration.

1.5 Thesis Summary

The work presented here is organised as follows: **Chapter 1 - Introduction**: Presents the motivation, challenges and main contributions of the thesis.

Chapter 2 - Background: This section discusses the key elements in understanding the motivation for this research, and relevant related work pertaining to the study of trust in automated systems. More precisely, this section:

1. Presents a brief overview of the field of HCI and the context in which collaborative systems were created as well as the challenges they face.

2. Discusses related work on trust in automation and how most studies measure it.

3. Goes into details about which elements have the most impact on the development of trust, including the systems' performance and uncertainty in the context of interaction.

4. Summarises the research goals of this thesis in relation to previous work and gaps in literature.

Chapter 3 - Methodology: Presents the interactive game framework used in all studies presented in this thesis. This section details the motivation for using such a framework, including related studies employing similar means. This section also presents technical details about the inner working of the game, including how agents are designed, what information the framework records and how relevant metrics such as task performance and reliance are computed.

Chapter 4: Agent Reliability & Predictability: Presents the initial study focused on how different levels of agent predictability (how easy it is to guess the agent's next actions) and reliability (how good the agent is) affect human-agent collaboration.

Chapter 5: Agents' Errors: Presents the second study focused on the way agents make errors (according to previously defined frameworks of error types) and how agent behaviours affect users differently, at the same level of agent reliability.

Chapter 6: Visual Environmental Uncertainty: This third study investigates how different types of uncertainty that impair vision in various ways (static or moving) affect the human-user relationship, including situational awareness.

Chapter 7: Visual Help: This fourth and final study investigates how different types of visualisations of the agent's reasoning process affect the human-agent relationship.

Chapter 8 - Discussion: In this chapter, a summary of our findings is presented for each independent variable investigated in our studies. In addition, recommendations are made for the design of future interactive agents.

Chapter 9 - Conclusion: This chapter provides answers to our main research questions and conclude the thesis.

1.6 Publications

Most of the work presented in this thesis was previously published at the following peer-reviewed conferences or journals:

1. Daronnat, S., Azzopardi, L., Halvey, M., and Dubiel, M. Human-agent collaborations: trust in negotiating control. *CHI 2019* (2019) [39]

2. Daronnat, S., Azzopardi, L., Halvey, M., and Dubiel, M. Impact of agent reliability and predictability on trust in real time human-agent collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (2020), pp. 131–139 [40]

3. Daronnat, S., Azzopardi, L., and Halvey, M. Impact of agents' errors on performance, reliance and trust in human-agent collaboration. In *Human Factors and Ergonomics Society Annual Meeting* (2020), pp. 1–5 [37] (**Recipient of the 2020 CSTG Mark Resnick Best Paper Award**)

4. Daronnat, S. Human-agent trust relationships in a real-time collaborative game. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play* (2020), pp. 18–20 [36]

5. Daronnat, S., Azzopardi, L., Halvey, M., and Dubiel, M. Inferring trust from users behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration. *Frontiers in Robotics and AI 8* (2021), 194 [41]

6. Daronnat, S., Azzopardi, L., and Halvey, M. Investigating the impact of visual environmental uncertainty on human-agent teaming. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2021), vol. 65, SAGE Publications Sage CA: Los Angeles, CA, pp. 1185–1189 [38]

Chapter 2

Background

Computer systems are implemented to make work easier for users. Automated systems are implemented in various domains and environments, from optimising the management of a nuclear power-plant with multi-agent systems [80] to booking holiday trips using a virtual assistant [176]. Studies focusing on the interactions between users and computer systems feature different implementations of computerised systems, from embodied robotic systems to disembodied software-based agents that act for the benefit of users or on behalf of them. Depending on the context of interaction, users can either interact with one or multiple systems in tasks where decisions have to be made continuously, over time. In the Human-Computer Interaction literature, different terms have been employed to describe the type of automated systems users are interacting with, going from “Machine” [13], to “Computer” [177, 196] and later, “Agents” [109]. The use of these terms is always influenced by their context of interaction and reflects the evolution in the way automation is perceived by both researchers and end-users. In this Chapter, we look at different domains of interaction and paradigms in which human-agent interaction takes place. We then present relevant work that explores these concepts and highlight gaps in knowledge. At the end of this Chapter, we introduce the research goals that motivated our research questions (presented in Section 1.3) leading, subsequently, to our user studies presented in Chapters 4, 5, 6 and 7.

2.0.1 Human-Computer Interaction

Human-Computer Interaction (HCI) encompasses all scenarios where automated systems provide support to users via software (virtual) and/or hardware capabilities (robots). Human-Computer Interaction and its potential benefits for task performance and increased safety have been theorised since before the inception of personal computing, and is described in the work of Licklider [111] in 1959 as a being somewhere between a “mechanically extended man” and “Artificial Intelligence”, where computer systems, “if introduced effectively [...] would improved or facilitate human thinking and problem solving in an important way” [111]. This notion of symbiotic interaction between humans and machines was further developed in the 1970s, with companies like Xerox pioneering innovations such as the "Desktop metaphor" or "Direct Manipulation Interfaces" which allowed non-expert users to interact more seamlessly with computerised systems. Extensive reviews of studies in the the fields of mobile HCI [96] or HCI game research [19] offer insights regarding the most commonly used research methods to study users' interaction with systems. In general, HCI studies focused on user-system interactions are conducted either in a "natural" (field studies) or artificial (simulation) environment where users interact with one or more systems. The purpose of these studies, as described in the work of Kjeldskov et al. [96], can be described as the following:

- **understand** phenomena through data analysis,
- **engineer** new solutions or improve existing systems,
- **evaluate** the benefit or impact of a theory on users,
- or **describe** desirable properties that a system should have.

These research paradigms are helpful if we are to understand the motivation of a specific type of research, and will inform our choice of experimental design. Within HCI research, automated systems can take many forms depending on the context of interaction and type of studies undertaken. There is a key distinction between studies involving physical, embodied robots (Human Robot Interaction - HRI) and virtual, disembodied virtual agents (Human Agent Interaction - HAI). A study by Kramer et al [100] compared theories and empirical results derived from interactions with robots, agents or other humans. In their work, Kramer et al. explain that while HRI and HAI studies are both concerned with understanding users' feelings, thoughts and motivations when interacting with systems, HRI studies tend to be better at creating studies focused on affect and emotions due to the wider range of anthropomorphic acts possible for a physical robot.

As this thesis does not take place in settings with physical, embodied agents, our focus is mostly on HAI research literature, while also incorporating ideas and findings emanating from broader HCI or HRI studies. Furthermore, this thesis focuses on *evaluating* methods to study human-agent collaboration and *describing* negative and positive properties that facilitate interactions with collaborative systems. Human-Agent Collaboration (HAC) or Human-Agent Teaming (HAT) is a subset of HCI and HAI research that focuses on scenarios and environments where decision-making is shared between systems and users. In these scenarios, humans and systems constitute a team where parties collaborate to solve issues together.

2.0.2 Human-Agent Collaboration

The benefit of Human-Agent collaboration resides in combining the inherent strengths of both humans and automated agents while making up for their individual shortcomings. Nowadays, the availability of computer processing power and sensors often leads to situations where it is impossible for users alone to make sense of available data and complete their tasks efficiently and effectively. A solution to this information-overload problem is often found in the use of collaborative or fully-automated agents [5, 35] which human operators can rely on in order to complete tasks successfully. Automated agents are often designed as a way to reduce cognitive load on human operators, help maintain a certain level of performance and allow for clearly defined roles and responsibilities between users and agents. This has led to automated agents being implemented in a number of environments, from disaster response scenarios where agents help with time-constrained decision making [143] to scenarios involving scheduling [187] and monitoring problems [142].

2.0.3 Challenges of Human-Agent Collaboration

A number of challenges have to be overcome before decision-aid systems can evolve from automated tools to intelligent collaborative agents [98]. The introduction of collaborative agents led to important changes in the way human operators complete tasks in scenarios supported by automation. For instance, early work on the impact of industrial robots on users elicited benefits in human productivity and decision-making while also outlining concerns about complacent behaviours due to the increase in general downtime [8]. Complacency, defined by Moray et al. [126] as “self-satisfaction which may result in non-vigilance based on an unjustified assumption of satisfactory system state”, has become an ever-present problem in scenarios involving automated aid. These problems became apparent in safety critical environments where catastrophic complacency-induced accidents occurred, a fact which motivated numerous studies on the topic of complacency and reliability [132, 137].

In addition to complacency, the implementation of more decision-support systems made apparent issues related to disuses (under-utilisation), misuses (over-reliance) and abuses (no regard for consequences). Parasuraman’s seminal work [134] detailed how these 3 types of wrongful use of automation could be mitigated via, for instance, appropriate training, but also required system-makers to focus on system transparency and clearly delimiting the responsibilities and capabilities of humans and automated systems in collaborative environments [134].

While the aforementioned issues are linked to how users perceive automated systems, other key problems have to do with the actual capabilities of systems. Users’ mental models of automated systems and the way users interact with them change over time [103]. As early as the 1980s [127], research on the usage of automated systems found reported trust to be an important variable regarding how effective human-agent interaction could be. The work of Muir et al. stressed that “a decision aid (system), no matter how sophisticated or ‘intelligent’ it may be, may be rejected by a decision maker who does not trust it” [127]. Based on past work in HCI and particularly HAI, we elicited specific elements that are of particular importance regarding the quality of an interaction between a system and a user, namely:

- **Trust in automation.** The explicitly reported trust in a system by a user.

- **Reliance on the agent.** As operationalised given a task and captured via a context-specific proxy.

- **Performance and Reliability.** As assessed via explicit task goals clearly defined in the context of interaction. Performance is often used to describe users' success at a task while reliability is used to describe the system's capabilities or perception thereof by users.

- **Uncertainty.** This concept can be used to describe either the lack of transparency regarding a system's actions or a lack of information regarding the context of interaction.

In the next Sections, we present the construct listed above with relevant studies that assessed their importance in HCI, and particularly HAC research.

2.1 Trust

Trust is defined as an important factor for understanding *how* and *why* humans are willing to interact with other people [168] or computerised systems [85]. The development and evolution of trust is influenced by numerous factors such as experience, self-confidence and a general propensity to rely on another party [103,133,167]. While defining trust depends on the domain of interaction, application and task, it is generally accepted that trust represents a willingness to act based on the decisions, words or actions of "another" [106]. This other can either be a system, automated agent or human operator. The similarity between trust in automation and people is reflected in the work of Lee and See, and in particular their proposed definition of trust:

> "the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability [...] an agent can either be an automated system or another person that actively interacts with the environment on behalf of the person" Lee and See. [103, p. 2].

This above definition and the rest of the work by Lee and See are of particular interest as they highlight that trust:

1. Neither differs between team members nor differentiates between humans and automated systems.

2. Involves collaboration and cooperation between team members.

3. is task dependent.

4. Evolves over time and through interactions.

Past research has found that trust is a key component for effective human-human and human-agent collaboration in a wide range of domains and scenarios. In the following sections, we present studies about *interpersonal trust* and *trust in automation* that explored the development, evolution, loss and repair of trust in a variety of scenarios.

2.1.1 Interpersonal Trust

Most studies on interpersonal trust have focused on how trust is developed, maintained and lost between individuals or groups of individuals in the fields of Psychology and Organisational Studies. For instance, the work of Six et al. [168] examined organisations that explicitly favoured the development of interpersonal trust via internal policies compared to similar organisations that did not. To conduct their work, the authors relied on Eisenhardt's method to build theory based on case-studies [47], which involved looking at both theories and empirical findings. In their findings, they present four effective policies that help develop interpersonal trust:

1. Promoting a culture in which building relationships and showing care are valued.

2. Facilitating relational signalling (communication) between colleagues, no matter their rank in the company.

3. Socialising newcomers to teach them which values the company favours.

4. Actively managing and developing employees' skills and competencies.

From their recommendations, we can notice how important *transparency* and *communication* are in the development and calibration of trust over time, and how people have to understand their roles and responsibilities to maintain a positive trust relationship towards a company and its employees.

The development of trust in organisational studies has often been investigated alongside methods to best maintain trust, or repair trust in the likely event of a trust-damaging violation. While studies on trust repair are relatively recent in human-agent interaction, they were more thoroughly investigated in organisational studies and help us to understand what can negatively affect trust and how it can be regained. The work of Lewicki and Bunker [105], for instance, presents a series of recommendations to repair trust, consisting of

1. Acknowledging that a violation has occurred.

2. Determining the cause of the violation.

3. Admitting that the action was destructive.

4. Accepting responsibility.

Again, it is clear that transparent communication and a clear delineation of roles and responsibilities play an important part in regaining trust. In a related study, Gillespie et al. [66] investigated how trust repair is attempted after an organisational failure. They designed a framework of how internal factors (within the company) and external factors (outside of the company) influence employees' perception of the company's trustworthiness. This framework include lists of actions that can *contain the development of distrust*, such as imposing sanctions for individuals that breached trust, and other actions aim at *favouring the development of trust*, such as trust-enhancing communications of new regulations, or the diagnosis of previously found problems.

As we have seen, interpersonal trust is a complex construct that is impacted by many factors; some positive such as transparency and clear communication, and some negative such as violation acts that instil distrust and loss of confidence. Trust relationships between people or within organisations will fluctuate over time, and must therefore be maintained, repaired when needed, and explicitly valued in order to create good environments of interaction for all parties involved.

2.1.2 Trust in Automation

While interpersonal trust and trust in automation share similarities, for instance in the way transparency and clear communication positively affect their development, they also differ in key areas. To understand how interpersonal trust and trust in automation differ from one another, one can consult the work of Hoff et al. [85], in which the authors present a framework based on numerous prior studies on trust in automation. The framework describes some of the most important components that influence the evolution of trust prior to, and during interactions with automated agents. The model highlights how a system's features as well as users' cultural backgrounds can determine whether a user will trust the decisions of an automated system. In this work, trust is divided into three main categories: "dispositional trust, situational trust, and learned trust" [85]. All components of trust are subject to change, not only during the interaction but also outside of it. As a whole, the work of Hoff et al. emphasises that despite the complexity of the interplay of components influencing the evolution of trust, it is nevertheless meaningful to focus on single elements, and evaluate how they affect the trust relationship over time.

Some studies have attempted to isolate and study how specific users' attitudes affect their trust in automation. In the work of Singh et al. [167], an aviation monitoring task was undertaken by non-expert users to investigate the impact of complacent behaviours on task performance. Their results show that it is hard to predict complacency behaviours based on individual characteristics alone, and that "other individual and social factors may play a role, particularly in work environments" [167, p. 17].

Trust in automation is difficult to assess, monitor and measure. This difficulty is elicited in an article by Hoffman et al. [86] where challenges related to trust in automation in modern systems are presented [86]. The work by Hoffman et al. underlines that there is more than one way of measuring trust, and that each method has to take into account its context of interaction to allow for the meaningful capture of trust-related components. Most methods of studying trust and the evolution of its components consist in survey instruments administered before or after a task, such as the "Checklist for Trust between People and Automation" by Jian et al. [90] which was designed to assess both *trusting* and *distrusting* behaviours in order to get a single score reflecting participants' attitude toward automation. These methods of assessing trust, while non-context dependent, fail to take into account the ever-evolving nature of trust, as they require participants to divert their attention from a task and reflect on their past interactions. This modality of trust assessment can prove problematic, for instance, in fast-paced scenarios that require the constant monitoring of multiple systems, or when operators must remain alert in a safety critical environment.

Understanding, measuring and monitoring trust is, however, important to appreciate how human operators are willing to cooperate with automated agents, as it sheds light on the evolution of the human-agent relationship itself. The work of Merritt et al. [125] focused on the issue of "Trust Calibration", with a task requiring participants to look at luggage X-rays and decide whether they are suspicious and require further investigation or not. In their experiment, participants were helped by automated agents that provided recommendations regarding which decision to take, as well as information regarding the accuracy of the agent's recommendations. Task performance was evaluated as the number of correct decisions taken at each turn while trust was operationalised by looking at participants' perceptual accuracy, sensitivity and trust sensitivity. In their findings, the authors found that giving information regarding an agent's performance was highly correlated with positive reported trust in the agent and helped reduce the development of complacent or distrustful behaviours. The combined operationalisations of trust calibration, however, was found to be poorly correlated with task performance, indicating that there is more work to be done on the role of trust in human-agent relationships. In addition, while factors likely to predict trust may be inferred from individual user capabilities and informed by prior related work, they can ultimately prove to be bad predictors of actual reported trust in the system.

As we have seen in the work of Merritt et al. [125], the study of reported trust is paramount to help reduce the likelihood of *complacent* and *distrusting* user behaviours. In the context of human-agent interaction, many factors can lead to incidents with potentially extremely negative implications in safety-critical environments. For instance, the catastrophic accident related to the Boeing 737 MCAS system [161] was caused by software giving pilots wrong instructions after having suffered sensor malfunctions. In this situation, pilots followed the system's recommendations based on their past experience and training, but were given an erroneous mental model of the situation. Subsequent attempts to override the system ultimately failed, which led to the incident.

In past work, most studies looked at the effects of an agent's reliability on users. For instance, the work of Fan et al [57] investigated how different levels of agent reliability affects users' reliance and trust in an agent during a simulated battlefield command-and-control scenario. In their findings, the authors indicate that while higher agent reliability leads to higher reliance on the agent, users' individual skills heavily influence their attitude toward the system, with expert users remaining more cautious and less prone to complacent behaviours that novice users. In a related study about agent reliability, Hussein et al. [87] investigated how participants relied on agents in a dispatching/foraging scenario. Their findings highlight that while increasing agents' accuracy leads to lower completion times and better overall performance, it also leads to a significant decrease in users' ability to reject agents' decisions correctly, further emphasising the need to understand how to manage nascent complacent behaviours.

While the relationship between agent reliability, trust and reliance is not straightforwarded but dynamic and heavily context-dependent as highlighted by the work of Fan et al. [57] or Hussein et al. [87], recent work has focused on other factors influencing the human-agent relationship. A study by Jensen et al. [89], for instance, investigated how users' emotional experience relate to trust in agents, whereas the study of Correia et al. [34] focused on the impact that added transparency in the agent's actions can have on the user after experiencing automation failure.

As we have seen in this Section, reported levels of trust in an agent are never static, and evolve over time and through interactions [125]. Most past work on trust in automation has focused on turn-based tasks where the users and agent interact asynchronously and where reported trust in agents is measured with pre and/or post-hoc instruments. As trust is dynamic and context-dependent, more work is needed that investigates both its measurements (which methods to employ) and purpose (how does trust relate to other elements in human-agent interaction). This thesis seeks to further our understanding of trust in agents by studying how different kinds of agent behaviours (see Chapters 4 and 5), environments of interaction (see Chapters 6) and information regarding the agent (see Chapter 7) can positively or negatively affect users' reported trust in agents. The relevant literature motivating our research focus is presented in each chapter detailing our empirical studies (see Chapters 4, 5, 6 and 7).

2.2 Reliance

In HAI research, reliance is a purely behavioural construct defined by Ross et al. as a "tendency to employ automation to replace manual control" [148]. In Human-Agent studies, the concept of reliance usually describes the voluntary act of following the recommendations of an agent or not. In the influential work on trust in automation by Lee and See [103], the authors describe reliance as “a discrete process of engaging or disengaging [on automation]”, and acknowledge that this definition is a simplification to explain certain relationships with trust more clearly. In their work, Lee and See further noted that “trust guides but does not completely determine reliance,” [103]. And indeed, the mixed findings from many related studies in human-agent scenarios such as the ones presented in Section 2.1 by Jensen et al. [87] or Fan et al. [57] would seem to bear this out. For instance, while both the Fan and Jensen studies highlight that high agent reliability induces better task performance and higher reliance on the agent, Jensen’s study [89] reports a greater failure to correct false positive errors as agent reliability increases. Due to the relationship between trust and reliance, most work has investigated both in parallel, seeking to study how reliance develops under various levels of system transparency or system reliability.

Reliance on automation and the overall propensity for users to rely on an agent are affected by many factors coming from either the users themselves or the environment of interaction. A study by Sanchez et al. [151] investigated elements likely to have an important impact on reliance on automation such as experience, age (users) or error type and error distribution (agents). In their studies, the authors employed a framework requiring participants to confirm or reject an agent’s decision in an agricultural setting. The automated aid suffered from different types of errors and participants were grouped according to their age and experience with agricultural engines. In their findings, Sanchez et al. found that older participants took longer than younger participants to adapt and properly respond to the system when needed. Older participants, however, continued to check alarms more consistently than younger participants, and were therefore better able to respond to sudden automation failure. As we can see, individual factors pertaining to either users (age, experience) or the system (error types) will influences attitudes toward reliance. While these characteristics are important, they are not sufficient to explain how reliance evolves over time. In their conclusion, the authors noted that *all* participants, no matter their age or experience, tried to adjust their behaviours to collaborate more effectively with the automated aid, as they grew more used to interacting with the system. The findings of Sanchez et al. are reminiscent of the concept of “appropriate reliance”, presented in the work of Lee and See [103], which describes the link between a system’s capabilities and users’ perception of them.

In a related study by Dzindolet et al. [46] on the role of trust in automation reliance, the authors focused on how explaining an agent’s behaviour affects reliance in a human-agent setting. The authors found that giving reason as to why an agent might fail tends to lead to higher reliance on the agent, even when compared to scenarios including less reliable agents, and that the type of error experienced by participants (false negative or false positive) did not significantly change their attitudes toward the system. These findings are different from the ones presented by Sanchez et al, and indicate that the context of interaction is of paramount importance in understanding reliance, as its measurement is purely behavioural.

As we have seen so far, reliance on automated systems can be studied in a number of ways, but no matter the modality, understanding users' roles, experience and the context of interaction is important to operationalise reliance and measure it in a meaningful way. For instance, when the agent provides help in the form of explicit feedback, reliance can be studied in terms of the likelihood of users adopting the agent's help in their decision-making process. When the user and agent have similar decision-making capabilities, reliance can be studied as the amount of corrections the user issues, with fewer corrections indicating more reliance on the agent. Most studies relied on the manipulation of automation failures (error types and distribution [46,151]) to understand how users interact and adapt to automation in various domains of interaction.

In an important work about reliance on automation, Parasuraman and Riley [134] detailed the different types of attitudes toward reliance, and categorised them as follows:

- **Misuses:** over-reliance on automation.
- **Disuses:** under-utilisation of automation.
- **Abuses:** "inappropriate" application of automation, from a system designer perspective.

This categorisation is useful if we are to understand the intent behind different types of inappropriate reliance on a system. Like trust, reliance is calibrated over time and through interaction, and for an appropriate reliance to develop, systems must be designed to help users while not leading to complacent attitudes.

The work of Garnick et al [65] sought to study reliance by inciting users to take into account an agent's recommendations. Their experiment took place in a controlled game-like environment where reliance was measured as a behavioural factor (how many times participants followed the agent's advice), and where agent and user performance could be compared or aggregated into one measure of team performance. In their results, the authors underlined the importance of designing frameworks and controlled environments to study a construct as context-specific as reliance. Changes in agent reliability, error type or mission goal can have a major impact on the development of reliance over time. The use of controlled game-like environments also allows us to quantify and model reliance via task-specific metrics, as demonstrated by the work of Boubin et al. [14] which uses the framework of Garnick et al. [65] to model users' interaction and understand appropriate reliance and trust calibrations.

As we have seen throughout this section, reliance is a behavioural factor that is linked to trust but does not necessarily share a straightforward, linear relationship with it. The study of reliance is context-specific, and most studies investigated relationships between reliance and other constructs by varying agents' errors [151] or the degree of information given to users about the system [46]. Interactive frameworks such as the one found in the study of Garnick et al. [65]

and Boubin et al. [14] constitute a good, albeit abstract, avenue to contextualise, measure and study reliance on agents. In a similar fashion, this thesis measures reliance within a game-based environment, where corrections issued by the participants are recorded and operationalised as our main measure of reliance.

2.3 Task Performance and Agent Reliability

According to the work of Wiebe et al. [190], performance is generally understood as the measure of an outcome in cognitive tasks. Similarly to reliance, performance is a context-specific construct. In HAI studies, performance is often understood as how well the human-agent team can succeed at a task presenting clearly defined goals, given different levels of human expertise and agent reliability. In an article by Lewis [109] on "Designing for Human-Agent Interaction", the author shows that most studies vary the level of information and feedback about the system's actions and reliability to see resulting changes in users' task performance, reliance and trust.

In HAI studies, while task *performance* represents how good the users and/or agent are at meeting specific goals, agent *reliability* represents the accuracy of the help given by the agent. In general, it is assumed that agent reliability is key to understanding how human operators perform and calibrate their trust in agents. In a work by Hoc et al. [83], the authors present a framework where both agents and users are considered as separate agents pursuing their own goals in a driving task. In their framework, the impact of both the user's performance and the agent's level of reliability can be evaluated individually or collectively, as tasks require users to act on their own or rely on inputs given by the agent. For maintaining high task performance in a safety-critical scenario, the authors recommend that agents should be designed to support users, not replace them, and that the responsibilities of both users and agents should be clearly defined.

The relationship between users' performance and agents' reliability is, however, not always straightforward. In human-agent collaborative scenarios where user(s) and agent(s) form a team, performance is often evaluated as a single construct, which makes it harder to study the impact that either the user or agent has on overall team performance. The work of Fan et al. [57], outlined in Section 2.1, studies team performance in an experiment focusing on agent reliability and user expertise. The authors found that an increased knowledge of the agent's reliability level help mitigate the chances of false positive errors. Similarly, Chavaillaz et al., [24] tested the impact of different levels of agent reliability on trust, reliance and task performance in a turn-based X-ray scanning scenario. Their results showed that a decrease in agent reliability resulted in a decreases in users' reported trust in the agents. Furthermore, Chavaillaz et al. found that users' perception of the reliability of agents was more accurate when interacting with low performing agents, which led to complacent behaviours. As we can see from both Fan and Chavaillas' studies, added transparency regarding the agent's level of reliability significantly influences the way users trust and rely on the agent's input, which, in turn, positively affect overall team performance. In addition to studies focusing on different degrees of agent reliability in assessment tasks, the work of Shirado et al. [166] explored turn-based coordination problems, in a colour-selection game. In their work, the authors found that error-prone agents (up to 30% loss in accuracy) can be beneficial to collaborative task performance as they reduce the chance of users being complacent while interacting with the agent.

Given the evidence of past research, it is clear that agent reliability is one of the major factors that will influences both team performance and a user's propensity to trust and rely on the agent. Most studies manipulated the way agents made errors by having different levels of agent reliability, but other methods have been employed to vary the way an agent performs and makes errors. While some studies (as described in Section 2.1) introduced false-alarms (Merritt et al. [125]) or systematic biases (Fan et al. [58]), other studies focused on the *way* agents make errors, rather than on testing different levels of agent reliability. For instance, some studies experimented with agents that suddenly stop working, such as presented in the work of Correia et al. [34] that involves a robot and a user playing a collaborative card game. In their study, the robot would explain (or not) its faults to the user. With their findings, Correia et al. showed that minor faulty behaviours that are harmful to trust and task performance can be mitigated by simply having the robot acknowledge its own shortcomings. More serious loss in the robot's reliability, however, still resulted in inferior team performance, no matter the justification used by the robot.

In most cases, prior work delving into agent reliability has showed that participants react differently to various types of automation failures, and report higher trust in systems that can justify their behaviours. This represents a new avenue to not only understand how team performance can be degraded, but also mended back to suitable level.

Any type of agent failure can be loosely described as an"error". "Errors", however, is a term that fails to explain the detailed nuances of how an agent can stray towards undesirable outcomes. The work of Marinaccio et al. [118] presents 4 unique types of errors: mistakes, lapses, slips and violations. All of which are derived from human-human interactions studies [145] and are presented in table 2.1. These definitions of errors all come from studies focusing on *human errors*, where they were also conceptualised in the context of human-human interaction in healthcare in a study by Kim et al. [94] that investigated these types of errors alongside different repair mechanisms on individual and groups of people. In their findings, Marinaccio et al. explain how more work should be done to analyse how the particular features of a system and the way in which it errs influence human-agent relationships.

Overall, a number of studies empirically tested the effect of trust-damaging acts in both human-human and human-agent settings by manipulating the nature of an agent's violation. The work of Baker et al [11, 125] presents a comprehensive review of past research in the field of human-robot interaction with ideas for future research on maintaining and repairing trust in robotic agents. Among the 6 avenues for future research mentioned by the authors, one is of particular relevance for human-agent collaboration: "Adapt existing trust research to investigate how robot features affect trust" [11, p. 20]. As we have seen in previous work on agent reliability [118], one of the most important features of future research should be the way in which agent make errors: how to categorise them and how users perceive them.

In this section, we have looked at past work on human-agent interaction to understand how agent reliability is defined and varied to study its resulting impact on users' trust and performance. We have seen that task performance in HAI scenarios is often measured with task-dependent metrics related to the success of one particular activity, whereas agent performance (how good at the task the agent is), is controlled in terms of reliability [14, 58, 65]. However, as we have seen when looking at trust in Section 2.1, system reliability is only one factor that influences trust in agents and the overall human-agent relationship. In this thesis, in addition to reliability, we study the impact of different agents' behaviours in terms of predictability and intent, and their resulting impact on participants.

Table 2.1: Different types of errors presented in the work of Marinaccio et al. [118] and inspired by the work of Reasons [145].

Error Type (Reason, 1990)	**Examples**	**Violation Type (Kim et al. 2013)**	**Effective Repair (Kim et al. 2013)**
Slips - Errors of commission - when an intended action is wrongly executed	Flipping the wrong switch on an IV pump	Integrity-based	Denial
Lapses - Errors of omission - resulting in failure to carry out the action	Forgetting to administer medication	Competence-based if due to memory failure, integrity-based if attention failure	Context-dependent
Mistakes - Errors of planning or judgment	Prescribing an incorrect dosage	Competence-based	Apology
Violations - Intentional commission of an error	Prescribing an inappropriate medication because of sponsor loyalty	Integrity-based	Denial

2.4 Uncertainty

Uncertainty is a subjective construct that involves risk and opportunities, as defined by Rachev et al. in their work on uncertainty in the financial domain [140]. Uncertainty is most studied in the field of economics for risk management purposes, in order to predict potential negative outcomes or lessen their impact. In HAI research, uncertainty is inherent to most environments of interaction, and it is often the role of an automated agent to lessen uncertainty by processing data and anticipating potential outcomes. Uncertainty is harmful to any collaborative settings, whether they include only people [188] or people with agents [101], the latter being the focus of a number of important human-agent teaming studies assessing uncertainty in high-risk domains such as driving [101] or aviation [167]. As uncertainty comes from aspects either within or outside of the human-agent relationship, the following subsections will present prior work that studied uncertainty related both to agents and the environment of interaction.

2.4.1 Uncertainty in Agent Reasoning

In the work of Wu et al. [198], the authors proposed frameworks to manage uncertainty between entities (humans or automated agents). The constituents of the agents' reasoning are represented using the BDI (Belief, Desire and Intention) attributes where each agent is programmed with "beliefs" (what the agent knows),"goals" (desired outcome) and "plans" (how to achieve the desired outcome, step by step). This framework allows for reducing uncertainty by making evident what the agent knows or doesn't know, and by openly sharing its intentions with other agents or users. From the work of Wu et al., we see that both assigning clear roles and enforcing transparent intent in the decision-making process of all entities are paramount to reduce risk and uncertainty. In HAI studies, the concepts of "trust calibration" or "appropriate reliance" (described respectively in Sections 2.1 and 2.2) are commonly found and employed in an effort to increase transparency and reduce uncertainty. During calibration, by increasing the transparency of the agent's actions (in terms of how its inner reasoning is presented), users can judge the capability levels of the agent in order to adapt their trust and reliance accordingly.

The work of Kunze et al [101] investigated how to best communicate uncertainty in an HAI context. Their findings show that while displaying some degree of information about the agent's reasoning helps improve performance, prolonged attention and monitoring induce a higher cognitive load, which in turns leads to users not being willing to interact with the system. As demonstrated by Kunze's study, transparency may be helpful to reduce uncertainty but it comes at a cost, especially given the primary function of agents is to reduce rather than increase cognitive load. Related work from Sacha et al. [149] thoroughly reviewed prior studies that researched how to best communicate uncertainty and improve trust with visual analytic tools. In their findings, they highlight what they call "traps" concerning uncertainty, such as being clear about exactly what is meant by agent uncertainty, and having a system that overloads users with information that is not of immediate relevance. Such an information overload could even lead users to create erroneous mental models of the agent's inner workings, which in turns provokes more uncertainty. For instance, the study by Wang et al [186] created a framework for automated agents to automatically come up with an explanation for their reasoning for particular actions. As transparency is a key element in reducing uncertainty, the authors found that explanations did indeed lead to higher trust in the agent and better team performance. A side-effect, however, was that the added confidence in the system led participants to believe that they fully understood the robot's decision-making process, whereas the explanations provided by the system were *not* sufficient to make such claims [186].

No matter the context, we have seen throughout this section that uncertainty in the agent's decision-making process can be reduced by increasing transparency. This added clarity, however, comes at a cost: users increase their attention levels, which in turn affects cognitive load. To mitigate such problems, researchers have experimented with various visualisation techniques and issued some recommendations, such as:

- Reducing the attention required or the amount of information provided by the system [101].
- Taking into account the context of interaction and being clear about the kind of uncertainty that the system should manage [149].
- Not over-simplifying the information communicated, so as to prevent users from having erroneous mental models of the system's capabilities [186].

2.4.2 Uncertainty in the Environment

A major factor in the development of uncertainty is the context of interaction itself. In this Section, we will use the terms "environment" and "context of interaction" interchangeably. In HAI settings, uncertainty can manifest itself differently depending on the task and context of interaction. In these situations, Game Theory offers relevant frameworks and techniques to study and assess uncertainty, as it often involve optimisation problems and minimising risk. For instance, in a game scenario where information is incomplete, a common strategy for a player who seeks security would be to work towards a "secure equilibrium", which is a state of action that would lessen the impact of the worst predicted outcome [179]. In an HAI task, however, this secure equilibrium can be hard to define, as dynamic changes in the environment can alter the decision-making process of both users and agents in real-time, which makes planning difficult. A number of studies have focused on the visual (un)availability of information in the environment of interaction and its impact on users' decision-making in HAI settings.

Uncertainty coming from the environment of interaction is especially an issue in situations where task-sensitive information used to inform decisions is limited or unavailable. To help understand and mitigate this type of uncertainty, a study by Sarter et al [152] focused on the impact of different display aids for pilots in uncertain and time-sensitive scenarios, especially in the aviation domain. In their findings, the authors show that the type of display and the accuracy of the information were the most important components related to task performance. In addition, environmental uncertainty has been found to be mitigated by agents capable of giving information that lessens the impact of uncertainty. Such findings are presented in the work of Kunze et al. [101] where different types of visualisations and behavioural metrics were used to adapt and present information to users while not overloading them with data. While agents can help mitigate uncertainty, paying prolonged attention to explanations, however, could lead to lowered situational awareness and even missing real-time changes in uncertainty displays [101]. Beyond the unavailability of data, uncertainty comes into play in scenarios where predictions are being made based on the assessment of existing information. In these situations, prior information and the way it is constructed, shared and understood is crucial to how future decisions are made. A study by Herdener et al. [81] demonstrates that providing information about potential outcomes, with a sense of variability, does indeed have an impact on decision-making and cognitive load.

Most work in the field of Human-Agent interactions has focused on the uncertainty inherent to the decision-making capabilities and transparency of automated agents. As we have seen in Section 2.4.2, uncertainty can also come from the environment of interaction, and there is currently a knowledge gap regarding how users' behaviour evolves during a human-agent task in an uncertain environment. As a means of responding to this gap in knowledge, the current thesis will present work related to human-agent collaboration under visual uncertainty in Chapters 6 and 7.

2.5 Measuring behavioural information

Behavioural data are information captured during human-agent interactions that act as a proxy for some of the concepts detailed in this Chapter, such as reliance (see Section 2.2) or performance (see Section 2.3). These measures are all context-specific and are often collected by systems sensors and operationalised according to the explicit goal of the task, such as *task performance* and *reliance*. Every study conducted during this thesis and presented from Chapters 4 to 7 records and analyses behavioural data to understand and model user behaviour.

2.5.1 Task Performance

In human-agent collaborative scenarios, the outcome of a task is often measured in order to assess the quality of interaction, as detailed in Section 2.3. As performance is an important variable in any kind of goal-oriented task, it is important to measure it via adequate methods that are in accordance with the goal of the task. The work of Lebas [102] in the field of organisational sciences defines performance as "contextual both in terms of users and in terms of purpose." [102, p. 2]

Some studies, focusing on the acceptance of automated aid, report performance as the number of times participants relied on the agent when the agent was providing reliable inputs. Examples of these studies include the work of Fan et al. [57] or Hussein et al [87] presented in Section 2.3. Other researchers measured team performance as a whole (human and agent combined), which is only possible during collaborative scenarios where users and agents have similar capabilities and levels of agency, such as is shown in the work of Shirado et al. [166]

One of the most common metrics used to measure performance (notably in the field of Information Retrieval and Natural Language Processing) are Recall, Precision and F1 [136], which are based on the notion of false or true positives and negatives, and can be easily adapted to most goal-oriented tasks where agents provide help that can vary in terms of reliability. For the studies presented in this thesis, we will mostly rely on Recall, Precision and F1 for the assessment of individual and team performance. These metrics are further detailed in Section 3.4.3.1. It is also worth noting that we do not measure the rapidity at which participants complete the task as every session is timed and speed is not considered as a proxy for performance in our studies.

2.5.2 Reliance

Reliance and trust are often studied together in HAI and Human Factors studies [46, 120], as they share an evident, albeit complex, relationship as described in Sections 2.1 and 2.2.

In terms of behavioural measurement, reliance is often studied in conjunction with compliance, as elicited by the work of Chancey et al: "The operator responding when a signal is issued is referred to as compliance. The operator refraining from a response when the system is silent, or indicating normal operation, is referred to as reliance" [21, p. 1]. In HAI studies where the agent not only guides the user but also interacts on its own or on the behalf of the user, reliance is often studied as the amount of corrections the user issues, with fewer corrections indicating a greater reliance on the agent. An example of such a behavioural measure can be found in the work of Hussein et al. [87] where users' acceptance of the agent's recommendations can be recorded by the number of time they followed said recommendations.

While there is no consensus on which behavioural or physiological metric best represents a robust proxy for the measure of trust in automation, a few other physiological variables have been explored, such as Heart rate [101], post-neuronal activity (EEG [3]), gaze and electrodermal activity [185]. While the usefulness of some of these variables might be highly affected by the context of interaction, they still represent potential avenues for monitoring reliance in a context-free manner.

In this thesis, reliance will be measured using a context-specific variable: "user control time", or the amount of time for which users took over control from the agent. This variable is presented in Section 3.4.

2.6 Measurement of reported information

In user studies, reported information is often recorded as ratings given to either statements or questions on likert-scale instruments administered before, during and/or after a task. Survey instruments are commonly used to assess subjective concepts such as trust, cognitive workload or situational awareness that are otherwise hard or impossible to record using behavioural measurements. Every study presented in this thesis (from Chapters 4 to 7) relies on validated survey instruments to understand how users perceived automated agents, and how their perceptions relate to their behaviours while interacting with agents. Most of the original survey instruments used in this thesis are presented in Appendix E.

2.6.1 Trust

Trust is often measured using rating scales where users have to indicate how they agree with one or more statements, eliciting either an overall attitude toward the system or evaluating sub components linked to the development of trust.

2.6.1.1 Multi-item Instruments

One of the most widely used self-reporting scales is the “Checklist for Trust between People and Automation” by Jian et al. [90] which consists of 12 statements that participants have to rate on a 7-point Likert scale (with higher scores indicating greater agreements). The purpose of each statement is to elicit different attitudes toward the system linked to trust (reliability, honesty...) or distrust (suspicion, deception...). This instrument can be used before or after an experiment in order to measure how trust and distrust evolve. This 12-item instrument has been widely used in a wide range of studies owing to its context-free nature, from Satterfield’s work involving UAVs [153], to Erebak’s study on robots in the field of elderly care [55].

2.6.1.2 Single-item and Analogue Instruments

Multi-item survey instruments are best used to assess sub-components of multi-dimensional concepts, such as distrust in the work of Jian et al [90]. In many scenarios, however, multi-item survey instruments can prove too cognitively taxing to use, reducing the quality of the answers provided, since many studies not only assess trust but other relevant metrics such as cognitive load [2] or situational awareness [70]. As a result, many studies have resorted to the use of short and quick-to-deploy instruments for the measurement of trust in a system. Some are made and tailored to the need of particular studies, such as the one used by Wiczorek et al. [189] consisting of an analogue trust scale with verbal anchors ranging from “my trust is very strong” to “I barely trust the system” [189, p.7], where participants’ answers are then transformed using a scale from 0 to 100.

2.6.2 Cognitive Workload

While Cognitive Load theory is a concept that attempts to understand how people make decisions based on evolutionary theory [171], Cognitive Workload is a concept that has been most studied in human-machine interface research, and refers to the study of the “cost of accomplishing mission requirements for the human operator.” [75]

By far, the most common way of assessing Cognitive Load is via a pre- or post-hoc 6-item survey instrument named NASA Task Load Index (TLX) [76] that has been used and tested by numerous studies since its inception in 1988, including in an important meta-article by its original authors that analyses 20 years worth of studies that relied on the instrument. [75] In this thesis, cognitive load is measured in all studies using NASA TLX, and its score is usually reported using the Raw TLX technique [17].

2.6.3 Situational Awareness

As alluded to in Section 2.4.2, an important concept in understanding how users make decisions is Situational Awareness (SA). SA can be described as the capacity to locate and remember important information in order to make predictions about future outcomes [48]. There are three commonly used methods to measure situational awareness, namely "SAGAT", "SART" and "SPAM":

- **SAGAT** [49] (Situation Awareness Global Assessment Technique) consists in designing a set of SA related queries related to a specific task. During the experiment, the task is frozen, giving time for participants to answer the queries. A numerical value is then derived from each question and aggregated into an overall score evaluating participants' overall SA.

- **SPAM** [45] (Situation Present Assessment Method) is similar to SAGAT but puts an emphasis on quick questions focused on locating information in the environment. The response time is taken into account as an indication of SA.

- **SART** [173] (Situation Awareness Rating Technique) consists of a pre-determined set of general queries related to Situational Awareness.

All methods have pros and cons. While some require freezing of the current task (SAGAT), others require multi-tasking (SPAM), or the completion of a post-hoc survey (SART). A recent comparison of all these methods [52] found the SAGAT method was the least intrusive and least harmful in terms of users' task completion performance. The SART method, however, offers a post-hoc context-free survey instrument that could potentially be adapted to any scenario. In this thesis, Chapters 6 and 7 investigate the impact of different visual conditions on situational awareness and make use of the SAGAT and SART methods.

2.7 Research Goals

The work of Baker et al. [11] discusses important challenges for studies on trust in automated systems and robots, and makes recommendations on which issues are important to tackle in future work. Of the 6 recommendations elicited, 2 are of particular significance for human-agent collaboration: "Adapt existing trust research to investigate how robot features affect trust" and "Develop & validate measures of human-robot trust". Following on from these recommendations, and based on previous work in the field of human-agent interaction, this thesis seeks to contribute to our understanding of trust in collaborative automated agents during real-time scenarios, both through self-reported survey instruments and behavioural information directly captured from human-agent interactions. More specifically, this work seeks to address the following gaps in the HAI literature by:

1. Experimenting with new ways to elicit and record human-agent interactions.
2. Assessing both reported subjective metrics and recorded behavioural information related to the human-agent interaction.
3. Studying how trust in automated agents develops and which elements influence it the most.
4. Experimenting with different types of agent behaviours and environments of interaction.

While our focus is on the development of trust in automated agents, we sought to study the development of related elements based on previous relevant work. In the study we conducted, we are measuring the following dependent variables:

1. Trust. With self-reported metrics.
2. Reliance. By recording and analysing task-specific variables.
3. Performance. According to the tasks' goals and performance metrics.
4. Cognitive Load. Via validated self-reported questionnaires, such as Cognitive Load.
5. Situational Awareness. Via different validated methods including validated surveys and task-specific prompts.

By taking into account the gap in knowledge identified in this section, and the research goal elicited above, we present the following research questions, also presented in Section 1.3:

RQ1 How do changes in agent predictability (how easy it is to guess its next actions) and reliability (how good the agent is at the task) impact the human-agent relationship?

RQ2 How do different types of agent errors defined from previous related work such as slips, mistakes and lapses affect the human-agent relationship?

RQ3 How do different types of environmental conditions (static or moving), which impair vision and induce uncertainty, affect the human-agent relationship?

RQ4 How do different types of visual help (designed to elicit different levels of situational awareness) influence the human-agent relationship?

Chapter 3

Methodology

3.1 Motivation

Past human-agent interaction studies assessing trust, reliance and cognitive load often consist of either competitive (user VS agent(s)) [128] or collaborative (user AND agent(s)) [34, 91] scenarios where users interact with *one* [169] or *multiple* agent(s) [147]. In these studies, agent(s) either provide *information* aimed at helping users make a decision [27] or directly interact with the task as another autonomous entity [135]. In most HAI studies, the human-agent relationship is assessed via validated, general-purpose survey instruments which focus on different reported variables, such as trust [90], cognitive load [76], situational awareness [49] or system usability [108].

To answer our Research Questions presented in Section 1.3, we needed a human-agent collaborative framework that meets the following requirements:

1. Provides a collaborative scenario where human and agents have to work together in a **real-time task**.

2. **Can record information related to the human-agent interaction task itself**, such as performance or reliance on the agent.

3. Allows the usage of **standardised survey instruments** to measure reported measures such as trust, situational awareness or cognitive load.

4. Can allow the modification of **both the agent's attributes** (behaviours, reliability) and **the task's features** (difficulty, information provided or excluded).

Finding a framework that allows for the elicitation and collection of meaningful information while keeping participants engaged is an on-going problem in many research domains. For more than two decades [20, 200], HCI studies have reliably used commercially available games (such as "Minecraft" [32]) as well as purpose-made games (Atomic Orchid by Fischer et al. [61]) in order to study participants' behaviours in controlled environments. Based on our needs and after carefully reviewing relevant work, we decided to use a well-known and simple arcade game named “Missile Command” [10] and adapt its gameplay to a human-agent teaming scenario. We chose this particular game as its gameplay requires users to make quick decisions in time-limited scenarios where the difficulty of the tasks and reliability of a collaborative agent can be easily controlled. In comparison to other popular and accessible arcade games such as Space Invader or Pac-Man, Missile Command offers more flexibility regarding the addition of an agent, as our tasks strive to elicit fast decision-making, rather than focus on decision-planning. Other games were considered, such as DOTA 2 or Counter-Strike due to their anteriority and presence in related research [12, 141]. However, their inherent complexity, learning-curve and limited flexibility for custom-made agents meant that they failed to meet our study requirements.

Carter and collaborators [20] argue that games in HCI studies can be understood within 4 distinct research paradigms: “Operative”, “Epistemological”, “Practice” and “Ontological”. In the context of our work, we are interested in the **Epistemological paradigm**, as our research seeks to use a game-like activity to generate insights on a broader topic, namely *how humans trust and collaborate with an automated system.*. We also focus on the **Ontological paradigm**, as our work strives to inform the design of future interactive systems by focusing on the elements that make up the game’s interface or mechanics.

In order to satisfy all our requirements, we created an interactive human-agent collaborative task in the form of a 2D collaborative game. In this game, the user and agent have to cooperate to protect assets by destroying a series of incoming missiles. The agent can aim automatically, however only the user has the ability to fire projectiles. This framework offers the following advantages:

- It offers a real-time task where important behavioural constructs can be operationalised and recorded, such as task performance (number of missiles hit or shot fired) and reliance on the agent (amount of times the user corrected the agent).
- The task can be easily divided into blocks of various difficulty levels where survey instruments can be deployed to measure pre- and post-hoc concepts such as trust, cognitive load or situational awareness.
- It can be deployed either in a lab environment (with gamepad or keyboard controls) or exported online as a web-app where participants go automatically through all stages of the study.
- Flexibility, as elements such as the agent’s accuracy, the availability of information or the complexity of the task can be controlled and monitored to design immersive scenarios.
- The task is user-friendly and requires no prior knowledge. Controls are easy to learn, and the goal of the task is straightforward.

Using commercially-available games can be a quick and effective way to have participants engage in immersive tasks, especially when involving a multitude of users. For instance, the work of Schaekermann et al. [156] used the game “Destiny” to understand what motivates player and how this relates to their own personality. Results from such studies, while insightful, are heavily context-specific, and commercially available games are often not flexible enough to incorporate custom-made AI agents or the logging of specific behavioural information. In addition to commercial games, frameworks used in previous work often occurred in either domain-specific [57, 60, 61] or turn-based scenarios [157] where the human-agent relationship evolves asynchronously. Trust, as described in Section 2.1, evolves over-time and designing frameworks that allow for the study of its interactions with other elements would benefit our understanding of trust in automation. More recently, an increasing number of studies have used arcade-like games to study human-agent relationships, as noted in the work of Rapp et al. [144] who describe arcade games as being “inspiring” thanks to their “simplicity and immediateness” [144, p. 4].

3.2 Missile-Command Framework

In this section, we present the overview of our framework, mostly inspired by the 1980 “Missile Command” arcade game [10]. The goal of our real-time interactive task consists in aiming at and destroying missiles appearing from the top of the screen in order to protect 4 cities situated at the bottom of the screen. To achieve this goal, participants can move a crosshair across the screen and fire projectiles in the direction of their choosing. **In our studies, agents help participants by moving the crosshair automatically. At any moment, however, participants can choose to override the agents’ input and manually move the crosshair.** In all scenarios, only participants can fire projectiles to destroy incoming missiles. As in related studies where reliance is defined as “the tendency to employ automation to replace manual control” [148], we decided to ensure that there was a way for participants to *correct* the agent’s decision by taking over the controls and *confirm* theirs or the agent’s actions by assigning firing capabilities to participants. This design decision also ensures that participants have to keep on monitoring the situation, in order to keep a maintain the necessary level of engagement with the task.

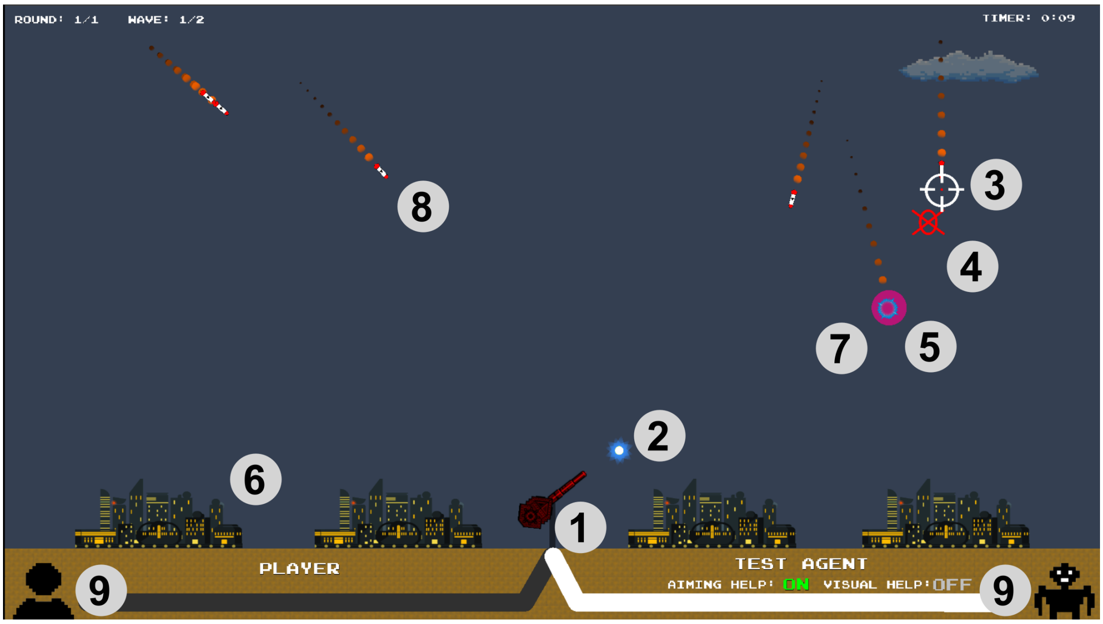


Figure 3.1: Screen-capture of the latest version of the missile command game framework (as used in Chapter 7), where elements of interest are annotated and described below, in Section 3.2.

Figure 3.1 shows our interactive game in action, where the main elements are numbered and described as follows:

1. **Gun-turret**: controlled by either the participant or the agent in order to aim and target incoming missiles. All projectiles are fired from the turret.

2. **Projectile**: fired by the participant, it travels at a fixed speed (1250 pixels/second) until it explodes in a small circular area. If a missile lies within this area, it is destroyed. The speed of the Projectile was constant and hard-coded for each study after initial informal pilots were conducted to tailor the different elements in the game.

3. **Crosshair**: provides a visual indication of where the participant or agent is aiming. The crosshair changes colour depending on who is controlling it: yellow for the participant, white for the agent, and dark-grey for neither. Crosshair colours were selected to be easily distinguishable depending on whether the user or the agent was controlling the crosshair. In addition, we ensured that each colour (white and yellow) represented a significant contrast to the dark blue background of the game for increased accessibility. The speed at which the agents or the users can move the crosshair was the same throughout each study (800 pixels / second).

4. **Red Indicator Area**: appears when a projectile is fired to show participants the area where the projectile will explode.

5. **Projectile's explosion/halo**: In order to be destroyed, missiles have to enter the radius of such an explosion.

6. **City**: Assets that the participants are tasked with protecting.

7. **Missile Impact**: when a missile reaches a city, it produces an orange/red explosion with smoke emanating from the city.

8. **User and Agent panels**: The participant's panel (on the bottom left of the screen) and the agent's panel (on the bottom right) light up in green when one of them is moving the crosshair.

9. **Enemy missile**: progresses at a fixed speed and angle depending on the task difficulty. At the end of a session (with or without an agent), participants are shown how many missiles they hit and/or missed. All missiles missed eventually hit a city. Depending on the study, the number of missiles appearing (spawning) and their individual speed was varied in order to test different levels of task complexity.

3.2.1 Godot Engine

After conducting a review of the different tools dedicated to the creation of interactive software, we decided to use a game engine [110] as game engine specialise in the prototyping and deployment of highly interactive software while taking into account hardware constraints and the need for speed of execution, all of which are paramount to the success of seamless, real-time and immersive human-agent interactions. Multiple engines were considered, such as Unity [74], Unreal [54] and Godot [69]. All of these engines have been used in HCI studies, for instance to simulate visual impairments via Virtual Reality [107] or to promote e-learning [31]. After carefully examining the pros and cons of each game engine, we decided to choose the newer Godot Engine to create our interactive game as it is:

- A highly mature 2D engine with a very simple Python-like programming language called Godot Script, which makes prototyping much easier that with, say, Unity which uses C++. This drastically cut-down the learning curve associated with designing prototypes.

- Much lighter than Unity or Unreal, which reduces hardware requirements for both developers and end-users.

- Completely open-source, which could facilitates the reproduction of results and dissemination of the game.

- Easily exportable to WebGL and HTML, which was important to conduct online studies running on different types of hardware via web browsers. Such exports are possible under the Unity or Unreal Engine, but have an impact on the performance of 2D games (as of 2021).

All studies presented in this thesis were designed using the aforementioned Godot Engine in its versions 3.0 (for the study presented in Chapter 4) and 3.2 (for the studies presented in Chapters 5, 6 and 7).

3.2.2 Interaction & game controls

For participants to play the game, they have to assume control of the gun tower located at the bottom of the screen to fire at incoming missiles (see element 1 in Figure 3.1). In our lab-based studies, presented in Chapters 4 and 5, participants were interacting with the game via an Xbox360 controller using the analog stick to move and the (A) button to fire projectiles. Subsequent studies presented in Chapters 6 and 7 were conducted entirely online, where participants interacted with the game using their keyboard's directional arrows to move the crosshair and space bar to fire projectiles. The controls used in all studies are displayed in Figures 3.2 and 3.3.

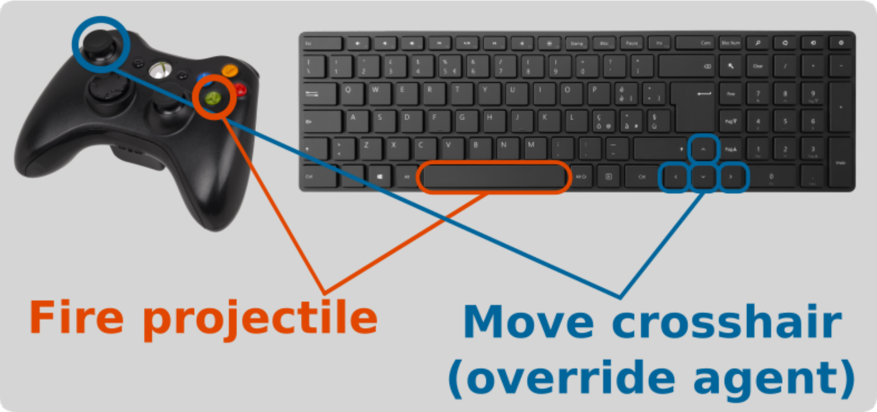


Figure 3.2: Controls used in our studies. For lab-based experiments, a controller was used, whereas online studies relied on keyboard inputs.

In all studies, participants alone were able to fire projectiles, which fact we used as a way of measuring the extent to which participants were explicitly confirming agents' actions. When a projectile is fired, the following events happen:

1. A red target/cross is placed at the location where the crosshair is located. This red target indicates where the projectile is heading and will remain there until a projectile has reached it.

2. A projectile is travelling from the gun-tower to the red target at a fixed speed hard-coded into the game. The speed of the projectile was calibrated after initial pilots to adjust game elements.

3. Upon reaching the previously spawned red target, the projectile will disappear and spawn in the same location an "explosion" in the form of a blinking circle with a fixed radius. The blinking circle will remain in place for a set duration.

During sessions with the agents, participants could decide to let the agent guide the crosshair, or take over the controls and correct the agent manually. We recorded the number of times participants took control and the duration of each takeover as a way of measuring explicit reliance on the agents.

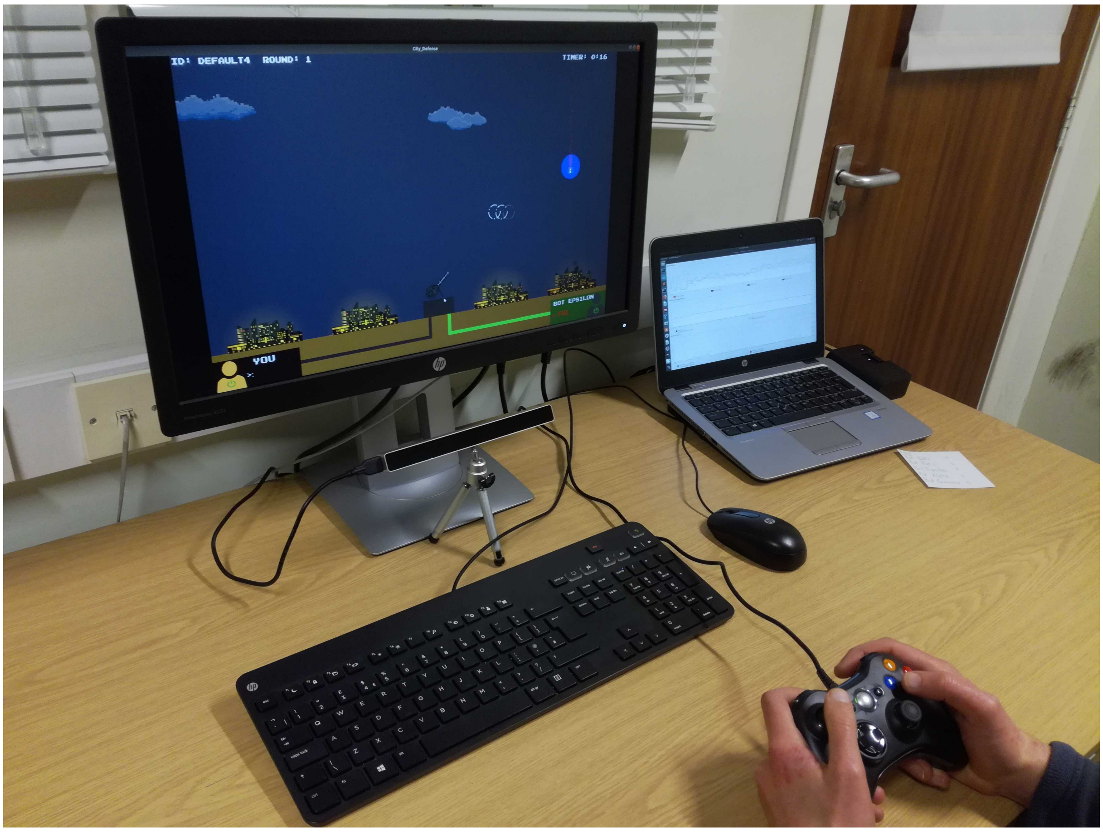


Figure 3.3: Experimental setup used in our lab-based studies presented in Chapters 4 and 5

3.2.3 Interface

While the majority of the elements in our game remained unchanged throughout our studies, some visual changes and improvements were incorporated by making use of participants' feedback. The most noticeable changes were brought to the game interface (see Figure 3.4 used in the first study, and 3.5 used in all subsequent studies) where we decided to simplify three items:

1. The "lines" indicating the agent or the user was controlling the crosshair. We decided to make them bigger and use colours (yellow for users and white for agents) depending on which one was in charge of aiming.

2. Names and textbox messages. In the first study (see Figure 3.4), we used simple messages when certain actions were conducted (for instance "get ready..." when the task was about to start, "fire now!" when the agent was in position). We decided to remove them for subsequent studies as the message could have distracted participants and anthropomorphised agents in ways that would be hard to assess and control.

3. Labels AIMING HELP (agent aiming automatically) and VISUAL HELP (agent providing visual information). These only appeared in the study presented in Chapter 7.

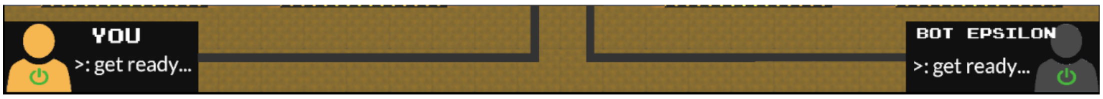


Figure 3.4: Early interface located at the bottom of the screen and used for our first study focusing on agent reliability and predictability.

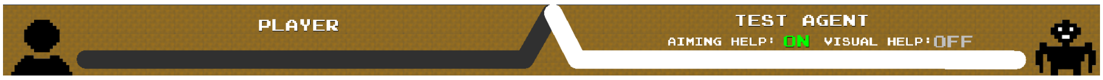


Figure 3.5: Most up-to-date interface located at the bottom of the screen and used in studies 2, 3 and 4.

3.2.4 Task Difficulty

In our game, we controlled the task difficulty by modifying the following variables, also presented in Figure 3.6:

- Spawning rate of missiles (how often missiles appear from the top of the screen).
- Speed of each missile (in pixel/seconds, with faster speed requiring quicker decision-making to destroy).
- Type of missile (threat or non-threat), which was only added for our last study presented in Chapter 7. When describing missiles' types, we use the term "Threat" to indicate when a missile is heading toward a city and "Non-Threat" when it is not.

As our game is inspired by the 1980 arcade game "Missile Command" [10], we decided to use equivalent features to control the pace and difficulty of the task. Our decision to incorporate different difficulty levels was motivated by previous HCI studies that demonstrated how different levels of task difficulty can affect dependant variables such as cognitive load or task performance [1]. In our game, the difficulty of each level is defined by the delay between each missile spawning (spawn-rate) as well as their individual speed. A lower spawn-rate and faster missiles lead to a higher degree of difficulty. For consistency reasons, the delay mentioned above was not altered within a single round, but rather between studies in order to craft immersive and challenging scenarios. While these changes don't allow for between-studies comparisons in terms of difficulty levels, we always included a "no agent" (as in, *user-only*, without any agent) session in order to judge users' skills on their own. This assessment allowed us to monitor the individual level of performance of each user without the help of a collaborative agent, which in turn provided information about the perceived difficulty of the task. In the studies presented from Chapter 4 to Chapter 7, we often qualify each level of difficulty as "easy", "medium" or "hard" in order to compare results more easily.

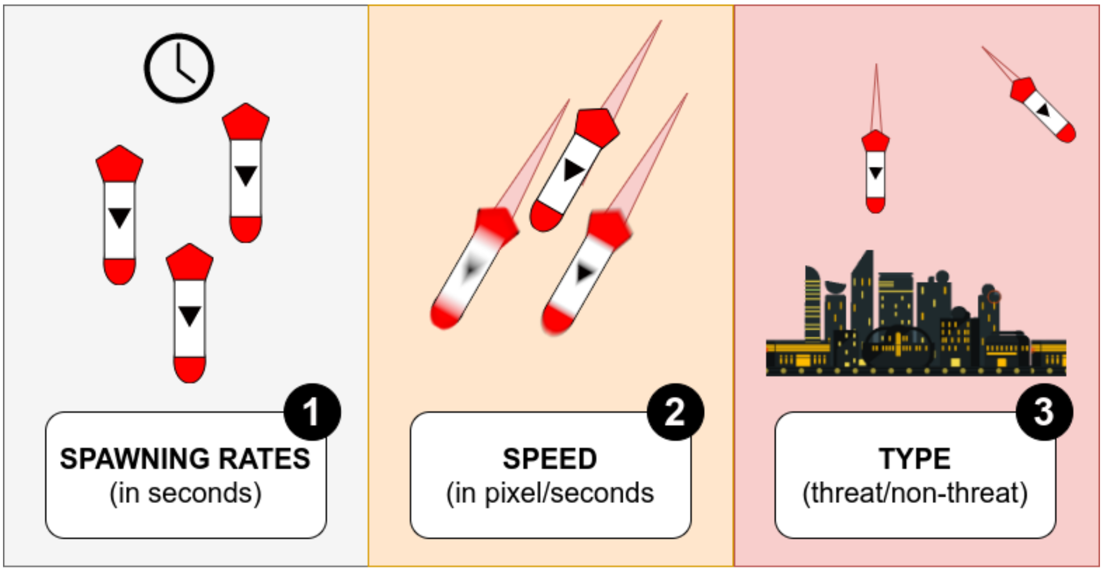


Figure 3.6: Main variables controlling the task difficulty. Each variable can be modified to induce different levels of task complexity.

3.3 Collaborative Agents

3.3.1 Purpose & Interaction

In our experiments, we tested the impact of collaborative agents that varied in terms of their reliability and behaviour in an explicit human-agent collaborative task. Agents helped with the decision-making process by either guiding the crosshair towards targets (see Chapters 4, 5 and 6 and 7) or by providing visual information regarding the agent's reasoning and/or the environment of interaction (see Chapter 7).

We developed a series of agents each displaying different behaviours in order to study the resulting impact on the human-agent relationship and performance. As our studies focus exclusively on agents' behaviours through their in-game decisions, we decided not to include anthropomorphic elements such as human-like avatars or voice-enabled agents. Furthermore, findings from studies with anthropomorphic systems are usually "highly sensitive to the human individual in the loop" [p. 7] [59], which makes generalisation based on the results harder in studies with a limited amount of participants. Table 3.1 presents a summary of all different agents used in our studies. Depending on the individual study and its respective focus, variables were modified to affect one or more of the following elements:

1. The Agent's accuracy (how reliable it is, see Section 3.3.2).
2. The Agent's error pattern (how easy it is to anticipate its next errors, see Section 3.3.2).
3. The Agent's behaviour (the kind of overall behaviour that affects its capacity to detect and target missiles, see Section 3.3.3).

4. The Agent's error types (the amount of False Positive and False Negative errors the agent makes) (only used in the study presented in Chapter 7, see Section 3.3.3).

Table 3.1: Summary of all agents used in our different studies with their associated levels of reliability and differences in the way they behaved and made errors.

Study	Agent Codenames	Baseline Reliability	Error Pattern	Temporary Error Behaviour	Balanced Type 1&2 Errors
1	Agent 1.a	70%	Systematic Biases	None	False
	Agent 1.b	30%	Systematic Biases	None	False
	Agent 1.c	70%	Random Biases	None	False
	Agent 1.d	30%	Random Biases	None	False
	Agent 1.e	100%	None	None	False
2	Agent 2.a	70%	Random Biases	Lapses	False
	Agent 2.b	70%	Random Biases	Slips	False
	Agent 2.c	70%	Random Biases	Mistakes	False
	Agent 2.d	70%	Random Biases	None	False
3	Agent 3.a	80%	Random Biases	None	False
4	Agent 4.a	80%	Random Biases	None	True

3.3.2 Aiming & accuracy

Despite their differences in terms of reliability and behaviours, the process that each agent went through in order to acquire targets and aim at them was kept the same throughout all studies. The agent aiming accuracy or behaviour was modified depending on the research focus of each experiment, for instance by introducing different levels of reliability and predictability in Chapter 4 or varying the way the agents make errors in Chapter 5. This section presents the targeting process of each agent. The following steps take place when an agent registers a target:

1. Retrieves information about the target including its *position and speed.*
2. Retrieves information about the *gun-tower's position* and its *projectiles' speed.*

Information regarding positions are taken as 2D Vectors (with X and Y values), whereas information regarding speed are computed as integers (speed of missiles as pixels per second). Using the aforementioned information, the agent will then compute where best to aim in order to hit the target. A pseudo-script presented in Algorithm 1 details the procedure.

Algorithm 1 How an agent computes an optimal targeting position. The algorithmn is largely based on Pythagoras' theorem, as calculations are done in a 2D space.

```
procedure getOptimalAimingPosition(target,targetSpeed,projSpeed,projRotation,delta)
    Get the future position of the target at the next delta using its speed and rotation.
    Get the distance to the future position of the target.
    Compute the travel time for a projectile to hit the target given the distance to the future
    position of the target and the speed of the projectile.
    Return the position to move the crosshair to hit the target given the future position of
    the target and the travel time of the projectile.
```

Once the optimal target to hit a specific missile has been computed, a fixed bias can be applied in order to skew the agent's aiming in a random or biased direction, which in turn controls the agent's performance in targeting accuracy. This variance in the agent's performance was calculated using a random Gaussian distribution. Using a Gaussian distribution (also called normal distribution) ensures that the accuracy of the agent is controlled throughout the task in a consistent fashion. When computing new coordinates for the agent to move to, a fixed σ (randomly determined using a Gaussian distribution) is used to determine the level of the agent's performance. The greater the variance/σ, the less accurate the agent's aim is, resulting in lower aiming accuracy. Figure 3.7 illustrates these changes.

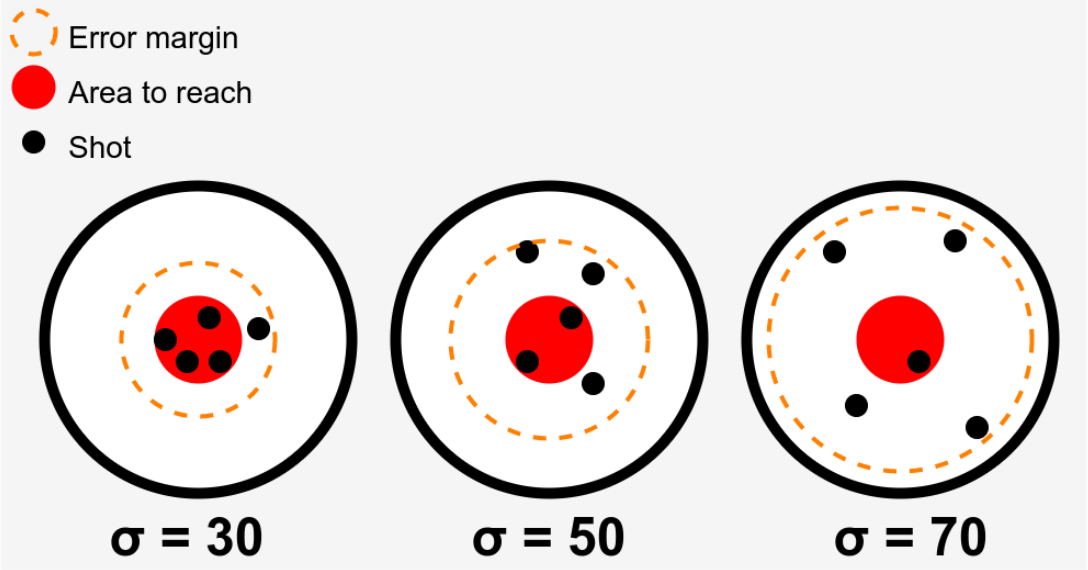


Figure 3.7: Simplified representation of how different sigma values (noted σ) skew the agent's error margin, resulting in different levels of accuracy. The greater the value, the lower the agent's performance.

3.3.3 Target Awareness & Prioritisation

In addition to an agent's aiming accuracy, another key component of the agent's performance is its ability **to register and assign priorities to targets of importance**. For studies where all targets represented a threat (see Chapters 4, 5 and 6), the agents' errors were manufactured by varying the accuracy of each agent, but not their capacity to detect targets. In our final study (see Chapter 7), we introduced False Positive and False Negative errors, where the agent would register targets as non-threats when they were threats (False Negative error) and threats when they were non-threats (False Positive error). Figure 3.8 shows how these types of errors are distributed given a probability of 80% of missiles spawning as threats and 20% as non-threats.

Assigning priorities to targets is the other critical component of the agent's targeting system, and directly influences how reliable and predictable the agent is perceived to be by users. The first study (see Chapter 4) was designed with a simple priority system where the agent would target *the oldest targets* as each missile spawned one at a time, travelled at the same speed and was always aimed at a single city within one round. While this simple prioritisation system worked for the first study, it had to be modified in subsequent studies to take into account the presence of multiple missiles with varying speed.

For other studies (see Chapters 5, 6 and 7), multiple missiles spawned at the same time with varying speed, which called for changes in the way the agent prioritised targets. At this point, we implemented an algorithm that would assign priorities to targets based on their distance from the crosshair and their distances from the cities.

For the final study (presented in Chapter 7), we introduced false positive (missiles that do not hit cities). We decided to introduce an option to control how frequently (or not) the agent would aim at a false positive target. More details are available in Chapter 7. Below is a summary of all options that influenced the targeting awareness and prioritisation strategies of the agent, along with the Chapter where these options were present.

- Multiple missiles spawning at once (see Chapters 5, 6 and 7).
- Semi-randomised missiles' speed and targets (see Chapters 6 and 7).
- Non-threatening missile (i.e. not heading toward a city, to add False Positives to the task) (see Chapter 7).

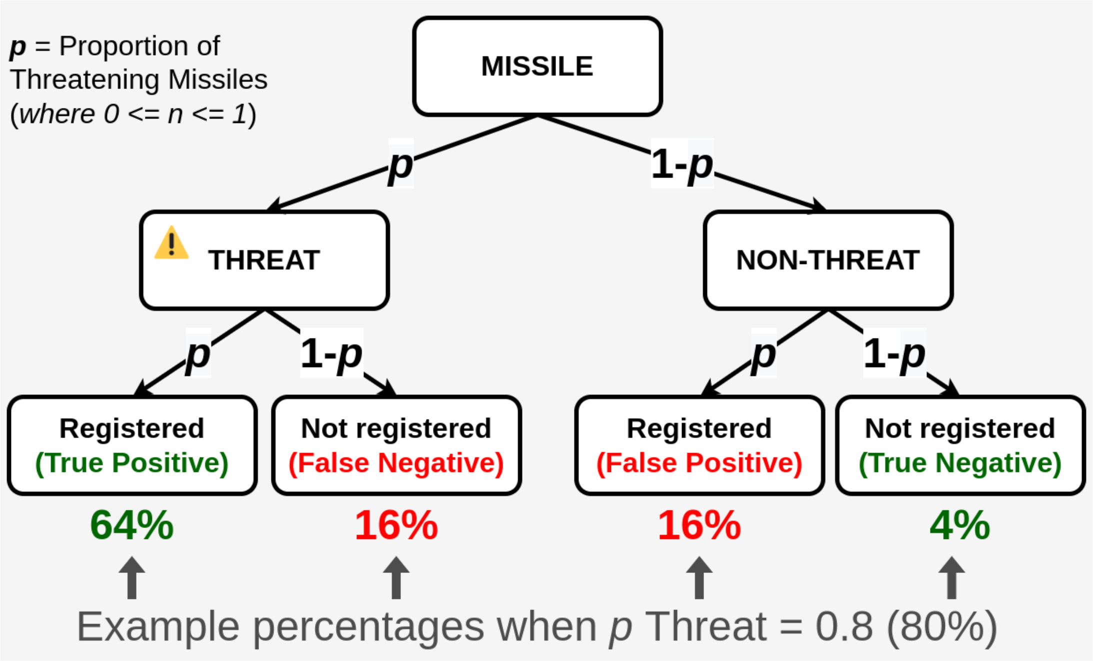


Figure 3.8: Distribution of agent's errors per missile when manually inputting a target number of False Positive and False Negative errors. Here, the examples shows a balanced distribution of 50% False Negative and 50% False Positive errors.

3.4 Evaluation

3.4.1 Experimental Procedure

In this thesis, we used different dependent metrics to assess human-agent collaborative scenarios in real-time by modifying the difficulty of the task(s) and the behaviours of agents. We also defined dependent variables to monitor and record changes in the human-agent relationship via log-based metrics (recorded during the task) or survey instruments. Figure 3.9 presents an overall diagram of the procedure used for all studies presented in this thesis. In all our studies (presented in Chapters 4, 5, 6 and 7), participants were briefed on the experiment and asked to provide consent prior to undertaking study. After completing a demographic questionnaire and questions related to their gaming experience and trust in automation, participants were first given a short tutorial on how to play the game. In all studies, participants were instructed that their goal was to protect cities by destroying all incoming missiles and that agents were there to assist with aiming. Participants were informed that they could always correct the agents' aim if they so desired. To reduce the learning effect, the sequence in which participants interacted with each agent was randomised using a William Square design [193] in order to ensure consistency in all studies. At the end of the study, participants were compensated for their time with either a shopping voucher for lab-based studies or online payment for remote studies.

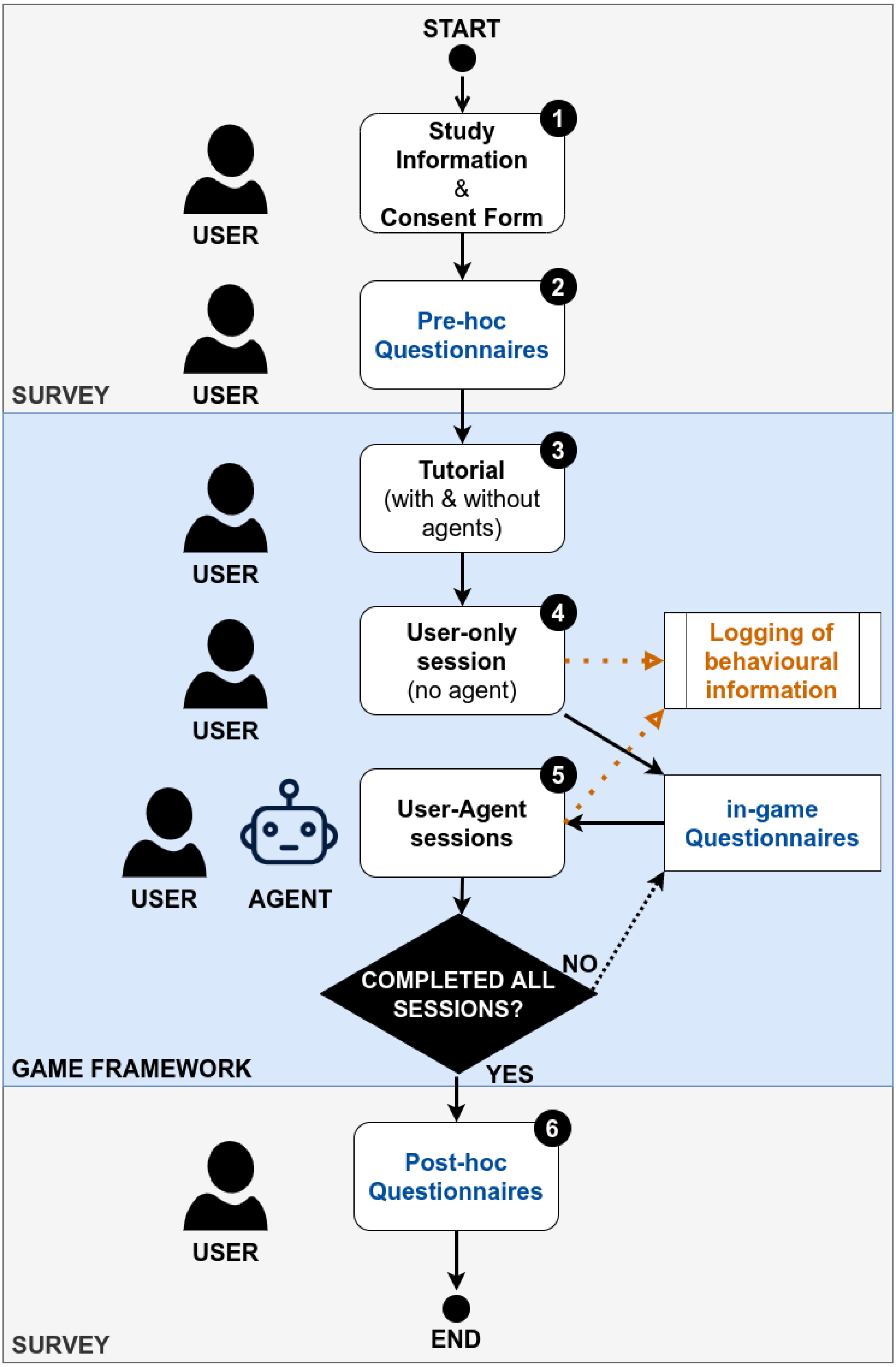


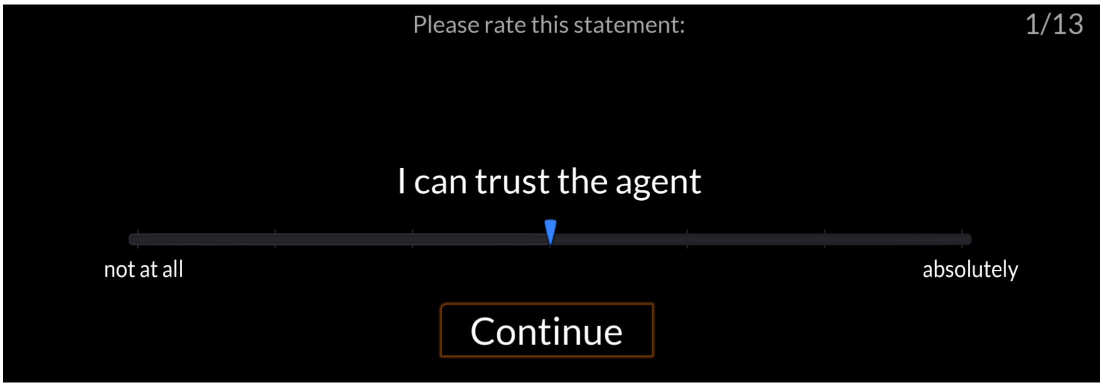
Figure 3.9: Experimental procedure using this framework. Both survey instruments (in blue) and the logging of behavioural metrics (orange) were used to assess the human-agent relationship at different points in the study.

3.4.2 Survey-based assessment

Most reported measures designed to assess the quality of the human-agent relationship were collected using an integrated survey system. Figure 3.10 provides an example of a question displayed during a task. The template can be modified and adapted to fit any other rating-scale based instrument. More details on the nature of analogue scale instruments are available in Section 2.6.1.2.

3.4.2.1 Trust

To measure reported trust in agents in our studies, we mostly relied on the questions designed by Jian et al. [90], available in Appendix E.2. In this survey, a series of statements has to be rated by participants in order to elicit different attitudes relating to trust in automation. In our work, we used either parts of Jian's scale (see Chapter 4 and 5) or a single item scale (see Chapter 6 and 7). In all studies, statements were presented at the end of tasks and required participants to report their trust on Likert scales with verbal anchors that denoted a "complete distrust in the agent" at worst (left) and a "total trust in the agent" at best (right).



Please rate this statement:

I can trust the agent

Figure 3.10: Example of the in-game prompt used for self-reported measures (in the example: trust). When prompted, questions are displayed (one at a time) and participants have to rate statements according to verbal anchors. Numbers and extra labels can be added when needed. The pointer has to have been clicked at least once before the user can click on Continue.

3.4.2.2 Cognitive Workload

To evaluate cognitive workload, we used the validated NASA TLX survey instrument, available in Appendix E.3, which consists of 6 individual rating scales relating to Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration. Each scale can be rated from 0 (low) to 21 (high). This method is the most widely used means of measuring cognitive workload [75]. In all studies, we focused on reporting the "RAW TLX" score, which consist of aggregated Nasa TLX transposed on a 100 point scale. This technique was shown to be as useful as weighted scores in a meta-review article by Cao et al. [18].

3.4.2.3 Situational Awareness

Situational Awareness can be described as the capacity to locate and remember important information to carry out a task and make predictions about future outcomes [48]. There are three commonly used methods to measure situational awareness, namely:

1. **SAGAT** (Situation Global Assessment Technique) [48],

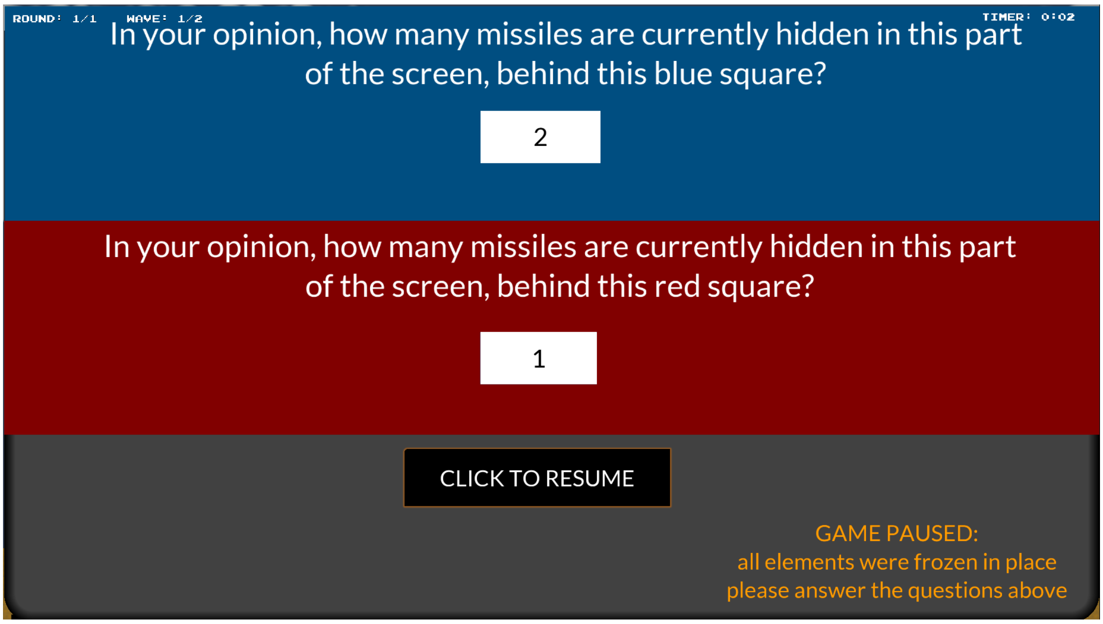
2. **SART** (Situation Rating Technique) [173]

3. **SPAM** (Situation Present Assessment Method) [45].

All methods have pros and cons. While some require freezing the task (SAGAT), others require multi-tasking (SPAM), or the completion of a post-hoc survey (SART). A recent comparison of all these methods [52] found the SAGAT method to be the least intrusive, and least harmful in terms of users' task completion performance. In this thesis, we used the SAGAT method in Chapter 6 and SPAM method in Chapter 7. More details are presented in the relevant Chapters (6 and 7) regarding their choice and implementation.

SAGAT [49] consists in designing a set of SA related queries related to a specific task. During the experiment, the task is frozen, giving participants time to answer the queries. A numerical value is then derived from each question and aggregated into an overall score evaluating participants' overall SA. Figure 3.11 shows the SAGAT instrument as used in our framework.

SART [173] consists in administering a pre or post-hoc survey designed by Taylor et al. [173]. An example of the survey is available in Appendix E.4. The survey is composed of 10 questions spread across three domains: "Attentional demand" (3 questions), "Attentional supply" (4 questions) and "Understanding" (3 questions). Participants have to rate each statement on a 7-point Likert scale from "Low" to "High". There is also a 3-item version of SART called "3D SART" which comprises a single statement for each of the three dimensions.



In your opinion, how many missiles are currently hidden in this part of the screen, behind this blue square?

2

In your opinion, how many missiles are currently hidden in this part of the screen, behind this red square?

1

GAME PAUSED:
all elements were frozen in place
please answer the questions above

Figure 3.11: Implementation of the SAGAT survey instruments to evaluate Situational Awareness at the first level (perception of data). When prompted, the task is immediately frozen and hidden in the background.

3.4.3 Behavioural metrics

3.4.3.1 Performance

During gameplay, various game-specific variables are monitored to record task-performance, both during sessions with an agent (team performance) and without (user-only performance). The variables we are interested in are computed with the amount of shots fired, missiles destroyed and shots missed. These variables are directly linked to the success of the task, with more missiles hit leading to a greater overall performance. Below are the metrics used to study performance with Recall, Precision and F1.

$$\textbf{Precision} = \frac{\#MissilesDestroyed}{\#ShotsFired} \tag{3.1}$$

$$\textbf{Recall} = \frac{\#MissilesDestroyed}{\#IncomingMissiles} \tag{3.2}$$

$$\textbf{F1} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.3}$$

Higher precision indicates greater accuracy (fewer attempts to hit a target), while higher recall indicates greater task performance (less damage being sustained by the cities). F1 is the harmonic means of precision and recall and provides a combined measure of performance. The user control time was computed as the number of seconds participants were controlling the crosshair during each round (a greater user control time indicates less reliance on the agent).

3.4.3.2 Reliance

In our framework, reliance is measured via user control time, which is the amount of time for which the user controlled the crosshair in a scenario where an agent was present. The greater the time, the lower the reliance on the agent. Similar ways of measuring reliance have been used in other human-automation and human factors studies [14].

3.4.4 Data Visualisation and Statistical Testing

This section discusses the methodology and choices for the visualisation of quantitative data and statistical analysis performed in all of our experimental studies. Analysis was conducted using the Python 3 programming language [181] and several statistical libraries such as statsmodel [158], Scipy [184] and Pingouin [178].

Throughout our experimental studies, from Chapter 4 to 7, figures and tables were used to present quantitative results. In order to make data visualisation easier, boxplots were used to present quantitative results as they excel at clearly showing differences between experimental conditions [195]. An example of a boxplot with added explanations of its components is presented on Figure 3.12.

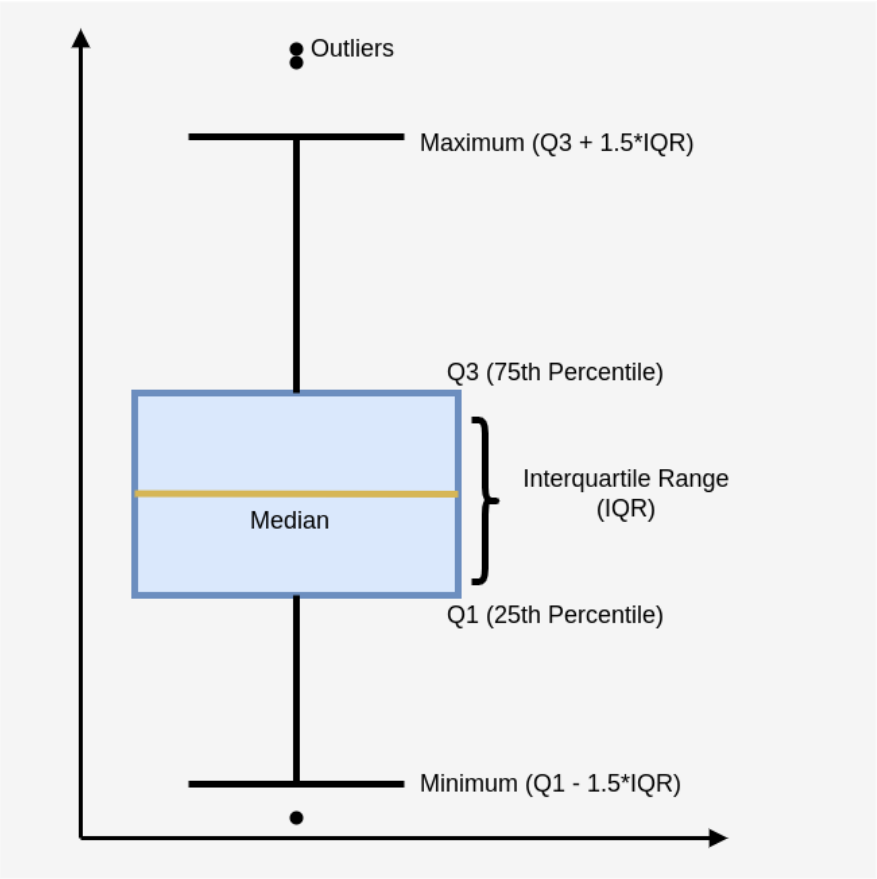


Figure 3.12: Example of a boxplot with a description of each of its components.

The remainder of this section presents the statistical methods used to compare experimental conditions. Before conducting any statistical comparisons, the Shapiro-Wilk test was used [162] to test the null hypothesis that samples came from a normally distributed population with a threshold for significance set at $p = 0.05$.

In the case of a normally distributed population, the following tests are conducted depending on the design of the study. For a between-groups design, a Levene's test [104] is used to determine whether the population is homoscedastic (equal variance) or not. If the homoscedasticity test is positive, an ANOVA [67] test is performed with follow-up pairwise Tukey-HSD tests [175], whereas if the homoscedasticity test is negative, a Welch ANOVA test [112] is conducted with follow-up pairwise Games-Howell [64] comparisons. For Within-group studies, repeated-measures ANOVAs [67] are used with follow-up pairwise T-tests. For mixed design studies (within and between groups), Mixed ANOVAs [67] are used with follow-up pairwise T-tests [199].

In the case of a non-normally distributed population, the following tests are conducted depending on the study design employed. For comparisons involving 2 groups, Wilcoxon signed-rank [192] tests are used in the case of repeated measures, and Mann-Whitney U tests [123] in the case of non-repeated measures. For comparisons involving more than 2 groups, Friedman tests [164] were employed for repeated-measures, and Kruskal-Wallis tests [122] for non-repeated measures. In both cases, follow-up pairwise comparisons were realised using Wilcoxon signed-rank tests for paired samples [199].

Regardless of how studies were designed, subsequent pairwise comparisons were only conducted when statistically significant results were found during previous statistical modelling. When statistically significant results are reported, Bonferroni corrections [9] were always applied. In addition, effect sizes were always reported using The Common Language Effect Size (CLES) for both parametric and non-parametric comparisons [121].

3.4.5 Coding

During most of our user studies (see Chapters 5, 6 and 7), we asked participants for feedback at the end of the experiment. We collected this feedback in accordance with the Critical Incident Technique [62] where participants were asked to describe one or more positive and negative aspects of their interaction in the study. Participants' written feedback was then given to three different PhD researchers to perform coding analysis on. For each coding task, specific codes were created by the lead researcher in accordance with the research questions and focus set for the study. When presenting results of coding analysis, Kappa scores [183] were used to assess the internal agreement between coders.

3.5 Study Modalities

3.5.1 Lab-based

Lab-based studies were conducted in the University of Strathclyde's CIS department, and involved one-on-one meetings with participants in offices equipped with computers to run the framework on, and xbox360-type controllers to permit interaction with the framework. Studies presented in Chapters 4 and 5 were conducted in a lab. For lab-based studies, a researcher was present to verbally explain and respond to any questions a participant might have, in addition to indicating when they had to answer certain post-hoc survey instruments.

3.5.2 Online

Online studies were conducted due to the worldwide COVID19 pandemic and our ensuing inability to conduct lab-based studies. Online studies involved participants recruited on the online crowd-sourcing Prolific© platform [33]. The Prolific platform was chosen thanks to its large participant pool and strict regulations to promote high quality data collection and attention to the task. Every online study was designed to be taking place without the need for a researcher to be present. For these studies, the game was exported to WebGL which most modern browsers can load and execute locally (via HTML5 and Javascript) by connecting to an executable of our study apparatus located on the Strathclyde's CIS servers. Contrary to lab-based studies where the same hardware was used for all participants, online-based studies dictated that participants' use their own computers, which led us to implement hardware checks to ensure consistent experimental conditions between all participants. A one minute benchmark-test was provided for free before advising participants to take part in the study. Only participants who experienced more than 24 frames per seconds (FPS) at a resolution of 720p (1280 x 720 pixels) or above were kept for further data analysis in our online studies. The frame-rate threshold was set based on audiovisual standards [191] while the resolution cap was set based on the first draft of the framework (see Chapter 4), which was designed at a signal format of 720p.

Compared to the lab-based studies, the online version required a more streamlined process where participants could easily navigate between the different steps of the experiment, as researchers were not present to assistant participants in tackling the study. Most of the time, the duration of experiments was shortened in order to foster better quality of data, and a benchmark was added to allow participants to test whether their machine could run the game properly, before taking part in the study. In addition, in-game performance was recorded to make sure that all participants experienced the game in comparable settings. Table 3.2 describes the main differences between the lab-based and online-version of our study.

Table 3.2: Comparison of the lab-based and online modalities of the framework.

Feature	**Lab-based**	**Online**
Access to pre,post-hoc and in-game survey	YES	YES
Access to the same game scenario	YES	YES
Logging of in-game performance	YES	YES
Benchmarking tool	NO	YES
Online logging of participants data	NO	YES
Researcher available to answer question(s)	YES	NO
Semi-structured interviews	YES	NO
Xbox controls	YES	NO
Keyboard controls	NO	YES

Part II

Investigating Collaborative Human-Agent relationships

Chapter 4

Agent Reliability and Predictability

4.1 Motivation

In this chapter, we explore how different levels of agent predictability and reliability influence users in a real-time task. As we have seen in Section 2.1, many features influence the propensity of a human operator to trust and rely on an automated agent. Past work has shown that an agent's performance (in terms of reliability) as well as an agent's behaviour (in terms of predictability) are positively correlated with trust [42, 129]. However, such studies have largely been conducted in turn-based settings [34, 131] where operators and agents interact asynchronously. Human-agent teams often work together in real-time scenarios where the trust relationship evolves over time and is affected by various factors such as task performance and agent behaviours [85]. Currently, there is a limited amount of work exploring the relationship between performance, predictability and trust when agents and humans work together in real-time collaborative settings. The study presented in this chapter answers our first Research Question, formulated in Section 1.3: **How do changes in agent predictability (how easy it is to guess its next actions) and reliability (how good the agent is at the task) impact the human-agent relationship?** Specifically, this work seeks to address the following sub research questions: **How, at the same level of agent reliability, do changes in the agent's predictability affect**:

- **RQ1.a:** the users' task performance when interacting with agents?
- **RQ1.b:** the users' reliance on the agent?
- **RQ1.c:** the users' cognitive workload when interacting with the agent?
- **RQ1.d:** the users' reported trust in the agent?

The work presented in this chapter is based on a previously published article entitled *Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration* [41].

Using the collaborative framework detailed in Chapter 3, we conducted a lab-based 5 (agents) x 3 (levels of difficulty) within-subjects study with 30 participants who interacted with five agents exhibiting different levels of reliability (more or less reliable) and predictability (more or less predictable) in tasks of varying difficulty (easy, medium and hard). We found that, at the same level of agent reliability, the more predictably the agent makes errors, the greater the reported trust and task performance, and the lower the cognitive load.

4.2 Related Work

This Section presents relevant past work that has studied systems biases and trust in automated systems. An extensive number of studies have explored ways of measuring trust in systems (see the work of Schaefer et al. [155] for a comprehensive review). As we have seen in Section 2.1 and 2.3, most prior work has focused on assessing trust in relation to the agent's reliability. Less attention, however, has been paid to examining the effects of agent reliability *and* predictability in real-time human-agent collaborative tasks.

In HAC scenarios, agents are generally introduced to reduce users' cognitive workload, while trying to improve users' situational awareness and overall task performance [44, 56, 92, 170]. Past work has shown that an agent's *reliability* and its task *performance* heavily influence users' willingness to trust and rely on it. In their seminal work on the trustworthiness of command and control systems, Sheridan et al. [165] posit that the "effectiveness" (and subsequent reliability) of a system will influence how the operator perceives and trust a systems, and that it is possible to quantitatively measure the development of trust. The work of Robinette et al [146] showed that, in the case of human-robot coordination tasks, even a single error from a robot can strongly impact the development of a person's trust, regardless of how the robot fails at the task. A different study by Hoc et al. [84] on human-agent cooperation in a driving task showed that the type of automation error as well as its timing will have the most impact on how likely operators are to understand, trust and subsequently rely on the system. In their study, Hoc et al stress the importance of understanding the cognitive process involved in cooperating with automation, and how a system's failure can alter it.

In another command and control task focused on threat assessment, Fan et al. [57] tested different levels of agent variability (using systematic biases). They found that informing participants about the agent errors helped users to calibrate their trust accordingly, which led to higher task performance. However, too much information related to the agent's errors can quickly overload users. In a related work, Chavaillaz et al. [24] investigated different levels of agent reliability on trust, reliance and overall task performance in a turn-based X-ray scanning scenario. Their results showed that, as agent reliability decreased, so too did trust in agents. Furthermore, they found that perceived reliability (how much a person is willing to rely on an agent's input) is also affected by the capabilities of the automated system. In their studies, users' perception of the reliability of agents was more accurate when interacting with low-performing agents, compared to high-performing ones. In addition to studies focusing on different degrees of reliability, the work of Shirado et al. [166] explored turn-based coordination problems and found that error-prone agents (up to 30% loss in accuracy) could be beneficial to collaborative performance as their presence reduced the probability of a complacent attitude developing towards the agent.

Given the evidence of past research, it is clear that the performance of an agent (its reliability) as well as the agent's behaviour (its predictability) impact trust, but it is unclear how precisely both of these concepts influence it. Past work in human-agent interaction [24, 57] linked higher predictability and agent reliability to higher reported trust in the agent. In this Chapter, based on these previous insights, we are going to study how human-agent collaboration evolves based on the following working hypotheses:

1. **H1** At the same level of agent reliability (performance), agents exhibiting systematically biased behaviours (errors committed in a more predictable and consistent fashion) will be trusted more than agents exhibiting randomly varied behaviours (errors that are unpredictable and committed in an inconsistent way).

2. **H2** As seen in previous HAI work, we hypothesise that it is possible to use behavioural data from human-agent interactions to model and infer users' perceived trust in agents.

The main contribution of our work lies in testing the impact of different degrees of agent reliability on the human-agent trust relationship in real-time scenarios. We use interaction data to model and determine how accurately reliance, agent reliability and performance can help us predict trust in automation.

4.3 Method

To study agent reliability and predictability, we used our interactive human-agent framework described in Chapter 3 in a lab-based study the modality of which is outlined in Section 3.5.1. As the focus of this study is on the explicit impact of predictability and reliability, we designed five different agents that each varied in their *targeting behaviours*. These behaviours were controlled to create different levels of reliability and predictability. The resulting experiment is a 2x2 within subjects design with 2 levels of agent reliability (low and high) and 2 levels of agent predictability (more or less predictable). We also included "no agent" and "perfect" agent conditions serving respectively as the baseline (no support from an agent) and upper bound (highest agent reliability) for the experiment.

4.3.1 Agent Reliability

Agents were designed with different level of reliability, which directly affected how good they were at hitting targets. All agents had a certain degree of variance in their aiming accuracy such that, for a given target, a certain degree of error would be applied to the targeting. This variance in the agent's performance was calculated using a random Gaussian distribution with a fixed σ (sigma) for each level of the agent's performance. The greater the variance (and thus the σ), the less accurate the agent's aim, leading to worse reliability and task performance (see Figure 4.1).

4.3.2 Agent Predictability

In addition to variance, some agents had their aim systematically biased towards a particular direction: (i) always above and to the right of their target, (ii) always below and to the left, (iii) always above and to the left, (iv) always below and to the right. The direction of the systematic bias was randomly selected for each participant at the beginning of the experiment and kept constant during the session. By randomly selecting the direction of the bias for each participant, we ensured that our findings were not constrained by a specific type of systematic bias. By setting specific fixed biases, participants could learn to anticipate the error committed by the agents. This systematic bias impacted the agents' targeting behaviours, but *not* their performance, which were only impacted by random variance.

Agent performance was calibrated using simulations where agents completed the task by themselves (e.g. the same task without users). In these simulations, we calculated the agent's performance based on the Recall scores described in Section 3.4.3.1. We then used T-tests to ensure that the performance of agents with a similar level of reliability were not significantly different. This was to make sure that high or low degrees of predictability would not impact agents' reliability, thus allowing comparisons. While comparing the Recall scores of agents Alpha and Gamma (low performing agents), a t-test yielded $p > 0.05$. Similarly, t-tests performed using the Recall scores of agents Beta and Delta also yielded $p > 0.05$. Agents Beta and Delta were tuned to be high performance (approx. 0.7 Recall scores or 70% of the targets being hit), while agents Alpha and Gamma were tuned to be low performance (approx. 0.3 Recall scores, or 30% of the target beings hit).

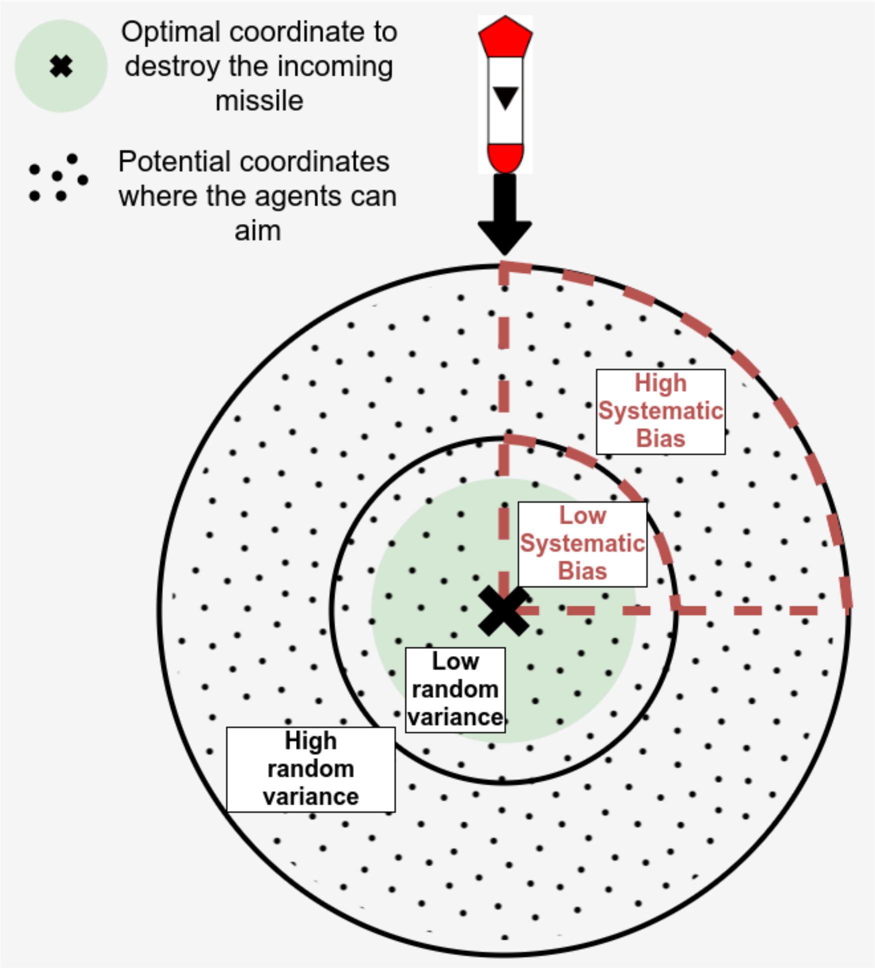

Optimal coordinate to destroy the incoming missile
Potential coordinates where the agents can aim
High Systematic Bias
Low Systematic Bias
Low random variance
High random variance

Figure 4.1: Visualisation of the different biases applied to the agents in the study (not to scale). The greater the bias, the lower the accuracy of the agent. For the systematic bias, a quadrant is randomly chosen for each participant at the beginning of a session. Low systematic bias and low random variance or high systematic bias and high random variance result in the same performance output.

4.3.3 Agent configurations

Using different combinations of levels of agent reliability (low and high) and predictability (more or less predictable), we designed five different agents. Agent names were introduced to make it easier for participants to refer to a particular agent. Agents *Alpha* and *Beta* were designed to be more predictable with respectively a high (Alpha) and low (Beta) level of performance. Agents *Gamma* and *Delta* were designed to be less predictable with respectively a high (Gamma) and low (Delta) level of performance. In addition to the aforementioned agents, we also included a *perfect* agent: "Epsilon" which exhibited no bias and no variance – and thus had the highest reliability and predictability out of all of the agents.

Figure 4.2 shows the different combinations of agents used, which we refer to as: Alpha, Beta, Gamma, Delta and Epsilon (A,B,C,D,E). Here is a summary list of the agent configurations used in this study:

- No Agent.
- **Agent Alpha**: high performance, low predictability.
- **Agent Beta**: low performance, low predictability.
- **Agent Gamma**: high performance, high predictability.
- **Agent Delta**: low performance, high predictability.
- **Agent Epsilon**: Highest performance and predictability.

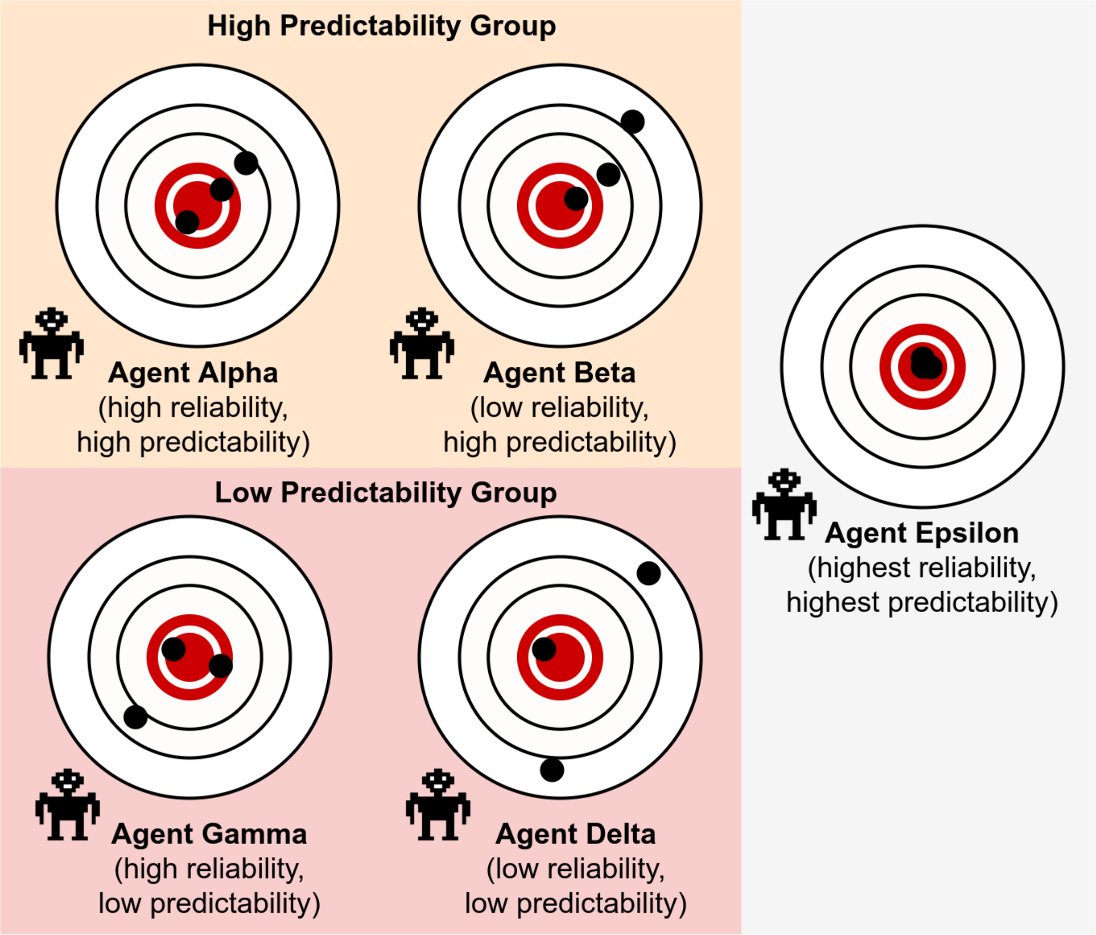


Figure 4.2: Simplified representation of all agents used in this study and their respective aiming patterns. Both agents displaying either HIGH or LOW performance are programmed with the same accuracy, regardless of their aiming patterns.

4.3.4 Task Difficulty

During each interaction with an agent, participants went through three rounds which lasted for 90 seconds each. This duration was set so that participants had enough time to familiarise themselves and adapt to the agents, while ensuring that the experiment could be completed within an hour (reducing participants' fatigue). Each round increased in difficulty (going through

"Easy", "Medium" and "Hard" levels of difficulty). Here are the details about task difficulty in this study:

- In the "Easy" level, missiles spawned every 4 seconds at a speed of 100 pixels per second for a total of 22 missiles.
- In the "Medium" difficulty level, missiles spawned every 2 seconds with a speed of 150 pixels per second for a total of 45 missiles.
- In the "Hard" difficulty level, missiles spawned every second with a speed of 200 pixels per second for a total of 90 missiles.

The speed and number of missiles in each level were calibrated during pilot testing with 10 participants, to make sure that changes in difficulty were noticeable without completely overwhelming participants (see Section 4.3.5 for a more detailed description of the pilot study).

4.3.5 Piloting

Before conducting the main study, a formal pilot experiment was carried out. Ten participants were recruited from our local Computer Sciences department. This pilot experiment focused on calibrating the single player (no agent) experience, as well as core gameplay elements such as the controls, visuals and overall difficulty of the game.

To evaluate participants' performance, we used F1 scores. F1 is a metric related to participants' overall task performance and is computed using the number of missiles participants hit, the number of shots fired and the total number of missiles present in each level. For more information, all of the performance metrics are detailed in Section 3.4.3.1. F1 scores varied between 0.88 for the "Easy", 0.77 for the "Medium" and 0.46 for the "Hard" difficulty levels. As the difference in terms of participants' performance between the "Easy" and "Medium" levels was low (a loss of 0.11 for F1 scores), we decided to increase the speed of missiles in the "Medium" difficulty level to intensify its difficulty.

4.3.6 Independent and dependent variables

In this section, we summarise the independent and dependent variables used in this study and as motivated by our experimental method and research questions.

Our independent variables are the following:

- *Task difficulty*, as defined by the amount and speed of missiles in each level.
- *Aiming agent reliability*, how accurate the agent is in its predictions (low, high or near-perfect reliability).
- *Aiming agent predictability*, how predictable the agent's actions are (low or high).

Our dependent variables are the following:

- *Task Performance*, in terms of missiles hit, shots fired and missile missed.
- *Reliance*, expressed by the duration for which participants relied on the aiming agent's help.
- *Trust*, as reported by participants.
- *Cognitive Workload*, as reported by participants.

4.3.7 Experimental Procedure

Ethics approval for this study was obtained from the University of Strathclyde's Department of Computer and Information Sciences (Approval No. 793). The study took place in a lab located in the University of Strathclyde Computer Sciences Department. The experimental procedure is outlined in Section 3 and in Figure 3.9. The study lasted for approximately an hour, and participants were compensated for their time with a shopping voucher worth £10. After being briefed on the experiment and asked to provide consent, participants went through the following steps:

1. Demographic questionnaire and pre-hoc surveys. (5 minutes).
2. Tutorial with and without an agent. (2 minutes)
3. Session without an agent, to record individual performance. (4 minutes 30 seconds).
4. One session with each agent (Alpha, Beta, Gamma, Delta, Epsilon). (4 minutes 30 seconds each).
5. Post-hoc surveys. (5 minutes).

The sequence in which participants interacted with each agent was randomised using a William Square design in order to mitigate learning effects [193]. During each session, participants played through three rounds of *low* to *medium* and *high* levels of difficulty. At the end of each round, participants were asked to rate their trust in the agents using a subset of the "Checklist for trust in Automation" created by Jian et al [90]. At the end of each session, participants were asked to complete the NASA Task Load Index (TLX) questionnaire [75]. Survey instruments are present in more detail in Section 3.4.2.

4.3.8 Demographics

Participants were recruited through mailing lists and flyers posted on our university campus. We recruited a total of 30 participants (14M, 16F) with ages ranging from 19 to 38 years old ($M = 27 \pm 5.19$). Most participants were enrolled as postgraduate students. When asked how often they played video-games on a 6 point Likert scale from 0 (never) to 6 (Very Frequently), participants scored on average 3.63 (± 1.85), indicating that they played "occasionally". Ratings from the Complacency Potential Rating Scale (CPRS) [167] were used to evaluate general attitudes toward automation. CPRS scores ranged between 55.57 and 90.84 ($M = 72.55 \pm 9.3$) which indicates that our sample consisted of participants who were more likely to rely on automation than not [167]. Overall, the distribution of scores was homogeneous enough that our sample could *not* be divided into different groups representing distinct attitudes toward automation. The pre- and post-hoc survey instruments presented to participants on the Qualtrics platform are available in Appendix A.

4.4 Results

In this section, we present results regarding task performance, users' reliance on agents, reported trust in agents and cognitive workload. We used the overall scores participants obtained at the end of each session, across all levels of difficulty. More details on the inclusion of different difficulty levels are available in Section 3.2.4. The statistical methods we used to compare and report results are detailed in Section 3.4.4.

Table 4.1: Metrics related to performance (Recall, Precision and F1, higher scores = better performance) and reliance (User control time (in seconds) higher = less reliance on the agent). Superscript letters next to the results indicate which agents yielded significantly worse scores ($p < 0.05$). Highest values are highlighted in bold.

	No Agent	**Agent Alpha** **rel+/pred+**	**Agent Beta** **rel-/pred+**	**Agent Gamma** **rel+/pred-**	**Agent Delta** **rel-/pred-**	**Agent Epsilon** **highest reliability**
Recall	0.64 ± 0.03	0.82 ± 0.02	0.60 ± 0.03	0.72 ± 0.02	0.58 ± 0.03	**0.98 ± 0.01**
Precision	0.57 ± 0.02	0.60 ± 0.02	0.50 ± 0.03	0.53 ± 0.02	0.47 ± 0.02	**0.86 ± 0.01**
F1	0.60 ± 0.02	0.68 ± 0.02	0.54 ± 0.03	0.60 ± 0.02	0.51 ± 0.03	**0.91 ± 0.01**
User Ctrl Time	25.12 ± 0.96	5.34 ± 0.83	24.18 ± 1.16	10.61 ± 1.09	**27.68 ± 1.29**	1.02 ± 0.43

4.4.1 Performance

To measure task performance, we computed Recall, Precision and F1 scores based on the number of shots fired, missiles hit and total missiles present in each level of our experiment. Recall, Precision and F1 scores are detailed in Section 3.4.3.1. Table 4.1 and Figures 4.3, 4.4 and 4.5 show the average task performance achieved by participants with each individual agent. These scores are averaged over all three levels of difficulty. If we consult Table 4.1, we can see that participants achieved better Recall scores while interacting with high reliability agents (Alpha and Gamma) than on their own (without an agent). Participants performing poorly in the no agent session benefited the most from this increase in performance. As expected, participants performed the best with agent Epsilon (highest reliability and predictability) than with any other agents, across all measures. When interacting with Alpha (high reliability, high predictability) and Gamma (high reliability, low predictability), participants were able to achieve higher precision scores than on their own (no agent), but performed worse with Beta (low reliability, high predictability) and Delta (low reliability, low predictability) across all performance measures, yielding lower Recall, Precision and F1 scores. A repeated-measure ANOVA yielded significant results for Recall scores ($p < 0.0001$, $F = 51.26$, $np^2 = 0.64$), Precision scores ($p < 0.0001$, $F = 60.67$, $np^2 = 0.68$) and F1 ($p < 0.0001$, $F = 59.01$, $np^2 = 0.67$). Follow-up pairwise comparisons using pairwise T-tests yielded significant results between Alpha (high reliability, high predictability) and Gamma (high reliability, low predictability) for Recall ($p < 0.0001$, $T = 6.36$, $CLES = 0.72$), Precision ($p < 0.0001$, $T = 4.63$, $CLES = 0.64$) and F1 ($p < 0.0001$, $T = 5.42$, $CLES = 0.67$).

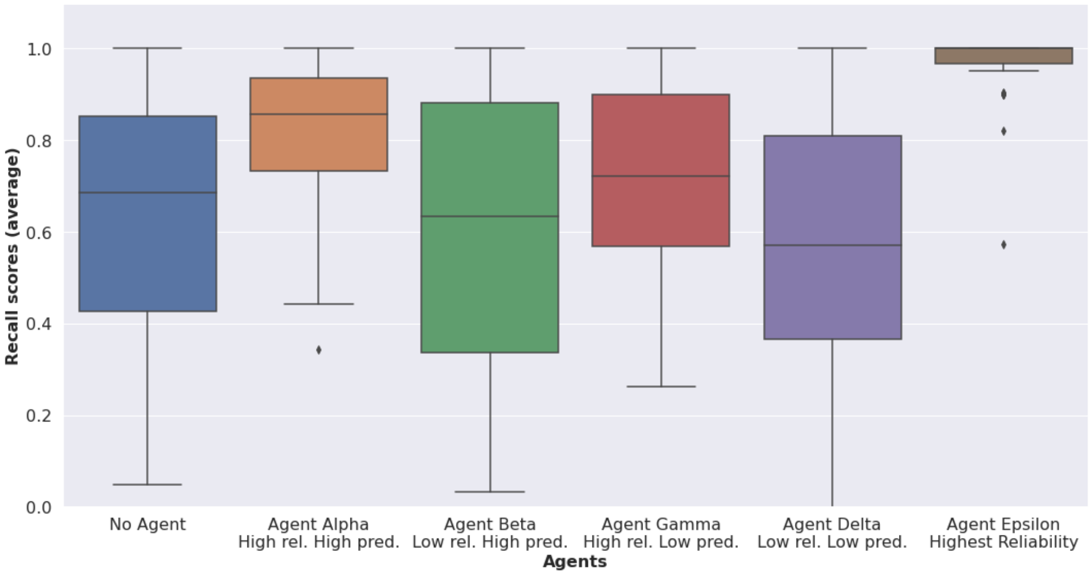


Figure 4.3: Recall scores for each session. A higher score indicates a better performance. Recall scores give a measure of how many missiles participants and agents hit. We can see that agent Alpha (high predictability and reliability) led participants to achieve higher median Recall scores than agent Gamma (low predictability and high reliability).

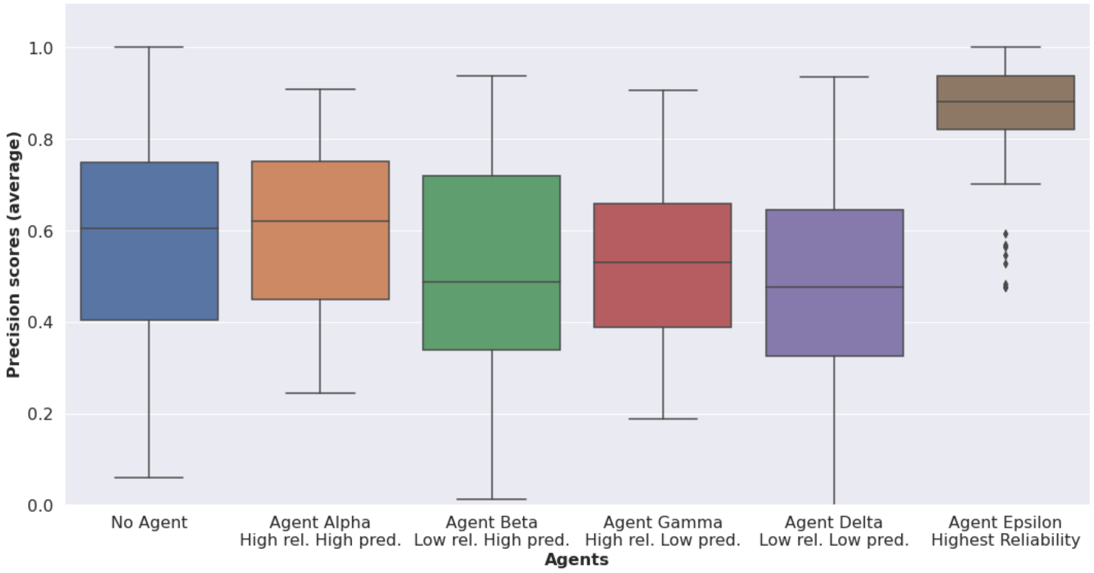


Figure 4.4: Precision scores for each session. A higher score indicates a better performance. Precision scores give a measure of how effective, in terms of shots fired per missile hit, participants were at the task. Overall, precision scores are higher for more predictable agents.

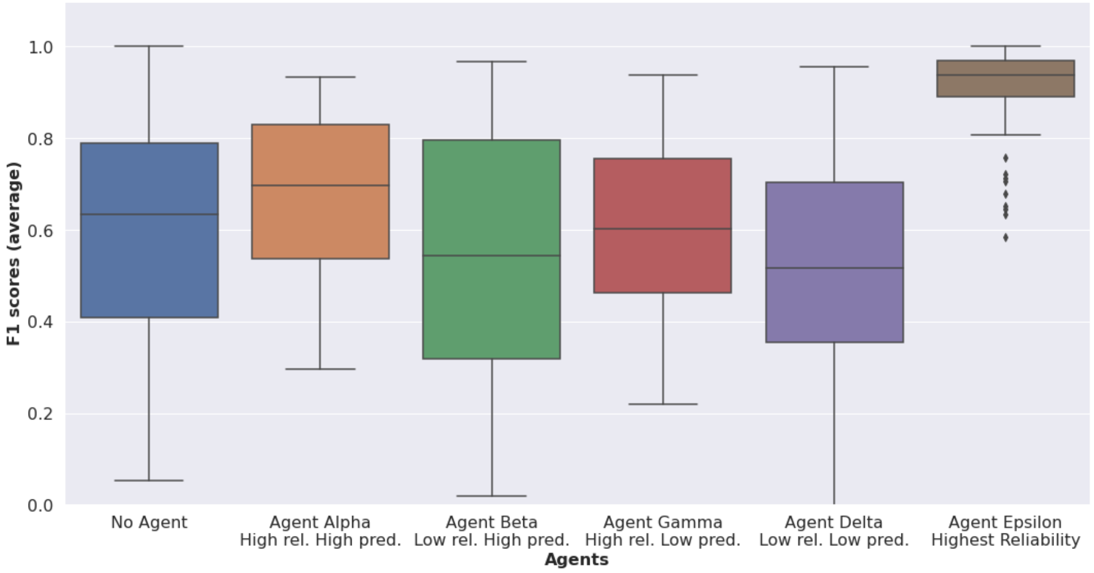


Figure 4.5: F1 scores for each session. A higher score indicates a better performance. F1 is a harmonised mean of Recall and Precision scores. Overall, F1 scores are higher for high predictability agents.

4.4.2 Reliance

To measure to what extent participants relied on an agent, we computed the duration for which each participant controlled the crosshair. Participants controlling the crosshair for a longer period of time suggested that they relied on the agents less (and vice versa). Table 4.1 and Figure 4.6 show the average amount of time (in seconds) participants spent in control of the crosshair (denoted as *User Ctrl Time*). As expected, we observed that participants spent less time controlling the crosshair when working with Epsilon (highest reliability and predictability) compared to any of the other conditions, with or without agents. In addition, participants spent more time controlling the crosshair when collaborating with low reliability agents (Beta and Delta) compared to high performance agents (Alpha and Gamma). A Friedman test yielded significant results ($p < 0.0001$, $W = 0.85$) when comparing the overall user control time between sessions. Follow-up non parametric pairwise T-tests indicated that participants relied on the agent significantly more when interacting with agent Alpha (high reliability, high predictability) than with agent Gamma (high reliability, low predictability) with $p < 0.0001$, $U = 250$, $CLES = 0.28$.

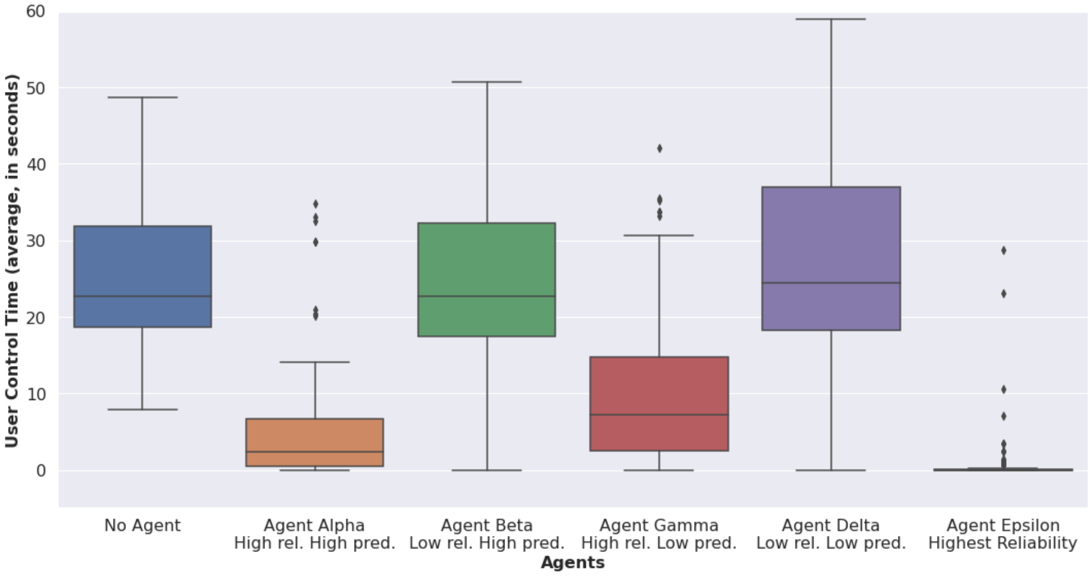


Figure 4.6: Amount of time each participant spent correcting the agents. A longer duration indicates less reliance on the agents. At at high level of agent reliability, participants corrected the more predictable agent (Alpha) less than the less predictable agent (Gamma).

4.4.3 Trust

To measure trust, we asked participants to rate their perceived trust in the agent with a series of trust-related questions graded on a scale of 1 to 11, with a lower score indicating a lower reported trust in the agent. While multiple elements were used, we focused our analysis of trust on the ratings associated with the statement "I can trust the agent" presented in Table 4.2 and Figure 4.7. On consulting the results we notice that, on average, participants trusted agent Epsilon (highest reliability and predictability) more than any of the other agents, which was expected as its reliability was the highest. In addition, trust ratings of agents with low reliability (Beta and Delta) were on average much lower than agents with high reliability (Alpha and Gamma). When comparing answers pertaining to the reported trustworthiness of agents, a Friedman test yielded significant results ($p < 0.0001$, $W = 0.90$). While performing follow-up comparisons using Wilcoxon signed-rank tests, we found that participants rated Alpha (high reliability, high predictability) significantly higher than Gamma (high reliability, low predictability) with $p < 0.0001$, $U = 650$, $CLES = 0.72$. However, no significant results were found when comparing Beta (low reliability, high predictability) with Delta (low reliability, low predictability). These results indicate that, at the same high level of agent reliability, participants were more trustful of an agent with high predictability (Alpha) than an agent with low predictability (Gamma).

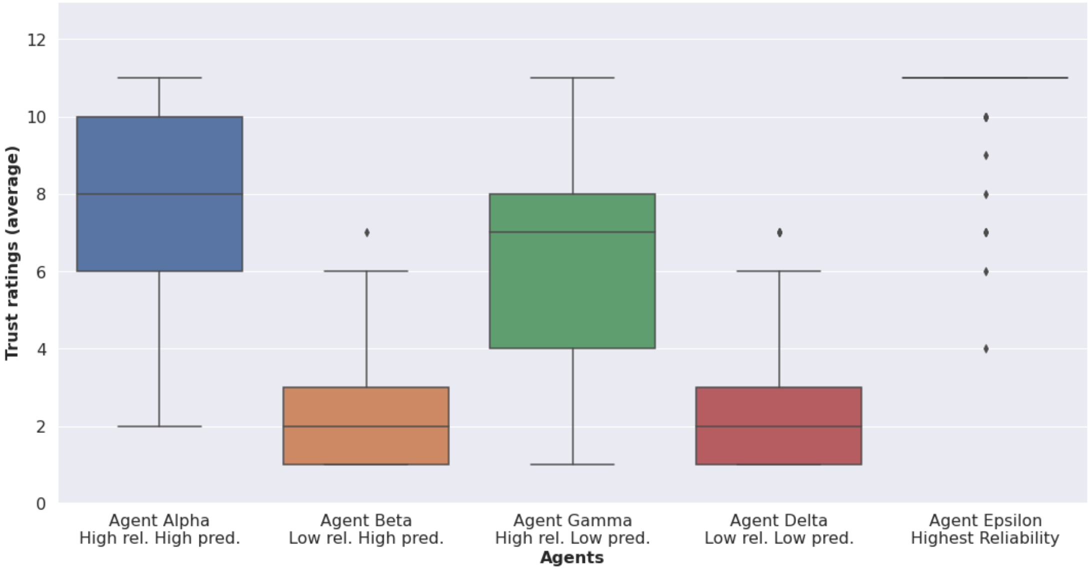


Figure 4.7: Average reported trust in the agents. Higher scores indicate greater trust in the agents. At a high level of agent reliability, participants reported a higher trust in the more predictable agent (Alpha) than the less predictable one (Gamma).

Table 4.2: Metrics related to reported trust and cognitive load. Higher scores indicate greater agreement with the statements. Highest values are highlighted in bold.

Question	Bot Alpha	Bot Beta	Bot Gamma	Bot Delta	Bot Epsilon
I can trust the system	4.80 ± 0.26	1.27 ± 0.10	3.57 ± 0.27	1.47 ± 0.16	**6.70 ± 0.13**
I enjoy interacting with the system	5.20 ± 0.26	1.87 ± 0.22	4.53 ± 0.29	2.03 ± 0.26	**6.13 ± 0.27**
I am suspicious of the systems intent, action or outputs	3.37 ± 0.32	5.13 ± 0.37	3.83 ± 0.33	**5.67 ± 0.35**	2.30 ± 0.42
The system's actions will have a negative outcome	3.13 ± 0.28	**5.70 ± 0.32**	4.00 ± 0.27	6.20 ± 0.19	1.90 ± 0.34
The system provides security	4.30 ± 0.24	1.37 ± 0.13	3.80 ± 0.28	1.53 ± 0.16	**6.37 ± 0.24**
The system is reliable	4.13 ± 0.30	1.33 ± 0.11	3.47 ± 0.30	1.70 ± 0.23	**6.67 ± 0.14**
The system is very unpredictable, I never know how it's going to act from one moment to another	3.33 ± 0.33	**5.80 ± 0.34**	4.17 ± 0.30	5.53 ± 0.36	2.07 ± 0.35
How mentally demanding was the task?	51.90 ± 4.43	71.27 ± 3.99	62.86 ± 3.68	**79.05 ± 2.81**	23.02 ± 3.24
How hurried or rushed was the pace of the task?	51.43 ± 3.95	73.02 ± 2.77	60.32 ± 3.05	**78.25 ± 2.36**	34.44 ± 4.78
How successful were you in accomplishing what you were asked to do?	65.56 ± 3.50	44.29 ± 5.26	57.46 ± 4.08	41.11 ± 4.95	**92.38 ± 2.59**
How hard did you have to work to accomplish your level of performance?	52.38 ± 3.65	**76.83 ± 3.25**	62.06 ± 2.73	80.79 ± 2.31	24.60 ± 4.10
How insecure, discouraged, irritated, stressed and annoyed were you?	34.60 ± 3.68	66.51 ± 5.22	42.86 ± 4.80	**66.67 ± 5.29**	19.52 ± 4.51
Overall RAW TLX score	51.17 ± 2.37	66.38 ± 2.15	57.11 ± 2.22	**69.17 ± 1.99**	38.79 ± 2.67

4.4.4 Cognitive Load

To measure Cognitive load, we used the Nasa TLX survey instrument detailed in Section 3.4.2.2. Higher scores indicate a greater reported workload. As presented in Table 4.8 and Figure 4.2, we observed that participants reported a much lower cognitive load (NASA TLX scores) when interacting with agent Epsilon (highest reliability and predictability) than with any of the other agents. Furthermore, participants reported a much higher cognitive load when interacting with low reliability agents (Beta and Delta) than with high reliability ones (Alpha and Gamma). When comparing overall Raw Nasa TLX scores, a repeated-measure ANOVA yielded significant results ($p < 0.0001$, $F = 85$, $np^2 = 0.75$). While performing pairwise T-test comparisons, we found that participants perceived the high reliability, high predictability agent (Alpha) as significantly less cognitively taxing than the high reliability, low predictability agent (Gamma) with $p < 0.0001$, $T = -2.96$, $CLES = 0.40$. In addition, participants found the agent with low reliability and low predictability (Delta) to be significantly more cognitively taxing than the agent with low reliability and high predictability (Beta) with $p < 0.0001$, $T = -2.07$, $CLES = 0.40$.

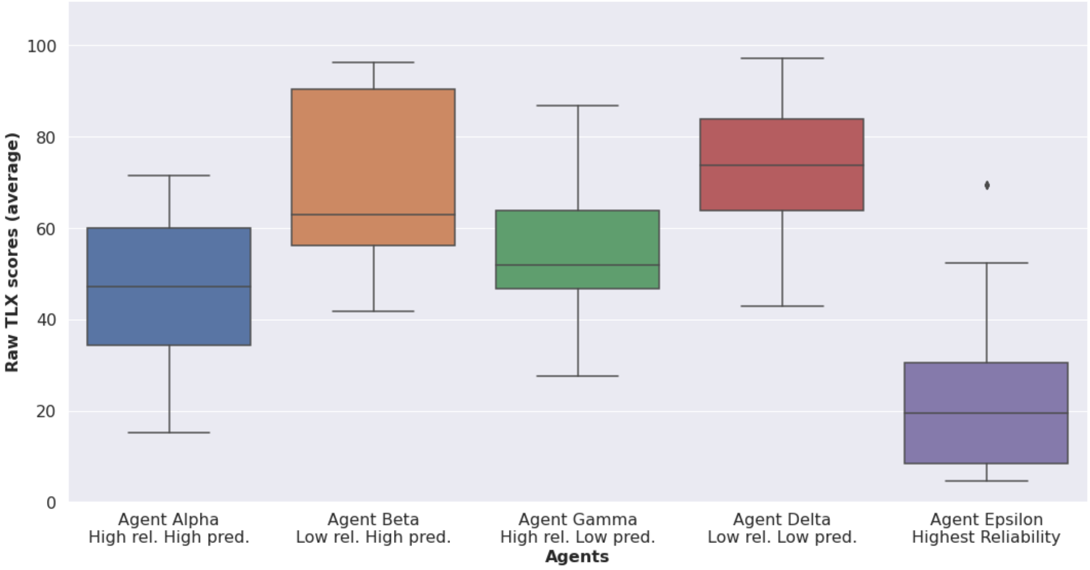


Figure 4.8: Raw NASA TLX ratings for each session with agents. Higher scores indicate greater cognitive loads. Raw TLX scores are lower for participants that interacted with the high reliability and high predictability agent (Alpha) compared to the high reliability and low predictability agent (Gamma).

4.4.5 Predicting Trust

In our Research Questions presented in Section 4.1, we sought to understand how different variables influenced reported trust. According to our working hypothesis presented in Section 4.2 which posited that trust can be predicted using behavioural information, we analysed correlations between trust ratings, task difficulty, the reliance metric (user control time), cognitive workload (NASA TLX scores) and performance metrics (Precision, Recall and F1 scores). From Table 4.4, we can see that participants' reliance on the agents (as measured by user control time) led to the highest correlation ($\rho = -0.801$, $p < 0.001$) followed by Cognitive Load (Raw TLX scores) with $\rho = -0.730$, $p < 0.001$, whereas performance metrics (Recall, F1 and Precision) resulted in lower correlations ranging from 0.50 to 0.61.

In addition to analysing correlations between our main variables, we created multiple linear regression models to determine which combinations of factors led to the best predictions of users' trust ratings. The selection criteria for the variables used in our models were based on the work of Hoff et al. [85], where elements related to the development of trust are categorised according to their impact on trust prior or during the interaction with an agent. Table 4.3 shows the combination of factors, mean square error, and adjusted correlation coefficients for each models. Our results show that the best performance for predicting trust ratings ($R^2 = 0.915$) were achieved by combining measures related to reliance (user control time), performance (the number of shots fired, missiles destroyed and misses), task complexity and information related to the participants' age and reported gender. These results corroborate the findings from [85] where elements captured during the interaction (such as performance and reliance related to "Dynamic Learned Trust" [85]) coupled with elements captured prior to the interaction (such as age and gender related to "Dispositional Trust" [85]) help us understand and be more accurate in our prediction of reported trust in the agent, even though each variable does not contribute equally to the overall quality of the predictive models.

Table 4.3: Linear regression results when predicting participants' trust ratings by using contextual (difficulty) and behavioural measures (performance and reliance). Only the most important results are presented. A higher R^2 value indicates more accurate predictions.

Parameters	**Total Mean Square Error**	**Adjusted R^2**
User Ctrl Time, Precision, Recall, F1 Difficulty, Raw TLX, Gender, Age	2491.9	0.915
User Ctrl Time, Precision, Recall, F1, Difficulty, Raw TLX	3244.1	0.894
User Ctrl Time, Precision, Recall, F1, Difficulty	3890.0	0.893
User Ctrl Time, Precision, Recall, F1	4717.9	0.867
Recall	17253.2	0.793
F1	16994.9	0.781
Precision	16666.6	0.766
Age	14927.5	0.686
Gender	13634.2	0.626
Difficulty	12517.2	0.575
Raw TLX	7796.8	0.357
User Ctrl Time	1830.2	0.082

Table 4.4: Spearman's correlation tests between behavioural or reported metrics and trust ratings. A higher ρ scores indicates greater correlation.

Parameter 1	**Parameter 2**	ρ	**p-value**
User Control Time	Trust ratings	-0.801	<0.001
Raw TLX	Trust ratings	-0.730	<0.001
Recall	Trust ratings	0.614	<0.001
F1	Trust ratings	0.553	<0.001
Precision	Trust ratings	0.501	<0.001
Age	Trust ratings	0.080	0.092
Difficulty	Trust ratings	-0.079	0.094
Gender	Trust ratings	-0.018	0.698

4.5 Discussion

In this study, we have explored how agent predictability and reliability influence users' perception of agents in terms of cognitive workload and trust, as well the implications on task performance. By conducting this study, we have provided a better understanding of the respective roles of agent reliability *and* predictability in human-agent collaboration and we have added to past work that mostly focused on testing different levels of systems reliability in collaborative tasks. This work helps us to answer our first Research Question: **How do changes in agent predictability (how easy it is to guess its next actions) and reliability (how good the agent is at the task) impact the human-agent relationship?** from the overall research questions elicited in Section 1.3. We sought to answer the sub-research questions presented in Section 4.1: **How, at the same level of agent's reliability, do changes in the agent's predictability affect**:

- **RQ1.a:** the users' task performance when interacting with agents?
- **RQ1.b:** the users' reliance on the agent?
- **RQ1.c:** the users' cognitive workload when interacting with the agent?
- **RQ1.d:** the users' reported trust in the agent?

With this work, we have found that interacting with more predictable agents at a high level of agent reliability positively contributes to the human-agent relationship in terms of task performance, reliance on the agent, reported trust in the agent and cognitive load. Additionally, we have also explored how different metrics correlate with reported trust in the agent. The next sections discuss the implications of our findings related to the sub-research questions outlined in Section 4.1.

4.5.1 Predictability in the agent's actions

We hypothesised that, at the same level of agent reliability, more predictable agents would be perceived as more trustworthy than less predictable ones. We found this hypothesis to be true, but only for agents with high levels of reliability. When looking at task performance, we found that interacting with a nearly perfect agent (agent Epsilon) led participants to achieve higher performance and to view the agent more positively in a general sense, which was to be expected from a highly reliable agent. When comparing the rest of the agents, however, clear differences in users' behaviours and perceptions were found.

With our sub-research questions presented in Section 4.1, we set out to explore how agent predictability impacted performance, reliance, workload and trust. When comparing the agents, we noticed that participants interacting with low reliability agents (Beta and Delta) yielded poor overall task performance, even worse than when participants carried on with the task without any agent (which informs RQ1.a). These results were the lowest across all performance indicators: F1, Recall and Precision (see Table 4.1). Moreover, participants had to compensate more for the agents' inaccuracy, as is evidenced by higher user control times, greater reported workload and lower trust ratings (which informs RQ1.b and RQ1.d). Nevertheless, when comparing agent Beta (low reliability, high predictability) to agent Delta (low reliability, low predictability), we found that participants performed slightly better with agent Beta, in addition to spending slightly less time correcting the agent and reporting significantly lower cognitive workloads (which informs RQ1.c), even-though this was not statistically significant. This suggests that when an agent makes errors in a systematic, predictable way, participants are able to compensate for its inaccuracy better.

When comparing agent Alpha (high reliability, high predictability) to agent Gamma (high reliability, low predictability), we found that participants achieved significantly higher performance with Alpha. They also corrected agent Alpha significantly less and reported significantly lower workload. These results further suggest that when an agent's behaviour is more predictable, participants could not only better compensate for the agents' imprecision, but also *adapt* and *work* with the agent better, resulting in an overall better task performance.

Overall these findings suggest that, in the case of imperfect automation, predictability in the way an agent makes errors is important. That is to say, even if it makes a number of errors, an agent with high predictability allows users to adapt better and quicker to its behaviour, which results in higher reported trust in the agent, as well as better task performance and reduced cognitive load.

4.5.2 Factors influencing trust

We further hypothesised that it is possible to infer trust in an agent using information collected during human-agent interactions. To investigate this area, we first sought to determine which factors were the most important to predict participants' perceived trust in agents. Table 4.4 shows correlations between trust ratings and other variables monitored in our study. While previous work hypothesised that performance is the most important predictive factor regarding users' trust in agents [85], our results show that the different performance indicators used in our study (F1, Recall and Precision) correlate only moderately with trust ratings. Moreover, our findings reveal that reliance, expressed by the amount of time users spent correcting the agents, was the metric most correlated with trust, which is in line with previous work [46,103]. However, we found that cognitive load (in the form of Raw Nasa TLX scores) was more strongly correlated with users' reported trust in the agents than task performance. This finding is consistent with other work that focused on predictive decision-making, where cognitive load was found to be affected by trust, reliance and the overall difficulty of the task [6, 201]. To further explore which combinations of factors could predict trust ratings best, we performed several multilinear regressions. We achieved the best results (see Table 4.3) by using data related to users' reliance on the agents, performance scores and the difficultly of the task. These findings suggest that it is important to consider both performance and reliance metrics in order to infer users' trust in an agent more effectively. Moreover, we demonstrated that it is possible to predict users' trust ratings with a high degree of accuracy.

4.6 Conclusion

In this study, we set out to explore the relationship between trust, agent predictability and agent reliability in a real-time collaborative scenario. To achieve this, we designed a within-groups study where participants completed a series of aiming tasks with the help of different collaborative agents. Our findings were aimed at answering our first Research Question: "How do changes in agent predictability (how easy it is to guess its next actions) and reliability (how good the agent is at the task) impact the human-agent relationship?"

We found that, at the same level of performance, participants reported higher levels of trust in agents that were more predictable than less predictable agents. However, as the agents' reliability decreased, participants were less trusting of them, regardless of their predictability. In addition, participants achieved better performance and reported lower cognitive load with systematically biased agents compared to agents with more variance, especially when agents' performance level was high.

These findings further highlight the importance of predictability and consistency in the design of potentially error-prone agents, and how it impacts human-agent collaboration in real-time. Furthermore, our study investigated whether it was possible to infer trust ratings based on participants' interactions. Our findings show that while performance indicators are important, in the context of real-time collaboration, participants' reliance on agents is a better predictor of trust. These findings suggest that it is possible to develop methods that can monitor trust in automation over time, and that such methods could be used by agents to better adapt to individual users. However, to develop trust-aware agents, we need to find ways of recording real-time operationalised metrics of reliance, and explore what other factors, beyond system reliability, affect users' perceptions of agents.

Chapter 5

Agent Errors and Behaviours

5.1 Motivation

In this Chapter, we are focusing on how different types of agent errors can influence human-agent collaboration. We designed a study where participants interacted with different automated agents that performed similarly, but made errors in different ways. We then analysed how different types of agent errors, at the same level of agent reliability, influence the human-agent partnership.

Our understanding of imperfect agent behaviours and its impact on trust is limited. As we have seen in Chapter 4, the behaviour of an agent and in particular the predictability of its actions can greatly influence how users will perceive and be willing to engage with it. When studying agent behaviours, most past work has focused on agent reliability and its impact on task performance, but less on the different ways in which an agent can fail at a task. The work of Marinaccio et al. [118] defined different error types, describing how each was identified and investigated in past work, and how best to mitigate their negative effect on users. Largely inspired by their taxonomy, we decided to run a follow-up study to the work presented in Chapter 4 and explore how different agent behaviours impact the human-agent relationship within our human-agent collaborative framework. More details on our usage of their taxonomy is available in Section 5.3.1.

With this work, we address our second Research Question, presented in section 1.3: **How do different types of agent errors defined from previous related work such as slips, mistakes and lapses affect the human-agent relationship?** As in our study on agent reliability and predictability in Chapter 4, we sought to investigate how different types of agent errors influence:

- **RQ2.a:** the users' task performance when interacting with agents?
- **RQ2.b:** the users' reliance on the agent?
- **RQ2.c:** the users' cognitive workload when interacting with the agent?

- **RQ2.d:** the users' reported trust in the agent?

  The work presented in this chapter is an extension of the paper entitled *Impact of Agents' Errors on Performance, Reliance and Trust in Human-Agent Collaboration* [37].

We conducted a lab-based 4 (types of agent errors - no error, slips, mistakes and lapses) x 2 (difficulty levels - easy or hard) within-subjects experiment with 24 participants. The main contribution of this work lies in testing whether, at the same level of agent reliability, different types of agent errors have a noticeable impact on users. Our results show that, when agents perform the same, agent errors are perceived differently and change the way participants interact with agents. For instance, slips and mistakes are more harmful to performance than lapses while slips are more harmful to reliance than mistakes.

5.2 Related Work

5.2.1 Agent Reliability and Errors in HAI studies

This Section presents relevant past work that has studied trust in agents and, in particular, how to categorise the way automated systems can fail at a task. In most human-agent collaboration scenarios, particularly safety-critical scenarios, it is assumed that agent reliability is one of the most important factors that will contribute to an individual's propensity to trust and rely on an agent, alongside other factors such as "stress levels" or "task complexity", as described by Grawbowski et al. [73]. The work of Hoc et al. [84], investigated agent support in a driving scenario and stated that "if human–machine interference is perceived as costly, without gain in terms of performance, the driver may unconsciously by-pass automation" [82]. The relationship between agent performance and trust has been investigated by manipulating agent behaviours in several ways. For instance, while Correia et al. experimented with an agent that suddenly stop working in a collaborative game assignment [34], other studies experimented with false-alarms errors in a human-agent collaborative X-ray scanning task [125], or introduced systematic biases in an agent's decision-making capabilities during a collaborative control-and-command scenario. [57]. Despite taking place in different environments with dissimilar populations, both the aforementioned studies discussed the idea of "trust aware" agents, where transparency goes both ways and where the user and agent calibrate their expectations iteratively, over time. However, little attention was brought to specific types of agent failure, and how they impact the human-agent team decision-making process.

5.2.2 Defining errors

Any flaw in the decision-making capabilities of an agent that results in loss of performance can be understood as an "error". The work of Salem et al. [150] found that the type of error an agent makes (breach of privacy, violations etc.) has an impact on the way users perceive the agent, and will affect the extent to which users are willing to rely on it in subsequent interactions. In the context of Human-Agent interactions, "error" is a broad term as agents can err in different ways and for different reasons. The work of Marinaccio et al. proposes four distinct types of errors: mistakes, lapses, slips and violations [118] (see Table 2.1) derived from human-human interactions studies [145] in which different types of errors are defined by the violation they represent. While these errors all stem from studies focusing on human-human interactions, they were also conceptualised in the context of human-agent interactions, notably in healthcare by Kim et al. [94]. The framework proposed by Marinaccio et al. [118] is an important step towards a better understanding of users' relationships with automation systems for the following reasons:

- It provides definitions and examples of known automation failures based on prior work in both human-human and human-agent settings.
- It defines automation error in a context-specific manner, which provides researchers with a template for defining other agent-related incidents in different domains.
- It lists potential effective trust "repair techniques" to prevent further damage caused by automation failures. Each repair technique is justified by prior work, such as the healthcare study by Kim et al. [94].

In the HAI literature, a number of studies have empirically tested the effect of automation errors on human participants in an effort to try and understand how to best define and prevent their failures. The work of Baker et al. [11] provides a comprehensive review of the potential future avenues in HRI and HAI, and makes use of the framework by Marinaccio et al. [118] to highlight the necessity of studying context-specific definitions of automation errors and effective repair mechanisms. In their article, Baker et al. make relevant recommendations for future research, including "Adapting existing trust research to investigate how robots features affect trust" [11, p. 20]. Following on from this recommendation, we believe that more research should look at the explicit impact of different kinds of agent errors on the human-agent relationship, and how users react and adapt to these errors, over time. Our work seeks to investigate this area by using a framework that allows us to manipulate agent errors and monitor changes in reliance and performance during human-agent collaborative tasks.

Given evidence from past work that defined and assessed different kinds of agent errors as well as methods to repair their damage [118, 145], we expect error types to have an influence on users' behaviour and perception of the agent despite comparable levels of agent reliability. More precisely, we are going to study human-agent partnership under the following working hypotheses:

1. **H1** Agent errors will always have a negative impact on task performance, no matter their type.

2. **H2** Agent errors that are harder to detect (mistakes) will have a greater impact on task performance and reliance than agent errors that are easier to anticipate (lapses).

The main contribution of our work lies in testing different types of agent errors and assessing their influence on human-agent collaboration in terms of behavioural changes and perception of the agents.

5.3 Method

As with all studies conducted in this thesis, we used the human-agent collaborative framework presented in Section 3 in its lab-based form, as described in Section 3.5.1 where we modified and created new agent behaviours and errors to answer our research questions. In this iteration of our framework, each participant interacted with 4 different agents (within-groups study) in 2 levels of difficulty, where each agent displayed one particular type of erroneous behaviour defined as slips, lapses, mistakes or none. More information about agent errors is available in Section 5.3.1).

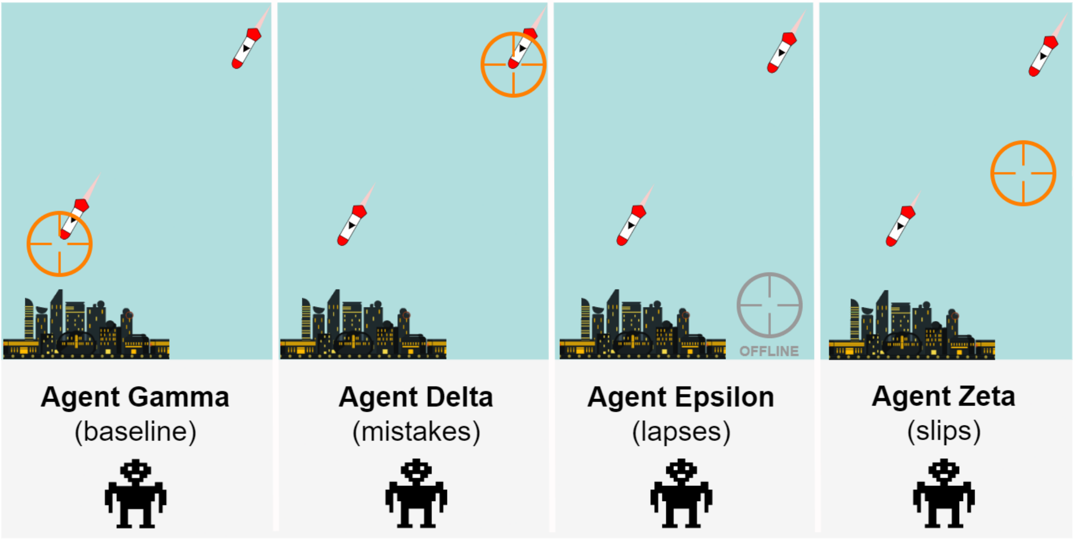


Figure 5.1: Abstract representation of the different errors each agent made. Mistakes are errors of priority, Lapses result in no response from the agent and Slips are experienced as an extreme loss of accuracy.

5.3.1 Agent Errors

The focus of this study is in testing the effect of agents displaying different behaviours and making different errors defined as "slips", "mistakes" and "lapses", all derived from the taxonomy of errors provided in Table 2.1, which is inspired by the work of Marinaccio et al. [118]. As further detailed in Section 5.2, we rely on the taxonomy of agent error by Marinaccio et al. which consists of 4 main error types: "Slips", "Lapses", "Mistakes" and "Violations". We decided against creating a "violation" type of error, as our task takes place in an explicit human-agent collaborative scenario, where the agent is assumed to be acting in the best interests of the user.

All agents created for this study were derived from a baseline agent named "Gamma" that did not display any specific errors, and achieved an average of 80% Recall on the overall task. The 20% loss in Recall was achieved by adding random biases in its aim, as presented in Section 3.3.2. For the rest of the error-prone agents presented below (agents Delta, Epsilon and Zeta), errors were triggered once every two rounds for a total duration of 30 seconds (half the duration of a level) so that they would be noticeable. To ensure consistency, all agents triggered their errors at the same time, for each participant. We verified that the types of error each agent displayed did not result in changes in reliability by ensuring that they did not significantly affect agents' Recall scores during simulations. During comparisons of Recall scores, an ANOVA yielded $p = 0.68$ when comparing the performance of all error-prone agents in the Easy difficulty, and $p = 0.15$ in the Hard difficulty levels, indicating that there was no significant difference between error types in terms of agent reliability. A simplified diagram of the different types of agent behaviour is available in Figure 5.1.

5.3.1.1 Mistakes - Agent Delta

To test the effect of mistakes, we designed agent "Delta". When the behaviour of this agent is triggered, the agent becomes incapable of focusing on one target at a time and instead "bounces" back and forth between available targets while never completely reaching them. This behaviour was created to simulate an explicit error of "planning", most commonly defined as "mistakes" [118, 145].

5.3.1.2 Lapses - Agent Epsilon

To test the effect of lapses; we designed agent "Epsilon". When the behaviour of this agent is triggered, the agent simply becomes unresponsive and stops working, showing no sign of activity at all. This behaviour simulates an error of "omission", commonly linked to "lapses" [118, 145].

5.3.1.3 Slips - Agent Zeta

To test the effect of slips; we designed agent "Zeta". When the behaviour of this agent is triggered, this agent becomes extremely inaccurate, barely capable of hitting any target as its aim would always be too far off. This behaviour simulates an error of "commission", commonly linked to "slips" [118, 145].

5.3.2 Task Difficulty

Each participant played with all agents (within-groups study) in sessions lasting for 60 seconds per level. Task difficulty in terms of the number of missiles to hit and their speed was fixed for the "Easy" and "Hard" levels across all sessions, with or without agents. In all of the sessions, the first levels (1 and 2) were set to have an "Easy" difficulty while levels 3 and 4 were set to have a "Hard" difficulty (with a total of 2 minutes per difficulty level). Here are details about the difficulty settings used in this study:

- In the "Easy" level, 2 missiles spawned every 6 seconds at a speed of 60 pixels per second for a total of 30 missiles.
- In the "Hard" level, 2 missiles spawned every 3 seconds with a speed of 120 pixels per second for a total of 60 missiles.

Each error-prone agent triggered their respective erroneous behaviour (slips, lapses or mistakes) during one of the "Easy" and "Hard" levels from T=30s to T=60s. In order to mitigate the learning effect, a William Square design [193] was used to control the order in which participants interacted with the agents.

5.3.3 Independent and dependent variables

In this section, we summarise the independent and dependent variables used in this study and as motivated by our experimental method and research questions.

Our independent variables are the following:

- *Task difficulty*, as defined by the amount and speed of missiles in each level.
- *Aiming agent reliability*, which remained high (80% accuracy) for each agent when no error was triggered.
- *Aiming agent error types*, which were all triggered at the same time and were defined as "slips", "mistakes" and "lapses".

Our dependent variables are the following:

- *Task Performance*, in terms of missiles hit, shots fired and missiles missed.
- *Reliance*, expressed by the duration for which participants relied on the aiming agent's help.
- *Trust*, as reported by participants.
- *Cognitive Workload*, as reported by participants.

5.3.4 Experimental Procedure

The study was approved by the University of Strathclyde Computer and Information Sciences Departmental Ethics Committee (Ethics Application No. 1029). The experiment took place in a lab, located in the University of Strathclyde's Computer and Information Sciences Department. We used a lab-based 4 (types of agent errors - no specific error (baseline 80% accuracy), slips, mistakes and lapses) x 2 (difficulty levels - easy or hard) within-subjects design including 2 baseline conditions: (1) no agent (individual performance), and (2) agent with constant performance (upper bound). The general experimental procedure is outlined in Section 3 and in Figure 3.9. At the end of the study, participants received a £10 shopping voucher. After consenting to take part in the study, participants went through the following steps:

1. Pre-hoc surveys and demographic questionnaire. (5 minutes) (see Appendix B).
2. Tutorial with and without an agent. The agent purposely stopped working so participants could understand that they might need to watch out for errors and take control when required (2 minutes).
3. One session without any agent (4 minutes).
4. One session with each agent (4 minutes each).
5. Post-hoc surveys (5 minutes) (see Appendix B).

After each level, participants were presented with a subset from the Checklist for Trust between People and Automation questionnaire [90] which contained questions pertaining to trust, dependability, reliability, deceptiveness, wariness, confidence and security. In addition, after each session, participants had to complete NASA TLX rating scales [76]. More details about each survey instrument can be found in Section 3.4.2. Prior to the experiment, participants were told that each agent had different behaviours, but no further detail was provided in order to avoid biases. The information sheet can be consulted in Appendix B.

5.3.5 Demographics

24 participants (13 M, 11 F) took part in our lab-based study, with ages ranging from 18 to 40 years old ($M = 26 \pm 5.01$). Most participants (13) had completed a bachelor's degree, while 6 had obtained a Master's degree and the remaining 5 were high school graduate or equivalent. When asked about how often participants played video-games on a 6 points Likert scale from 0 (never) to 6 (Very Frequently), participants scored on average 4 (± 1.82, indicating that they played "occasionally". The pre and post-hoc survey instruments presented to participants on the Qualtrics platform are available in Appendix B.

5.4 Results

In this section, we present results regarding task performance, users' reliance on agents, reported trust in the agents and cognitive workload. We used the overall scores participants obtained at the end of each session, across all levels of difficulty. More details on the inclusion of different difficulty levels are available in Section 3.2.4. The statistical methods we used to compare and report results is detailed in Section 3.4.4.

Table 5.1: Metrics related to task performance and reliance on the agents. Higher Recall, Precision and F1 scores indicate a better performance while higher User Control Time indicate less reliance on the agent. Highest values are highlighted in bold for each element.

	No Agent	**Gamma (baseline)**	**Delta Mistakes**	**Epsilon Lapses**	**Zeta Slips**
Recall	0.79 ± 0.02	**0.89 ± 0.01**	0.85 ± 0.01	0.87 ± 0.01	0.86 ± 0.01
Precision	0.63 ± 0.01	0.60 ± 0.01	0.60 ± 0.01	**0.64 ± 0.01**	0.61 ± 0.01
F1	0.69 ± 0.02	0.71 ± 0.01	0.70 ± 0.01	**0.73 ± 0.01**	0.71 ± 0.01
User Control Time	**24.88 ± 0.61**	10.40 ± 1.04	15.25 ± 0.91	13.15 ± 0.78	13.44 ± 0.85

5.4.1 Performance

Figures 5.2, 5.3, 5.4 and Table 5.1 present summary statistics for each session, where Recall, Precision and F1 scores were used to assess the performance of the participants and agents during each session as described in Section 3.4.3.1. Looking at Figure 5.2 and Table 5.1, we can see that participants performed better while interacting with agents than when completing the task themselves, a fact made especially apparent with Recall scores. In addition, Recall scores are noticeably higher for sessions with an agent not displaying any particular type of error, but these differences are not significant. A Friedman test yielded significant results for Recall ($p < 0.0001$,$W = 0.34$) scores, but further comparisons using Wilcoxon signed-rank tests did not yield any significant results. When performing a repeated measure ANOVA test on F1 scores, significant results were found ($p < 0.0001$, $F = 3.917$, $np^2 = 0.08$). Subsequent pairwise T-test comparisons indicated that F1 scores with agent Gamma (mistakes) were significantly lower than Epsilon (lapses) with $p = 0.0479$, $T = 2.08$, $CLES = 0.63$, while F1 scores with agent Delta (mistakes) were also found to be significantly lower than with agent Epsilon (lapses) with $p = 0.044$, $T = -2.12$, $CLES = 0.37$. Tests on Precision scores are not included as they did not yield important information that tests on Recall or F1 scores do not already provide.

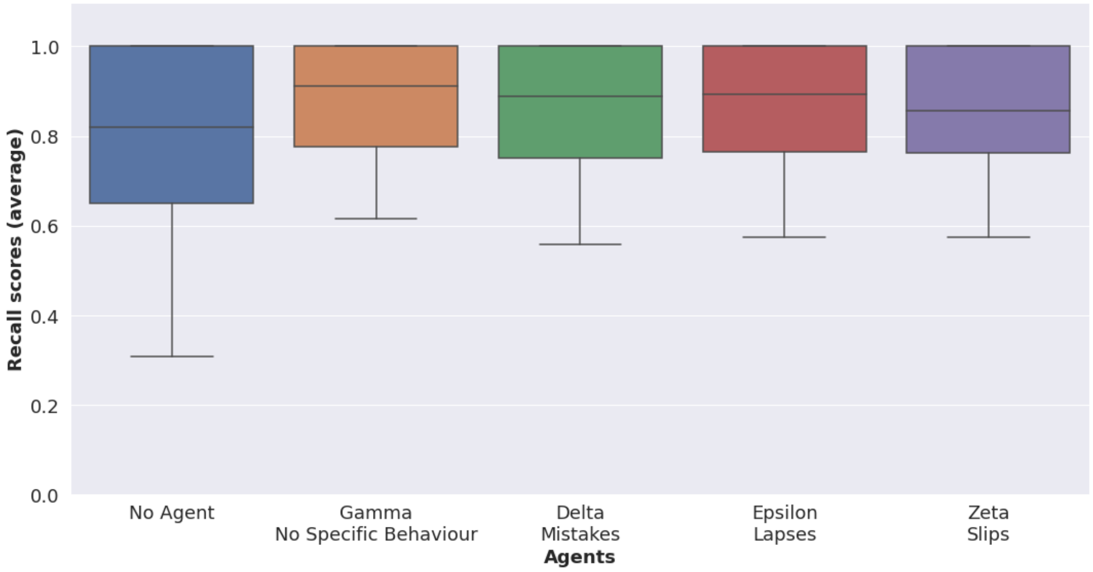


Figure 5.2: Recall scores for each session with or without agents. A higher score indicates better performance. Overall, participants benefited from the addition of an agent, regardless of its error type.

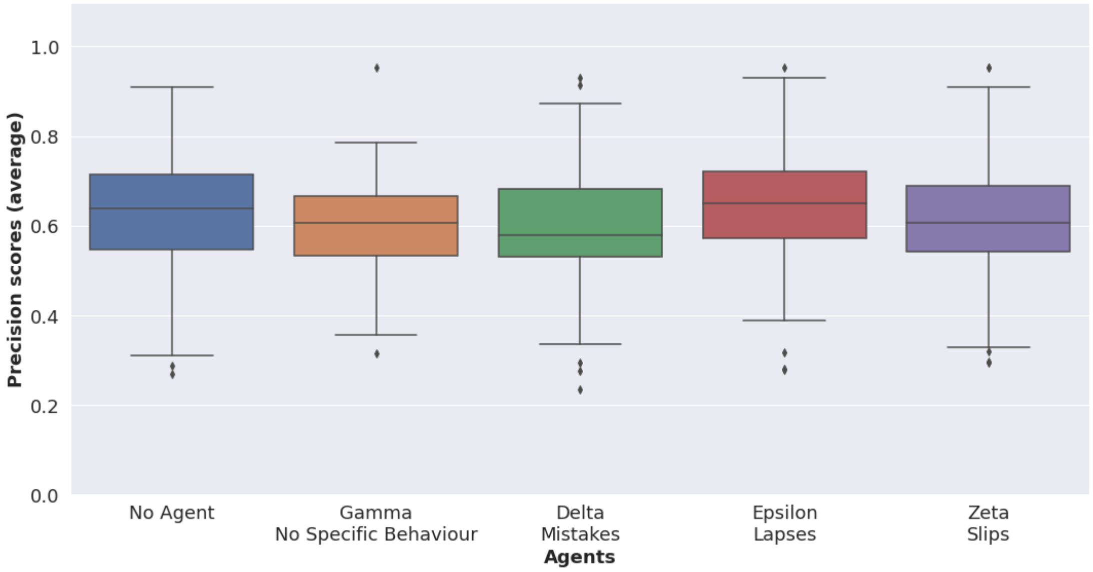


Figure 5.3: Precision scores for each session with or without agents. A higher score indicates better performance. Overall, the presence or absence of an agent, regardless of its error type, did not affect participants' Precision scores in any important way.

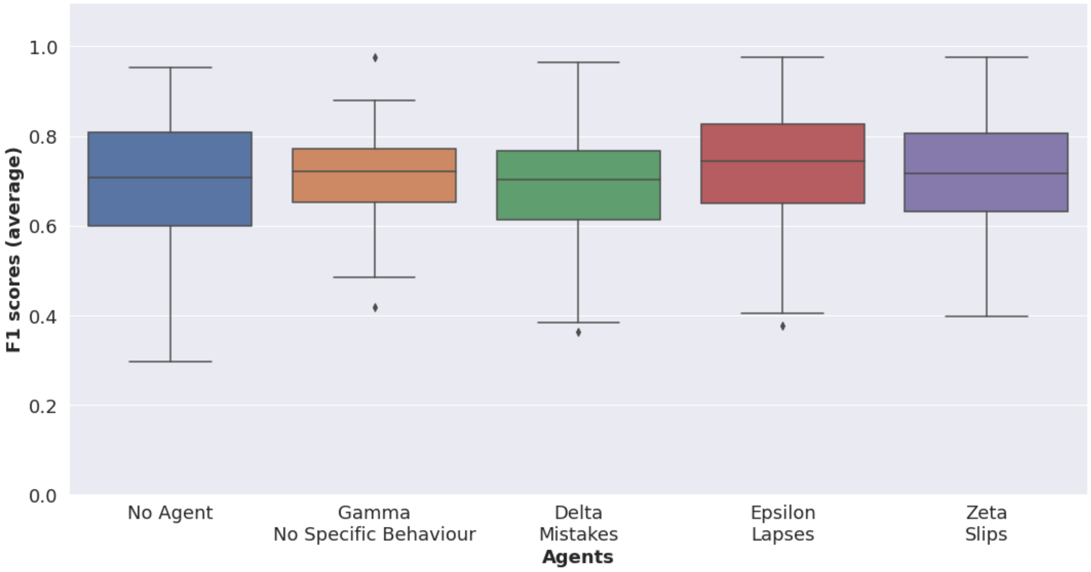


Figure 5.4: F1 scores for each session with or without agents. A higher score indicates a better performance. Participants interacting with agent Epsilon (lapses) managed to achieve higher performance than in any other condition.

5.4.2 Reliance

Participant control times were compared, that is, the amount of time participants chose to control the crosshair themselves instead of leaving the agents in charge of aiming, correcting the agents' inputs. The longer the participant control time, the lower the reliance on the agents. From consulting Figure 5.5 and Table 5.1, we notice that participants, unsurprisingly, controlled the crosshair much less when interacting with agents than without, no matter whether the agent displayed errors or not. Furthermore, agent Gamma (no error) was corrected less than any other erroneous agent, with agent Delta (mistake) being corrected more than agent Epsilon (lapses) and agent Zeta (slips). When analysing User Control Times, a repeated-measure ANOVA test yielded significant results ($p < 0.0001$, $F = 61.63$, $np^2 = 0.74$). Follow-up pairwise T-tests indicated that participants relied on agent Delta (mistakes) significantly less than they did on agent Gamma (no error type) with $p < 0.0001$,$T = 5.52$,$CLES = 0.71$, Epsilon (lapses) with $p =$,$U =$,$CLES =$ or Zeta (slips) with $p < 0.0001$,$T = 2.22$,$CLES = 0.61$. In addition, participants relied on agent Zeta (slips) significantly less than they did on agent Gamma (no error type) with $p < 0.0001$,$T = -4.32$,$CLES = 0.33$.

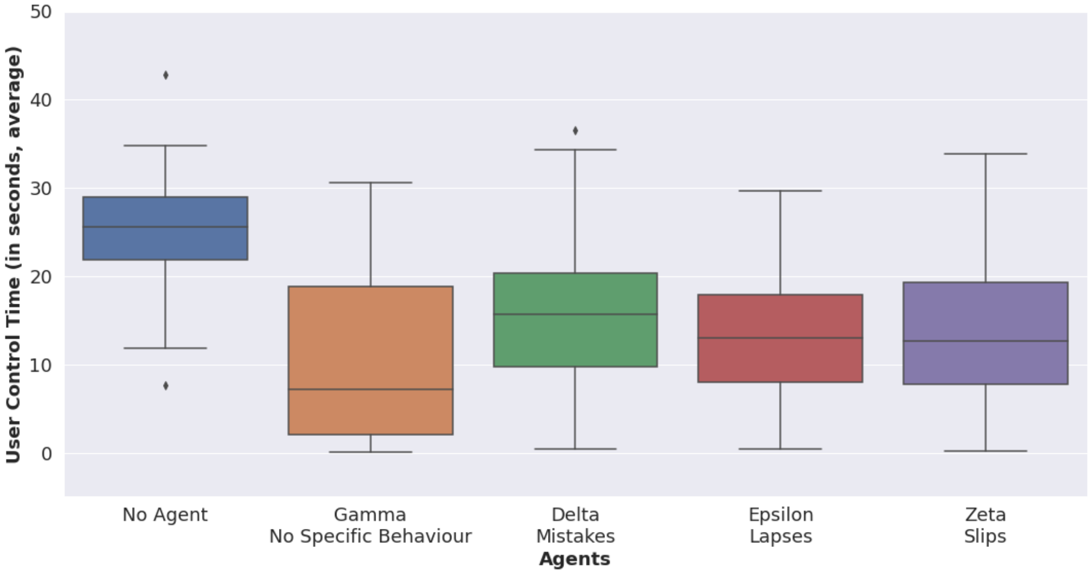


Figure 5.5: User Control Time for each session with or without agents. A higher score indicates lower reliance on the agent. As expected, the agent without any error type (Gamma) induced more reliance on its inputs than any other agent.

Table 5.2: Ratings given to statement related to cognitive load and reported trust in agents. Higher values indicate greater agreement with the statement or higher reported rating for questions. Highest values are highlighted in bold.

Question	**No Agent**	**Gamma (baseline)**	**Delta Mistakes**	**Epsilon Lapses**	**Zeta Slips**
I can trust the agent	n/a	**64.15 ± 2.35**	44.65 ± 2.72	44.19 ± 2.69	46.67 ± 2.47
The agent is deceptive	n/a	44.96 ± 4.58	48.50 ± 5.74	**58.96 ± 5.06**	47.46 ± 4.17
I am wary of the agent	n/a	53.42 ± 5.03	59.04 ± 5.58	**61.17 ± 5.54**	60.88 ± 4.90
I am confident in the agent	n/a	**52.92 ± 4.67**	34.62 ± 4.36	32.96 ± 3.06	37.54 ± 4.02
The agent provides security	n/a	**55.71 ± 4.31**	40.38 ± 5.30	36.50 ± 4.35	42.21 ± 5.16
The agent is dependable	n/a	**61.29 ± 3.13**	40.27 ± 3.51	40.35 ± 3.21	44.96 ± 3.43
The agent is reliable	n/a	**63.00 ± 3.26**	40.33 ± 3.49	44.06 ± 3.46	43.79 ± 3.41
How mentally demanding was the task?	**62.90 ± 4.70**	54.96 ± 5.20	57.34 ± 6.12	52.98 ± 5.83	51.19 ± 5.02
How physically demanding was the task?	**45.83 ± 6.12**	41.27 ± 6.36	47.42 ± 6.37	44.05 ± 6.21	41.87 ± 6.31
How hurried or rushed was the task?	**74.60 ± 4.81**	55.16 ± 5.13	61.90 ± 5.59	56.55 ± 5.47	62.30 ± 5.18
How successful were you in accomplishing your level of performance?	64.29 ± 4.35	58.93 ± 3.84	**71.63 ± 4.15**	65.08 ± 4.05	63.10 ± 4.22
How hard did you have to work to accomplish your level of performance?	**71.83 ± 3.76**	56.55 ± 4.66	61.71 ± 4.77	59.72 ± 4.73	61.90 ± 4.05
How insecure, discouraged, irritated stressed, and annoyed were you?	48.81 ± 5.36	50.00 ± 5.02	**58.93 ± 6.05**	51.79 ± 6.10	55.16 ± 5.28
Overall RAW TLX score	**61.38 ± 3.54**	52.81 ± 3.72	59.82 ± 4.54	55.03 ± 4.40	55.92 ± 3.97

5.4.3 Trust

At the end of each session or level, participants had to rate the agents based on their perceived trustworthiness, dependability, whether they provided security or if participants perceived the agent as being deceptive. From looking at Figure 5.6 and Table 5.2 and the ratings given to the statement “I can trust the agent”, we can clearly see that participants, on average, perceived agent Gamma (no error type) as more trustworthy than any of the other error-prone agents. In addition, agent Zeta was rated as the most trustworthy of all error-prone agents. When analysing the ratings given to the statement “I can trust the agent”, a Friedman test yielded significant results ($p < 0.0001, W = 0.30$). Follow-up non parametric pairwise T-tests indicate that participants perceived agent Gamma as significantly more trustworthy than any other agents with $p < 0.0001$ yielded during each test. Table 5.3 presents correlations between reported trust ratings and various other behavioural factors linked to reliance, task performance or reported cognitive load. Overall, correlation scores are weak, with Recall scores (measure of performance in terms of target hit) having the highest ρ value with 0.36 points. In close second is Cognitive Workload, with a correlation to Raw TLX scores of -0.32 points.

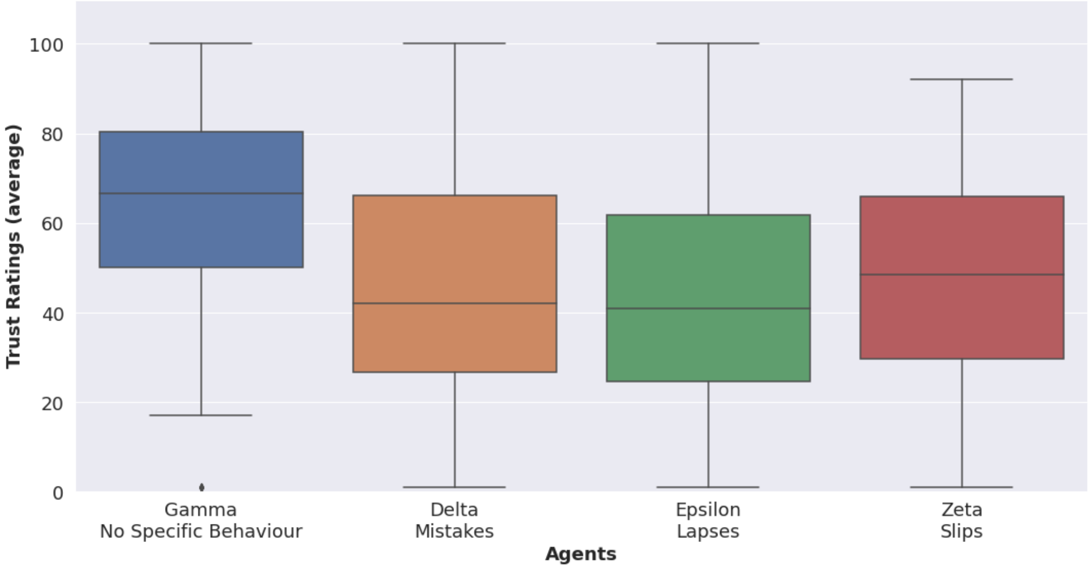


Figure 5.6: Participant ratings for the statement “I can trust the agent” for each session with agents. A higher score indicates greater trust in the agent. Participants trusted the agent without any error type (Gamma) the most, followed by the agent with the “slips” error type (Zeta).

5.4.4 Cognitive Load

At the end of each session, participants had to fill in a Nasa TLX workload questionnaire. Raw NASA TLX scores are used in this study as they were proven to be as effective as their weighted counter-part [75]. From consulting Figure 5.7 and Table 5.2, we can observe that, on average, interacting with an agent was perceived to be less cognitively taxing than without any agent at all, no matter the behaviour displayed by the agent. When comparing the agents displaying erroneous behaviours, collaborating with agent Epsilon (lapses) seemed to result in a comparatively low cognitive load, while collaborating with agent Gamma (no error) resulted in the lowest cognitive load of all. When comparing Raw TLX scores, a repeated ANOVA test yielded significant results with $p = 0.01$, $F = 3.52$, $np^2 = 0.13$. Follow-up pairwise

T-tests found that participants reported a significantly lower cognitive load when interacting with agent Gamma (no error) than with agent Delta (mistakes) with $p = 0.0194$, $T = 2.51$, $CLES = 0.59$.

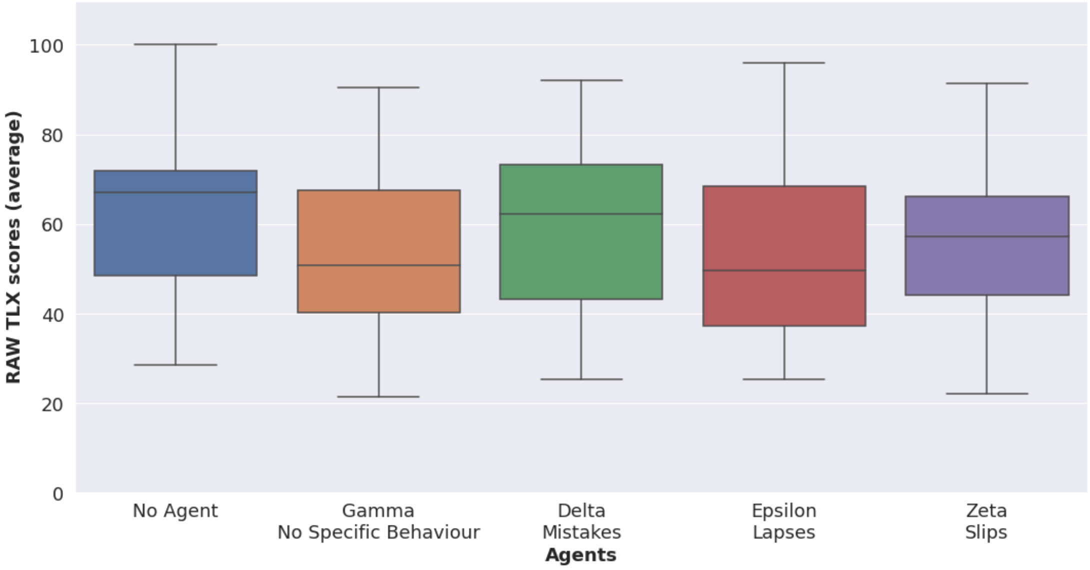


Figure 5.7: Raw TLX score for each session. A higher score indicates a greater reported cognitive load. Among the error-prone agents, the agent with the "mistakes" error type led to the highest reported cognitive load.

Table 5.3: Spearman's correlation tests between behavioural or reported metrics and trust ratings. A higher ρ scores indicates a greater correlation.

Parameter 1	**Parameter 2**	ρ	**p-value**
Recall	Trust ratings	0.3638	<0.001
RAW TLX	Trust ratings	-0.3247	0.0012
F1	Trust ratings	0.2715	<0.001
Task Difficulty	Trust ratings	-0.2604	<0.001
User Control Time	Trust ratings	-0.2515	<0.001
Precision	Trust ratings	0.1758	<0.001
Age	Trust ratings	-0.1514	0.0029
Gender	Trust ratings	-0.0361	0.4808

5.4.5 Participants' Feedback

At the end of each session, in accordance with the Critical Incident Technique [62], participants were asked to write about the positive and negative aspects of each agent, and what improvement(s) they would suggest. We then performed qualitative coding to understand how agent errors were perceived by participants. More details about coding analysis are available in Section 3.4.5. After being presented with definitions of "lapses", "mistakes" and "slips", three PhD students in Strathclyde Computer Sciences department were recruited and given the task of coding the "positive", "negative" and "improvement" feedback given by each participant regarding agents. Internal agreement scores (Kappa scores [183]) were used to interpret participants' feedback in relation to agent behaviour. The Kappa score obtained for coding feedback related to "mistakes" was 0.5243 (perceived as "Moderate" [183]), with an internal agreement of 78.73%. The agent most frequently associated with mistakes was, correctly so, Delta, with the most common occurrences being "target(s)" (91 occurrences), "prioritise" (24 occurrences), "focus" (23 occurrences), "confused" (21 occurrences) or "struggle" (14 occurrences). The Kappa score obtained for "lapses" was 0.5493 (perceived as "Moderate"), with an internal agreement of 91.84%. The agent most frequently associated with lapses was, correctly so, Epsilon with the most common occurrences being "sometimes stops" (37 occurrences), "not working" (17 occurrences) or "stopped" (10 occurrences). The Kappa score obtained for "mistakes" was 0.2109 (perceived as "fair"), with an internal agreement of 77.34%. The agent most frequently associated with slips was, correctly so, Zeta, with the most common occurrences being "aim" (23 occurrences), "accurate" (21 occurrences) or "accuracy" (18 occurrences). Overall, these results indicate that participants clearly and correctly recognised the "lapses" and "mistakes" error types while "slips" was harder to differentiate from the rest.

5.5 Discussion

In this study, we examined the effect of different erroneous agent behaviours designed using past work in human-human and human-automation research [118, 145]. This study sought to answer our second Research Question presented in Section 1.3: **How do different types of agent errors defined from previous related work such as slips, mistakes and lapses affect the human-agent relationship?**. We sought to answer the sub-research questions presented in Section 5.1, namely **how different types of agent errors influence**:

- **RQ2.a:** the users' task performance when interacting with agents?
- **RQ2.b:** the users' reliance on the agent?
- **RQ2.c:** the users' cognitive workload when interacting with the agent?
- **RQ2.d:** the users' reported trust in the agent?

Our analysis of the results showed that, at the same high level of agent reliability, different types of errors (namely: "slips", "lapses" and "mistakes") affected users' reliance on the agents, as well as task performance, cognitive workload and trust in different ways. From looking at the results presented in Section 5.4, it is clear that participants preferred to interact with agent Gamma (no error), as Gamma was deemed more trustworthy than any other agent in terms of reported trust (which informs RQ2.d) while resulting in a lower cognitive workload than any of the error-prone agents. These results were to be expected, as the agent maintained a constant and predictable level of performance throughout the whole study. Error-prone agents, however, influenced the human-agent relationship in more varied ways.

5.5.1 Lapses, Errors of omission

We expected that agent errors that were harder to detect (mistakes) would have a greater impact on task performance and reliance than agent errors were easier to anticipate (lapses). This hypothesis was correct, as lapses led to most changes in reliance and performance compared to the slips and mistakes error types. When looking at the results, it is apparent that participants' interactions with the agent displaying *lapses* (error of omissions) led to the biggest changes in the human-agent relationship. While agent Gamma (no error) led participants to achieve the best task Recall scores (see Figure 5.2), participants interacting with agent Epsilon (lapses) managed to achieve higher Precision (see Figure 5.3) and F1 scores (see Figure 5.4) on average, indicating that they performed with higher accuracy and needed fewer attempt to hit targets when interacting with an agent making errors of omission (which informs RQ2.a). In addition, we also observed that participants corrected agent Epsilon (lapses) less frequently than Gamma (no error) in terms of individual user corrections (see Table 5.1), thus relying more on agent Epsilon (which informs RQ2.b). However, on average, participants corrected agent Epsilon (lapses) for significantly longer periods of time than they did agent Gamma (see Figure 5.5) (which informs RQ2.b). We hypothesise that, in order to match Gamma's performance (no error), participants had to involve themselves more in the aiming process. This increased involvement resulted in fewer corrections (see Table 5.1) than with the baseline agent Gamma (no error), but for longer periods of time (see Figure 5.5), as the errors caused by Epsilon (lapses) were easier to spot and, in turn, fix. As a result, participants actually managed to perform better with Epsilon than with Gamma (no error). However, while doing so, participants reported a significantly higher cognitive workload (see Figure 5.7) due to the higher number of actions they were obliged to carry out. Participants' overall perception of Epsilon (lapses) was also significantly more negative in all of the reported measures (see Section 5.4.5), which informs RQ2.c. Agent Epsilon (lapses) was still perceived as being helpful, as participants reported significantly lower cognitive workload when interacting with it, compared to sessions without any agent. This result could indicate that participants were more tolerant of an agent committing lapses, as this type of error seems to be easier to notice. When looking at participants' feedback for each of the agents, we can also observe that the agent most frequently associated with "lapses" is agent Epsilon (lapses), with many references to its tendency to "stop" or to suddenly "stop working".

5.5.2 Slips and Mistakes, Error of commission and intention

We expected agent errors to have a negative impact on task performance, no matter their type. This hypothesis was proved incorrect for the "slips" type of agent error. Among the other error-prone agents, noticeable differences were found when studying participants' reactions to agent Delta (mistakes) and agent Zeta (slips). In terms of performance, Delta's F1 score (see

Figure 5.4), the lowest among all the agents, is significantly lower than Epsilon's F1 score (which address RQ2.a). This decrease in performance goes with a decrease in reliance, as participants corrected agent Delta (mistakes) for significantly longer periods of time (see Figure 5.5) than any of the other agents; error-prone or not (which informs RQ2.b). These findings indicate that an agent making mistakes is harder for participants to correct than an agent having lapses. Interactions with agent Delta resulted in the worst performance and reliance scores, which in turn led to higher cognitive workloads (see Figure 5.7, the highest among all agents and sessions), with a statistically significant difference when compared to the cognitive workload associated with agent Gamma (no error) (which informs RQ2.c).

5.5.3 Perception of the agents

From the qualitative coding (see Section 5.4.5), most of the negative descriptions associated with agent Delta were adequately coded as resulting from "mistakes", with recurring terms such as "confused", "targets" or "prioritise", highlighting how indecisive the aim of the agents was perceived to be. These observations suggest that participants were well aware of how and when Delta was making mistakes, which led them to rely on it less, and to rate it as less dependable and less reliable than any of the other error-prone agents. The scores obtained by participants interacting with agent Zeta (slips) seem to place it in the middle of the other error-prone agents in terms of performance (F1 scores, see Figure 5.4) and participant reliance (user control time, see Figure 5.5). Nonetheless, the number of participant corrections for agent Zeta (slips) was both the highest among all the agents and significantly higher than for Epsilon (lapses). Participant control time was also found to be significantly lower than for Delta (mistakes) and higher than for Gamma (no error).

5.6 Conclusion

In this study, we explored the impact of different types of agent errors and behaviours on users in a real-time collaborative scenario. To do so, we designed different kinds of agent errors based on a taxonomy from the work of Marinaccio et al. [118], namely: "lapses" (errors of omission), "slips" (errors of commission) and "mistakes" (errors of prioritisation). Our findings were aimed at answering our second Research Question: "How do different types of agent errors defined from previous related work such as slips, mistakes and lapses affect the human-agent relationship?"

While all error types adversely affected the way users interacted with agents, each had a different impact on the dependant variables we examined. For instance, "lapses" made participants more alert and even led them to perform better than the baseline agent, while "slips" were perceived as harder to correct, and "mistakes" had the overall worst impact on both perception of the agent and task performance. These results indicate that participants felt the need to correct the agent making slips (Zeta) more than any other agent, however these corrections were shorter than those for the agent making mistakes (Delta). Overall, the increase in user corrections for Zeta (slips) came with an increase in cognitive workload, as Zeta was found to be the most cognitively taxing agent. Nonetheless, the slips displayed by Zeta were still perceived as less cognitively taxing than the mistakes displayed by Delta. When looking at participants' feedback, agent Zeta was most often coded as displaying "slips", albeit with the lowest Kappa score (0.21) of all the other agent errors. Participants still perceived differences in the way agent Zeta behaved, mainly mentioning issues in the "accuracy" displayed by the agent. In order to make up for Zeta's accuracy, participants had to constantly adjust the aim themselves. However, out of all the error-prone agents, Zeta was rated as the most dependable, as providing the most security and as being the most trustworthy agent.

Our findings suggest that clearly defining the type of error made by an agent can be useful in anticipating the impact it can have on users. As a result, we posit that when designing collaborative agents likely to give imperfect input, it is best to avoid indecisiveness, and that, in most circumstances, a total lack of input is preferable to indecisive or inaccurate information.

Chapter 6

Visual Uncertainty

6.1 Motivation

In this chapter, we focus on the environment of interaction and its impact on human-agent collaboration. More specifically, we investigate whether limiting the amount of visual information required to successfully complete a task can have an effect on the human-agent partnership. As we have seen in Chapters 4 and 5, the behaviour and reliability of an agent largely affect how users are willing to interact with the system. However, the environment of interaction also plays an important role on the effectiveness of the human-agent collaboration, as restricted access to information can directly affect the outcome of a task, which may be disruptive to the human-agent partnership as well.

Previous research on the topic of trustworthy automation has often focused on the design of the agents themselves, by studying the effect of their reliability on users [57] or by adding more information about their actions and reasoning [93, 124], as an attempt to help mitigate adoption issues. In a study about human-agent teamwork, Van et al. [182] suggest that task complexity changes users' attitudes towards agents in terms of trust and fairness of judgement. As we have seen in Section 2.1, many elements influence how users perceive an agent. While most studies have researched how agent features influence the perceived trustworthiness of a system and how to reduce the uncertainty inherent in interacting with an agent [174,198], fewer works have researched how uncertainty stemming from the *environment of interaction* affects the development of trust, over time.

This study seeks to answer our third Research Question presented in Section 1.3: **How do different types of environmental conditions (static or moving), which impair vision and induce uncertainty, affect the human-agent relationship?** Specifically, we sought to answer the following sub-research questions: how Visual Environmental Uncertainty that restricts, occludes or hides visual information from the user influences:

- **RQ3.a:** how well the human-agent team performs on a task.
- **RQ3.b:** how much the user trusts and relies on the agent.

- **RQ3.c:** situational awareness in relation to the users' trust in the agent.

The work presented in this chapter is an extension of the paper entitled *Investigating the Impact of Visual Environmental Uncertainty on Human-Agent Teaming* [38].

Using the collaborative framework detailed in Chapter 3, we conducted an online-based 4 (types of visual uncertainty) x 2 (levels of difficulty) mixed-design experiment with 96 participants. In this study, four different types of visual environmental uncertainty covering the screen and hiding information (within-subjects) were used as conditions, between two different levels of task complexity (between-subjects). Participants interacted with the same agent throughout the experiment to complete a goal-oriented task in our interactive human-agent framework. Our findings indicate that human-agent collaboration can be greatly influenced by different types of visual uncertainty. For instance, participants trusted the agent more and relied on it more when visual uncertainty was at its highest, despite a loss in task performance. On the contrary, some visual uncertainty that reduced the amount of time participants had to react actually increased task performance.

6.2 Related Work

In HAI, uncertainty is often studied in terms of the transparency of an agent's actions [198], but uncertainty can also arise from the environment in which the human-agent interaction takes place. As previously discussed in Section 2.4.2, *uncertainty* refers to a lack of certainty about a piece of information or a result. As evidenced by a study on trust under visual clutter by Sacha et al. [149], while a significant amount of work has gone into evaluating different techniques to manage visual uncertainty [160], very few studies have focused on how visual impairments, in terms of visual clutter or occlusions, affect users' attitude towards automated systems [101].

To understand and define visual uncertainty, it is useful to look at other disciplines. For instance, past work in Geographical studies has thoroughly defined and researched artefacts that impede users' vision, such as "blurry surfaces" or "dark patterns", as defined by MachEachren et al. [116]. All these patterns, lead, in turn, to the creation of visual occlusions. These occlusions are composed of a shape and a position the combination of which can, in turn, create "occlusions" as defined by VanLier et al. [180]. The resulting occlusions create a "virtual structure" when two or more shapes merge together [180], effectively preventing vision in a particular part of the visual space. Occlusions and the overall masking of visual information can have different effects on people. For instance, an occlusion that restricts peripheral awareness in such a way that attention is only focused on a specific area is called *tunnel vision* [79, 115, 172]. Tunnel vision, as explained in the work of Ma et al. [115], reduces information input and increases the amount of effort required to gain information, as people affected by tunnel vision have to scan the visual space while being impaired by a narrower field of view. Tunnel vision is a phenomenon that can either be artificially produced by occluding the field of view, as seen in a study by Ma et al. [115] or Mackworth et al. [115, 117], or be the result of deteriorating cognitive or physical capabilities [194].

HAI can occur in many different scenarios of various task complexity where information required to perform the task successfully can be unavailable or partially missing. Such settings often consist of search-and-rescue or monitoring tasks, often involving multiple agents such as UAV survey assignments [4, 29]. In these contexts, visual uncertainty can take many forms, whether it stems from a lack of information or the obstruction of visual information. The concept of Situational Awareness (SA), as described in Section 2.6.3, is useful in evaluating how well users can perceive information to inform their decisions. Situational Awareness is used to evaluate how much information users understand, and how well they can use this information to anticipate future outcomes [49] and carry out their task. Ideally, an effective collaborative agent should increase reported SA by increasing users' task knowledge while not increasing their cognitive workload [26]. In HAI, evaluating SA represents a good way of studying how visual uncertainty derived from occlusions can affect users and their attitude towards an agent.

As little work has investigated the impact of visual uncertainty in a human-agent collaborative task, we would like to address this gap by employing an interactive human-agent framework in which different types of visual uncertainty are present, and study the resulting impact on the human-agent relationship. Given the evidence of past research, it is likely that visual impairments will heavily impact on user performance. It is less clear, however, how the same impairments will influence trust and reliance on an agent. In this Chapter, we look at how human-agent collaboration evolves in the context of visual uncertainty with the following working hypotheses:

- **H1** According to past work in trust on automation and transparency, visual uncertainty will negatively affect users' trust in the agent.
- **H2** The types of visual uncertainty that hide the most amount of information will have the most significant impact on situational awareness and trust.
- **H3** Participants will rely and trust the agent more when a type of visual uncertainty forces them to react more quickly.

The main contribution of our work lies in testing the impact of different types of visual uncertainty on the human-agent relationship in a real-time scenario.

6.3 Method

This study was conducted using our interactive human-agent collaborative framework described in Section 3. As opposed to studies presented in Chapters 4 and 5, this study was conducted online. Modalities for online-based experiments are presented in Section 3.5.2.

6.3.1 Visual Uncertainty

We designed four types of visual uncertainty (one visual uncertainty type per condition). Every type of visual uncertainty obstructs visibility and visual information in a different way. Each impairment lasted for the same amount of time (half the duration of one level) and all were triggered at the same time, for each participant. Figure 6.1 presents an abstract representation of the visual uncertainty used in our study while Figure 6.2 showcases the actual implementation as seen in our framework. We named these visual occlusions as follows:

- **Top Disruption:** 50% of the upper part of the screen is hidden (see Figure 6.2c). This type of uncertainty is *static*, and was designed to place participants in a situation where waiting to see the missiles that had recently spawned would reduce the amount of time they had to react.
- **Bottom Disruption:** 50% of the lower part of the screen is hidden (see Figure 6.2d). This *static* type of uncertainty was designed to force participants into memorising and making predictions on the future position of missiles based on where they spawned and at which speed.
- **Moving Clouds:** the bottom and top of the screen are alternatively hidden in a continuous manner. Both of the clouds present in Figure 6.2a move across the sky from left to right, always hiding 25% of the top and 25% of the bottom of the screen (i.e. 50% in total). This type of uncertainty is more *dynamic* in nature, and users were able to achieve a higher level of situational awareness if they decided to wait in order to obtain better visibility.
- **Darkness:** the totality of the screen is obscured (see Figure 6.2b), and only a specific area around the cross-hair is made visible to give users a minimum amount of feedback on their actions. This type of *dynamic* uncertainty was designed to constrain the user's focus on one single spot on the screen, creating a "tunnel vision" effect. This condition was designed to test whether participants would completely rely on what the agent was showing them or would try to scan the area manually to increase their situational awareness.

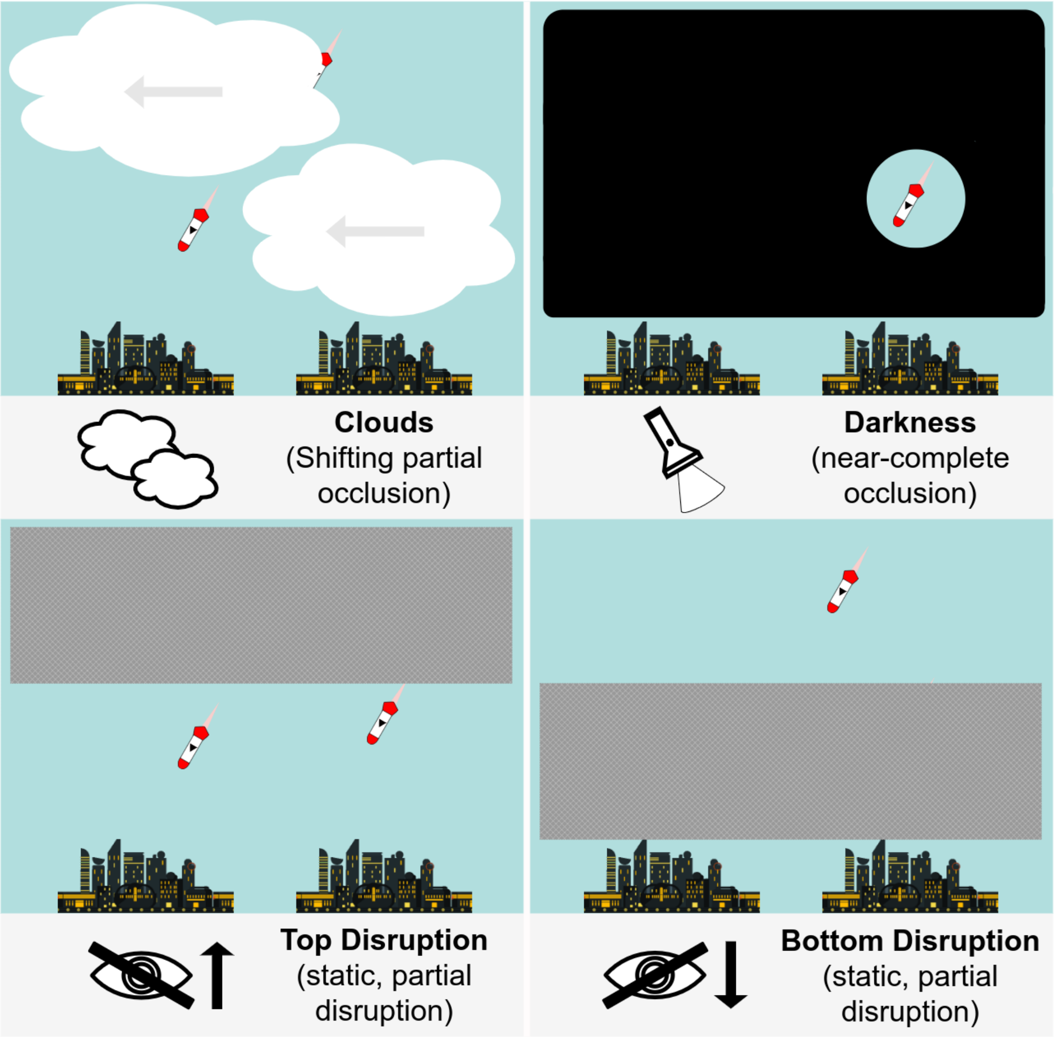
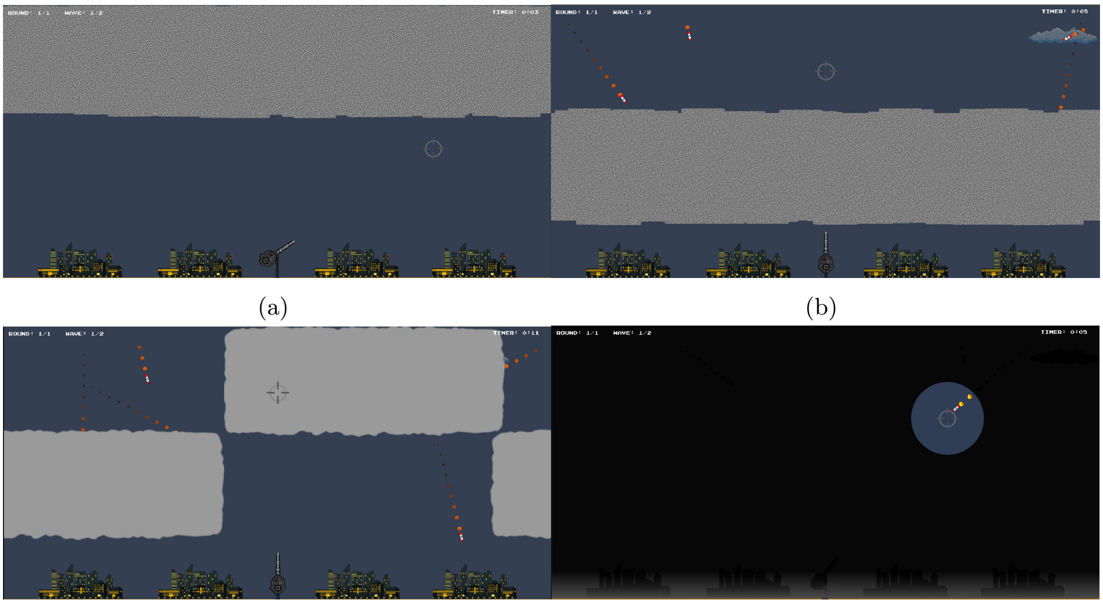


Figure 6.1: Abstract representation of the different types of visual obstruction used in this study. Each obstruction results in occlusions that are either near-complete or partial and static or fixed.



(a) (b)

(c) (d)

Figure 6.2: All of the different types of visual uncertainties used in this study. (a) "Top Disruption" and (b) "Bottom Disruption" occlude, respectively, the upper and bottom-most parts of the screen while (c) "Clouds" occludes the top and bottom alternatively and (d) "Darkness" obscures the totality of the screen except for the immediate proximity of the cross-hair, producing tunnel-vision.

6.3.2 Task Difficulty and Complexity

Two groups were recruited for this study. The first group experienced a “low” level of task complexity (fewer and slower missiles over both difficulty levels) while the second experienced a “high” level of task complexity (more and faster missiles in both difficulty levels). We chose to include these two levels of complexity as past studies have shown that situational awareness is affected by the number of elements that need to be monitored [49, 52]. Each participant played through all four visual uncertainty conditions in sessions lasting for 90 seconds per difficulty level (defined as “easy” and “high” - relative to the task complexity participants were recruited for). Similarly to previous studies, participants were able to adjust the crosshair themselves and were responsible for firing projectiles. Within each group (low or high complexity), we controlled the *difficulty* of the task based on 2 variables: (1) the speed of each missile and (2) the delay between the spawning of each missile. Here are more details about the difficulty settings used in this study for both task complexity groups:

1. Low task complexity group:

   (a) In the “Easy” levels, 1 missile spawned every 6 seconds at a speed of 60 pixels per second (15 missiles in total).

   (b) In the “Hard” difficulty levels, 1 missile spawned every 3 seconds with a speed of 140 pixels per second (22.5 missiles in total).

1. High task complexity group:

   (a) In the “Easy” levels, 3 missiles spawned every 5 seconds at a speed of either 30 or 50 pixels per second (random selection) for a total of 54 missiles.

   (b) In the “Hard” difficulty levels, 3 missiles spawned every 4 seconds with a speed of either 60 or 80 pixels per second (random selection) for a total of 67 missiles.

The number of missiles spawning at once as well as their speed was calibrated using pilot studies, in order to make sure that the task was perceived as “easier” or “harder” depending on the level of task complexity and difficulty. In addition, the performance of the agent was calibrated through test simulations where the agent played by itself to ensure that, no matter the level of complexity, it would always get an average Recall score of 0.7 over the “easy” and “hard” levels of difficulty. Each level (either difficulty) lasted for 90 seconds. This duration was set so that participants had enough time to interact with the agent in each condition, while ensuring that the entirety of the experiment could be completed in about 45 minutes, thus reducing participants’ fatigue and fostering higher attention levels throughout the online-based study.

6.3.3 Agent Performance

The exact same agent was used in all conditions where an agent was present, no matter the type of uncertainty, level of difficulty or complexity of the task. The agent's level of reliability was set to an average accuracy of 80%, similar to the agent used in Chapter 5.

6.3.4 Independent and dependent variables

In this section, we summarise the independent and dependent variables used in this study and as motivated by our experimental method and research questions.

Our independent variables are the following:

- *Task difficulty*, as defined by the amount and speed of missiles in each level.
- *Task Complexity*, which was mainly defined by the amount of missiles spawning at once during a level. We divided the recruitment of participants into two groups, one experiencing a lower and the other a higher level of task complexity.
- *Aiming agent reliability*, which remained high during all conditions.
- *Visual uncertainty*, which were occluding visual information and were divided into four different entities referred to as "clouds", "darkness", "top disruption" and "bottom disruption".

Our dependent variables are the following:

- *Task Performance*, in terms of missiles hit, shots fired and missile missed.
- *Reliance*, expressed by the duration for which participants relied on the aiming agent's help.
- *Trust*, as reported by participants.
- *Cognitive Workload*, as reported by participants.
- *Situational Awareness*, as inferred using participants' attempts to remember how many important elements (in this study, missiles) were present in the top and bottom half of the screen before it was occluded (see Section 3.4.2.3 for more explanations).

6.3.5 Experimental Procedure

This study was approved by the University of Strathclyde Computer and Information Sciences Departmental Ethics Committee (Ethics Application No. 1177). Recruitment was limited to people residing in the UK, where our experimental apparatus was hosted. Limitations on the minimum hardware required to take part in the study was put in place to ensure that participants experienced equivalent experimental conditions. Only data from participants that experienced an average frame-rate above 24 frames per second with a resolution of at least 720p were kept in our dataset for further analysis (more details available in Section 3.5.2). A one minute bench-marking test was provided before users consented to take part in the study. This test was designed in order to filter out participants that did not meet hardware requirements. Final recruitment was conducted online on the Prolific©web based experiment platform, where participants received £5.50 for undertaking the experiment (45 minutes in length). Once registered, each participant went thought the following steps:

1. Demographic and pre-hoc survey. (five minutes) (available in Appendix C).
2. Tutorial aimed at understanding the controls of the game and how to interact with the agent. (two minutes).
3. Session with an agent. (three minutes).
4. Session without an agent. (three minutes).
5. Four sessions, with each one having a unique type of environmental visual uncertainty (the order of which was randomised using a latin square design [15]) in which cognitive workload, trust and situational awareness surveys were presented. (three minutes each).
6. Post-hoc survey collecting participants' feedback. (five minutes) (available in Appendix C).

During the study, participants completed NASA TLX rating scales, which are 6-item survey instruments extensively used to measure cognitive workload [75]. In this study, RAW TLX [18] scores are reported. To measure trust in the agent, we used a single statement at the end of each round: "I can trust the agent" graded on a 7 point Likert scale from 1 (complete distrust in the agent) to 7 (total trust in the agent) adapted from the work of Jian et al. [90]. To measure Situational Awareness, we used the Situation Awareness Global Assessment Technique (SAGAT) [49], which involves freezing the task, hiding all elements present on the screen and asking participants different questions related to the location of items of interest. Due to the nature of the study, we chose to assess "Level 1" situational awareness only, which consists of knowing *where* to find information [53]. In our study, we asked participants how many missiles were present in the top- and bottom-most parts of the screen. More information about the survey instruments detailed here can be found in Section 3.4.2.

6.3.6 Demographics

96 participants (53M, 41F, 2 Non-Binary) were recruited for this study, with ages ranging from 21 to 31 years old ($M = 25 \pm 2.4$). In terms of education levels, most participants ($n = 44$) held at least a bachelor's degree. When asked how often they played video-games, most participants ($n = 29$) reported playing "Occasionally".

6.4 Results

In this Section, we present results regarding task performance, users' reliance on agents, reported trust in the agents, cognitive workload and situational awareness. We used the overall scores participants obtained at the end of each session, across all levels of difficulty and across the two level of task complexity. More details on the inclusion of different difficulty levels are available in Section 3.2.4, while more explanations regarding the inclusion of different task complexity levels can be found in Section 6.3.2. The statistical methods we used to compare and report results is detailed in Section 3.4.4.

Table 6.1: Metrics related to performance and reliance recorded during the task. Higher recall, precision and F1 scores indicate a better performance while a higher user control time indicates lower reliance on the agent. Highest values are highlighted in bold.

	No Agent	**Agent**	**Agent Clouds**	**Agent Darkness**	**Agent Top Dis.**	**Agent Bottom Dis.**
Recall	0.75 ± 0.02	0.77 ± 0.01	0.74 ± 0.01	0.72 ± 0.02	**0.79 ± 0.01**	0.72 ± 0.02
Precision	**0.65 ± 0.02**	0.57 ± 0.01	0.50 ± 0.01	0.56 ± 0.02	0.57 ± 0.02	0.50 ± 0.02
F1	**0.69 ± 0.02**	0.64 ± 0.01	0.58 ± 0.01	0.62 ± 0.02	0.66 ± 0.01	0.58 ± 0.02
User Ctrl Time	**34.47 ± 0.57**	24.01 ± 0.87	23.71 ± 0.95	22.51 ± 0.88	25.13 ± 0.95	22.34 ± 0.94

6.4.1 Performance

To calculate performance, we used Recall, Precision and F1 scores which are all computed using metrics related to the success of the task as described in section 3.4.3.1. Tables 6.1 and Figure 6.3 present an overview of the performance scores recorded in the study. As we can see from consulting Table 6.1 and Figure 6.3 participants performed better while interacting with an agent in terms of Recall scores and slightly worse in terms of F1 scores. When visual uncertainty was added, participants' performance dropped slightly, with the exception of the "Top Disruption" condition which actually saw an increase in performance.

A Mixed ANOVA test for Recall scores yielded significant results at all levels: complexity ($p < 0.0001$, $F = 7.11$, $np^2 = 0.07$), visual uncertainty type ($p < 0.0001$, $F = 9.07$, $np^2 = 0.09$) and interaction effect ($p < 0.0001$, $F = 1.34$, $np^2 = 0.01$). Follow-up pairwise T-tests showed that Recall scores during session with the "Top Disruption" were significantly higher than during sessions featuring other types of visual uncertainty: "Bottom Disruption" ($p < 0.0001$, $T = -6.37$, $CLES = 0.38$), "Clouds" ($p < 0.0001$, $T = -6.23$, $CLES = 0.39$) and "Darkness" ($p < 0.0001$, $T = -4.71$, $CLES = 0.41$).

A Mixed ANOVA test for F1 scores yielded significant results at all levels: complexity ($p < 0.0001$, $F = 0.77$, $np^2 = 0.15$), visual uncertainty type ($p < 0.0001$, $F = 20.42$, $np^2 = 0.78$) and interaction effect ($p < 0.0001$, $F = 2.66$, $np^2 = 0.10$). as with Recall scores, Follow-up pairwise T-tests indicate that participants yielded significantly higher Precision scores under "Top Disruption" than under any other type of visual uncertainty: "Bottom Disruption"

($p < 0.0001$, $T = -6.49$, $CLES = 0.37$), “Clouds” ($p < 0.0001$, $T = -6.65$, $CLES = 0.39$) and “Darkness” ($p < 0.0001$, $T = -2.33$, $CLES = 0.45$).

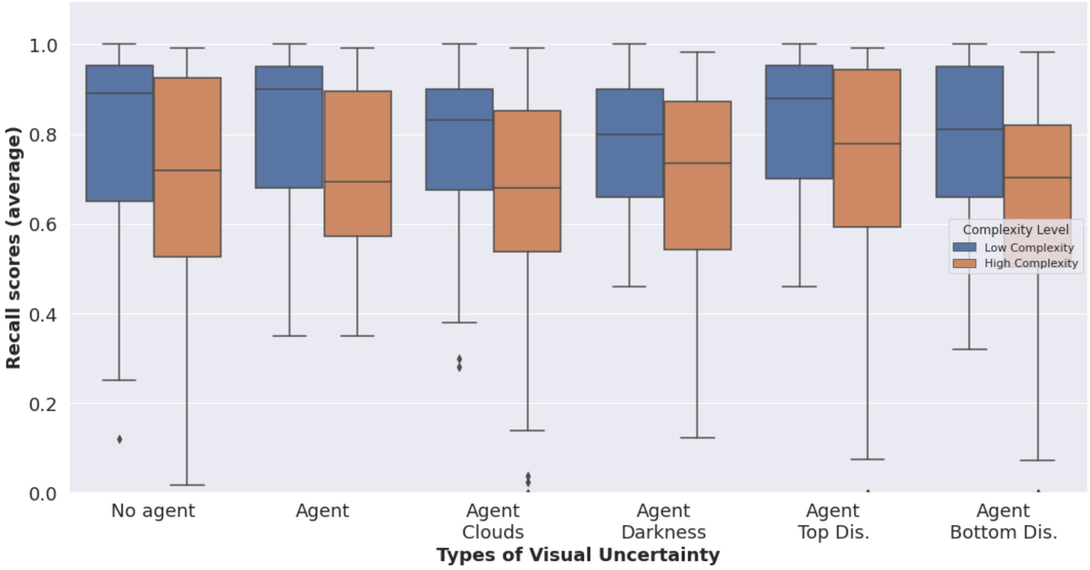


Figure 6.3: Recall scores for each session with or without agents. A higher score indicates better performance and a higher number of missiles hit. While a higher complexity level led to worst overall performance, participants in the “Top Disruption” condition performed better than in any other condition with visual uncertainty.

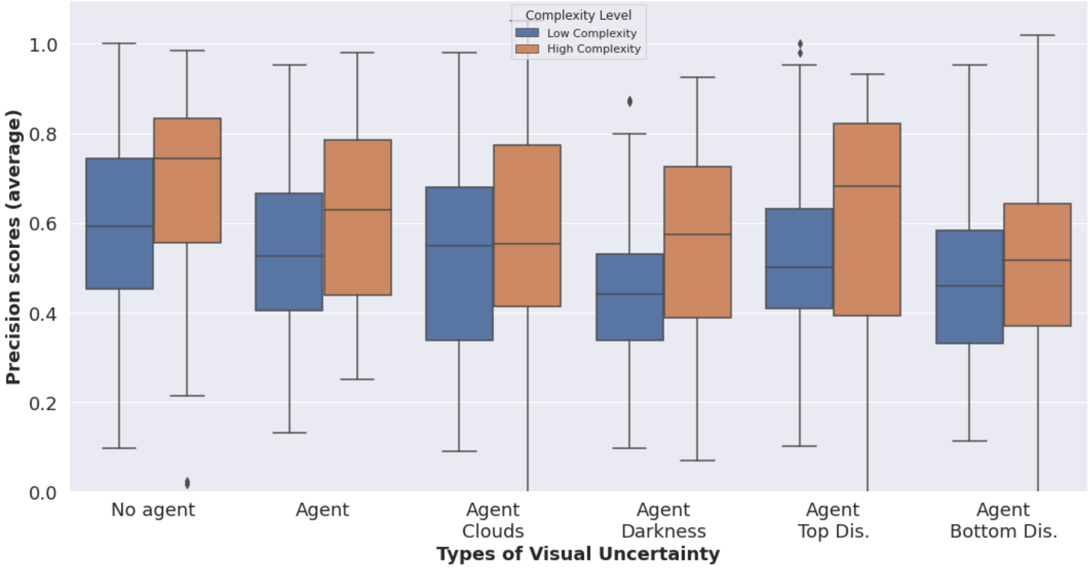


Figure 6.4: Precision scores for each session with or without agents. A higher score indicates a better performance and that fewer attempts were required to hit missiles. Similarly to Recall scores, participants in the “Top Disruption” condition performed better than in any other condition with visual uncertainty.

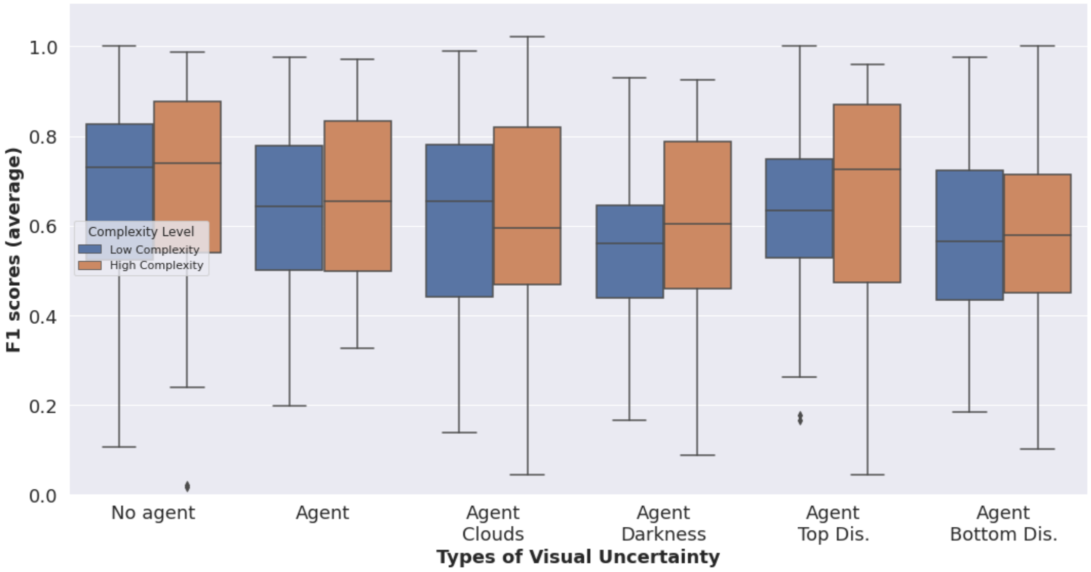


Figure 6.5: F1 scores for each session with or without agents. A higher score indicates a better performance. Overall, the presence of an agent actually led to overall worse performance in terms of F1 scores compared to sessions without any agent.

6.4.2 Reliance

User control time was recorded as the amount of time for which participants corrected agents for (in seconds). A higher user control time indicates lower reliance on the agent. Tables 6.1 and Figure 6.6 present user control times for all sessions. As anticipated, participants controlled the crosshair longer when no agent was present. In general, participants seem to have corrected the agent for much longer periods of time when experiencing any kind of visual uncertainty, with a peak being reached under the "Top Disruption" condition at a high level of task complexity.

A Mixed ANOVA test for user control times scores yielded significant results at all levels: complexity ($p < 0.0001$, $F = 9.78$, $np^2 = 0.11$), visual uncertainty type ($p < 0.0001$, $F = 35.24$, $np^2 = 0.30$) and interaction effect ($p < 0.0001$, $F = 2.26$, $np^2 = 0.03$). Follow-up pairwise T-tests showed that participants in the "Top Disruption" condition relied on the agent significantly less than in the "Darkness" ($p < 0.0001$, $T = -2.73$, $CLES = 0.42$) and "Bottom Disruption" ($p < 0.0001$, $T = -2.38$, $CLES = 0.43$) conditions.

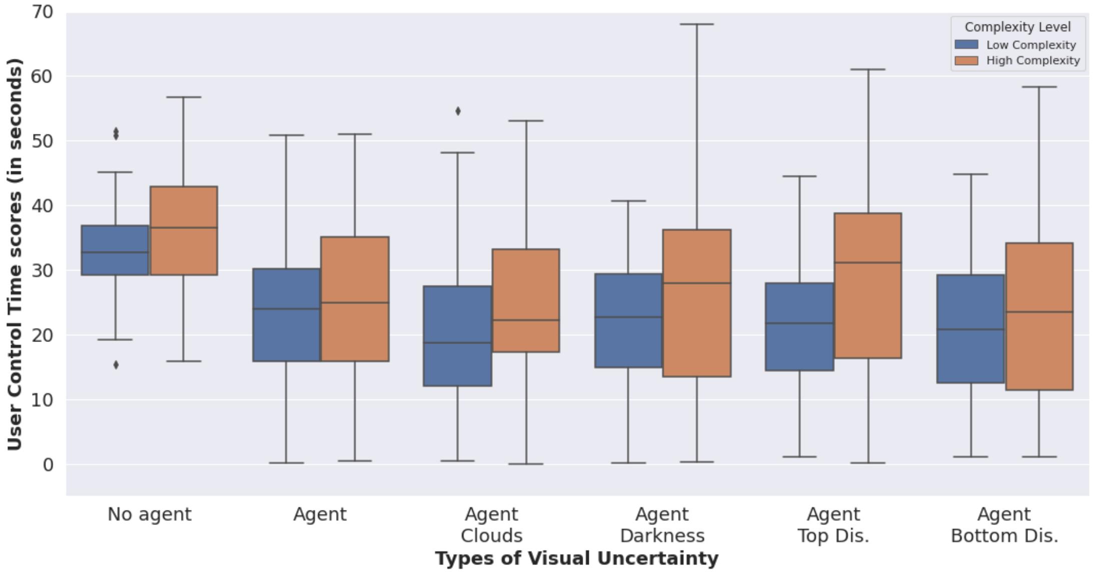


Figure 6.6: User Control Time for each session. A higher score indicates lower reliance on the agent. Among all of the sessions with an agent, participants relied on the agent the least in the "Top Disruption" one.

Table 6.2: Average scores related to reported Trust and Cognitive Workload. Highest values are highlighted in bold.

Question / Statement	No Agent	Agent	Agent Clouds	Agent Darkness	Agent Top Dis.	Agent Bottom Dis.
I can trust the agent	n/a	3.49 ± 0.13	3.83 ± 0.12	**4.11 ± 0.13**	3.90 ± 0.13	3.64 ± 0.12
How mentally demanding was the task?	14.08 ± 0.46	14.02 ± 0.45	14.46 ± 0.48	**14.67 ± 0.49**	14.12 ± 0.48	14.48 ± 0.49
How physically demanding was the task?	45.24 ± 2.52	49.12 ± 2.72	**54.46 ± 2.75**	53.13 ± 2.72	50.15 ± 2.55	54.17 ± 2.82
How hurried or rushed was the task?	70.63 ± 2.06	71.33 ± 1.95	**71.73 ± 2.10**	69.97 ± 2.14	69.05 ± 2.09	71.13 ± 2.15
How successful were you in accomplishing your level of performance?	39.73 ± 2.57	42.96 ± 2.52	38.79 ± 2.48	35.84 ± 2.42	**43.70 ± 2.71**	36.51 ± 2.46
How hard did you have to work to accomplish your level of performance?	**72.22 ± 1.69**	66.52 ± 2.16	68.60 ± 2.14	68.57 ± 2.09	67.41 ± 1.91	69.10 ± 2.00
How insecure, discouraged, irritated, stressed and annoyed were you?	58.33 ± 2.78	61.95 ± 2.71	**65.67 ± 2.68**	62.41 ± 2.61	59.82 ± 2.67	64.53 ± 2.50
Overall Raw TLX score	57.23 ± 1.17	58.38 ± 1.30	**59.85 ± 1.36**	57.98 ± 1.27	58.03 ± 1.17	59.09 ± 1.26

6.4.3 Trust

Between each level and after each session, participants were asked to indicate how much they trusted the agent by rating statements graded from 0 (low trust) to 7 (high trust). These statements and associated ratings are presented on Table 6.2. For the study of trust, we are referring to ratings related to the following statement: "I can trust the agent". Table 6.2 and Figure 6.7 present ratings given to the agent for each session. We can see that trust in both low and high levels of task complexity seems to be higher under visual uncertainty than without any type of visual uncertainty. In addition, Table 6.3 presents correlations between trust ratings (under the column"Parameter 2") and other variables recorded in this study (under the column "Parameter 1"). By consulting Table 6.3, we can observe that, overall, correlations between dependant variables and trust ratings are low. The behavioural proxy for reliance (recorded as "user control time") holds the best negative correlation score with trust ratings with a ρ score of -0.30.

A mixed ANOVA yielded significant results for the following levels: visual uncertainty type ($p = 0.0004$, $F = 5.14$, $np^2 = 0.004$) and interaction effect ($p = 0.0130$, $F = 3.20$, $np^2 = 0.05$) but not for task complexity ($p = 0.53$, $F = 0.38$, $np^2 = 0.03$). Follow-up pairwise T-tests indicated that participants in the "Darkness" condition reported significantly more trust in the agent than in the "Bottom Disruption" condition ($p = 0.001$, $T = 3.37$, $CLES = 0.58$).

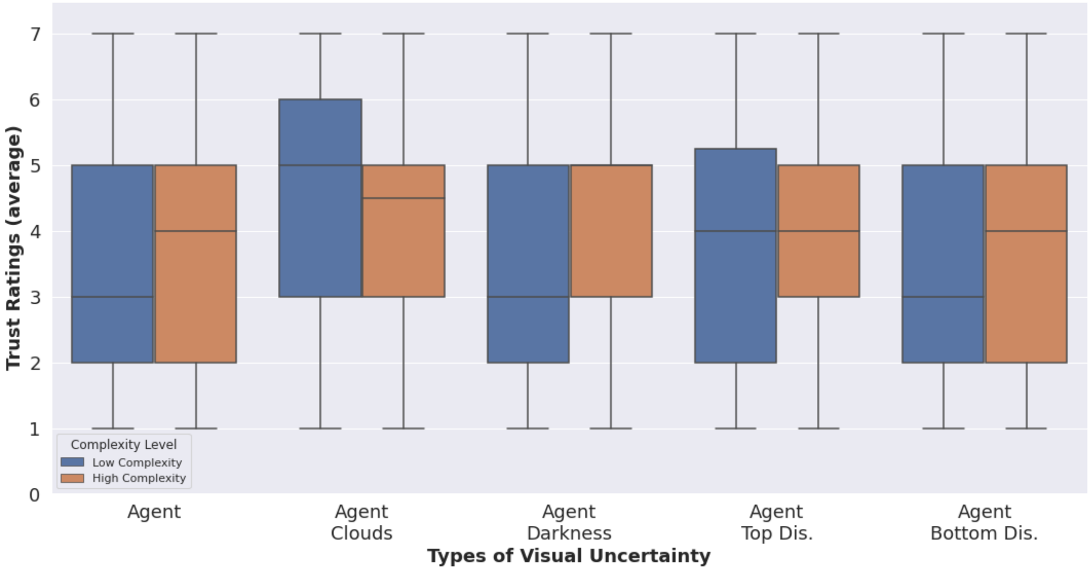


Figure 6.7: Reported Trust scores per session, where a higher score indicates greater reported trust in the agent. Participants reported higher trust levels in the "Clouds" condition compared to any other type of visual uncertainty.

6.4.4 Cognitive Workload

After each session, participants were asked to fill in a NASA TLX questionnaire [75] aimed at understanding their cognitive workload. Tables 6.2 and Figure 6.8 present RAW TLX scores, which are aggregated scores used to estimate the overall cognitive load a person reported after completing a task. Looking at Figure 6.8 we can see that participants reported a higher cognitive workload when interacting with an agent under visual uncertainty.

A mixed ANOVA yielded significant results for the following levels: task complexity ($p = 0.0096$, $F = 6.99$, $np^2 = 0.07$), interaction effect ($p = 0.0091$, $F = 3.10$, $np^2 = 0.02$) but not the visual uncertainty type ($p = 0.1807$, $F = 1.52$, $np^2 = 0.03$). No statistically significant differences were found in follow-up pairwise comparisons.

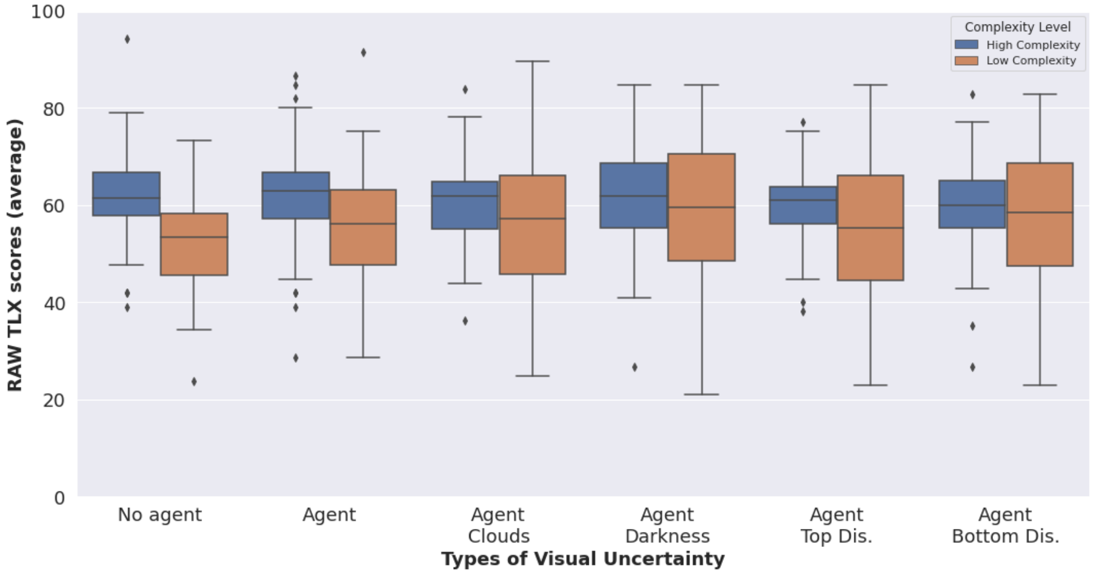


Figure 6.8: Reported Cognitive Load per session. A greater Raw TLX score indicates a more cognitively taxing experience. Overall, the level of task complexity affected reported Raw TLX scores the most.

6.4.5 Situational Awareness

To study situational awareness, we froze the task mid-way through and asked participants to report how many missiles were present in each half of the screen. We then subtracted their answers from the actual number of missiles present on the screen. Thus, a score of 0 means that they guessed the exact number of missiles, while a score of -2 or +2 means that they, respectively, underestimated or overestimated the missile count by 2. The total difference between guesses and the actual number of missiles is presented in Figure 6.9. We can observe that participants were more likely to make accurate guesses when interacting with an agent than when playing without one. However, the type of error participants made was influenced by the type of environmental uncertainty they encountered; for instance, participants operating under the "Darkness" condition recorded the largest range of under or overestimations. Figure 6.9 presents situational awareness results related to the bottom half of the screen only. We can observe that the "Bottom Disruption" and "Top Disruption" scores are very different, with the "Top Disruption" leading to more overestimations than the "Bottom Disruption", which led to more under-estimations.

A mixed ANOVA test yielded significant results on the following levels: task complexity ($p = 0.0117$, $F = 6.59$, $np^2 = 0.05$), visual uncertainty type ($p = 0.0068$, $F = 3.24$, $np^2 = 0.01$) but nothing significant on interaction effect ($p = 0.5571$, $F = 0.79$, $np^2 = 0.01$). Follow-up pairwise T-tests showed that participants underestimated the number of missiles present in the bottom half of the screen significantly more when the bottom of the screen was hidden under the "Bottom Disruption" ($p = 0.0106$, $T = 2.60$, $CLES = 0.65$)

condition than when the top of the screen was hidden under the "Top Disruption" condition ($p = 0.0106$, $T = 2.60$, $CLES = 0.65$).

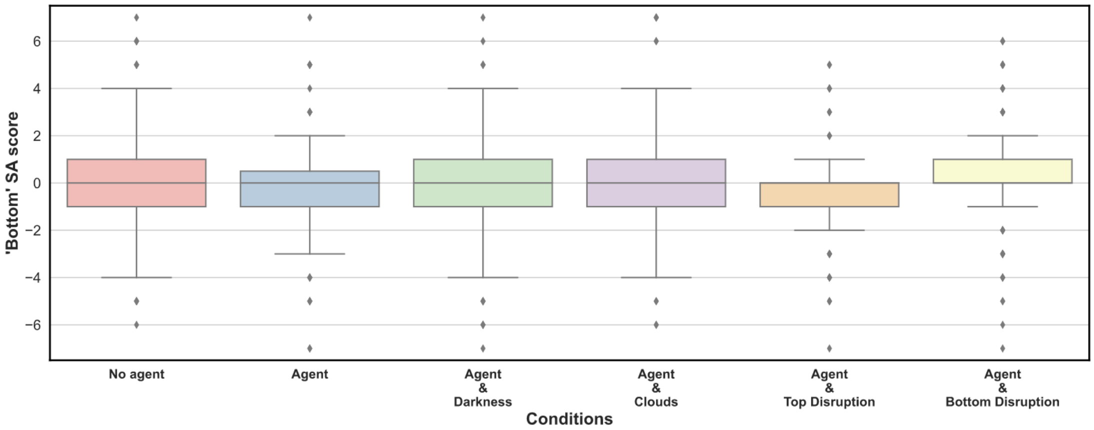


Figure 6.9: Situational Awareness scores. The closer a score is to 0, the better the participants' situational awareness is. The top-most plot presents overall situational awareness scores while the bottom-most plot presents SA scores related to elements located in the bottom half of the screen. The "Top Disruption" led to more under-estimations while the "Bottom Disruption" led to more over-estimations.

Table 6.3: Spearman's correlation tests between behavioural or reported metrics and trust ratings. A higher ρ scores indicates greater correlation.

Parameter 1	**Parameter 2**	ρ	**p-value**
User Control Time	Trust ratings	-0.3058	<0.001
F1	Trust ratings	-0.1497	<0.001
Task Difficulty	Trust ratings	-0.1455	<0.001
Precision	Trust ratings	-0.1355	<0.001
Gender	Trust ratings	-0.1179	<0.001
Raw TLX	Trust ratings	-0.1003	0.0284
Age	Trust ratings	-0.096	0.007
Recall	Trust ratings	-0.0783	0.0153
SAGAT Total Difference	Trust ratings	-0.0773	0.0166

6.4.6 Participants' Perceptions

At the end of the experiment, participants were presented with a description of each type of visual uncertainty they encountered and were then asked to report how each condition affected their "ability to play and rely on the agent". Qualitative coding was performed on the resulting dataset by 3 independent coders with no previous involvement in the study. Codes were created to look for particular changes in the way participants perceived their task performance or reliance on the agent. More details on coding analysis are available in Section 3.4.5. The codes used in this study were the following: "Increased Reliance on the agent", "No Change to Reliance", "Decreased Reliance on the agent", "Better Performance", "No Change to Performance" "Lower Performance". Internal agreement scores (Kappa scores [63, 183]), agreement scores and the number of references used for each code per visual uncertainty condition are presented in Table 6.4.

If we consult Table 6.4, we can observe that the "Increased Reliance on the agent" code was the most commonly used, with the highest number of references across all conditions. In addition, its associated kappa scores were always considered "fair to good", as they were found to be consistently above 0.41 [63]. This denotes that people were aware of being more reliant on the agent to varying degrees for each visual uncertainty they faced. The other code heavily employed is "Lower Performance", with a much higher frequency than any other code pertaining to performance (with "fair to good" Kappa scores), indicating that each visual uncertainty heavily impacted participants' perception of how successful they were at completing the task. When comparing participants' feedback for each session, we can observe differences in terms of the codes most used. For instance, the "Clouds" type of visual uncertainty resulted in the highest amount of references used than any other visual uncertainty type, denoting a more varied reaction.

Table 6.4: Results from qualitative coding analysis on post-hoc surveys data asking participants to report how each condition changed their reliance on the agent and ability to play. The highest Kappa scores for each visual uncertainty type are in bold.

Visual Uncertainty	**Code**	**References**	**Agreement score**	**Kappa score**
Darkness	Increased Reliance on agent	213	89.34%	**0,54**
	Lower performance	118	91.15%	0,48
	Better performance	31	94.59%	0.18
	No change to reliance	22	95.89%	0.04
	Decreased reliance on agent	27	96.39%	0.16
	No change to performance	12	95.89%	0.37
Clouds	Increased Reliance on agent	153	87.86%	0.38
	Lower performance	102	92.03%	**0.51**
	Decreased reliance on agent	62	93.49%	0.29
	Better performance	37	96.27%	0.24
	No change to reliance	34	94.49%	0.02
	No change to performance	30	96.21%	0.16
Top Disruption	Increased Reliance on agent	108	93.30%	0.41
	Decreased reliance on agent	72	92.31%	0.34
	Lower performance	68	93.73%	**0.45**
	No change to reliance	49	95.19%	0.25
	Better performance	43	95.63%	0.35
	No change to performance	36	96.31%	0.40
Bottom Disruption	Increased Reliance on agent	147	88.20%	0.39
	Lower performance	121	88.72%	0.44
	No change to reliance	43	94.38%	0.15
	Decreased reliance on agent	39	94.39%	0.07
	Better performance	24	95.99%	0.25
	No change to performance	21	96.84%	**0.46**

6.5 Discussion

In this study, we explored how *Visual Uncertainty* was impacted users during an interactive human-agent collaborative task. We designed an experiment comprising four different types of visual uncertainty, each occluding visual information in different ways, and tested their impact on the human-agent relationship. This study sought to answer our third Research Question: **How do different types of environmental conditions (static or moving), which impair vision and induce uncertainty, affect the human-agent relationship?** from our overall research questions highlighted in Section 1.3. With our sub-research questions, we were concerned about the role of visual uncertainty regarding:

- **RQ3.a:** how well the human-agent team performs at a task.
- **RQ3.b:** how much the user trusts and relies on the agent.
- **RQ3.c:** situational awareness in relation to the users' trust in the agent.

Our results indicate that visual uncertainty, in addition to having a negative impact on task performance, can also alter the way users are willing to rely on and trust an agent in a collaborative task.

6.5.1 Trust under Visual Uncertainty

By referring to prior work related to trust in automation and transparency, we anticipated that uncertainty in the environment would have a negative effect on how participants trusted the agent (see hypotheses presented in Section 6.2). We found evidence to suggest that this is not the case in all situations. When comparing trust ratings (see Figure 6.7), we found that participants reported significantly higher levels of trust in the agent under the "Darkness" type of uncertainty than any other type of uncertainty (which informs RQ3.b). This finding is surprising, as trust is a construct that is calibrated through interactions [43, 85], and should benefit from the added transparency found in an environment with perfect visibility. We believe that participants trusted the agent more in such conditions because of the extensive nature of the "Darkness" type of uncertainty, which effectively covered all the whole screen except for a "halo" around the crosshair. In a situation where participants wanted to obtain greater situational awareness, they could have bypassed the agent's aim completely and manually "scanned" the screen for missiles. Instead they decided to rely, to a greater or lesser extent on what the agent was showing them. In the end, lowering reliance on agents was shown to be an ineffective way of dealing with this type of uncertainty, as performance (as demonstrated by recall scores) was found to be significantly lower in "dark" conditions than in any other context.

6.5.2 Situational Awareness under uncertainty

We posited that the types of uncertainty that hid the most information would have the most significant impact on situational awareness and consequently reduce participants' trust the most (see hypotheses presented in Section 6.2). We found that to be the case for one type of uncertainty (see Figure 4.7). In our study, two types of environmental uncertainty were designed to be dynamic and to obstruct the most on-screen elements: “Darkness”, where the whole screen was hidden, and “Clouds” where the top and bottoms halves of the screen were hidden dynamically at different points. We found that despite a tendency for participants to make more erroneous guesses under the “Darkness” condition, there were no statistically significant differences related to these “Darkness” and “Clouds” in terms of situational awareness (which informs RQ3.c). However, we found that participants reported significantly higher trust ratings under the “Darkness” condition than when faced with any other type of uncertainty. As the “Darkness” condition hid the most elements from participants, our findings seem to indicate that forcing participants to focus on a specific area could be enough for them to forget what is happening elsewhere, in the same environment of interaction. This finding highlights how easily complacency can set in when environmental visual uncertainty hides crucial information, despite an agent’s performance and reliability remaining the same.

6.5.3 Adapting under uncertainty

With our third hypothesis (see Section 6.2), we expected participants to be more likely to rely and trust agents when they were forced to react more quickly. We found that uncertainty types negatively affected reliance, but not reported trust. In the design of our uncertainty conditions, two types of uncertainty were similar in terms of size and shape but different in their locations on the screen: the “Top” and “Bottom” disruptions, which hid, respectively, the top half and the bottom half of the screen. We found that participants relied significantly less on the agent when the top half of the screen was hidden than when the bottom half of the screen was hidden. In this task, missiles spawn from the top of the screen, making that a very important area to attend to, as it allows participants to assess incoming targets and predict their path using their current bearings. It was surprising to see that despite a lack of visual feedback on this crucial part of the screen, participants were actually significantly more likely to take control of the agent’s aim. However, when comparing trust ratings, no significant differences were found between the “Top Disruption” and “Bottom Disruption” conditions, which might indicate a level of cognitive dissonance between how users felt about the agent (same level of trust) and how they interacted with it (clear distrust in one set of conditions compared to the other).

6.6 Conclusion

In this Chapter, we examined the impact of different types of visual uncertainty on participants in a human-agent interactive collaborative task. We used both interaction logs and survey instruments to infer and study how and why participants changed their behaviours when confronted with visual environmental uncertainty. We found that the type of uncertainty has a direct impact on how likely participants are to rely on an agent, assuming the agent’s level of performance remains the same. More precisely, we found that a high level of uncertainty (studied through the “Darkness” condition in this paper, producing a *tunnel vision* effect) led participants to trust and rely significantly more on the agent compared to uncertainties that were more static in nature, and only obscured parts of the screen (namely, our “Bottom” and “Top Disruption” conditions). In addition, we found that this complacent behaviour led participants to achieve a significantly lower performance, which was ultimately detrimental to the human-agent collaboration. Nonetheless, some types of uncertainty actually mitigated this complacency and kept users more focused on their tasks. We found that hiding the top half of the screen (“Top Disruption” condition), where important information were shown, led participants to perform significantly *better* than, even, in sessions without any type of visual uncertainty. However, overall situational awareness remained fairly constant throughout the study, albeit this could be explained by the inherent difficulty of the tasks, and not the type of uncertainty experienced.

Our findings show that there is a need to understand exactly how users process information in order to plan future actions and make decisions when this same information is missing, as well as how an agent can help mitigate this environmental uncertainty.

Chapter 7

Visual Explanation and Agent Transparency

7.1 Motivation

In this chapter, we look at how visual agents that display information about the task or an aiming agent's actions can affect the human-agent partnership. We intended for the design of each visual agent to enhance specific SA levels and assess their impact on trust, reliance, task performance, cognitive load and situational awareness. This work was motivated by the work of Chen et al. [26], who proposed the "Situation Awareness-based Agent Transparency model" (SAT) which aims to support users' SA via different visualisation levels intended to help users understand a situation or system's decisions, and prepare for the future outcomes of their interactions. The framework defined three Situational Awareness levels. The first focused on understanding "what" is happening, the second one focused on understanding "why" something is happening and the third on understanding what will happen next.

This study seeks to answer the fourth Research Question defined in Section 1.3: **How do different types of visual help (designed to elicit different levels of situational awareness) influence the human-agent relationship?** More specifically, we aim to answer the following research questions:

- **RQ4.a:** How beneficial is the introduction of visual agents when users are by themselves (no aiming agent)?
- **RQ4.b:** How beneficial is the introduction of visual agents when users are supported by an aiming agent?
- **RQ4.c:** Which visual agent provides the best overall support?

Using the interactive framework detailed in Chapter 3, we conducted an online-based 6 (types of visual explanations) x 2 (levels of difficulty) mixed-design study where 180 participants interacted with an agent in tasks of various levels of difficulty (within-factor) with different types of visual agents (between-group factor) based on the SA framework by Chen et al. [26]. In addition to SA level, the design of each visual agent was intended to provide information in a descriptive (highlighting important information) or prescriptive ("telling" users what to do) manner. Our results indicate that participants did not significantly benefit from the addition of visual agents in any of the metrics we recorded, with even lower differences in scenarios where an aiming agent was introduced.

7.2 Related Work

7.2.1 Visualisation modalities

In the following paragraphs, we discuss some of the most common paradigms when it comes to communicating information via visualisations in HAI and other domains. Different modalities of providing relevant task-specific information have been designed and employed to provide better decision-making support in a HCI scenario. In a study focused on the design of Head-up Display (HUD), Charissis et al. state that "a successful human-centred interface should enhance human actions [...] senses [...] and judgement [...]. Furthermore, it should guide the user rather than constrain his/her [...] abilities" [23, p. 2]. While Charissis' work was focused on Human-Machine Interfaces (HMI) in an automotive environment (see Figure 7.1), guiding a user without hindering their abilities is an obligation for successful HCI interactions, especially when users are required to understand changes in the environment and respond to them appropriately and in a timely manner. A range of visualisation techniques have been tested in various studies to communicate information as quickly and efficiently as possible [23, 138, 163].

Alphanumeric (alphabetical and numerical) symbols are one of the oldest and most commonly used ways of presenting information to the user. In a work classifying different types of visual representations, Lohse et al. [114] describe how employing numeric elements is often perceived as "unattractive" when used to emphasise parts of a specific representation, and can often lead to confusion by overloading the user [23]. As a more compact way of displaying information, other systems rely on *icons* or *symbols*, which assign an unambiguous meaning to a picture [163]. Icons are used when the meaning of the icon is apparent to the target audience; for instance, signs with an exclamation mark "!" are commonly used to indicate a potentially hazardous area or incoming danger. Icons were found to be interpreted much more quickly by human operators in fast-changing scenarios, where "iconic displays led to response times three times shorter than responses time with alphanumeric displays" [163, p. 5]. In fast-paced tasks, icons can also be used as "attention indices" [138, 139], where symbols serve only as "pointers to attention processes" so that users know what to focus their attention on.

Most visualisation techniques rely on "alphanumeric" symbols and/or "icons" to display changes on an Human-Machine Interface. While these modalities remain the same, their implementation can vary greatly depending on the task. For instance, some interaction scenarios will require the future state of one or more elements to be displayed. In these situations, a technique named “Conformal Symbology” [68] can be useful, as it highlights elements of importance by overlapping them on the environment of interaction itself, providing a more seamless integration of visual elements. In other, longer, surveillance tasks, attention does not have to be sustained at all times but only in short bursts. In these scenarios, “dead reckoning” [163] techniques are often used to provide navigational information by displaying headings (usually in the form of arrows) above elements, hinting at their most likely future positions and letting the user decide whether they deserve further investigation or not.

There are, in short, many different types of visualisation techniques, all of which can be deployed in different ways, and all of which boast their own strengths and weaknesses. For instance, while icons can provide information quickly, they lack the transparency of alphanumeric information. And while alphanumeric symbols are an excellent means of accurately conveying information, they lose in terms of cognitive workload demand and their potential for cognitive capture. To better understand *when and on which criteria* a visualisation should be chosen, one can refer to the literature related to situational awareness [26, 50, 51], where visualisations are assessed based on the type of information displayed and requirements of the task.

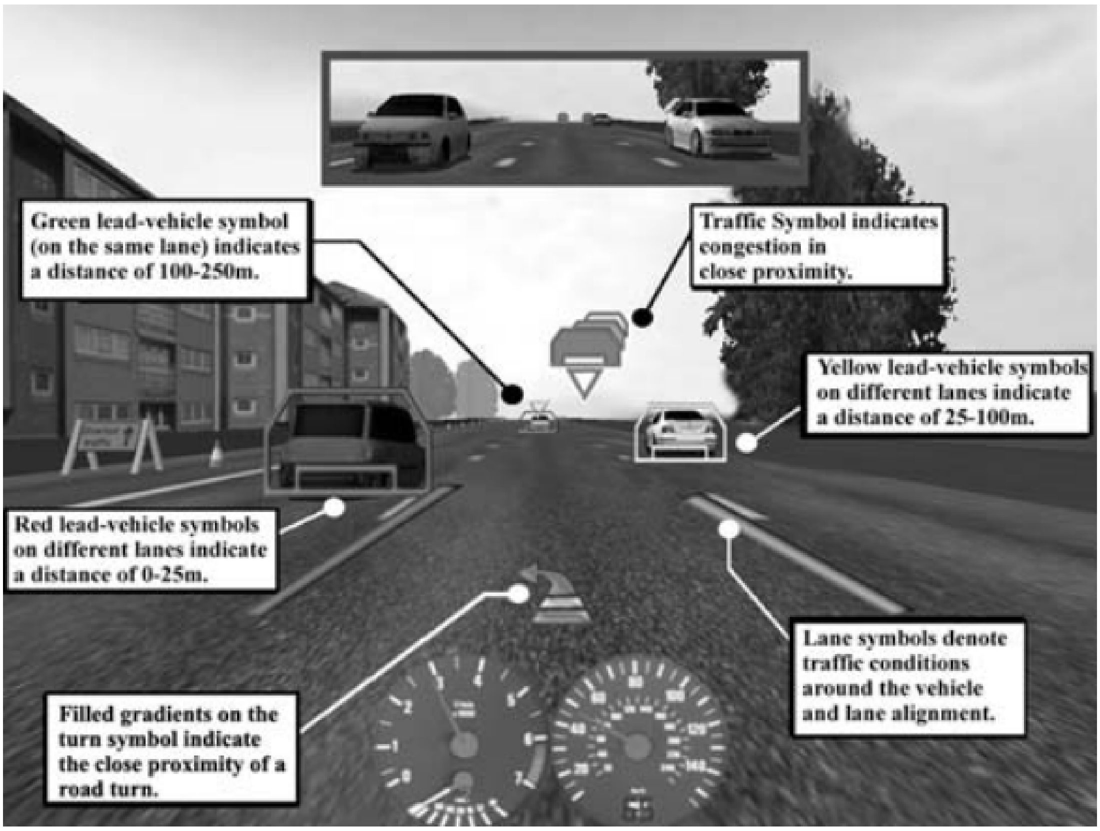

Green lead-vehicle symbol (on the same lane) indicates a distance of 100-250m.
Traffic Symbol indicates congestion in close proximity.
Yellow lead-vehicle symbols on different lanes indicate a distance of 25-100m.
Red lead-vehicle symbols on different lanes indicate a distance of 0-25m.
Lane symbols denote traffic conditions around the vehicle and lane alignment.
Filled gradients on the turn symbol indicate the close proximity of a road turn.

Figure 7.1: Example of an Head-Up Display (HUD) interface from a study by Charissis et al. [23]. A good HUD interface must manage to present task-specific information without obstructing the user’s field of view.

7.2.2 Visualisation & Situational Awareness

In a comprehensive report by Chen et al. [26], Situational Awareness-based agents are presented as likely to have a positive impact on trust in agents, as they could improve trust development by providing more information about the system's inner workings in a simplified form, as hinted by Lee and See's work [103]. Situational Awareness is usually studied on 3 different levels [26]:

- **SA level 1:** "The what": The agent conveys information about the current situation.
- **SA level 2:** "The why": The agent shows its reasoning process and explains its constraints.
- **SA level 3:** "The what next": The agent indicates what could happen next based on current limitations and/or trends.

Previous work in situational awareness and visualisation techniques has either theorised or empirically tested the benefits of numerous visualisation modalities that aim to present information about an agent or the environment of interaction back to a user. However, finding the "best" type of visualisation to display information is challenging, as it largely depends on the type of task and broader context of interaction. In order to investigate the strengths and weakness of different types of visualisation modalities, we make use of the SAT framework formalised by Chen et al. [26] which classifies visualisation modalities based on the type of support they provide.

Based on previous research related to visualisation and situational awareness, the study presented in this chapter investigates the impact of different types of visual agents on the human-agent relationship in a collaborative scenario. Each visual agent employs different visualisation modalities and was both designed according to past studies related to Situational Awareness levels [26] and informed by the empirical implementation of visualisations from past HAI and Human Factor research. In addition to the SAT framework, we further divide our visualisations into two categories: "descriptive" and "prescriptive". Visualisations that are *descriptive* are intended tp focus on highlighting important elements and letting the users make sense of the information while *prescriptive* visualisations are intended to process more of the information for the user.

In this Chapter, we study how human-agent collaboration evolves when supported by different types of visual agents intended to provide different levels of Situational Awareness support. Specifically, we are starting out with the following hypotheses:

- **H1** As greater transparency is linked to better task performance [101], visual agents will have a noticeable impact on the human-agent collaboration no matter whether aiming agents are present or not.

- **H2** Visual agents that provide prescriptive information (telling the users "what to do") will have a more positive impact on reliance and task performance than agents that provide descriptive information (highlighting important elements in the environment), as they require less time to be processed.

The main contribution of this work lies in testing different types of visual agent (and subsequent visualisations) on the human-agent partnership during a real-time task. Below we outline the method used to design visual agents and conduct this study.

7.3 Method

This study was conducted using our interactive human-agent collaborative framework described in Section 3. Like the study presented in Chapter 6, this experiment was conducted online. Modalities for online studies are described in Section 3.5.2.

7.3.1 Visual Agents

Previous studies presented in this thesis (see Chapters 4, 5 and 6) used agents designed to help users aim at targets. In the work presented in this chapter, we incorporated agents displaying visualisations to users. Here, these agents are referred to as *visual agents*, as opposed to previous agents (presented in Chapters 4, 5 and 6) that were only helping with the aiming process and will consequently be referred to as *Aiming Agents* throughout this Chapter. We developed a total of six different types of visual agent based on the situational awareness framework proposed by Chen et al. [26]. Each visual agent is focused on giving the user more information about either what the agent is doing (SA level 1), the agent's reasoning and prioritising process (SA level 2), or what the agent will do next (SA level 3). These 3 levels of Situational Awareness are further detailed in Section 7.2. In addition, we intended for the design of each visual agent to be either "prescriptive" (telling users what to do) or "descriptive" (letting the user make sense of the information). To make all designs comparable, each visualisation only displayed information about 5 targets at once. Below, we provide descriptions of the visual agents developed for this study (please see Figure 7.2 for an abstract representation). Contrast between colours used by visual agents (especially for the "Threat Shape" visual agent) were designed to be high enough to account for users suffering from deuteranopia [71], the most common form of colour blindness.

7.3.1.1 SA Level 1

- **Threat Shapes**, presented in Figure 7.2 and implemented as shown in Figure 7.3a. This visualisation indicates which target(s) the agent recognises as threats (red triangle), or non-threats (green polygon). This visualisation focuses on SA level 1: understanding "what" is happening in the current situation. The visualisation was based on notes and implementations found in the work of Pylyshyn et al. [139], Shekkhar et al. [163] and Mercado et al. [124]. This visualisation can be described as “prescriptive”, as the visual agent processes most of the information for the user (which targets are important or not).

- **Priority Numbers**, as presented in Figure 7.2 and implemented as shown in Figure 7.3b. This visualisation displays the results of the agent’s prioritisation process via numbers, indicating which missiles the user is advised to focus on. This design was inspired by descriptions and implementations found in the work of Lohse et al. [114] and Chen et al. [26], and can be described as “descriptive”, as it describes the order in which the agent should aim at targets.

7.3.1.2 SA Level 2

- **Agent’s Prioritisation**, presented in Figure 7.2 and implemented as shown in Figure 7.3c. This visualisation focuses on SA level 2 and understanding “why” certain actions are being recommended to the user or undertaken by the agent. With this visualisation, targets deemed as threats are highlighted with red squares of different sizes and opacity (the bigger and more opaque, the more important the target, according to the agent) to indicate the priority in which participants are recommended to deal with them. This visualisation combines priority and threat assessment to communicate *why* users have to deal with targets in a certain order. The design was mostly informed by the work of Kilgore et al. [93] who empirically tested an interface where the size and transparency of icons were changed to highlight specific elements. This type of visualisation can be described as “prescriptive”, as it parses most of the information for the user (which targets are important and in which order they should be hit).

- **Missile paths**, presented in Figure 7.2 and implemented as shown in Figure 7.3d. This visualisation focuses on SA level 2 and understanding“why” the agent aims at certain targets based on their current paths. This visualisation consists in displaying the paths of missiles and their trajectories. This design was informed by the work of Iordanescu et al. [88]. This visualisation can be described as “descriptive” at it display missiles’ trajectories.

7.3.1.3 SA level 3

- **Agent’s Plan display**, presented in Figure 7.2 and implemented as shown in Figure 7.3e. This visualisation displays paths between targets in the order in which the agent is aiming at them. This gives an explanation as to *why* the agent is heading in a particular direction. The design was inspired by the work of Ramchurn et al. [143]. This visualisation can be described as “prescriptive”, as it provides the user with a path that they can choose to follow or not.

- **Performance Graph**, presented in Figure 7.2 and implemented as shown in Figure 7.3f. This visualisation support SA level 3 and gives more general information about the current level of performance as well as the current trend (whether the team is getting better - with a green arrow or worse - with a red arrow). This visualisation can be described as "descriptive" as it displays information about the evolution of task performance over time.

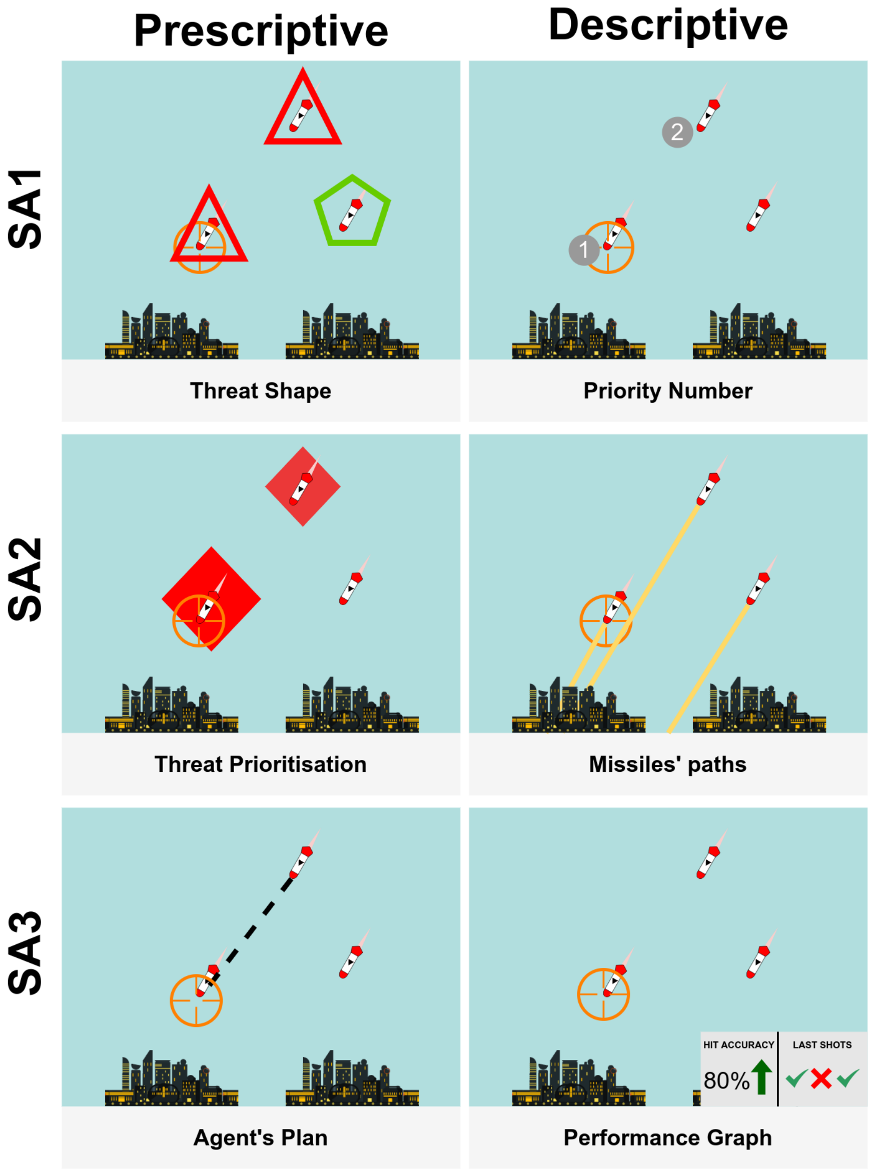
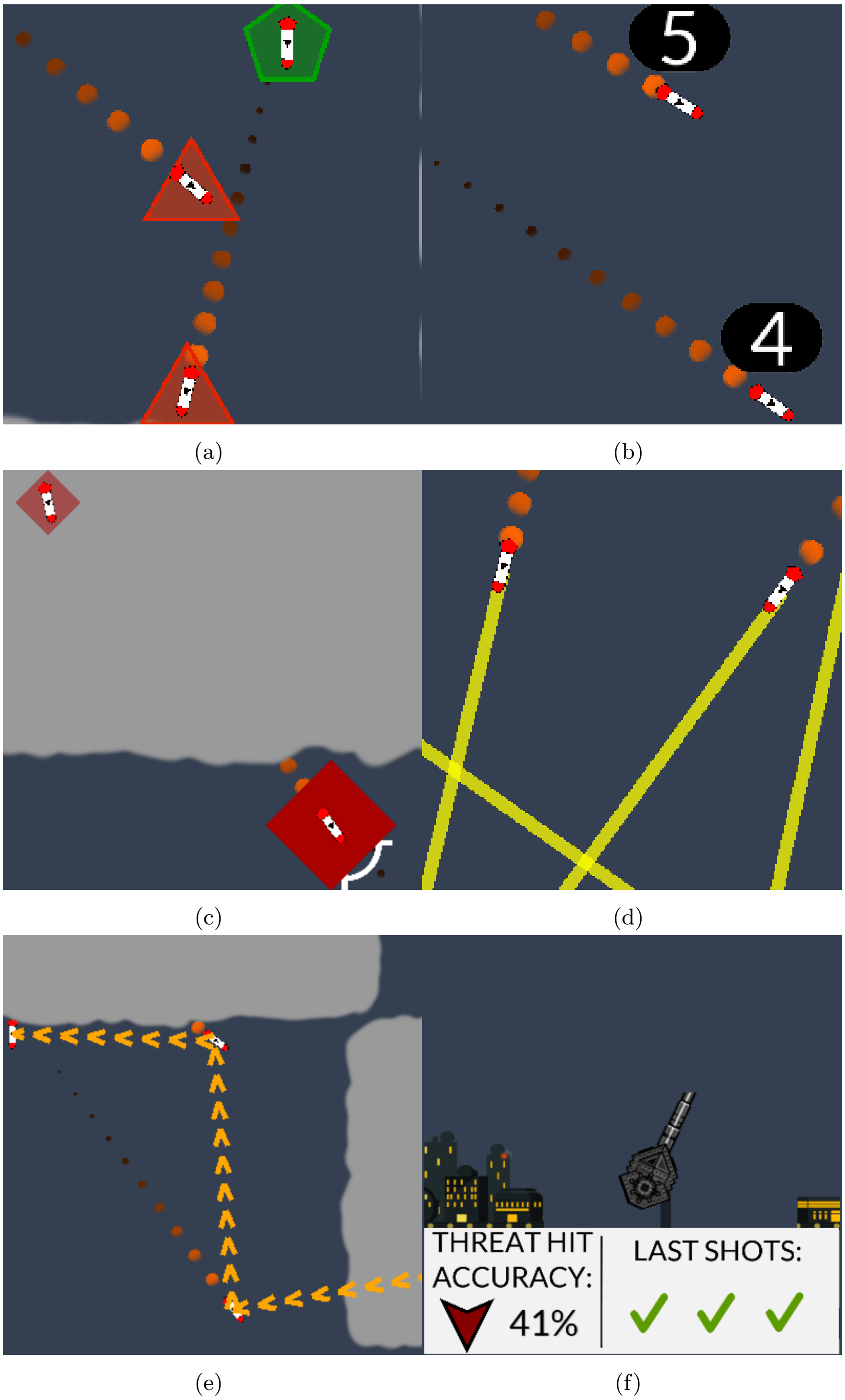


Figure 7.2: Abstract representation of the visual agents developed for this study. Each visualisation supports a different level of Situational Awareness.



(a) (b)

(c) (d)

(e) (f)

Figure 7.3: All visual agents intended to support different SA levels: (a) represents SA1 Threat Shape, (b) Priority Number, (c) Threat Prioritisation, (d) Agent Plan display, (e) Missiles' paths and (f) Performance Graph.

7.3.2 Task Difficulty

Each participant, no matter whether an aiming agent and/or a visual agent was present, played in sessions composed of two levels of difficulty and lasting for 120 seconds per level of difficulty.

Task difficulty was considered in terms of the number of missiles to hit. Their speed was fixed for the "Easy" and "Hard" levels across all sessions, with or without agents. The details regarding difficulty settings are presented below:

- In the "Easy" level, 3 missiles spawned every 5 seconds at a speed of either 30 or 60 pixels per second (random selection) for a total of 54 missiles. 30% (16 missiles) of the missiles spawned were "False Positives" (not heading toward cities).
- In the "Hard" difficulty level, 3 missiles spawned every 4 seconds with a speed of either 60 or 80 pixels per second (random selection) for a total of 67 missiles. 30% (20 missiles) of the missiles spawned were "False Positives" (not heading toward cities).

Contrary to previous studies presented in Chapters 4, 5 and 6, where participants were asked to hit as many missiles as they could, the experiment presented in this chapter relies on testing different visual agents, supporting different SA levels that should help to identify relevant and non-relevant information. In our study, this difference is reflected by the introduction of True Positive (missiles that are going to hit a city) and False Positive (missiles that are NOT going to hit a city). As a result, 30% of all missiles spawned in each difficulty level (Easy and Hard) were False Positives. For this particular study, therefore, our metrics for task performance will be computed differently to take into account the inclusion of False Positives (missiles not colliding with cities highlighted as a threat by the agent) and False Negatives (non-threatening missiles highlighted as threats by the agent). The following equations detail these changes to the original ones present in Section 3.4.3.1.

$$\textbf{Threat Precision} = \frac{\#ThreateningMissilesDestroyed}{\#TMissilesDestroyed + \#NTMissilesDestroyed} \tag{7.1}$$

$$\textbf{Threat Recall} = \frac{\#ThreateningMissilesDestroyed}{\#ThreateningMissilesSpawned} \tag{7.2}$$

$$\textbf{Threat F1} = 2 * \frac{ThreatPrecision * ThreatRecall}{ThreatPrecision + ThreatRecall} \tag{7.3}$$

In addition, this study also makes use of "Relative metrics", which are defined as the relative gain or loss in a single session when compared to a reference one. As this experiment contains a between-groups design, computing Relative metrics helps understand whether a visual agent resulted in a relative improvement or loss. For instance, if participants scored a Recall of 0.6 by themselves and 0.8 with a visual agent, the relative gain would be 0.2. This process is illustrated by Equation 7.4 below:

$$\textbf{Relative Metric} = ScoreSessionA - ScoreSessionRef \tag{7.4}$$

7.3.3 Agent Reliability

Contrary to previous studies (see Chapters 4, 5 and 6), the aiming agent's reliability was not determined by the accuracy of its aim, but by the type of error it made. For each level, as in previous studies, the agent would have a reliability of exactly 80% with the difference from previous studies being that the remaining 20% error-rate comprised only False Negatives (the agent not aiming at a missile that is going to hit a city) and False Positives (the agent aiming at a missile that is going off-screen, outside the viewport) errors.

7.3.4 Independent and dependent variables

In this section, we summarise the independent and dependent variables used in this study and as motivated by our experimental method and research questions.

Our independent variables are the following:

- *Task difficulty*, as defined by the amount and speed of missiles in each level.
- *Aiming agent reliability*, which remained high during all conditions.
- *Situational Awareness (SA) groups*, where each visual agent was intended to provide information regarding SA level 1, 2 or 3.
- *Prescriptive or Descriptive visual agents*, each visual agent, regardless of its SA group, was designed to be either "prescriptive" (intended to guide users toward specific decisions) or "descriptive" (intended to let users make sense out of the information presented).

Our dependent variables are the following:

- *Task Performance*, in terms of missiles hit, shots fired and missile missed.
- *Reliance*, expressed by the duration for which participants relied on the aiming agent's help.
- *Trust*, as reported by participants.
- *Cognitive Workload*, as reported by participants.
- *Situational Awareness*, as reported by participants.

7.3.5 Experimental Procedure

This online study was approved by the University of Strathclyde CIS Departmental Ethics Committee (Ethics App. No. 1395). The study was completed entirely online using participants' own hardware and via the Prolific platform (for more information about the platform, see Section 3). The experimental procedure was similar to the one described in Section 3. Only participants that experienced an average frame-rate above 24 frames per second were kept in our dataset for further analysis (to ensure "playable" experimental conditions). A minute long bench-marking test was provided for free before participants consented to take part in the study in order to filter out those whose hardware did not meet our requirements. Participants received £5.50 for undertaking the experiment (approximate duration of 45 minutes). Once registered, each participant went though the following steps:

1. Demographic and pre-hoc survey. (five minutes) (see Appendix D).
2. Tutorial aimed at understanding the controls of the game and getting used to interacting with the agent. (two minutes).
3. Session with an agent. (four minutes).
4. Session without an agent. (four minutes).
5. Session with visual agent only. (four minutes).
6. Session with an agent and visual agent. (four minutes).
7. Post-hoc survey collecting participants' feedback. (five minutes) (for further details, see Appendix D).

The order of the sessions described above (see items 3 to 6) was randomised using a Latin Square [15] to reduce the learning effect. A presentation of each visual agent was provided at the beginning of sessions with examples in order for participants to understand how to make sense of the information communicated to them. In terms of survey instruments, participants completed NASA TLX rating scales, which are 6 item survey instruments widely used to measure cognitive workload [75]. In this study, RAW TLX [18] scores are reported. To measure trust in the agent, we used a single statement at the end of each round: "I can trust the agent" graded on a 7 point Likert scale from 1 (complete distrust in the agent) to 7 (total trust in the agent) adapted from the work of Jian et al. [90]. To measure Situational Awareness, we used the 3 item Situation Awareness Rating Technique (SART) [159] also called "3D SART", which comprises 3 questions eliciting different elements related to SA such as "Attentional Demand", "Attentional Supply" and "Understanding". Further details on the survey instruments used in this study are presented in Section 3.4.2.

7.3.6 Demographics

180 participants (93M, 87F) participated in this study, where most people ($n = 104$) indicated being aged from 18 to 24 years old while the remaining participants ($n = 76$) were aged 25 to 34 years old. In terms of level of education, most participants reported having a Bachelor's degree ($n = 77$) while the rest reported having a College degree ($n = 33$), High school diploma ($n = 32$) or Master's degree ($n = 20$), or other ($n = 38$). The "Confidence" dimension

from the "Revised Computer Game Attitude Scale" [22] was used to evaluate how confident participants felt about their video-game skills with self reported rating scales ranging from 0 (denoting low confidence) to 5 (denoting high confidence). Participants' average rating was 3.85 ± 0.9, denoting a population used to playing games and confident in its abilities.

7.4 Results

In this section, we present results pertaining to task performance, users' reliance on aiming agent, reported trust in the aiming agent, cognitive workload and situational awareness. We report results obtained by participants at the end of each session, across all levels of difficulty. More details on the inclusion of different difficulty levels are available in Section 3.2.4. The statistical methods we used to compare and report results is detailed in Section 3.4.4. We are using different metrics from those employed in Chapters 4, 5 and 6 to obtain insights on how users performed with and without an aiming agent or visual agent. These changes are explained in Section 7.3.2. When describing results, we distinguish between scores obtained during sessions without aiming agents and sessions with aiming agents. This distinction helps us to focus on the influence of a visual agent on the human-agent collaboration with and without the assistance of an aiming agent.

7.4.1 Relative Performance

Figures 7.4, 7.5 and Tables 7.1, 7.2 and 7.3 present scores related to performance and relative performance. Overall, participants scored higher for all performance metrics in sessions where an aiming agent was present, compared to sessions where participants were only assisted by a visual agent, regardless of its SA level. Relative Threat Recall and Relative Threat Precision scores provide insights into the number of important targets (true positive - missiles heading toward cities) that participants hit during the tasks.

7.4.1.1 Visual Agent without Aiming Agent

Overall, participants' performance was lower when no aiming agent was present. Looking at the Relative Threat Recall scores on Figure 7.4, we can observe that participants benefited from having the support of a visual agent in all groups except "Threat Shape" (SA1) and "Agent Plan" (SA3), which display the most variance in Relative Threat Recall scores. Threat Precision scores, however, exhibit more significant differences when compared to all other performance indicators. Across all six groups, participants scored high Threat Precision scores when no aiming agent was present compared to sessions with an aiming agent. This difference is particularly important for Relative Threat Precision scores in the "Priority Number" group (SA1) and "Missile Path" group (SA2).

While performing between-groups comparisons on sessions with a visual agent and without an aiming agent, a Welch ANOVA yielded significant results for Relative Threat Recall scores ($F = 11.78$, $p < 0.0001$, $np^2 = 0.14$). Further pairwise comparisons using Games-Howell tests indicated that participants in the "Agent Plan" (SA3) group performed significantly worse than participants in the "Priority Number" (SA1) group ($T = 5.56$, $p = 0.001$, $CLES = 0.76$), "Missile Path" (SA2) group ($T = 5.21$, $p = 0.001$, $CLES = 0.75$), "Threat Prioritisation" (SA2) group ($T = 7.16$, $p = 0.001$, $CLES = 0.82$) and "Performance Graph" (SA3) group ($T = -93$, $p = 0.0021$, $CLES = 0.31$). In addition, participants in the "Performance Graph" (SA3) group scored significantly lower in terms of Relative Threat Recall than participants in the "Threat Prioritisation" (SA2) group ($T = 4.3$, $p = 0.001$, $CLES = 0.71$).

While performing between-groups comparisons on sessions with a visual agent and without an aiming agent, an ANOVA yielded significant results for Relative Threat Precision scores ($F = 8.41$, $p < 0.0001$, $np^2 = 0.11$). Further pairwise comparisons using Tukey tests indicated that participants in the "Priority Number" (SA1) group performed significantly better than participants in the "Threat Shape" (SA1) group ($T = 4.65$, $p = 0.001$, $CLES = 0.73$), "Agent Plan" (SA3) group ($T = 4.59$, $p = 0.001$, $CLES = 0.72$) and "Threat Prioritisation" (SA2) group ($T = 3.65$, $p = 0.004$, $CLES = 0.68$). In addition, participants scored significantly higher Relative Threat Precision scores in the "Missile Path" (SA2) group compared to participants in the "Threat Shape" (SA1) group ($T = -4.28$, $p = 0.001$, $CLES = 0.29$) and in the "Missile Path" (SA2) group compared to participants in the "Agent Plan" (SA3) group ($T = 4.22$, $p = 0.001$, $CLES = 0.71$).

7.4.1.2 Visual Agent and Aiming Agent

Overall, participants scored higher Relative Threat Recall scores in sessions where an aiming agent was present (with or without visual agent) compared to sessions with only a visual agent and no aiming agent. Looking at Relative Threat Recall scores (see Figure 7.4) we can observe increases in sessions with an aiming agent *and* a visual agent compared to sessions with only an aiming agent for the "Priority Number" and "Threat Shape" (SA1) visual agents groups. All other groups (SA2 and SA3) present lower Relative Threat Recall scores when a visual agent and an aiming agent are present, compared to sessions with only an aiming agent. For Relative Threat Precision scores (see Figure 7.5), participants in sessions with an aiming agent and visual agent scored higher than in sessions with only an aiming agent for the "Threat Shape" (SA1) group.

While performing between-groups comparisons on sessions with a visual agent and an aiming agent, a Welch ANOVA test on Relative Threat Recall ($F = 2.49$, $p = 0.0334$, $np^2 = 0.04$) yielded significant results, but no further significant results were found during pairwise comparisons.

While performing between-groups comparisons on sessions with an aiming agent and a visual agent, an ANOVA test on Relative Threat Precision ($F = 7.85$, $p < 0.0001$, $np^2 = 0.10$) yielded significant results while pairwise comparisons using Tukey tests indicate that participants in the "Threat Shape" (SA1) group performed significantly worse in terms of Relative Threat Precision than participants in the "Missile Path" (SA2) group ($T = -5.0$, $p = 0.001$, $CLES = 0.26$), "Threat Prioritisation" (SA2) group ($T = -4.57$, $p = 0.001$, $CLES = 0.27$), "Agent Plan" (SA3) group ($T = -4.38$, $p = 0.001$, $CLES = 0.28$), "Performance Graph" (SA3) group ($T = -5.54$, $p = 0.001$, $CLES = 0.23$) and "Priority Number" (SA1) group ($T = 3.93$, $p = 0.0014$, $CLES = 0.69$).

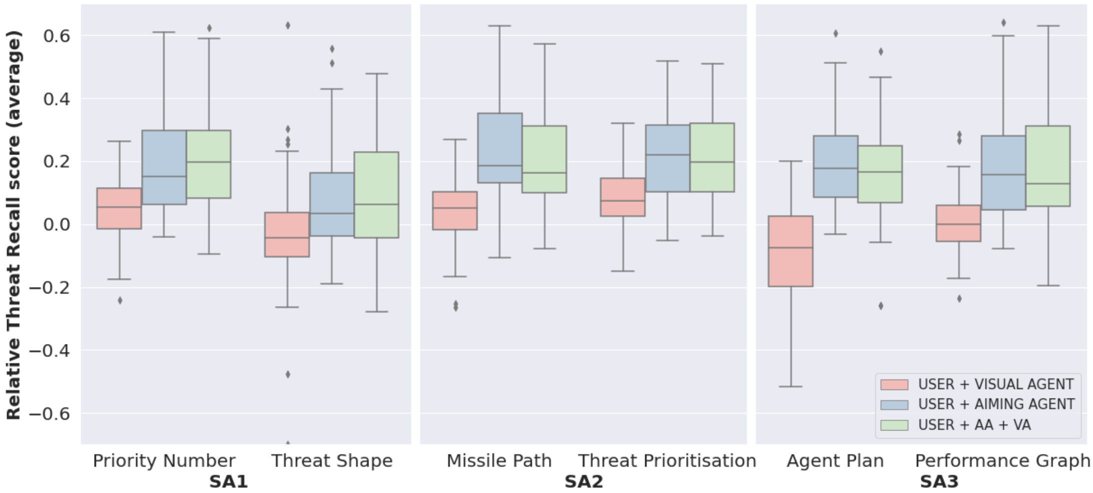


Figure 7.4: Relative Threat Recall scores for each visualisation, organised by Situational Awareness (SA) levels. A positive Relative Threat Recall score indicate a better performance in terms of Threat missiles hit (True Positive) compared to the baseline session without the help of any agent or visualisations. Overall, the addition of an aiming agent was the main reason for increases in Relative Threat Recall scores.

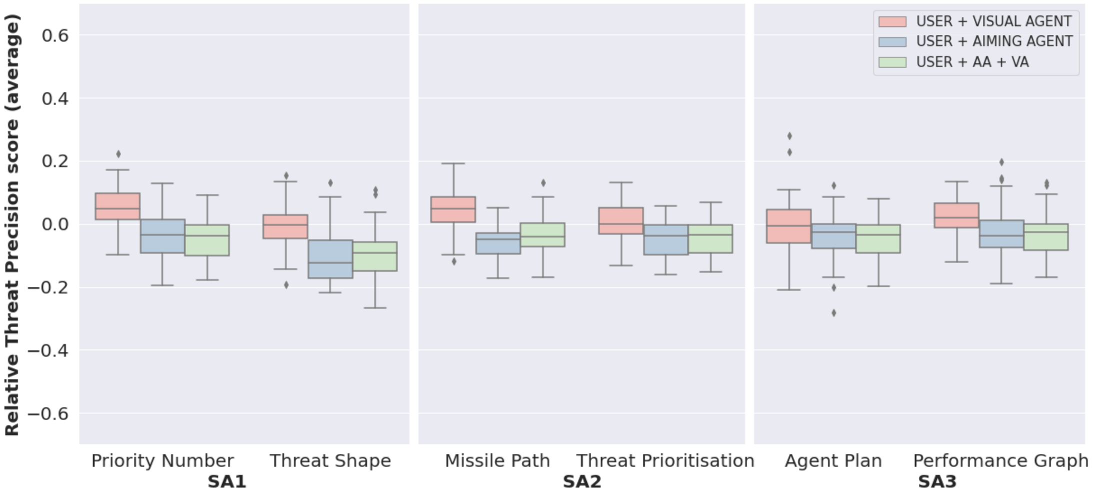


Figure 7.5: Relative Threat Precision scores for each visualisation, organised by Situational Awareness (SA) levels. A positive Relative Threat Precision score indicates a more efficient ratio of missiles hit (True Positive) to shots fired compared to the baseline session without the help of any agent or visualisations. As opposed to Relative Threat Recall scores, the addition of an aiming agent led to comparatively poorer Relative Threat Precision scores.

7.4.2 Reliance

Figures 7.6 and Tables 7.1, 7.2 and 7.3 present scores related to user control time, which are used as a proxy for measuring reliance on the aiming agent. User control time is measured as the duration for which participants manually controlled the crosshair. When an aiming agent is present, a higher user control time indicates lower reliance on the agent. As user control time is a useful means of understanding reliance on the aiming agent, following description of results will focus on conditions where an aiming agent was present.

7.4.2.1 Visual Agent and Aiming Agent

From consulting Figure 7.6, we can observe that participants controlled the crosshair less when aided by an aiming agent, with or without a visual agent. Nonetheless, differences in reliance were observed when comparing groups in sessions where an aiming agent was present. For instance, participants relied on the aiming agent more in sessions with a visual agent for the "Priority Number" (SA1) and "Agent Plan" (SA2) groups. All other visual agents (from SA1 to SA3), however, induced a lower reliance on the aiming agent.

While performing between-groups comparisons on sessions with an aiming agent and a visual agent, a Kruskal Wallis tests yielded significant results for user control time ($H = 21.99$, $p = 0.0005$). Further pairwise comparisons using paired T-TESTS indicate that participants relied on the aiming agent significantly more in the "Agent Plan" (SA3) group than in the "Threat Shape" (SA1) group ($U = 2246$, $p = 0.0004$, $CLES = 0.69$), "Missile Path" (SA2) group ($U = 2113$, $p = 0.0015$, $CLES = 0.67$) and "Threat Prioritisation" (SA2) group ($U = 2030$, $p = 0.0039$, $CLES = 0.66$). In addition, participants relied on the aiming agent significantly more in the “Performance Graph” (SA3) group than in the “Threat Shape” (SA1) group ($U = 2120$, $p = 0.0026$, $CLES = 0.66$).

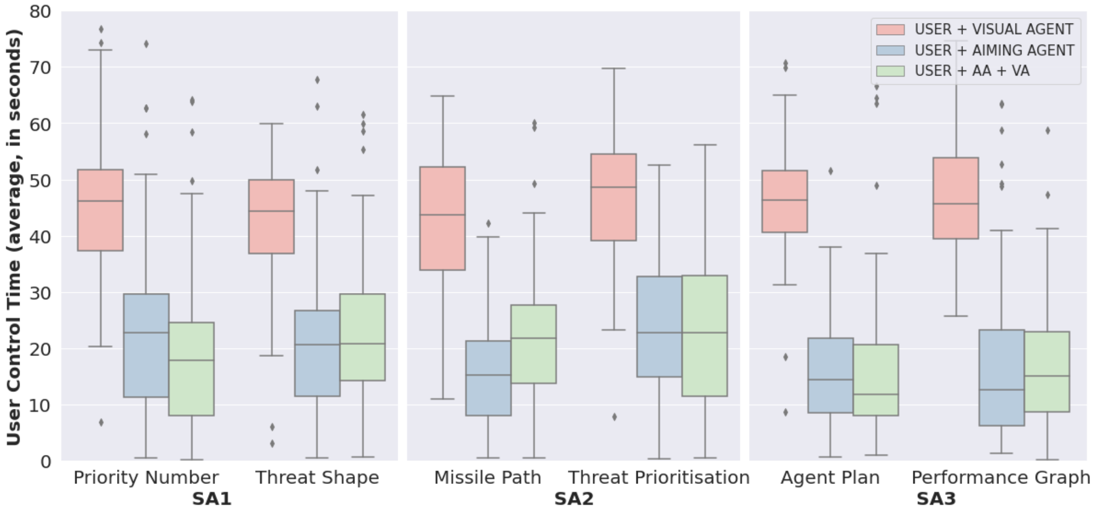


Figure 7.6: Amount of time spent controlling the crosshair during each session, organised by Situational Awareness (SA) levels. Positive scores indicate more manual control and lower reliance on the agent. Unsurprisingly, the addition of an aiming agent led to overall lower user control times.

Table 7.1: Average scores for Performance and Reliance metrics yielded by participants in each group related to Situational Awareness level 1. Scores are presented for each session with or without an aiming agent and/or with or without a visual agent (noted as “VA” in the table). Higher scores indicate a better performance or lower reliance on the agent.

	No Visual Agent		**Threat Shape (SA Level 1)**		**Priority Number (SA Level 1)**	
	User Only	**User + AA**	**User + VA**	**User + AA + VA**	**User + VA**	**User + AA + VA**
Recall	0.56 ± 0.02	0.73 ± 0.01	0.55 ± 0.02	0.73 ± 0.01	0.57 ± 0.02	**0.77 ± 0.01**
Precision	0.64 ± 0.01	0.74 ± 0.01	0.67 ± 0.03	**0.77 ± 0.02**	0.60 ± 0.01	0.75 ± 0.01
F1	0.60 ± 0.01	0.73 ± 0.01	0.60 ± 0.02	0.74 ± 0.01	0.58 ± 0.02	**0.76 ± 0.01**
Relative Recall	0.00 ± 0.00	0.17 ± 0.01	-0.01 ± 0.02	0.17 ± 0.02	0.01 ± 0.01	**0.21 ± 0.02**
Relative Precision	0.00 ± 0.00	0.11 ± 0.01	0.02 ± 0.03	0.12 ± 0.03	-0.03 ± 0.01	**0.13 ± 0.01**
Relative F1	0.00 ± 0.00	0.13 ± 0.01	-0.02 ± 0.02	0.12 ± 0.02	-0.01 ± 0.01	**0.18 ± 0.01**
Threat Recall	0.71 ± 0.02	0.85 ± 0.01	0.72 ± 0.03	0.85 ± 0.01	0.72 ± 0.02	**0.88 ± 0.01**
Threat Precision	0.89 ± 0.01	0.82 ± 0.00	**0.92 ± 0.01**	0.83 ± 0.01	0.90 ± 0.01	0.80 ± 0.00
Threat F1	0.78 ± 0.01	0.83 ± 0.01	0.79 ± 0.02	0.83 ± 0.01	0.78 ± 0.02	**0.84 ± 0.00**
Relative Threat Recall	0.00 ± 0.00	0.14 ± 0.02	-0.01 ± 0.03	0.11 ± 0.03	0.04 ± 0.01	**0.20 ± 0.02**
Relative Threat Precision	0.00 ± 0.00	-0.07 ± 0.01	-0.01 ± 0.01	-0.10 ± 0.01	**0.05 ± 0.01**	-0.05 ± 0.01
Relative Threat F1	0.00 ± 0.00	0.05 ± 0.01	-0.04 ± 0.02	0.01 ± 0.02	0.05 ± 0.01	**0.10 ± 0.02**
User Control Time	**45.62 ± 1.13**	22.42 ± 1.46	43.22 ± 1.61	23.81 ± 1.87	45.42 ± 1.68	19.61 ± 2.13
Relative User Control Time	**0.00 ± 0.00**	-23.97 ± 1.30	-0.24 ± 1.31	-20.27 ± 1.65	-1.70 ± 1.15	-28.96 ± 1.58
User Correction	n/a	11.12 ± 0.63	n/a	**12.87 ± 0.77**	n/a	8.72 ± 0.84

Table 7.2: Average scores for Performance and Reliance metrics yielded by participants in each group related to Situational Awareness level 2. Scores are presented for each session with or without an aiming agent and/or with or without a visual agent (noted as "VA" in the table). Higher scores indicate a better performance or lower reliance on the agent.

	No Visual Agent		**Threat Prioritisation (SA Level 2)**		**Missile Path (SA Level 2)**	
	User Only	**User + AA**	**User + VA**	**User + AA + VA**	**User + VA**	**User + AA + VA**
Recall	0.54 ± 0.02	**0.78 ± 0.01**	0.62 ± 0.02	0.78 ± 0.01	0.53 ± 0.02	0.73 ± 0.01
Precision	0.60 ± 0.01	0.76 ± 0.01	0.64 ± 0.02	**0.78 ± 0.01**	0.65 ± 0.02	0.75 ± 0.02
F1	0.56 ± 0.01	0.76 ± 0.01	0.62 ± 0.02	**0.78 ± 0.01**	0.57 ± 0.02	0.73 ± 0.01
Relative Recall	0.00 ± 0.00	**0.23 ± 0.01**	0.05 ± 0.01	0.21 ± 0.02	0.00 ± 0.01	0.21 ± 0.02
Relative Precision	0.00 ± 0.00	0.16 ± 0.01	0.05 ± 0.02	**0.19 ± 0.02**	0.05 ± 0.02	0.15 ± 0.02
Relative F1	0.00 ± 0.00	**0.20 ± 0.01**	0.05 ± 0.01	0.20 ± 0.02	0.02 ± 0.01	0.18 ± 0.02
Threat Recall	0.65 ± 0.02	0.87 ± 0.01	0.74 ± 0.02	**0.88 ± 0.01**	0.67 ± 0.03	0.84 ± 0.01
Threat Precision	0.85 ± 0.00	0.79 ± 0.00	0.84 ± 0.01	0.80 ± 0.01	**0.90 ± 0.01**	0.82 ± 0.01
Threat F1	0.72 ± 0.01	**0.83 ± 0.00**	0.77 ± 0.02	0.83 ± 0.00	0.75 ± 0.02	0.83 ± 0.01
Relative Threat Recall	0.00 ± 0.00	**0.23 ± 0.01**	0.07 ± 0.01	0.22 ± 0.02	0.03 ± 0.01	0.21 ± 0.02
Relative Threat Precision	0.00 ± 0.00	-0.05 ± 0.00	0.01 ± 0.01	-0.04 ± 0.01	**0.05 ± 0.01**	-0.04 ± 0.01
Relative Threat F1	0.00 ± 0.00	0.11 ± 0.01	0.04 ± 0.01	0.11 ± 0.01	0.04 ± 0.01	**0.12 ± 0.02**
User Control Time	**48.57 ± 1.04**	19.55 ± 1.22	47.48 ± 1.41	23.78 ± 2.05	42.67 ± 1.62	22.73 ± 1.76
Relative User Control Time	**0.00 ± 0.00**	-29.55 ± 1.14	-2.81 ± 1.18	-28.04 ± 1.86	-4.17 ± 1.07	-25.20 ± 1.87
User Correction	n/a	10.97 ± 0.58	n/a	11.57 ± 0.94	n/a	**12.92 ± 0.86**

Table 7.3: Average Scores for Performance and Reliance metrics yielded by participants in each group related to Situational Awareness level 3. Scores are presented for each session with or without an aiming agent and/or with or without a visual agent (noted as "VA" in the table). Higher scores indicate a better performance or lower reliance on the agent.

	No Visual Agent		**Agent Path (SA Level 3)**		**Performance Graph (SA Level 3)**	
	User Only	**User + AA**	**User + VA**	**User + AA + VA**	**User + VA**	**User + AA + VA**
Recall	0.58 ± 0.02	**0.77 ± 0.01**	0.49 ± 0.03	0.74 ± 0.01	0.58 ± 0.02	0.77 ± 0.01
Precision	0.63 ± 0.01	0.70 ± 0.01	0.49 ± 0.02	0.68 ± 0.02	0.69 ± 0.02	**0.76 ± 0.01**
F1	0.59 ± 0.01	0.73 ± 0.01	0.49 ± 0.02	0.71 ± 0.01	0.62 ± 0.02	**0.76 ± 0.01**
Relative Recall	0.00 ± 0.00	**0.19 ± 0.01**	-0.08 ± 0.02	0.17 ± 0.02	-0.00 ± 0.01	0.19 ± 0.02
Relative Precision	0.00 ± 0.00	0.08 ± 0.01	-0.11 ± 0.02	0.08 ± 0.02	0.04 ± 0.02	**0.11 ± 0.02**
Relative F1	0.00 ± 0.00	0.14 ± 0.01	-0.09 ± 0.02	0.13 ± 0.02	0.01 ± 0.01	**0.16 ± 0.02**
Threat Recall	0.68 ± 0.02	0.86 ± 0.01	0.59 ± 0.03	0.84 ± 0.01	0.68 ± 0.03	**0.87 ± 0.01**
Threat Precision	0.83 ± 0.01	0.80 ± 0.00	**0.84 ± 0.01**	0.80 ± 0.00	0.84 ± 0.01	0.79 ± 0.00
Threat F1	0.73 ± 0.01	**0.83 ± 0.00**	0.66 ± 0.02	0.81 ± 0.01	0.73 ± 0.02	0.82 ± 0.00
Relative Threat Recall	0.00 ± 0.00	**0.19 ± 0.01**	-0.09 ± 0.02	0.16 ± 0.02	-0.00 ± 0.01	0.19 ± 0.02
Relative Threat Precision	0.00 ± 0.00	-0.03 ± 0.01	-0.01 ± 0.01	-0.04 ± 0.01	**0.02 ± 0.01**	-0.03 ± 0.01
Relative Threat F1	0.00 ± 0.00	**0.10 ± 0.01**	-0.08 ± 0.02	0.08 ± 0.01	0.01 ± 0.01	0.10 ± 0.02
User Control Time	**49.83 ± 1.01**	16.99 ± 1.31	46.12 ± 1.45	16.67 ± 2.01	47.02 ± 1.49	16.66 ± 1.69
Relative User Control Time	**0.00 ± 0.00**	-33.34 ± 1.32	-4.23 ± 1.29	-33.97 ± 1.73	-2.28 ± 1.04	-32.76 ± 1.74
User Switch	n/a	8.97 ± 0.55	n/a	8.75 ± 0.79	n/a	**9.10 ± 0.82**

7.4.3 Relative Reported Trust

Figure 7.7 and Tables 7.4, 7.5 and 7.6 present relative ratings regarding the statement "I can trust the agent" which was presented at the end of every session. The ratings reflect the reported trust in the aiming agent in every session. The relative trust scores presented here were computed by taking the trust ratings from the sessions with an aiming agent and a visual agent and subtracting them from the trust ratings from sessions with the aiming agent and no visual agent. Positive relative trust ratings indicate that the visual agent had a positive impact on the user-aiming agent interaction.

7.4.3.1 Visual Agent and Aiming Agent

If we consult the results, we can observe that participants' trust levels in visual agents changed the most in sessions where an aiming agent was present. In general, the "Priority Number" (SA1), "Threat Prioritisation" (SA2) and "Performance Graph" (SA3) groups did not report widely different levels of trust in the aiming agent when a visual agent was added. Nonetheless, the addition of a visual agent actually reduced Relative Trust levels for the "Threat Shape" (SA1) and "Missile Path" (SA2) groups, and slightly increased Relative Trust levels for the "Agent Plan" group (SA3). In addition to relative trust scores, table 7.7 shows correlations between reported trust and various behavioural and reported metrics such as task performance, reliance and cognitive load. Looking at the results, we can see that, overall, most independent variables have a low correlation with reported trust scores. Surprisingly, Overall 3D SART - our measure of general situational awareness in this study - has a slightly higher correlation with trust (ρ of 0.262) than Recall - a measure of performance (ρ of 0.257).

While performing between-groups comparisons on sessions with an aiming agent and a visual agent, a Kruskal Wallis tests yielded significant result for Relative Trust scores ($H = 12.31$, $p = 0.03$). Further pairwise comparisons indicate that participants in the "Threat Shape" (SA1) group trusted the aiming agent significantly less than in the "Agent Plan" (SA3) group ($U = 1263$, $p = 0.0039$, $CLES = 0.35$).

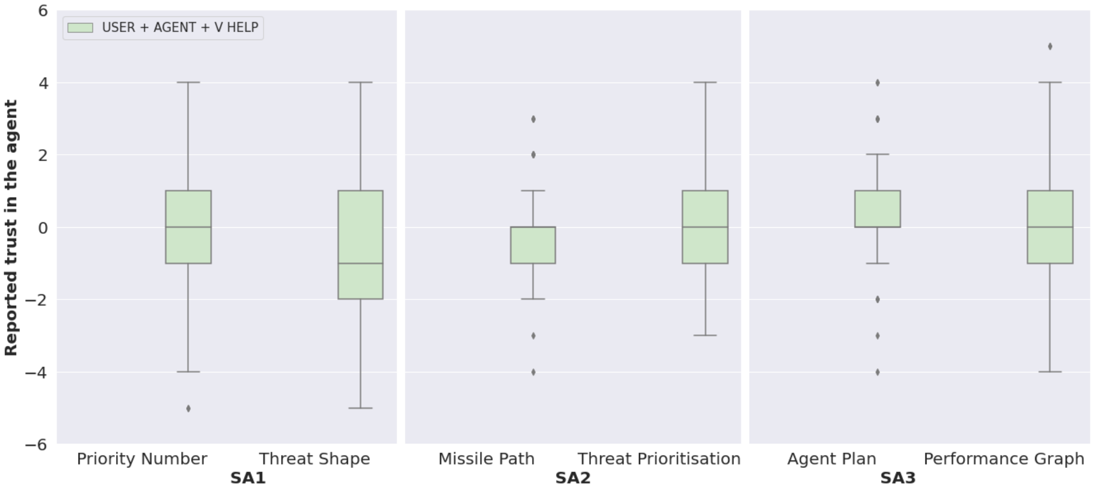


Figure 7.7: Relative ratings reported by participants based on the following statement: "I can trust the agent". Scores are organised by Situational Awareness levels. Higher scores indicate a higher reported trust in the agent. Overall, reported trust levels were the most inconsistent for the "Threat Shape" visual agent.

7.4.4 Relative Cognitive Workload

Figure 7.8 and Tables 7.4, 7.5 and 7.6 present relative Raw TLX scores for each session. Overall TLX scores represent the perceived cognitive workload of a session as perceived by participants.

Higher scores indicate higher relative cognitive workload for a given session, when compared to the baseline session which took place without an aiming or visual agent.

7.4.4.1 Visual Agent without Aiming Agent

As evidenced by Figure 7.8, participants reported slightly lower cognitive workload while supported by a visual agent when compared to baseline sessions without a visual or aiming agent in most groups, but this difference is not statistically significant. However, participants in the "Performance Graph" group (SA3) reported a higher relative workload with a visual agent and no aiming agent, but this difference is not statistically significant.

To further investigate differences in Relative Cognitive load between each group, we performed between-groups comparisons for Relative Overall Raw TLX scores. While performing between-groups comparisons on sessions with a visual agent and without an aiming agent, an ANOVA test on Relative Overall Raw TLX scores did not yield any significant results ($F = 0.43$, $p = 0.82$, $np^2 = 0.01$).

7.4.4.2 Visual Agent and Aiming Agent

If we consult Figure 7.8, we can see the addition of an aiming agent, no matter whether a visual agent was present or not, reduces reported relative cognitive load in all groups. When comparing relative Raw TLX scores in sessions where an aiming agent and a visual agent were both present, relative TLX scores decrease, particularly in the "Threat Prioritisation" (SA2) and "Performance Graph" (SA3) groups, but this difference is not statistically significant.

To further investigate differences in Relative Cognitive load between each group, we performed between-groups comparisons for Relative Overall Raw TLX scores. While performing between-groups comparisons on sessions with an aiming agent and a visual agent, an ANOVA test on Relative Overall Raw TLX scores did not yield any significant results ($F = 1.66$, $p = 0.14$, $np^2 = 0.04$).

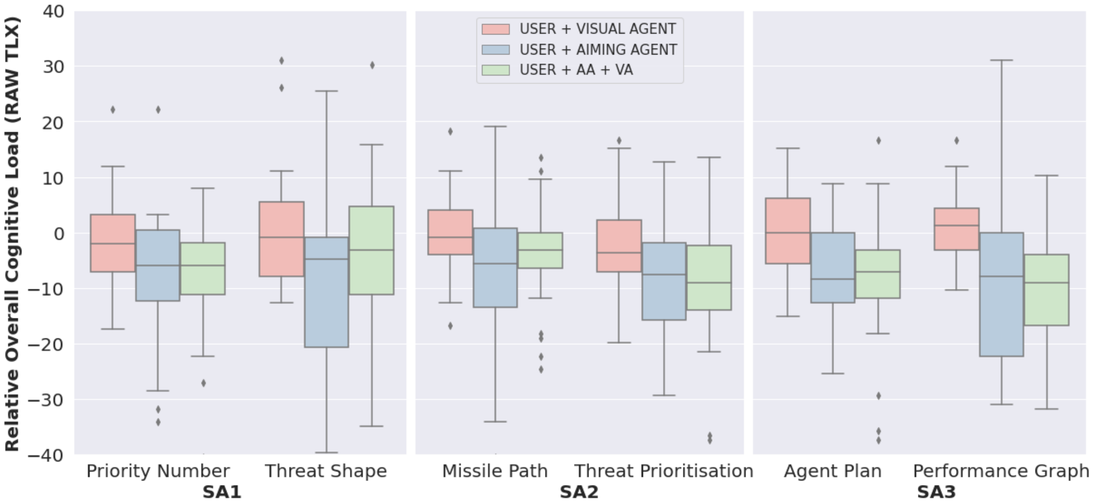


Figure 7.8: Relative scores calculated from the participants' ratings of the the NASA TLX survey to evaluate cognitive load. Relative overall Raw TLX scores are reported here, organised by Situational Awareness (SA) levels. Higher scores indicate a more cognitively taxing experience. Overall, reported Raw TLX scores were lower when an aiming agent was present. The "Threat Shape" (SA1) and "Performance Graph" (SA3) conditions saw the most variance in reported cognitive workload scores.

7.4.5 Relative Situational Awareness

Figure 7.9 and Tables 7.4, 7.5 and 7.6 present Relative 3D SART scores for each session. Positive scores indicate a better reported situational awareness when compared to baseline sessions (without an aiming agent or a visual agent), while negative scores indicate a lower reported situational awareness.

7.4.5.1 Visual Agent without Aiming Agent

Looking at Figure 7.9, we can observe that reported situational awareness varied widely between sessions and SA groups. Without the presence of an aiming agent, the introduction of a visual agent did not affect participants' reported SA in any major way.

To further compare reported overall relative situational awareness scores in each group, we performed between-groups comparisons on Relative Overall 3D SART scores. While performing between-groups comparisons on sessions with a visual agent and without an aiming agent, a Kruskal Wallis test on Overall 3D SART scores did not yield any significant results ($H = 6.51$, $p = 0.26$).

7.4.5.2 Visual Agent and Aiming Agent

For most groups, the addition of an aiming agent improved reported situational awareness. This is true for all groups except "Performance Graph" which actually saw a reduction in reported SA following the addition of an aiming agent (without visual agent). For "Priority Number" (SA1), "Threat Prioritisation" (SA2) and "Performance Graph" (SA3) groups, the addition of a visual agent *and* an aiming agent increased participants' reported SA when compared to sessions with only an aiming agent, but this difference is not statistically significant. For the "Threat Shape" (SA1), "Missile Path" (SA2) and "Agent Plan" (SA3) groups, however, visual agents actually reduced reported SA, most noticeably for the "Threat Shape" group, but this difference is not statistically significant.

To further compare reported overall situational awareness in each group, we performed between-groups comparisons on Relative Overall 3D SART scores. While performing between-groups comparisons on sessions with an aiming agent and a visual agent, a Kruskal Wallis test on Relative Overall 3D SART scores did not yield any significant results ($H = 9.64$, $p = 0.08$).

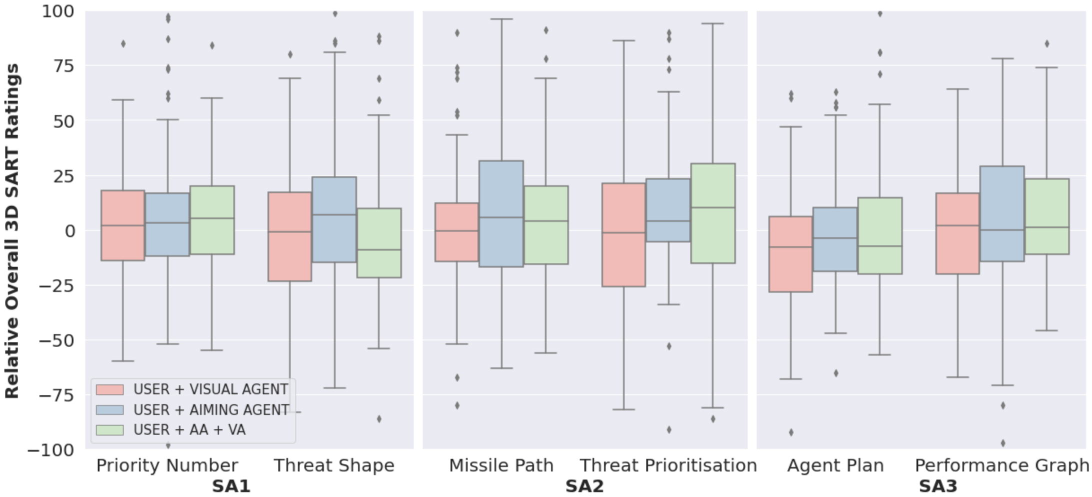


Figure 7.9: Relative ratings collected using the 3D SART survey to evaluate situational awareness. Relative ratings are organised by visualisations intended to support specific Situational Awareness (SA) levels. Higher scores indicate a better sense of situational awareness, as reported by participants. Overall, the addition of an aiming agent and/or a visual agent did not change reported SA scores in any significant way.

Table 7.4: Average ratings or scores reported by participants for every session in the Situational Awareness level 1 groups. Scores are presented for each session with or without an aiming agent and/or with or without a visual agent (noted as "VA" in the table). Depending on the dimensions, a lower or higher score indicates a better or worse experience.

	No Visual Agent		**Threat Shape (SA Level 1)**		**Prioritisation Number (SA Level 1)**	
Question / Statement	**User Only**	**User + AA**	**User + VA**	**User + AA + VA**	**User + VA**	**User + AA + VA**
I can trust the agent	n/a	**4.46 ± 0.14**	3.25 ± 0.21	3.88 ± 0.22	3.72 ± 0.23	4.45 ± 0.22
I can trust the agent (relative)	n/a	**0.00 ± 0.00**	-1.15 ± 0.27	-0.52 ± 0.25	-0.80 ± 0.28	-0.07 ± 0.24
How mentally demanding was the task?	17.05 ± 0.45	14.60 ± 0.49	15.37 ± 0.95	14.23 ± 0.83	**17.07 ± 0.57**	14.20 ± 0.79
How physically demanding was the task?	11.02 ± 0.80	8.57 ± 0.67	**11.40 ± 0.90**	10.50 ± 0.89	9.87 ± 1.23	7.33 ± 0.92
How hurried or rushed was the task?	17.14 ± 0.50	15.03 ± 0.47	15.63 ± 0.78	14.87 ± 0.65	**17.23 ± 0.64**	15.90 ± 0.71
How successful were you in accomplishing your level of performance?	14.14 ± 0.60	12.85 ± 0.56	12.43 ± 0.92	13.13 ± 0.89	**14.77 ± 0.82**	13.33 ± 0.83
How hard did you have to work to accomplish your level of performance?	15.90 ± 0.49	13.63 ± 0.54	15.00 ± 0.68	13.90 ± 0.76	**15.93 ± 0.68**	14.70 ± 0.82
How insecure, discouraged, irritated, stressed and annoyed were you?	13.02 ± 0.83	11.87 ± 0.64	**14.40 ± 0.88**	12.77 ± 0.89	12.67 ± 1.20	10.57 ± 1.15
Overall Raw TLX	**56.51 ± 1.55**	49.17 ± 1.44	54.66 ± 2.36	51.72 ± 2.03	55.93 ± 2.78	49.07 ± 2.41
Overall Raw TLX (relative)	**0.00 ± 0.00**	-7.65 ± 1.72	-1.23 ± 2.77	-4.05 ± 2.73	-1.56 ± 1.60	-8.41 ± 1.86
3D SART - Demand	70.57 ± 1.95	66.15 ± 2.01	65.33 ± 2.93	68.62 ± 2.53	**72.93 ± 2.37**	67.05 ± 2.73
3D SART - Supply	73.73 ± 1.66	74.17 ± 1.78	70.22 ± 2.96	72.88 ± 2.21	**79.05 ± 1.84**	73.28 ± 2.64
3D SART - Understanding	74.17 ± 1.75	75.47 ± 1.59	72.90 ± 2.40	72.62 ± 2.44	74.57 ± 2.59	**78.72 ± 2.00**
OVERALL 3D SART	77.33 ± 3.48	83.48 ± 3.23	77.78 ± 5.18	76.88 ± 4.36	80.68 ± 4.19	**84.95 ± 4.62**
OVERALL 3D SART (relative)	0.00 ± 0.00	6.29 ± 3.39	-1.86 ± 4.50	-3.25 ± 4.27	5.48 ± 4.52	**9.75 ± 4.76**
I know what the agent is trying to do	n/a	4.83 ± 0.14	4.00 ± 0.22	4.40 ± 0.20	4.28 ± 0.23	**5.25 ± 0.20**
I know why the agent is doing what it does	n/a	4.62 ± 0.15	3.83 ± 0.22	4.15 ± 0.20	4.30 ± 0.22	**4.75 ± 0.21**
I know what the agent is going to do next	n/a	3.79 ± 0.15	3.00 ± 0.18	3.25 ± 0.19	3.47 ± 0.22	**4.10 ± 0.23**

Table 7.5: Average ratings or scores reported by participants for every session in the Situational Awareness level 2 groups. Scores are presented for each session with or without an aiming agent and/or with or without a visual agent (noted as "VA" in the table). Depending on the dimensions, a lower or higher score indicates a better or worse experience.

	No Visual Agent		**Threat Prioritisation (SA Level 2)**		**Missile Path (SA Level 2)**	
Question / Statement	**User Only**	**User + AA**	**User + VA**	**User + AA + VA**	**User + VA**	**User + AA + VA**
I can trust the agent	n/a	**4.90 ± 0.13**	4.27 ± 0.22	4.87 ± 0.20	4.75 ± 0.19	4.67 ± 0.18
I can trust the agent (relative)	n/a	0.00 ± 0.00	-0.67 ± 0.23	-0.07 ± 0.19	-0.12 ± 0.24	-0.20 ± 0.19
How mentally demanding was the task?	17.15 ± 0.42	14.53 ± 0.58	17.13 ± 0.56	14.57 ± 0.88	**17.20 ± 0.55**	15.83 ± 0.58
How physically demanding was the task?	**11.97 ± 0.76**	9.47 ± 0.76	11.13 ± 1.07	10.00 ± 1.00	11.63 ± 0.99	11.60 ± 1.03
How hurried or rushed was the task?	17.33 ± 0.46	15.55 ± 0.53	17.00 ± 0.52	14.53 ± 1.01	**17.40 ± 0.70**	17.07 ± 0.65
How successful were you in accomplishing your level of performance?	**13.87 ± 0.66**	12.25 ± 0.61	13.80 ± 0.93	11.20 ± 0.79	12.63 ± 0.94	12.63 ± 0.81
How hard did you have to work to accomplish your level of performance?	**16.73 ± 0.38**	14.33 ± 0.50	16.43 ± 0.61	14.83 ± 0.85	16.57 ± 0.51	14.07 ± 0.61
How insecure, discouraged, irritated, stressed and annoyed were you?	14.35 ± 0.57	12.18 ± 0.71	14.43 ± 0.89	12.23 ± 0.79	**14.90 ± 0.96**	13.37 ± 0.94
Overall Raw TLX	**58.93 ± 1.36**	50.62 ± 1.70	57.78 ± 1.86	49.84 ± 2.20	58.04 ± 2.22	54.55 ± 1.99
Overall Raw TLX (relative)	**0.00 ± 0.00**	-8.31 ± 1.63	-1.51 ± 1.63	-9.44 ± 1.95	-0.53 ± 1.32	-4.02 ± 1.65
3D SART - Demand	74.01 ± 1.76	69.33 ± 1.69	71.98 ± 2.22	65.42 ± 2.98	73.08 ± 2.52	**76.25 ± 2.17**
3D SART - Supply	76.80 ± 1.69	77.71 ± 1.38	74.67 ± 2.93	74.65 ± 2.62	74.37 ± 2.21	**78.12 ± 1.95**
3D SART - Understanding	74.48 ± 1.73	77.50 ± 1.42	73.02 ± 2.33	77.48 ± 1.97	74.20 ± 2.34	**78.68 ± 2.05**
OVERALL 3D SART	77.28 ± 3.28	85.88 ± 2.58	75.70 ± 4.37	**86.72 ± 3.47**	75.48 ± 4.21	80.55 ± 3.80
OVERALL 3D SART (relative)	0.00 ± 0.00	**8.60 ± 3.17**	-2.87 ± 4.65	8.15 ± 5.11	-0.50 ± 4.22	4.57 ± 4.17
I know what the agent is trying to do	n/a	5.42 ± 0.12	4.62 ± 0.22	**5.48 ± 0.19**	5.20 ± 0.20	5.03 ± 0.19
I know why the agent is doing what it does	n/a	5.16 ± 0.14	4.27 ± 0.25	**5.17 ± 0.20**	5.15 ± 0.19	4.70 ± 0.20
I know what the agent is going to do next	n/a	4.30 ± 0.15	3.75 ± 0.21	**4.58 ± 0.21**	4.23 ± 0.21	3.73 ± 0.18

Table 7.6: Average ratings reported by participants on questions related to Trust, Cognitive Load (NASA TLX) and Situational Awareness (SART) in Situational Awareness level 3 groups. Higher trust ratings indicate greater trust in the agent, higher Raw TLX scores a more cognitively taxing experience and higher SART scores a better overall situational awareness.

	No Visual Agent		**Agent Path (SA Level 3)**		**Performance Graph (SA Level 3)**	
Question / Statement	**User Only**	**User + AA**	**User + VA**	**User + AA + VA**	**User + VA**	**User + AA + VA**
I can trust the agent	n/a	4.62 ± 0.15	4.05 ± 0.21	**5.07 ± 0.18**	4.30 ± 0.24	4.85 ± 0.22
I can trust the agent (relative)	n/a	0.00 ± 0.00	-0.63 ± 0.22	**0.38 ± 0.20**	-0.27 ± 0.31	0.29 ± 0.23
How mentally demanding was the task?	16.30 ± 0.53	14.17 ± 0.61	**17.97 ± 0.49**	14.07 ± 0.74	14.87 ± 0.80	12.60 ± 0.96
How physically demanding was the task?	10.38 ± 0.71	8.24 ± 0.77	**10.83 ± 1.05**	9.00 ± 0.96	8.83 ± 1.24	7.30 ± 0.89
How hurried or rushed was the task?	16.25 ± 0.56	14.32 ± 0.62	**17.03 ± 0.65**	14.67 ± 0.80	15.63 ± 0.77	13.37 ± 0.92
How successful were you in accomplishing your level of performance?	13.75 ± 0.59	13.00 ± 0.58	13.40 ± 0.90	11.70 ± 0.72	**16.20 ± 0.57**	12.83 ± 1.00
How hard did you have to work to accomplish your level of performance?	16.53 ± 0.36	13.03 ± 0.60	**16.80 ± 0.43**	14.27 ± 0.63	15.23 ± 0.80	11.37 ± 0.88
How insecure, discouraged, irritated, stressed and annoyed were you?	12.97 ± 0.77	10.54 ± 0.77	**14.67 ± 0.88**	13.07 ± 1.15	11.73 ± 1.19	9.33 ± 1.07
Overall Raw TLX	55.46 ± 1.37	46.93 ± 1.75	**57.72 ± 1.40**	49.76 ± 2.00	53.68 ± 2.02	43.02 ± 2.47
Overall Raw TLX (relative)	0.00 ± 0.00	-8.57 ± 1.67	-0.26 ± 1.38	-8.23 ± 2.12	**0.74 ± 1.13**	-9.92 ± 1.94
3D SART - Demand	73.35 ± 1.80	66.01 ± 2.00	**74.23 ± 2.38**	69.85 ± 2.53	64.25 ± 2.92	62.12 ± 3.26
3D SART - Supply	80.31 ± 1.64	75.54 ± 1.91	**82.48 ± 1.97**	80.67 ± 1.88	73.95 ± 2.92	75.65 ± 2.58
3D SART - Understanding	**80.30 ± 1.70**	78.63 ± 1.59	71.63 ± 3.04	78.67 ± 2.30	72.32 ± 2.90	78.07 ± 2.23
OVERALL 3D SART	87.26 ± 3.08	88.16 ± 2.89	79.88 ± 5.21	89.48 ± 4.37	82.02 ± 4.40	**91.60 ± 4.89**
OVERALL 3D SART (relative)	0.00 ± 0.00	0.50 ± 2.92	-10.15 ± 3.85	-0.55 ± 4.14	-2.47 ± 4.40	**7.12 ± 3.76**
I know what the agent is trying to do	n/a	5.24 ± 0.14	4.35 ± 0.24	5.22 ± 0.18	4.05 ± 0.25	**5.40 ± 0.18**
I know why the agent is doing what it does	n/a	4.75 ± 0.15	4.32 ± 0.23	4.85 ± 0.20	4.15 ± 0.25	**5.23 ± 0.20**
I know what the agent is going to do next	n/a	3.87 ± 0.15	3.63 ± 0.21	**4.67 ± 0.20**	3.67 ± 0.24	4.07 ± 0.24

Table 7.7: Spearman's correlation tests between behavioural or reported metrics and trust ratings. A higher ρ scores indicates a greater correlation.

Parameter 1	**Parameter 2**	ρ	**p-value**
Overall 3D SART	I can trust the agent	0.262	<0.001
Recall	I can trust the agent	0.2569	<0.001
Raw TLX	I can trust the agent	-0.2292	<0.001
F1	I can trust the agent	0.2258	<0.001
Threat Recall	I can trust the agent	0.2221	<0.001
Threat Precision	I can trust the agent	-0.2204	<0.001
Relative F1	I can trust the agent	0.2091	<0.001
Relative Precision	I can trust the agent	0.1994	<0.001
Relative Recall	I can trust the agent	0.1965	<0.001
Relative Threat Recall	I can trust the agent	0.1842	<0.001
Task Difficulty	I can trust the agent	-0.1744	<0.001
User Control Time	I can trust the agent	-0.1633	<0.001
Precision	I can trust the agent	0.1626	<0.001
Relative Threat F1	I can trust the agent	0.1234	<0.001
Threat F1	I can trust the agent	0.1089	<0.001
Extended Gamemode	I can trust the agent	0.0897	0.0032
Relative Threat Precision	I can trust the agent	-0.0852	0.0052
Gender	I can trust the agent	0.0257	0.4141
Age	I can trust the agent	0.0216	0.493

7.4.6 Participants' Feedback

At the end of each session, the "Critical Incident" Technique [62] was used, where participants were asked to write about the positive and negative aspects of each type of visual agent they interacted with. The resulting dataset was given to 3 independent coders who had no prior involvement with the study. More details about coding analysis are available in Section 3.4.5.

Coders were presented with the definition of Chen et al.'s levels of Situational Awareness [26] where level 1 is related to understanding *what* the system is doing, level 2 to "why" the system is acting in a specific way and level 3 to *what is going to happen next.* These codes were summarised as "What", "Why" and "Projection" and were used to code the comments left by all participants. Qualitative coding results are presented in Table 7.8.

Looking by Table 7.8, we can observe that the code "What" (SA level 1) was the most widely used for every type of visualisation, with more than 200 references for each visual agent. This finding is interesting, as even visual agents in the SA level 2 and 3 groups were coded as being helpful to understand "what" was happening, which is supposed to be a dimension supported by SA level 1 visual agents. Despite this anomaly, all visual agents also helped participants understand important elements regarding the current state of the task. However, the associated Kappa scores of the "what" code (SA level 1) were by far the lowest among all codes, which denotes a lack of agreement between coders.

Code "Why" had the second highest number of references throughout the dataset, with more than 100 for each visual agent group. In addition, its associated Kappa scores were found to be consistent around 0.36, which denotes a lack of agreement between coders.

Code "Projection" was the least used throughout the dataset, with around 50 references per visual agent group. Its associated Kappa scores, however, were found to be the highest in the visual agent Group "SA1 Priority Number". Interestingly, the visualisations intended to support SA level 3 (future states) were coded less frequently with the "Projection" code than SA level 1 visual agent groups.

Table 7.8: Results from qualitative coding analysis on post-hoc survey data asking participants to report one positive and one negative thing about each visual agent they encountered. 3 independent researchers then assigned relevant comments to individual SA levels (codes) as described in the framework by Chen et al. ("What", "Why", and "Projection").

visual agent Group	**Code**	**References**	**Agreement score**	**Kappa score**
SA1 Priority Number	*What*	227	57.48	0.05
	Why	127	84.41	0.36
	Projection	50	87.4	**0.67**
SA 1 Threat Shape	*What*	221	59.53	0.07
	Why	133	83.85	0.36
	Projection	52	87.18	**0.47**
SA 2 Missile Path	*What*	226	57.80	0.05
	Why	130	84.09	0.36
	Projection	49	87.15	0.04
SA 2 Threat Prioritisation	*What*	224	59.91	0.05
	Why	132	84.05	**0.36**
	Projection	44	88.65	0.07
SA 3 Agent Plan	*What*	224	59.13	0.06
	Why	135	84.16	**0.37**
	Projection	50	87.42	0.08
SA 3 Performance Graph	*What*	219	59.95	0.06
	Why	132	84.40	**0.37**
	Projection	47	87.78	0.05

7.5 Discussion

In this study, we created different visual agents intended to provide better transparency regarding the environment of interaction or an aiming agent's actions during a collaborative Human-Agent task. Each visualisation was intended to support a specific situational awareness level (1,2 or 3, see the work of Chen et al. [26] for more details) and was informed by previous work in Human Factors and HAI studies [124,139,163]. We analysed and compared the influence of each visual agent on task performance, reliance, trust, cognitive workload and situational awareness. This study sought to answer our fourth Research Question: **How do different types of visual help (designed to elicit different levels of situational awareness) influence the human-agent relationship?** and more specifically the following sub-research questions:

- **RQ4.a:** How beneficial is the introduction of visual agents when users are by themselves (no aiming agent)?
- **RQ4.b:** How beneficial is the introduction of visual agents when users are supported by an aiming agent?
- **RQ4.c:** Which visual agent provides the best overall support?

Our results indicate that different types of visual agent can reduce the uncertainty related to the agent's actions or the task itself, but also overload users and lead to sub-optimal behaviours.

7.5.1 Task Performance and Visual Agent

In this study we introduced new metrics that we named "Threat Recall, Threat Precision and Threat F1" to take into account changes related to the presence of False Positives (hitting missiles that are not going to hit a city) and False Negatives (not hitting missiles that are going to hit cities). In addition, we focused our analysis on "Relative metrics", which represent the relative gain or loss of points for a score in one session when compared against a baseline one (usually, the session without an aiming agent or visual agent).

Overall, changes following the introduction of a visual agent were mostly observed in sessions *without* an aiming agent, as the introduction of an aiming agent resulted in a clear improvement in most performance metrics throughout all groups, with or without the presence of a visual agent, which goes against our first hypothesis, and informs RQ4.b. However, visual agents resulted in some changes in users' performance and behaviours throughout most groups. We posited that visual agents intended to provide more prescriptive visualisations (instructing the user "what to do") would have a more positive impact on behavioural metrics, such as task performance. We found that the opposite may be true true for some visual agents. Overall, we found the most interesting results when analysing Relative Threat Precision scores (see

Figure 7.5), which was to be expected, as Precision scores measure how efficient the user is at a task, and our visual agents were all developed to help participants understand the agent's actions and, in particular, select relevant targets (threat VS non-threats). In terms of Relative Threat Precision scores, participants supported by a visual agent scored the highest in the "Priority Number" (SA1) group. These findings are interesting, as other types of visualisations processing more of the data *for* the participants (prescriptive visualisations - for instance, the "Threat Shape" (SA1) or "Threat Prioritisation" (SA2) visual agents) did not lead to a significant increase in Relative Threat Precision scores (which informs RQ4.a). It is important to note, however, than we did not check whether prescriptive or descriptive visualisations were perceived as such, which is important to bear in mind while interpreting our results. These findings are, surprisingly, at odds with our hypothesis as visual agents intended to help participants understand "what" was happening likely led participants to spend more time processing information and making decisions (for instance, "Priority Number" at SA1 and "Missile path" at SA2) which, in turn, led to increased performance when compared to other visualisations that presented information that was easier to act on (for instance, "Threat Prioritisation" at SA1 or "Agent Plan" at SA3). Nonetheless, descriptive visualisations led participants to gain a better understanding of targets which resulted in fewer false positive errors (higher Relative Threat Precision scores). Other visualisations that focused on processing more data for participants ("Threat Shape" at SA1 and "Threat Prioritisation" at SA2) gave more information regarding the agent's reasoning which induced better performance in terms of missiles hit (higher Relative Threat Recall scores) but made it harder for participants to distinguish between true and false positives (lower Threat Precision scores).

Overall, it seems that visualisations that encouraged participants to make their own decisions ("Missile Path", "Priority Number") as opposed to letting the agent process most of the data, led participants to obtain a more complete understanding of the task and thus perform better at it (which informs RQ4.c). The type of situational awareness level initially associated with each visual agent, however, did not seem to match the actual impact they had on participants, as the "Priority Number" (SA1) and "Missile Path" (SA2) visual agent groups mostly affected participants' understanding of what target to aim at, which appears to be linked to the SA level 1 from the work of Chen et al. [26].

7.5.2 Reliance, Trust and Visual Agent

We measured reliance as the duration for which participants corrected the agent, while trust was measured with rating scales indicating participants' self-reported propensity to trust an agent.

Overall, participants clearly relied on the aiming agent when its help was available, across all conditions, which could be the sign of a more complacent attitude toward the aiming agent where participants were often more likely to rely on its help than trying to improve on its decision-making capabilities. These results are congruent with past work describing complacent behaviours when interacting with automation [154]. In general, reliance did not change in any important way during sessions without an aiming agent, no matter whether visual agent were present or not. However, in sessions with agents, clearer differences could be noticed. In the "Missile Path" (SA2) group, participants relied less on the agent when they could see the trajectory of the targets, which indicates a better awareness of the situation as these changes led to significant increases in performance in terms of Relative Precision scores. The group "Priority Number" (SA1) witnessed the opposite change, with reliance increasing when priority numbers were displayed on each target, which also resulted in better overall Relative Precision scores. These clear changes in Reliance did not translate to changes in Trust when an aiming agent was added, as trust scores in the "Priority Number" and "Missile Path" groups remained constant across most conditions, except for a noticeable decrease in reported trust for the "Priority Number" group when only a visual agent was supporting participants. Overall, these changes show how different types of visualisations influence users' propensity to rely on the system, with a more descriptive visual agent (in our study, the one displaying the paths of missiles) leading to reduced reliance on the aiming agent. However, this decrease in reliance did not seem to alter participants' reported trust in the aiming agent. This difference in trust and reliance could be due to how trust ratings were collected in the study, with no differentiation between trust in the visual help or the agent when both were present, which could explain our inconclusive results.

7.5.3 Overloading Users

In this study, Cognitive Workload and Situational Awareness were measured via post-task survey instruments such as NASA TLX and the 3 item SART instruments. These metrics help us understand how complex the situation was perceived to be by participants, and how much they understood about it. Compared to metrics collected directly during the task (reliance and performance scores), no significant results were found when analysing reported ratings for any of the visual help groups. Nonetheless, these results still indicate different attitudes towards automation when supported by various types of visual agent.

Overall scores for Cognitive Workload did not undergo significant changes between groups, and noticeable differences were only seen between sessions with and sessions without an aiming agent, with or without visual agents. This was, however, expected as an aiming agent provided the most assistance with the task, as demonstrated in our previous studies (see Chapters 4, 5 and 6). Analysis of reported situational awareness on the other hand, while not yielding any significant results, highlighted more interesting changes. Overall, participants in the "Threat Shape" group reported better situational awareness (higher SART scores) when they were interacting with an aiming agent and without a visual agent, despite not being statistically significant. The same was true for the "Missile Path" group. These findings are surprising, as visualisations were intended to *support* different levels of situational awareness, and not harm them. In the case of the "Threat Shape" group, it is even more surprising as participants reported a lower overall cognitive load and a higher situational awareness in a session that did not include a visual agent despite, once again, not resulting in a statistically significant difference.

By performing qualitative coding, it became clear that most of the participants' descriptions of the visual agents did not match the situational awareness levels they were intended to support as represented in Chen et al. framework [26]. These results could indicate flaws in the conception of the visual agents, despite having informed their design from relevant past work assessing situational awareness. In addition, agreement scores were low for a lot of visual agents, particularly with the code concerning descriptions associated with "what" was happening on the screen. Our results could indicate than Chen's framework may not most suitable to assess SA in non-safety critical scenarios with a general, non-expert audience, as opposed to past SA work that mostly made use of military-oriented scenarios [26].

7.6 Conclusion

For the purpose of this study, we designed six visualisations based on previous work on situational awareness (SA) and tested their effect on human-agent relationships. Each visualisation was intended to be either *descriptive* (letting users make sense of the information) or *prescriptive* (telling users what to do) while supporting different levels of situational awareness (SA level 1,2 or 3) regarding the agent's actions or context of interaction.

When analysing findings, we found that descriptive visualisations (in our study, "Missile Path" and "Priority Number"), led participants to perform significantly better when no agent was present in terms of hitting true positive targets, compared to sessions where no visual or aiming agent was present. Another more prescriptive visualisation (in our study, "Threat Prioritisation") led participants to hit a significantly higher number of targets. Moreover, different types of visualisations led to significant differences in reliance, where a descriptive visual agent (in our study, "Priority Number") induced more reliance on the aiming agent and a prescriptive visual agent (in our study, "Missile Path") led to less reliance on the aiming agent.

In terms of reported trust, cognitive load and situational awareness, differences were observed between visualisation groups. Often, participants in sessions without aiming agents but supported by a visual agent reported having a lower situational awareness than in sessions without an aiming or visual agent. These findings are surprising, as the addition of a visual agent significantly improved performance in some groups ("Missile Path" and "Priority Number" groups for instance) but ultimately resulted in lower SA. However, no significant differences were found when analysing any reported metrics, indicating that participants' perception of the aiming agent or task itself did not dramatically change with the addition of different kinds of visualisation.

Overall, our findings indicate that users react differently to various visualisations intended to give them more information about a real-time human-agent collaborative task. We found that better performance can be achieved by presenting participants with visualisations that clearly describe a situation, without instructing users what to do. However, no clear differences were found when analysing participants' perception of the agents via survey instruments or qualitative coding, which suggest that the way we designed each visualisation did not result in improvements on areas that they were designed to support. Furthermore, these findings indicate that longer interactions might be required for participants to be aware of their changing behaviours towards an agent, and that there is a mismatch between participants' interactions within a task and their reported perception of it.

Part III

Final Discussion and Conclusion

Chapter 8

Discussion

In this chapter, we summarise and discuss the implications of our study findings and present a series of recommendations to support future HAI research and agent designers. As we have seen in Chapter 2, collaborative agents are being used more and more often in environments where users and automated agents have to make quick decisions under various levels of uncertainty [143, 149]. This increased use of agents calls for more research on the elements most likely to influence the development of human-agent collaboration. Some of these features have to do with either the agent itself (its level of reliability, behaviour) or the context of interaction (adversarial conditions, transparency of the agent's actions).

In this thesis, four user studies (presented in Chapters 4, 5, 6 and 7) were conducted using a collaborative human-agent aiming framework detailed in Chapter 3. In our framework, we varied agents' reliability, error patterns and visual elements to understand their effects on users' perception, behaviours and team performance. While the focus of this thesis was on the study of trust in agents, we also looked at behavioural metrics such as reliance and task performance, as well as other reported metrics such as cognitive workload and situational awareness.

8.1 Trust in agents

One of the key focuses of this thesis is the study of reported trust in agents and, in particular, how trust evolves when users interact with agents that display different levels of reliability or behaviours in environments with perfect or sub-optimal access to information. In all studies presented in Chapters 4, 5, 6 and 7, trust was measured via survey instruments composed of either multiple items (Chapters 4 and 5) or a single-item (Chapters 6 and 7). While using multiple-items surveys such as the "Checklist for Trust in Automation" by Jian et al. [90] is useful in understanding specific elements related to trust (such as perceived deceptiveness or reliability), we mostly relied on single-item instruments in order to reduce task interruption and provide a more seamless experience. In particular, we relied on a single-item rating scale derived from the work of Jian et al. [90] (the statement "I can trust the agent") which participants could rate from 1 (low trust) to 7 (high trust).

8.1.1 With regards to Agent Reliability and Error Types

In all studies in this thesis, we ensured that agent reliability (how "good" the agent is at the task) was controlled and maintained at fixed levels in terms of Recall (capacity to hit targets) or False Positive rate (capacity to hit relevant targets - which is only applicable for the study presented in Chapter 7). Unsurprisingly, we found that reported trust in an aiming agent was higher when the agent's reliability was high. This finding was apparent from our first user study (see Chapter 4) where high agent reliability with high agent predictability led participants to trust it significantly more than similarly performing but less predictable agents. These results imply that repetitive automation failures can give a sense of consistency to users, and allow them to better prepare for upcoming errors, which is similar to findings from the work of Fan et al. [58] who experimented with systematic errors in a multi-agent environment. In addition, reported trust scores were found to have a higher correlation with reported cognitive load than any of our performance metrics, indicating that the mental load of a task can be a better predictor of trust than even task performance, which tend to support previous work who found a high inverse correlation between reported trust and cognitive workload [2].

For our follow-up studies (see Chapter 5 onward), we chose to stop integrating "low reliability" agents. These, we reasoned, do not represent realistic use-cases for studying human-agent interaction as they are likely to be underused by users anyway (see results in Chapter 4). In our second user study (see Chapter 5), we tested the impact of different types of agent errors or "behaviours" and the way in which they changed users' willingness to trust an agent and rely on its help. We found that agents making aiming errors (defined as errors of "slips" - errors of commission) were perceived as more trustworthy than agents "forgetting" to aim at targets (defined as "lapses" - errors of omission) or agents that aimed at the wrong targets (defined as "mistakes" - errors of intention). Of course, all-error prone agents were perceived as being less trustworthy than the baseline 80% accuracy agent. These findings imply that users find an agent more trustworthy when the agent takes the initiative and clearly shows that it knows what to do (i.e. which target to aim at) even if it is ultimately unsuccessful at the task. Our findings seem to indicate that ascribing intent to an automated agent is something that users tend to do naturally when engaging with automation, even in a goal-oriented task where agents display no anthropomorphic indication of their motivations. Furthermore, our results add to a growing body of studies attempting to explain and/or leverage anthropomorphic features [95], often to favour a continuously calibrated level of trust in the agent, where users' expectations match the actual capabilities of the agent, as described in the work of Merritt et al. [125].

8.1.2 With regards to visual uncertainty and agent transparency

In the studies presented in Chapter 6 and Chapter 7, we tested different types of adversarial visual conditions on users. In both studies, aiming agents were set to have a 80% accuracy.

We designed adversarial conditions to induce uncertainty in the task by preventing users from clearly seeing targets. This was achieved by including either dynamic or static occlusions that were partial (part of the screen) or near-total (nearly the entire screen). We found that users trusted an agent the most when the occlusion was total, which was to be expected as the agent continued to function no matter the kind of visual uncertainty present. A more surprising finding, however, was that participants preferred to "blindly" trust the agent's recommendation under the highest levels of uncertainty, even though it led to poorer task performance. This finding was particularly interesting as users could, at any moment, correct the agent and get a better understanding of the situation before making a decision. Our results underline the need to promote visualisations that enhance agent transparency by informing about "mistaken uncertainties" (when an agent is misunderstood) and "unaware uncertainties" (when users are missing an important information), as presented in the work on uncertainty and trust in visual analytics by Sacha et al. [149].

In the study presented in Chapter 7, we added different types of "visual agent" which were intended to explain the aiming agent's actions through visualisations highlighting either the task (which elements are important) or the aiming agent's actions (why the agent is doing something). Each visualisations was inspired from past SA-related work and based on the framework by Chen et al. [26]. As this study used a between-groups design where every participant experienced one of six potential types of visual agent, we focused on "relative metrics" which describe the relative gain (or loss) that visual agents were responsible for during the human-agent interaction. Overall, we found that reported trust in the agent did not change when participants were supported by a visual agent. Nonetheless, we found that displaying the "path" of an agent (i.e. the agent's future plan of action) improved users' trust the most, while displaying which targets the agents thought were relevant or not actually reduced trust the most, however these results were not statistically significant. These findings are surprising, as more information and transparency about an agent's reasoning process should improve reported trust in it, or at least help users to better calibrate it, which contrasts with past work on automation transparency and uncertainty communication in HAI [101]. These results could be explained by the type of transparency provided to users, as participants tended to trust and rely on visualisations that presented information intended to be more transparent ("descriptive" visualisations) more than visualisations that were intended to simply tell users what to do ("prescriptive" visualisations). However, during analysis of qualitative feedback from participants, we found that most visualisations were not perceived as supporting their intended SA level as defined by the framework of Chen et al. [26], which could also indicate a flaw in the design of our visual agents, despite being based on previous HAI work.

8.2 Reliance on agents

Reliance, in addition to trust, is an important metric in understanding how users perceive automated agents. Where reported trust measures users' subjective perception of an agent, reliance, studied via a behavioural proxy, represents users' actual interaction with an agent [103]. In all our studies presented in this thesis, we studied reliance by recording the amount of time (in seconds) for which participants assumed responsibility for the controls when interacting with an aiming agent. This measure, that we call "User Control Time", is a direct behavioural proxy that gives an estimate of how much users actually relied on the aiming agents, where a higher user control time represents lower reliance on the agent, and vice-versa.

In our first study investigating the effect of agent reliability and predictability on users (see Chapter 4), we found that reliance was positively affected by more reliable *and* predictable agents. We observed that participants relied significantly more on agents with high levels of reliability and high levels of predictability than agents with high levels of reliability but low levels of predictability. These findings showed that greater reliance on an agent can indeed be correlated with higher trust, as seen in Section 8.1. It also means that when users can more easily predict the actions of an agent, they tend to correct it more efficiently as well as being more willing to rely on its future input, as they know how to anticipate its potential failures. These results are largely coherent with past work that linked a higher ability to predict an agent's actions with a better calibration of users' reliance [130].

In our follow-up study (see Chapter 5), we designed different types of agent behaviours and tested their impact on users' reliance. We found that errors types, much like levels of predictability, did have an important influence on reliance. Through our results, we learned that participants relied significantly less on agents that were committing errors of judgement (defined as "mistakes" in our study) and commission (defined as "slips") than agents committing errors of omission (defined as "lapses"). These results indicate that, as with trust, users ascribe intents to agents, and notice patterns in the way they make errors. In return, these differences influence the way participants rely on agents. For instance, we found that users were more likely to rely on agents that chose not to take any action ("lapses") than those that took a decision and made an error ("slips" or "mistakes"). This finding represents an interesting avenue to pursue if we are to understand reliance on faulty automation, since errors of omission seem to have a less destructive impact on users' perception of agents, and might be more easily salvageable than other kind of agent errors. Our findings seem to be in line with past work, such as a study by Sanchez et al. [151] which found that a high amount of "false alarms" (which is similar to the "slips" and "mistakes" error types in our experiment) is linked with decreases in reliance compared to other error types.

In the study presented in Chapter 6, we experimented with different types of "visual uncertainty" that occluded information from users to see how reliance on agents would be affected.

We observed that users tended to rely less on agents when crucial parts of the environment of interaction were occluded (in our study, the top of the screen, where targets appear from). This finding indicates that the introduction of an agent does not necessarily makes users more complacent, and that some kind of adversarial visual condition could even lead to improvements, as users performed better while relying less on the agent (in the session where the top of screen was hidden, for example). Conversely, other conditions where most elements were occluded did lead to changes in reliance albeit to the detriment of task performance, which underlines that *calibrated reliance* is just as important as *calibrated trust* in an agent is important for effective human-agent collaboration. These results add to past work that found reliance to be better calibrated in situations with visual uncertainty compared to situations with more readily available visual information [101].

In the study presented in Chapter 7 (and as opposed to the study presented in Chapter 6) we evaluated different types of visualisation intended to provide more transparency regarding the agent's actions or detailing important elements in the environment of interaction. Interestingly, while trust and performance were affected by the type of visualisation used, reliance evolved differently depending on the kind of visual agent present. For instance, participants that interacted with descriptive visual agents (intended to show more information about the task) tended to rely more on the aiming agent while the use of prescriptive visual agents (intended to tell participants "what to do") resulted in lower reliance on the agent. These findings could further highlight the situational and contextual nature of reliance, as reliance is heavily influenced by both the amount of information available during the task (used to make informed decisions) and the knowledge gained about the agent and its usefulness in a given situation. While further work is required to discuss the merits of specific visualisation types regarding reliance on an agent, our study contrast findings from previous work that found visualisations to significantly change (either positively or negatively) the way users rely on automation [101].

8.3 Task Performance

All of the lab-based and online user studies presented in this thesis were conducted using the framework described in Chapter 3. Despite differences in research questions and study goals, all studies were comprised of the same goal-oriented scenarios where participants had to protect cities from missiles that appeared at the top of the screen. As the task was goal oriented, we designed a set of metrics to evaluate participants' success in terms of performance. More precisely, we used the amount of shots fired, missiles hit and shots missed to compute Recall, Precision and F1 scores (more details in Section 3.4.3.1).

In our first experiment (see Chapter 4), we studied how agent reliability and predictability affect performance. Overall performance (in terms of missiles hit - Recall and ratio of shots fired to missile hit - Precision) was significantly better with more predictable agents, operating at a high level of reliability. Our findings indicate that an agent perceived as being more trustworthy and reliable was also linked to higher task performance. While it is hard to say precisely what influenced participants' perception of agents the most, higher task performance may have been a result of agents being easier to correct thanks to more predictable error patterns. When comparing correlations with performance metrics, Recall scores yielded a higher positive correlation with trust ratings than Precision score. Overall, these findings indicate that participants were perhaps more sensitive to the number of missiles hit (expressed with Recall scores) than the ratio of missiles hit to shots fired (expressed with Precision scores). These results are in line with past work that noted a strong correlation between trust and performance [25]. Nonetheless, in our experiment, we found that reported cognitive workload was more strongly correlated with trust ratings than any performance metrics.

In our second experiment (see Chapter 5), we compared the impact of different agent error types on participants. Of all the error-prone agents, we found that errors of omission (defined as "lapses" - where the agent simply does not do anything) actually had a *positive* impact on participants' performance in terms of Precision and F1 scores when compared against all other sessions, with error-prone agents *or not*. These results are surprising, as we expected errors to be a hindrance to participants, no matter their type or the way we designed them. Instead, it seems that errors of omission can lead users to be more focused on a task, as a closer monitoring of the agent's inaction is required. As a result, these lapses might have been more readily spotted and participants managed to react in a more timely manner which led to increased performance, especially in terms of accuracy with higher precision and F1 scores. Our findings add to a growing body of work on automation errors across a wide range of domains, such as Human-Robot Interaction where erratic automation errors (for instance, an unusual request or behaviour) has been shown to make users more careful with their future agent interactions [150].

In our third experiment, different types of visual uncertainties were designed and tested on human-agent collaboration (see Chapter 6). Unsurprisingly, we found that the higher the uncertainty (where most of the screen is occluded), the worse participants' performance. In general, hiding the top of the screen (which reduced participants' time to detect, aim at and hit targets) significantly *improved* performance. These results might indicate that, as we have seen in the context of users interacting with agents prone to errors of omission, better task performance can be fostered by reducing the amount of time users have to react and make decisions, which can in turn help maintain good attention levels and, as a result, lead to higher task performance. Our findings contrast with past work that has shown automation uncertainty or sensor-related uncertainty to always negatively affect task performance [78, 149].

In our fourth experiment (see Chapter 7) we designed and experimented with different visual agents displaying task-related information. We found that visualisations intended to be more descriptive (highlighting important elements) led to significant improvements in terms of relative performance (when compared to baseline sessions, without visualisation). These findings indicate that participants are likely to perform better while supported by visualisations intended to focus on increasing task transparency than while supported by visualisations intended to "tell them what to do". As with visual uncertainty, this kind of information transparency could be leveraged to foster increased levels of attention from users, while allowing them to perform better thanks to an added understanding of which elements require monitoring. These results add to an ever increasing body of work on visual analytics, where different solutions are proposed to support HAC in high-workload scenario where maintaining a high level of task performance is paramount [101].

8.4 Cognitive Workload

In all studies conducted and presented in this thesis, we used the NASA TLX rating scales [77] to evaluate reported cognitive workload. This survey instrument is the most widely used [75] method to evaluate cognitive workload thanks to its non context-specific nature and ease of administration. The higher the score reported, the more cognitively taxing the task is perceived to be. In all our studies, we focused on reporting *Raw TLX* which consists of an aggregate score (from 0 to 100) denoting the overall cognitive cost of a task.

In our first and second user studies (see Chapter 4 and 5), we tested different types of agent behaviour in terms of reliability, predictability (Chapter 4) and error types (Chapter 5). As expected, we found that higher agent reliability led to a significantly lower cognitive workload, and that the same was true for agent predictability, due to the ease with which users could predict errors and make adjustments accordingly. In addition, we found that reported cognitive workload tended to be higher when an agent makes a wrong decision (defined as "slips") than when it made no decision at all (defined as "lapses"). These results indicate that, assuming a high level of agent reliability, being able to better predict the actions of an agent will likely lead to a lower reported cognitive workload. Interestingly, in one of our studies (presented in Chapter 4), cognitive workload had the highest correlation with reported trust ratings, even higher than the task performance metrics (Recall, Precision F1). These findings highlight how important task complexity and the perceived complexity of interacting with an agent are regarding the development of cognitive workload, which is in line with past work that found a lack of agent transparency to have a negative impact on the development of cognitive workload, regardless even of task complexity [26].

In the third and fourth user studies presented in, respectively, Chapter 6 and 7, we tested different types of visual uncertainties on users and evaluated the benefits of visualisation displaying information about either the agent's actions or environment of interaction. More precisely, in our third user study, presented in Chapter 6, we tested different modalities of visual occlusions (dynamic, static, partial or total). Overall, we did not find any significant differences between them in terms of reported cognitive workload. These findings could be explained by the inherent complexity of the task, which led to high levels of reported cognitive workload even in sessions without any type of visual uncertainty. Similarly, in our fourth user study (see Chapter 7), we found that our different types of visualisation did not influence participants' reported cognitive workload. These findings were consistent throughout the study despite changes in either the intended type of visual agent (descriptive or prescriptive) or nature of visualisation used (see summary of the visual agents used in Section 7.3.1). Overall, and as with our study focusing on visual uncertainty (see Chapter 6), one could argue that participants' cognitive workload was most affected by task difficulty, and that visualisations did not significantly affect reported cognitive workload, despite other improvements in terms of reliance on the agent, trust or task performance.

8.5 Situational Awareness

In our studies, Situational Awareness (SA) represents the amount of information that a user can assimilate about what a system is doing (SA level 1), why it is acting in a specific way (SA level 2), and what it will likely do next (SA level 3) [26]. SA is an important concept in HCI studies as it allows for the measurement of users' understanding about a situation or an agent, and can be used to help understand the development of other related constructs such as task performance, reliance or reported trust. Situational Awareness was only assessed in two studies (see Chapter 4 and Chapter 5) and via two survey instruments: SAGAT and SART [53], which limits comparability between studies.

In our third user study (see Chapter 6), we used SAGAT to evaluate SA level 1 ("what" the user understands about a situation) by asking participants to report the number of missiles present in either the top or bottom half of the screen, and compare this to the actual number of missiles present. We found that there was no major discrepancy between conditions in terms of over- or underestimations, and that obscuring the bottom and top half of the screen led, respectively, to more under- and over estimations. Overall, these findings were not particularly insightful as regards to their relationship with other metrics (no clear correlations between SA and other behavioural or reported metrics were found). However, our findings served as a reminder that task complexity coupled with reduced access to important information will likely reduce situational awareness, no matter the kind of obstacle faced. Our findings seem to support past work that found limited access to information and high task complexity to negatively affect SA the most [135].

In our fourth study (see Chapter 6), we designed and experimented with different visual agents whose intended goal was to display more information about either the task or the aiming agent's actions. No matter the visualisations tested, we expected improvement in SA no matter what visualisations we tested, as task transparency is key to higher levels of reported SA [72]. We did not find any significant differences in terms of SART scores between sessions. Our results might indicate that visual agents were not perceived as being helpful by participants, even less so when an aiming agent was present, and that the addition of a visual agent could even be distracting and harmful to task completion. Additionally, our findings could reflect the downside of studying a concept often investigated in safety-critical environments (SART, SAGAT), which may not apply to a different domain of interaction (here, a game-based task). In our studies, we found that SART scores were ineffective at capturing (often small) differences in information gains between conditions. Our findings should serve as motivation to develop a greater range of non-context-specific methods to analyse the development of SA in more details.

8.6 Lab-based and Asynchronous Online Studies

Due to disruptions caused by the COVID19 pandemic throughout 2020 and 2021, we decided to switch from using a lab-based framework (as seen in studies presented in Chapters 4 and 5) to an asynchronous online framework (see Chapters 6 and 7) where participants took part in the experiment using their own computers. While the task remained the same, we would like to discuss the implications of switching from one framework to another for our results and future HCI work in general.

First, the quantity and quality of post-hoc qualitative feedback given by participants tended to be better during lab-based studies, which could have been the result of participants being able to ask questions directly to the lead researcher and referring more easily to particular moments during the experiments, compared to writing feedback on a post-hoc survey form.

Secondly, while asynchronous online studies make it easier to reach a higher number of people, we noticed that lab-based studies require less attention to details during the creation of the experimental flow and interface, as a researcher can be present to guide users throughout the experiment and assist them if needed. This added development time for online studies should be taken into account when weighting the pros and cons of remote and lab-based experiments, as more pilot testing is required to ensure that the flow of an online experiment remains straightforward enough for all participants to complete in time, as well as potentially designing attention checks.

Finally, we recommend for the network and performance of participants' computers to be assessed during and, equally importantly, *before* they take part in the study. To this end, we recommend creating a simple, scaled-down version of the experimental framework to test communication to servers (checking participants' ping to the experiment's servers could be useful) as well as a benchmarking tool (similarly to the one used in all of our online studies) to let prospective participants know if their machines are fit to complete the study or not. This is especially important when the framework is being ran directly from participants' computers.

This benchmark tool can be as simple as a short script using the same framework as the experiment and running a high workload where certain performance thresholds have to be met in order to allow prospective participants to take part in the study.

8.7 Limitations

8.7.1 Experimental Constraints

It should be noted that our work is not without limitations. In the studies presented in Chapters 4 an 5, we explored agent's behaviours and their influence on trust, reliance, task performance, cognitive workload and situational awareness in a goal-oriented, collaborative aiming task. While the selection criteria for the recruitment of participants were similar for all studies, our samples mostly consisted of students between the age of 18 and 30 and might not generalise to other, older, user groups. In addition, in order to ensure that studies could be completed within an hour for lab-based studies and 45 minutes for online studies, a number of constraints were set. We had to design our experiments with either four or five conditions, depending on the focus of each study (for instance, four different agents or types of visual occlusion). The duration for which participants interacted with each condition also had to be constrained. In most of our studies, participants spent between three and six minutes on each condition, which could have limited the amount of time participants had in order to get used to and adapt to different experimental conditions. In particular, it is possible that more time spent interacting with the agents could have helped participants calibrate their trust over time, which could coincided with changes in reliance and task performance. On the other hand, interactions that were too lengthy could have lead to complacency or complete distrust.

Additionally, of the four user studies conducted in this thesis, two took place in a lab environment (see Chapters 4 and 5) while two others were conducted online and asynchronously (see Chapters 6 and 7) due to disruptions caused by the COVID19 pandemic throughout 2020 and 2021. While we used the same framework for all studies, participants in remote online experiments took part using their own computers and keyboards as opposed to a controller in a lab-based environment. While participants' task performance in all studies was found to be comparable, we would like to acknowledge this as a limitation for the inter-comparability of our lab-based and remote studies. Furthermore, while we recorded software-related details about participants' machines such as frame-rate or screen resolution, hardware-level information were beyond the scope of our logging system. Screen-size, among other elements such as screen refresh rate, could be a relevant information to record as it has the potential to affect participants' ability to notice on-screen items and can change the way they perform during the task.

Longitudinal studies could be an interesting avenue that we left to future work. Such studies could focus on different aspects of the human-agent collaboration. For instance, when studying the impact of agent behaviours on users, such studies could assess the amount of time needed for users to feel comfortable enough with an agent, and how long it takes before potentially complacent attitudes set in. Findings from such experiments could further our understanding of the nature of trust calibration by linking it to the evolution of performance levels, reliance and/or reported trust. Furthermore, longitudinal studies including adverse environmental conditions would also be beneficial to the field of HAI. Such experiments could test a wide variety of adversarial conditions affecting the accuracy of the agent's help and/or the knowledge available to users. These studies could be easily contextualised by using scenarios most commonly affected by changes in the weather, terrain or location. While remaining relatively rare, these studies would prove useful to understand the impact of external factors on HAC, which have, for now, mostly been studied in war-oriented scenarios [99].

8.7.2 Domain-specific Constraints

Our framework was designed to assess human-agent collaboration in a task characterised by the following elements:

- A fast-paced task where participants had to rely on their reflexes to perform well.
- Multiple targets to monitor and track at once.
- A task divided into short sessions, the lengths of which depended on the constraints of each study.

We acknowledge that our findings may not generalise to different contexts of interaction, especially those where decisions will have an immediate or long-lasting impact on either the user interacting with the agent or other parties (for instance, civilians involved in the decisions of a bomb-seeking task [143] or practitioners in the medical domain [72]). Nonetheless, our findings and research framework could prove useful to other scientific fields concerned with collaborative decision-making such as autonomous vehicles where challenges involving uncertainty, appropriate reliance and explainability have been long-standing issues, as presented in the meta review of trust-related studies involving autonomous vehicles by Lockey et al. [113]. Indeed, our studies provided some insights regarding how, for instance, specific visualisations could support HAC. Our results could serve as the basis for follow-up studies where the validity of our findings and experimental design are put to the test in more realistic environments.

In addition, participants interacted with agents in tasks for which they had received little training, and which were necessarily short in nature to reduce fatigue. A completely different environment (for instance an asset tracking task [143]) where users' attention has to be sustained during hours, or even days (e.g. a submarine tracking task [28]) could yield different results. Longer interactions could allow users to have a better mental model of the agent and its decision-making, but could also increase the chance of complacent behaviours developing. Participants' expertise is another area that was under-explored in this thesis. While participants' skills and individual levels of performance were controlled by having them complete the task without the help of an agent first, our framework was new to all participants and previous expertise in related tasks was not assessed prior to undertaking the experiment. In more expert settings, the attitude users displayed towards the agent would likely be very different, as new tasks would require users to rethink their methods which could be, in turn, result in under-reliance on the agent, depending on the context of interaction.

We would like to note that the limitations mentioned above do not undermine the main findings of our studies, but we acknowledge that additional investigations are required to understand more precisely the relationship between the different variables linked to trust in agents, as well as how other types of tasks influence this relationship. We leave these avenues open for future work.

8.8 Recommendations

Based on the results gathered from studies conducted in this thesis along with past work in HAI and Human Factors, we present a series of four recommendations for the research and development of future collaborative agents. A summary of the areas we covered in this work as well as future research opportunities is provided in Figure 8.1.

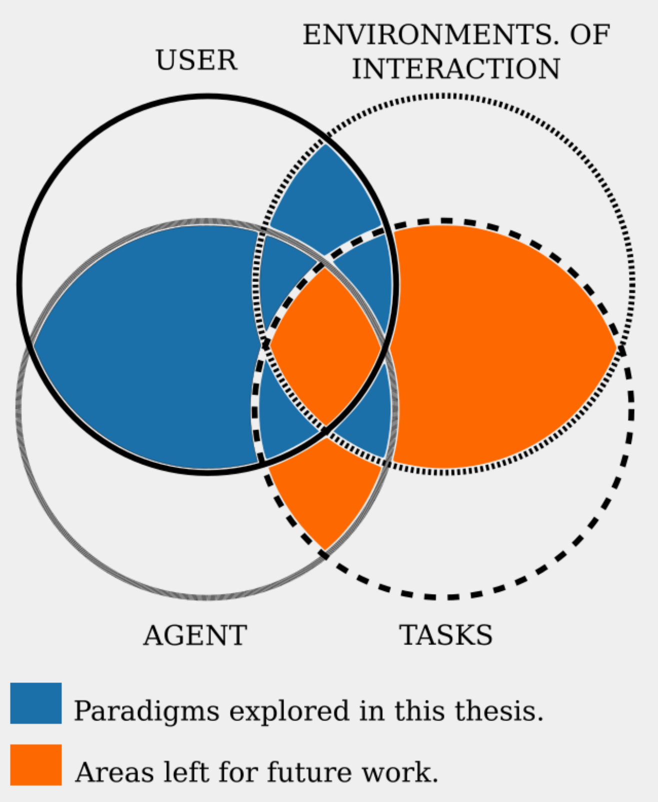

USER
ENVIRONMENTS. OF INTERACTION
AGENT
TASKS
Paradigms explored in this thesis.
Areas left for future work.

Figure 8.1: Venn diagram representing interactions between a user, agent, environment of interaction and type of task. Blue areas indicate which combination this work studied while orange areas are avenues left for future work.

8.8.1 Agent transparency and consistency can be prioritised over reliability

Regarding agent transparency and consistency, our recommendations can be summarised as follows:

- **Beyond a certain level of agent reliability, system makers should prioritise more predictable and easy to understand agent behaviours, rather than maximising reliability at all costs.**

- **Increasing transparency regarding the agent's actions and providing a better on-boarding process should result in better calibration of users' reliance and trust in the agent.**

Evidently, agent reliability is one of the core factors in users' adoption of any system, and this applies to human-agent collaboration: the agent must be reliable "enough" to carry a task successfully and be perceived as helpful [57, 103]. Quantifying how reliable the agent should be in order to be perceived as "good" is, however, difficult, as it largely depends on users' own expertise and the context of interaction, including the likely consequences of any decisions taken. In previous HAI studies, the threshold for acceptable performance was usually set at around 80% accuracy [21, 57], as we found that users tend to ascribe a high reliability to any agent whose input are correct (true positives) between 90 and 70% of the time. While high reliability is essential, we argue that beyond a certain threshold, further increases in reliability can yield diminishing returns, and even harm the Human-Agent team. These side effects could occur if they imply sacrifices in the transparency of the agent's actions or the consistency of its behaviours. As evidenced by our findings in Chapters 5 and 7, assuming agents' levels of reliability are the same, users tend to prefer more predictable. This preference for higher predictability and consistency is evidenced by higher reported trust in agents that would choose not to do anything (false negative error) rather than acting and failing in a situation where the likelihood of agent errors was high. We posit that increased agent consistency makes it not only easier to anticipate an agent's actions but also allows for more flexibility in the case of sudden changes in mission goals or the environment of interaction. More transparent and predictable agent actions would likely further reduce users' cognitive workload by allowing for a more calibrated mental model of the agent's reasoning process. Increasing predictability could also help mitigate adoption and on-boarding issues, as we found that users ascribe fewer negative intents to agents that are more consistent in their behaviours.

While high agent reliability, transparency and consistency are important, there is no alternative to a well calibrated, informed understanding of an agent's capabilities. Making users aware of the weaknesses of an agent, through repeated training and/or by displaying real-time information, is paramount to achieving effective human-agent collaboration and will foster more informed reliance on the agent.

8.8.2 Agent behaviours should be clearly categorised and defined

Regarding agent behaviours, our recommendations can be summarised as follows:

- **The perceived behaviour of an agent in terms of its reasoning process and perceived intent should be clearly elicited and defined in order to anticipate the impact on users' perception of the agent.**

- **A framework of agent behaviours, errors and likely consequences should be created. In particular, this framework should vary the context of interaction to incorporate a wide range of domain-specific findings.**

The type of support and level of autonomy of an agent will influence the way users develop a mental model of the inner workings of the agent [119]. While anthropomorphised agents will be more likely to have intents ascribed to them [7], intent will also be ascribed to agents that don't display human characteristics (such as speech, or a form of physical embodiment). This tendency for users to anthropomorphise systems should be taken into account, if not leveraged, by system makers. Wrongfully ascribing a specific intent to an agent (for instance, about the perception of its "real" intentions) could have harmful consequences. As an example, users could believe that a system is actually working against them which could likely lead to under-reliance issues and defeat the whole purpose of designing a collaborative agent. On the contrary, not being transparent about the limitations of an agent's decision-making process could lead to dangerous over-reliance in situations of high uncertainty.

In our studies, we mostly documented participants' impressions of agents' capabilities via trust-related ratings scales or open-ended post-hoc interviews. We found that not only does the type of agent behaviour have a clear impact on users' perception of the agents, it also influences the way users perform and rely on the agent. These differences are important to take into account, as our studies showed that certain types of error are perceived as being less harmful than others, while not resulting in improved task performance. The work of Marinaccio [118] and Reason [145] pioneered the elicitation and categorisation of agent errors and their consequences for interpersonal or human-agent interactions. We recommend that more work should focus on testing newly defined and existing automation errors in a variety of contexts, from general non-expert scenarios to safety-critical settings requiring expert knowledge. More information on potential differences and similarities of users' perceptions of errors across different domains could allow system makers to better anticipate and mitigate adoption issues, with a selection of repair mechanisms designed to account for an agent's shortcomings.

8.8.3 Taking into account the environment of interaction to better anticipate changes in human-agent collaboration

Regarding the environment in which HAC takes places, we summarise our recommendations as follows:

- **In future HAC studies, the environment of interaction and its features should be clearly controlled and accounted for independently of the task.**
- **As with agent behaviours, a framework of different types of environment of interaction and their features (availability of required information, noise, uncertainty) should be created to compare their influence across a wide range of HAC tasks.**

As we have seen with the behaviour of virtual agents, the environment of interaction is an often under-studied aspect of HAI experiments. In our studies, we found that the environment of interaction also impacts *users* in a significant way, regardless of the actual capabilities of the agent. For instance, in environments with high levels of uncertainty, users are likely to trust an agent more irrespective of its reliability, which subsequently induces more reliance on the agent's input. Additionally, we found that restricting the amount of information required to make informed decisions led participants to react more quickly and, as a result, perform better. We believe that human-agent training could benefit from such findings. In simulated environments, uncertainty and sub-optimal access to important information could be artificially induced in order to train participants to prioritise tasks more effectively and be more critical of the help provided by an agent.

As with past studies investigating and classifying different types of agent errors and their impact on users [118], there is, to the best of our knowledge, no work focused on assessing how different types of tasks, constraints, and environments of interaction in general affect human-agent collaboration. Such a framework could detail how the features of different environments of interaction (such as availability of information, noise or uncertainty level) would affect the outcome of a task as well as the general attitude of a user towards an agent. This work should also carefully account for the risks and consequences of each interaction scenario, as the outcome of a task (in safety critical environments for instance) are likely to influence users' decision-making and risk-taking potential.

8.8.4 Creating a framework to categorise transparent interface elements and their impact on situational awareness

Regarding the communication of an agent's intent and decision-making process, we recommend the following:

- **More reproduction studies on past SAT-based work.**
- **The development of more general SA assessment tools for use in non safety-critical contexts.**

Situational awareness (SA) and the related SAT framework are often used as a basis for qualifying and designing interfaces that support increased agent transparency. Most SA studies have been conducted in defence-oriented, military-focused environments where task require users to assess situations over extended periods of time (asset tracking, dispatch missions etc.). While insights from SA studies are a useful way to categorise the impact of different visual aids on SA in high-risk situations, they are less helpful when it comes to selecting specific types of visualisations for a given task. We recommend that Human Factors and HCI researchers reproduce past SA-based designs and develop new ones in order to test their effectiveness in more varied scenarios. In particular, we recommend that more studies are conducted in non safety-critical environments. This would allow for a better understanding of which characteristics inherent to visual aids impact SA the most, and which contexts of interaction make it most relevant to assess SA in the first place.

As a follow-up study, and using results from reproduction studies, we recommend that research efforts are centred around designing more context-free assessment tools for SA. Current SA assessment tools are either deployed in a context-specific manner (SAGAT [53]) or as a series of rating-scale instruments (SART [53]). Most tools have been designed and calibrated in safety-critical environments. However, and line with the results of our non-safety critical user studies, findings from previously designed SA assessment tools may not prove conclusive in tasks designed to be accessible to non-expert users. For instance, the question wording in the SART tool clearly refers to concept from aviation, where it was originally designed. These instruments may not resonate with a general audience. If SA is to be an important factor in assessing and monitoring a wide range of scenarios, a more general means of assessment is needed, one that (like NASA TLX [76] for cognitive workload, for instance) could be easily administered in a wide range of scenarios. These results could help create a framework where likely drawbacks and advantages of specific visualisations are presented, along with their relationship to different types of task.

We hope that these recommendations will speak to designers of human-agent systems and be tested in a wide range of interaction scenarios. For future work on trust in collaborative agents, we recommend that studies focus on the development of trust-aware agents, which includes the design of systems capable of *detecting loss of trust in real-time* and *repairing trust via a series of mechanisms* depending on the type of error committed and the context of interaction. The notion of “trust-aware agents” is most commonly found in multi-agents environments where trust is often operationalised as a score based on system’s performance and used to plan future decisions, such as in the context of AI-supported transportation systems [30]. It would be insightful for future work to extend the concept of trust awareness to agents involved in more social interactions where the detection and assessment of trust is usually more complex. These studies could then deploy new or previously studied [145] trust-repair mechanisms and assess their relevance in a variety of scenarios and tasks.

Chapter 9

Conclusion

Collaboration between human and automated agents is now commonplace in many applied settings, from safety environments to more casual, everyday activities. Due to the complex interplay between users, agents and the environment of interaction, it is important to study specific factors that are likely to influence how users trust an agent, as well as the implications for task performance, reliance and other reported metrics. In this thesis, we have presented four user-studies designed to test the impact of different agent behaviours and visualisations on users. To perform all of our experiments, we designed a game-like human-agent framework (see Chapter 3) where users and agents collaborated to complete goal-oriented tasks with various levels of difficulty. Using this framework, we were able to record behavioural metrics to assess task performance and reliance, as well as reported constructs with the use of verified survey instruments. We would like our results to inform the design of collaborative agents and motivate future research into key components of the human-agent partnership.

In this Section, we summarise our answers to the overall research questions defined in Section 1.3. Our first research questions was defined as follows: **How do changes in agent predictability (how easy it is to guess its next actions) and reliability (how good the agent is at the task) impact the human-agent relationship?** With our lab-based user study presented in Chapter 4, we showed that, when agent reliability is high, added predictability increases task performance, reliance on the agent, trust and reduces cognitive load.

Our second research question was defined as follows: **How do different types of agent errors defined from previous related work such as slips, mistakes and lapses affect the human-agent relationship?** With our lab-based user study presented in Chapter 5, we demonstrated that, when agent reliability is high, the type of error committed by an agent affects the way users interact with it. Errors of omission (defined as “lapses”) are the easiest to correct and can even improve task performance in terms of Precision by reducing the chance of users engaging in complacent behaviours. Errors of intention (“mistakes”) and commission (“slips”) were harder for users to correct, and resulted in the worst task performance as well as higher reported cognitive load.

Our third research question was defined as follows: **How do different types of environmental conditions (static or moving), which impair vision and induce uncertainty, affect the human-agent relationship?**. With our remote online user study presented in Chapter 6, we showed that conditions impairing vision and inducing visual uncertainty can affect the way users perform and rely on an agent. Near-total visual uncertainty led to higher reported trust and increased reliance on the agent while negatively affecting task performance. Visual occlusions that forced participants to react more quickly led to higher task performance, despite also resulting in a higher reported cognitive load.

Our fourth and final research question was defined as follows: **How do different types of visual help (designed to elicit different levels of situational awareness) influence the human-agent relationship?** With our remote online user study presented in Chapter 7, we showed that the type of visualisation used in a human-agent collaborative task has important implications for human-agent collaboration. Visualisations that are descriptive in nature (highlighting important elements) were perceived as more helpful and led to higher task performance than visualisations that were prescriptive in nature (telling the user “what to do”). Overall, visualisations designed to support SA level 1 and 2 had a more positive influence on users while an SA level 3 visualisation led to lower task performance and a higher cognitive load.

Bibliography

[1] Adler, R. F., and Benbunan-Fich, R. The effects of task difficulty and multitasking on performance. *Interacting with Computers 27*, 4 (2015), 430–439.

[2] Ahmad, M. I., Bernotat, J., Lohan, K., and Eyssel, F. Trust and cognitive load during human-robot interaction. *arXiv* (2019).

[3] Ajenaghughrure, I. B., Sousa, S. C., Kosunen, I. J., and Lamas, D. Predictive model to assess user trust: a psycho-physiological approach. In *Proceedings of the 10th Indian Conference on Human-Computer Interaction* (2019), pp. 1–10.

[4] Alan, A. T., Liu, C., Salisbury, E., Prior, S. D., Ramchurn, S. D., Wu, F., Tatlock, K., and Rees, G. Human-uav teaming in dynamic and uncertain environments. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* (2018), pp. 1791–1793.

[5] Aljukhadar, M., Senecal, S., and Daoust, C.-E. Using recommendation agents to cope with information overload. *International Journal of Electronic Commerce 17*, 2 (2012), 41–70.

[6] Alvarado-Valencia, J. A., and Barrero, L. H. Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior 36* (July 2014), 102–113.

[7] Araujo, T. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior 85* (2018), 183–189.

[8] Argote, L., and Goodman, P. S. The organizational implications of robotics. *Implementing advanced technology* (1985).

[9] Armstrong, R. A. When to use the b onferroni correction. *Ophthalmic and Physiological Optics 34*, 5 (2014), 502–508.

[10] Atari, I. *Missile Command.* Game [Atari 2600], July 1980. Atari, Inc, United States.

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

[12] Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[13] Bisbey, R., and Kay, M. The mind (management of information through natural discourse) translation system: a study in man-machine collaboration. Tech. rep., RAND CORP SANTA MONICA CALIF, 1972.

[14] Boubin, J. G., Rusnock, C. F., and Bindewald, J. M. Quantifying compliance and reliance trust behaviors to influence trust in human-automation teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 61*, 1 (Sept. 2017), 750–754.

[15] Bradley, J. V. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association 53*, 282 (1958), 525–528.

[16] Bradshaw, J. M., Feltovich, P., Johnson, M., Breedy, M., Bunch, L., Eskridge, T., Jung, H., Lott, J., Uszok, A., and van Diggelen, J. From tools to teammates: Joint activity in human-agent-robot teams. In *International conference on human centered design* (2009), Springer, pp. 935–944.

[17] Cao, A., Chintamani, K. K., Pandya, A. K., and Ellis, R. D. Nasa tlx: Software for assessing subjective mental workload. *Behavior research methods 41*, 1 (2009), 113–117.

[18] Cao, A., Chintamani, K. K., Pandya, A. K., and Ellis, R. D. NASA TLX: Software for assessing subjective mental workload. *Behavior Research Methods 41*, 1 (Feb. 2009), 113–117.

[19] Carter, M., Downs, J., Nansen, B., Harrop, M., and Gibbs, M. Paradigms of games research in hci: a review of 10 years of research at chi. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play* (2014), pp. 27–36.

[20] Carter, M., Downs, J., Nansen, B., Harrop, M., and Gibbs, M. Paradigms of games research in hci: A review of 10 years of research at chi. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play* (New York, NY, USA, 2014), CHI PLAY '14, Association for Computing Machinery, p. 27–36.

[21] Chancey, E. T., Bliss, J. P., Yamani, Y., and Handley, H. A. Trust and the Compliance-Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human Factors 59*, 3 (2017), 333–345.

[22] Chang, M., Kuo, R., and Liu, E. Z. F. Revised computer game attitude scale. *Proceedings of the 22nd International Conference on Computers in Education, ICCE 2014* (2014), 598–607.

[23] Charissis, V., and Papanastasiou, S. Human–machine collaboration through vehicle head up display interface. *Cognition, Technology & Work 12*, 1 (2010), 41–50.

[24] Chavaillaz, A., Schwaninger, A., Michel, S., and Sauer, J. Automation in visual inspection tasks: X-ray luggage screening supported by a system of direct, indirect or adaptable cueing with low and high system reliability. *Ergonomics 61*, 10 (Oct. 2018), 1395–1408.

[25] Chavaillaz, A., Wastell, D., and Sauer, J. System reliability, performance and trust in adaptable automation. *Applied Ergonomics 52* (2016), 333–342.

[26] Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. Situation awareness-based agent transparency. Tech. rep., Army research lab aberdeen proving ground md human research and engineering . . . , 2014.

[27] Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., and Barnes, M. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science 19*, 3 (Feb. 2018), 259–282.

[28] Chen, S., Loft, S., Huf, S., Braithwaite, J., and Visser, T. Static and adaptable automation in simulated submarine track management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2014), vol. 58, SAGE Publications Sage CA: Los Angeles, CA, pp. 2280–2284.

[29] Chen, T., Campbell, D., Gonzalez, L. F., and Coppin, G. Increasing Autonomy Transparency through capability communication in multiple heterogeneous UAV management. In *IEEE International Conference on Intelligent Robots and Systems* (2015).

[30] Cheng, M., Zhang, J., Nazarian, S., Deshmukh, J., and Bogdan, P. Trust-aware control for intelligent transportation systems. In *2021 IEEE Intelligent Vehicles Symposium (IV)* (2021), IEEE, pp. 377–384.

[31] Christel, M. G., Stevens, S. M., Maher, B. S., Brice, S., Champer, M., Jayapalan, L., Chen, Q., Jin, J., Hausmann, D., Bastida, N., et al. Rumbleblocks: Teaching science concepts to young children through a unity game. In *2012 17th International Conference on Computer Games (CGAMES)* (2012), IEEE, pp. 162–166.

[32] Cipollone, M., Schifter, C. C., and Moffat, R. A. Minecraft as a creative tool: A case study. *International Journal of Game-Based Learning (IJGBL) 4*, 2 (2014), 1–14.

[33] company, P. Prolific, 2021.

[34] Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., and Paiva, A. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Richland, SC, 2018), AAMAS '18, International Foundation for Autonomous Agents and Multiagent Systems, pp. 507–513.

[35] Costanza, E., Fischer, J. E., Colley, J. A., Rodden, T., Ramchurn, S. D., and Jennings, N. R. Doing the laundry with agents: a field trial of a future smart energy system in the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), pp. 813–822.

[36] Daronnat, S. Human-agent trust relationships in a real-time collaborative game. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play* (2020), pp. 18–20.

[37] Daronnat, S., Azzopardi, L., and Halvey, M. Impact of agents' errors on performance, reliance and trust in human-agent collaboration. In *Human Factors and Ergonomics Society Annual Meeting* (2020), pp. 1–5.

[38] Daronnat, S., Azzopardi, L., and Halvey, M. Investigating the impact of visual environmental uncertainty on human-agent teaming. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2021), vol. 65, SAGE Publications Sage CA: Los Angeles, CA, pp. 1185–1189.

[39] Daronnat, S., Azzopardi, L., Halvey, M., and Dubiel, M. Human-agent collaborations: trust in negotiating control. *CHI 2019* (2019).

[40] Daronnat, S., Azzopardi, L., Halvey, M., and Dubiel, M. Impact of agent reliability and predictability on trust in real time human-agent collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (2020), pp. 131–139.

[41] Daronnat, S., Azzopardi, L., Halvey, M., and Dubiel, M. Inferring trust from users behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration. *Frontiers in Robotics and AI 8* (2021), 194.

[42] de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., and Parasuraman, R. The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 56*, 1 (Sept. 2012), 263–267.

[43] de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., and Neerincx, M. A. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics 12*, 2 (Nov. 2019), 459–478.

[44] Demir, M., McNeese, N. J., and Cooke, N. J. Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research 46* (2017), 3–12.

[45] Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., and Manning, C. A. Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly 6*, 1 (Jan. 1998), 1–20.

[46] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. The role of trust in automation reliance. *International journal of human-computer studies 58*, 6 (2003), 697–718.

[47] Eisenhardt, K. M. Building theories from case study research. *Academy of Management Review 14*, 4 (Oct. 1989), 532–550.

[48] Endsley, M. R. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting* (1988), vol. 32, SAGE Publications Sage CA: Los Angeles, CA, pp. 97–101.

[49] Endsley, M. R. Situation awareness global assessment technique (sagat). In *Proceedings of the IEEE 1988 national aerospace and electronics conference* (1988), IEEE, pp. 789–795.

[50] Endsley, M. R. Toward a theory of situation awareness in dynamic systems. *Human factors 37*, 1 (1995), 32–64.

[51] Endsley, M. R. Direct measurement of situation awareness: Validity and use of sagat. In *Situational awareness*. Routledge, 2017, pp. 129–156.

[52] Endsley, M. R. A systematic review and meta-analysis of direct objective measures of situation awareness: A comparison of SAGAT and SPAM. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (Sept. 2019), 001872081987537.

[53] Endsley, M. R., Selcon, S. J., Hardiman, T. D., and Croft, D. G. Comparative analysis of SAGAT and SART for evaluations of situation awareness. *Proceedings of the Human Factors and Ergonomics Society 1* (1998), 82–86.

[54] Epic Games. Unreal engine.

[55] Erebak, S., and Turgut, T. Caregivers' attitudes toward potential robot coworkers in elder care. *Cognition, Technology & Work 21*, 2 (2019), 327–336.

[56] Fan, X., McNeese, M., Sun, B., Hanratty, T., Allender, L., and Yen, J. Human-Agent Collaboration for Time-Stressed Multicontext Decision Making. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 40*, 2 (Mar. 2010), 306–320.

[57] Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., and Endsley, M. R. The influence of agent reliability on trust in human-agent collaboration. In *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction* (2008), ACM, p. 7.

[58] Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., and Endsley, M. R. The influence of agent reliability on trust in human-agent collaboration. *Proceedings of the 15th European conference on Cognitive ergonomics the ergonomics of cool interaction - ECCE '08* (2008), 1.

[59] Fink, J. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics* (2012), Springer, pp. 199–208.

[60] Fischer, J. E., Greenhalgh, C., Jiang, W., Ramchurn, S. D., Wu, F., and Rodden, T. In-the-loop or on-the-loop? interactional arrangements to support team coordination with a planning agent. *Concurrency and Computation: Practice and Experience 33*, 8 (2021), e4082.

[61] Fischer, J. E., Jiang, W., and Moran, S. Atomicorchid: A mixed reality game to investigate coordination in disaster response. In *International Conference on Entertainment Computing* (2012), Springer, pp. 572–577.

[62] Flanagan, J. C. The critical incident technique. *Psychological Bulletin 51*, 4 (1954), 327–358.

[63] Fleiss, J. L., Levin, B., and Paik, M. C. *Statistical methods for rates and proportions.* john wiley & sons, 2013.

[64] Games, P. A., and Howell, J. F. Pairwise multiple comparison procedures with unequal n's and/or variances: a monte carlo study. *Journal of Educational Statistics 1*, 2 (1976), 113–125.

[65] Garnick, C. J., Bindewald, J. M., and Rusnock, C. F. Designing an automated agent to encourage human reliance. *Proceedings of the Human Factors and Ergonomics Society 2017-October* (2017), 1730–1734.

[66] Gillespie, N., and Dietz, G. Trust repair after an organization-level failure. *Academy of management review 34*, 1 (2009), 127–145.

[67] Girden, E. R. *ANOVA: Repeated measures.* No. 84 in Quantitative Applications in the Social Sciences. Sage, 1992.

[68] Gish, K. W., and Staplin, L. Human factors aspects of using head up displays in automobiles: A review of the literature. *U.S. Department of Transportation, National Highway Traffic Safety Administration* (1995).

[69] Godot Engine Development Team. Godot engine.

[70] Gombolay, M., Bair, A., Huang, C., and Shah, J. Computational design of mixed-initiative human–robot teaming that considers human factors: situational awareness, workload, and workflow preferences. *International Journal of Robotics Research 36*, 5-7 (2017), 597–617.

[71] Gordon, N. Colour blindness. *Public health 112*, 2 (1998), 81–84.

[72] Graafland, M., Bemelman, W. A., and Schijven, M. P. Game-based training improves the surgeon's situational awareness in the operation room: a randomized controlled trial. *Surgical endoscopy 31*, 10 (2017), 4093–4101.

[73] Grabowski, M., and Sanborn, S. D. Human performance and embedded intelligent technology in safety-critical systems. *International journal of human-computer studies 58*, 6 (2003), 637–670.

[74] Haas, J. A history of the unity game engine. *Diss. WORCESTER POLYTECHNIC INSTITUTE* (2014).

[75] Hart, S. G. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (2006), vol. 50, Sage publications Sage CA: Los Angeles, CA, pp. 904–908.

[76] Hart, S. G., and Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52. Elsevier, 1988, pp. 139–183.

[77] Hart, S. G., and Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52. Elsevier, 1988, pp. 139–183.

[78] Hawkins, K. P., Bansal, S., Vo, N. N., and Bobick, A. F. Anticipating human actions for collaboration in the presence of task and sensor uncertainty. In *2014 ieee international conference on Robotics and automation (ICRA)* (2014), IEEE, pp. 2215–2222.

[79] Heidlauf, P., Collins, A., Bolender, M., and Bak, S. Verification challenges in f-16 ground collision avoidance and other automated maneuvers. In *ARCH@ ADHS* (2018), pp. 208–217.

[80] Heo, J. S., and Lee, K. Y. A multi-agent system-based intelligent heuristic optimal control system for a large-scale power plant. *2006 IEEE Congress on Evolutionary Computation, CEC 2006* (2006), 1544–1551.

[81] Herdener, N., Clegg, B. A., Wickens, C. D., and Smith, C. A. Anchoring and Adjustment in Uncertain Spatial Trajectory Prediction. *Human Factors 61*, 2 (2019), 255–272.

[82] Hoc, J. M. From human – machine interaction to human – machine cooperation. *Ergonomics 43*, 7 (2000), 833–843.

[83] Hoc, J. M., Young, M. S., and Blosseville, J. M. Cooperation between drivers and automation: Implications for safety. *Theoretical Issues in Ergonomics Science 10*, 2 (2009), 135–160.

[84] Hoc, J.-M., Young, M. S., and Blosseville, J.-M. Cooperation between drivers and automation: implications for safety. *Theoretical Issues in Ergonomics Science 10*, 2 (2009), 135–160.

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

[86] Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. Trust in automation. *IEEE Intelligent Systems 28*, 1 (2013), 84–88.

[87] Hussein, A., Elsawah, S., and Abbass, H. A. Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human Factors 62*, 8 (2020), 1237–1248.

[88] Iordanescu, L., Grabowecky, M., and Suzuki, S. Demand-based dynamic distribution of attention and monitoring of velocities during multiple-object tracking. *Journal of vision 9*, 4 (2009), 1–1.

[89] Jensen, T., Khan, M. M. H., Albayram, Y., Fahim, M. A. A., Buck, R., and Coman, E. Anticipated emotions in initial trust evaluations of a drone system based on performance and process information. *International Journal of Human–Computer Interaction 36*, 4 (2020), 316–325.

[90] Jian, J.-Y., Bisantz, A. M., and Drury, C. G. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics 4*, 1 (Mar. 2000), 53–71.

[91] Jung, M. F., Difranzo, D., Stoll, B., Shen, S., Lawrence, A., and Claure, H. Robot Assisted Tower Construction - A Resource Distribution Task to Study Human-Robot Collaboration and Interaction with Groups of People. *arXiv* (2018), 1–18.

[92] Karikawa, D., Aoyama, H., Takahashi, M., Furuta, K., Wakabayashi, T., and Kitamura, M. A visualization tool of en route air traffic control tasks for describing controller's proactive management of traffic situations. *Cognition, Technology & Work 15*, 2 (2013), 207–218.

[93] Kilgore, R., and Voshell, M. Increasing the transparency of unmanned systems: Applications of ecological interface design. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8526 LNCS*, PART 2 (2014), 378–389.

[94] Kim, P. H., Cooper, C. D., Dirks, K. T., and Ferrin, D. L. Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes 120*, 1 (jan 2013), 1–14.

[95] Kim, T., and Song, H. How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics 61* (2021), 101595.

[96] Kjeldskov, J., and Graham, C. A review of mobile hci research methods. In *International Conference on Mobile Human-Computer Interaction* (2003), Springer, pp. 317–335.

[97] Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems 19*, 6 (2004), 91–95.

[98] Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems 19*, 6 (2004), 91–95.

[99] Kott, A. Challenges and characteristics of intelligent autonomy for internet of battle things in highly adversarial environments. *arXiv preprint arXiv:1803.11256* (2018).

[100] Krämer, N. C., von der Pütten, A., and Eimler, S. Human-agent and human-robot interaction theory: similarities to and differences from human-human interaction. In *Human-computer interaction: The agency perspective*. Springer, 2012, pp. 215–240.

[101] Kunze, A., Summerskill, S. J., Marshall, R., and Filtness, A. J. Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics 62*, 3 (Feb. 2019), 345–360.

[102] Lebas, M. J. Performance measurement and performance management. *International Journal of Production Economics 41*, 1-3 (1995), 23–35.

[103] Lee, J. D., and See, K. A. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society 46*, 1 (2004), 50–80.

[104] Levene, H. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling* (1961), 279–292.

[105] Lewicki, R. J., Bunker, B. B., et al. Developing and maintaining trust in work relationships. *Trust in organizations: Frontiers of theory and research 114* (1996), 139.

[106] Lewicki, R. J., Wiethoff, C., and Tomlinson, E. C. What is the role of trust in organizational justice. *Handbook of organizational justice* (2005), 247–270.

[107] Lewis, J., Brown, D., Cranton, W., and Mason, R. Simulating visual impairments using the unreal engine 3 game engine. In *2011 IEEE 1st International Conference on Serious Games and Applications for Health (SeGAH)* (2011), IEEE, pp. 1–8.

[108] Lewis, J. R. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction 34*, 7 (2018), 577–590.

[109] Lewis, M. Designing for human-agent interaction. *AI Magazine 19*, 2 (1998), 67–67.

[110] Lewis, M., and Jacobson, J. Game engines. *Communications of the ACM 45*, 1 (2002), 27.

[111] Licklider, J. C. Man-computer symbiosis. *IRE transactions on human factors in electronics*, 1 (1960), 4–11.

[112] Liu, H. *Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance.* Virginia Commonwealth University, 2015.

[113] Lockey, S., Gillespie, N., Holm, D., and Someh, I. A. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. *Hawaii International Conference on System Sciences* (2021).

[114] Lohse, G. L., Biolsi, K., Walker, N., and Rueter, H. H. A classification of visual representations. *Communications of the ACM 37*, 12 (1994), 36–50.

[115] Ma, Y., and Ghasemzadeh, H. Head-mounted sensors and wearable computing for automatic tunnel vision assessment. *Proceedings of the 2017 Design, Automation and Test in Europe, DATE 2017* (2017), 634–637.

[116] MacEachren, A. M. Visualizing Uncertain Information. *Cartographic Perspective 13*, 13 (1992), 10–19.

[117] Mackworth, N. H. Visual noise causes tunnel vision. *Psychonomic science 3*, 1 (1965), 67–68.

[118] Marinaccio, K., Kohn, S., Parasuraman, R., and De Visser, E. J. A Framework for Rebuilding Trust in Social Automation Across Health-Care Domains. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care 4*, 1 (jun 2015), 201–205.

[119] Matthews, G., Panganiban, A. R., Lin, J., Long, M., and Schwing, M. Supermachines or sub-humans: mental models and trust in intelligent autonomous systems. In *Trust in Human-Robot Interaction.* Elsevier, 2021, pp. 59–82.

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

[121] McGraw, K. O., and Wong, S. P. A common language effect size statistic. *Psychological bulletin 111*, 2 (1992), 361.

[122] McKight, P. E., and Najab, J. Kruskal-wallis test. *The corsini encyclopedia of psychology* (2010), 1–1.

[123] McKnight, P. E., and Najab, J. Mann-whitney u test. *The Corsini encyclopedia of psychology* (2010), 1–1.

[124] Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., and Procci, K. Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors 58*, 3 (2015), 401–415.

[125] Merritt, S. M., Lee, D., Unnerstall, J. L., and Huber, K. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors 57*, 1 (2015), 34–47.

[126] Moray, N., and Inagaki, T. Attention and complacency. *Theoretical Issues in Ergonomics Science 1*, 4 (2000), 354–365.

[127] Muir, B. M. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies 27*, 5-6 (1987), 527–539.

[128] Novitzky, M., Robinette, P., Benjamin, M. R., Gleason, D. K., Fitzgerald, C., and Schmidt, H. Preliminary interactions of human-robot trust, cognitive load, and robot intelligence levels in a competitive game. In *Companion of the 2018 ACM/IEEE international conference on human-robot interaction* (2018), pp. 203–204.

[129] Ogreten, S., Lackey, S., and Nicholson, D. Recommended roles for uninhabited team members within mixed-initiative combat teams. In *2010 International Symposium on Collaborative Technologies and Systems* (2010), IEEE.

[130] Ososky, S., Sanders, T., Jentsch, F., Hancock, P., and Chen, J. Y. C. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. *Unmanned Systems Technology XVI 9084* (2014), 90840E.

[131] Pak, R., Fink, N., Price, M., Bass, B., and Sturre, L. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics 55*, 9 (July 2012), 1059–1072.

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

[133] Parasuraman, R., and Miller, C. A. Trust and etiquette in high-criticality automated systems. *Communications of the ACM 47*, 4 (2004), 51–55.

[134] Parasuraman, R., and Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Human factors 39*, 2 (1997), 230–253.

[135] Petersen, L., Robert, L., Jessie Yang, X., and Tilbury, D. M. Situational awareness, driver's trust in automated driving systems and secondary task performance. *arXiv* (2019), 1–26.

[136] Powers, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).

[137] Prinzel, L. The relationship of self-efficacy and complacency in pilot-automation interaction (technical memorandum no. tm-2002-211925). *NASA Langley Research Center, Hampton, VA* (2002).

[138] Pylyshyn, Z. W. *Seeing and visualizing: It's not what you think.* MIT press, 2003.

[139] Pylyshyn, Z. W., and Storm, R. W. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision 3*, 3 (1988), 179–197.

[140] Rachev, S. T., Stoyanov, S. V., and Fabozzi, F. J. *Risk and uncertainty*, vol. 211. John Wiley & Sons, 2011.

[141] Rambusch, J., Jakobsson, P., and Pargman, D. Exploring e-sports: A case study of game play in counter-strike. In *3rd Digital Games Research Association International Conference:" Situated Play", DiGRA 2007, Tokyo, 24 September 2007 through 28 September 2007* (2007), vol. 4, Digital Games Research Association (DiGRA), pp. 157–164.

[142] Ramchurn, S. D., Fischer, J. E., Ikuno, Y., Wu, F., Flann, J., and Waldock, A. A study of human-agent collaboration for multi-uav task allocation in dynamic environments. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)* (2015), pp. 1184–1192.

[143] Ramchurn, S. D., Wu, F., Jiang, W., Fischer, J. E., Reece, S., Roberts, S., Rodden, T., Greenhalgh, C., and Jennings, N. R. Human–agent collaboration for disaster response. *Autonomous Agents and Multi-Agent Systems 30*, 1 (2016), 82–111.

[144] Rapp, A., Hopfgartner, F., Hamari, J., Linehan, C., and Cena, F. Strengthening gamification studies: Current trends and future opportunities of gamification research. *International Journal of Human Computer Studies 127*, November 2018 (2019), 1–6.

[145] Reason, J. *Human error.* Cambridge university press, 1990.

[146] Robinette, P., Howard, A. M., and Wagner, A. R. Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems 47*, 4 (2017), 425–436.

[147] Ross, J. M. *Moderators of trust and reliance across multiple decision aids.* university of central florida, 2008.

[148] Ross, J. M., Szalma, J. L., Hancock, P. A., Barnett, J. S., and Taylor, G. The effect of automation reliability on user automation trust and reliance in a search-and-rescue scenario. In *proceedings of the human factors and ergonomics society annual meeting* (2008), vol. 52, Sage Publications Sage CA: Los Angeles, CA, pp. 1340–1344.

[149] Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., and Keim, D. A. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 240–249.

[150] Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. 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 2015-March* (2015), 141–148.

[151] Sanchez, J., Rogers, W. A., Fisk, A. D., and Rovira, E. Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science 15*, 2 (2014), 134–160.

[152] Sarter, N., Schroeder, B., and McGuirl, J. Supporting decision-making and action selection under time pressure and uncertainty: The case of inflight icing. *39th Aerospace Sciences Meeting and Exhibit 43*, 4 (2001), 573–583.

[153] Satterfield, K., Baldwin, C., De Visser, E., and Shaw, T. The influence of risky conditions in trust in autonomous systems. In *Proceedings of the Human Factors and Ergonomics Society* (2017), vol. 2017-Octob.

[154] Sauer, J., Chavaillaz, A., and Wastell, D. Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics 59*, 6 (June 2016), 767–780.

[155] Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., and Hancock, P. A. A meta-analysis of factors influencing the development of trust in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society 58*, 3 (Mar. 2016), 377–400.

[156] Schaekermann, M., Ribeiro, G., Wallner, G., Kriglstein, S., Johnson, D., Drachen, A., Sifa, R., and Nacke, L. E. Curiously motivated: Profiling curiosity with self-reports and behaviour metrics in the game" destiny". In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (2017), pp. 143–156.

[157] Schaffer, J., O'Donovan, J., Marusich, L., Yu, M., Gonzalez, C., and Höllerer, T. A study of dynamic information display and decision-making in abstract trust games. *International Journal of Human Computer Studies 113*, December 2017 (2018), 1–14.

[158] Seabold, S., and Perktold, J. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference* (2010).

[159] Selcon, S., Taylor, R., and Koritsas, E. Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation. *Proceedings of the Human Factors Society Annual Meeting 35*, 2 (Sept. 1991), 62–66.

[160] Senaratne, H., Gerharz, L., Pebesma, E., and Schwering, A. Usability of spatio-temporal uncertainty visualisation methods. In *Bridging the geographic information sciences*. Springer, 2012, pp. 3–23.

[161] Sgobba, T. B-737 max and the crash of the regulatory system. *Journal of Space Safety Engineering 6*, 4 (2019), 299–303.

[162] SHAPIRO, S. S., and WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika 52*, 3-4 (Dec. 1965), 591–611.

[163] Shekhar, S., Coyle, M., Shargal, M., Kozak, J., and Hancock, P. Design and validation of headup displays for navigation in ivhs. In *Vehicle Navigation and Information Systems Conference, 1991* (1991), vol. 2, IEEE, pp. 537–542.

[164] Sheldon, M. R., Fillyaw, M. J., and Thompson, W. D. The use and interpretation of the friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International 1*, 4 (1996), 221–228.

[165] Sheridan, T. Trustworthiness of command and control systems. In *Analysis, Design and Evaluation of Man–Machine Systems 1988*. Elsevier, 1989, pp. 427–431.

[166] Shirado, H., and Christakis, N. A. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature 545*, 7654 (2017), 370–374.

[167] Singh, I. L., Molloy, R., and Parasuraman, R. Automation-induced" complacency": Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology 3*, 2 (1993), 111–122.

[168] Six, F., and Sorge, A. Creating a high-trust organization: An exploration into organizational policies that stimulate interpersonal trust building. *Journal of Management Studies 45*, 5 (2008), 857–884.

[169] Sosa, A., Stanton, R., Perez, S., Keyes-Garcia, C., Gonzalez, S., and Toups, Z. O. Imperfect robot control in a mixed reality game to teach hybrid human-robot team coordination. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (2015), pp. 697–702.

[170] Stowers, K., Kasdaglis, N., Rupp, M., Chen, J., Barber, D., and Barnes, M. Insights into human-agent teaming: Intelligent agent transparency and uncertainty. In *Advances in Human Factors in Robots and Unmanned Systems.* Springer, 2017, pp. 149–160.

[171] Sweller, J. Cognitive load theory. In *Psychology of learning and motivation*, vol. 55. Elsevier, 2011, pp. 37–76.

[172] Taha, Z., and Jizat, J. A. M. A comparison of two approaches for collision avoidance of an automated guided vehicle using monocular vision. *Applied Mechanics and Materials 145* (2012), 547–551.

[173] Taylor, R. M. Situational awareness rating technique (sart): The development of a tool for aircrew systems design. In *Situational awareness.* Routledge, 2017, pp. 111–128.

[174] Torre, I., Goslin, J., White, L., and Zanatto, D. Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society.* Association for Computing Machinery, Inc., 2018, pp. 1–6.

[175] Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114.

[176] Tulshan, A. S., and Dhage, S. N. Survey on virtual assistant: Google assistant, siri, cortana, alexa. In *International symposium on signal processing and intelligent recognition systems* (2018), Springer, pp. 190–201.

[177] Turkle, S. The intimate machine. *Science'84 5* (1984), 40–47.

[178] Vallat, R. Pingouin: statistics in python. *The Journal of Open Source Software 3*, 31 (Nov. 2018), 1026.

[179] Van Huyck, J. B., Battalio, R. C., and Beil, R. O. Coordination Games , Strategic Uncertainty , and Coordination Failure. *American Economic Association 80*, 1 (1990), 234–248.

[180] van Lier, R., van der Helm, P., and Leeuwenberg, E. Integrating global and local aspects of visual occlusion. *Perception 23*, 8 (1994), 883–903.

[181] Van Rossum, G., and Drake, F. L. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA, 2009.

[182] van Wissen, A., Gal, Y., Kamphorst, B., and Dignum, M. Human–agent teamwork in dynamic environments. *Computers in Human Behavior 28*, 1 (2012), 23–33.

[183] Viera, A. J., Garrett, J. M., et al. Understanding interobserver agreement: the kappa statistic. *Fam med 37*, 5 (2005), 360–363.

[184] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17* (2020), 261–272.

[185] Walker, F., Wang, J., Martens, M., and Verwey, W. Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. *Transportation research part F: traffic psychology and behaviour 64* (2019), 401–412.

[186] Wang, N., Pynadath, D. V., and Hill, S. G. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. *Aamas*, Aamas (2015), 997–1005.

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

[188] White, M. P., and Eiser, J. R. Marginal trust in risk managers: Building and losing trust following decisions under uncertainty. *Risk Analysis 26*, 5 (Oct. 2006), 1187–1203.

[189] Wiczorek, R., and Manzey, D. Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors 56*, 7 (2014), 1209–1221.

[190] Wiebe, E. N., Lamb, A., Hardy, M., and Sharek, D. Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior 32* (2014), 123–132.

[191] Wilcox, L. M., Allison, R. S., Helliker, J., Dunk, B., and Anthony, R. C. Evidence that viewers prefer higher frame-rate film. *ACM Transactions on Applied Perception (TAP) 12*, 4 (2015), 1–12.

[192] Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.

[193] Williams, E. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry 2*, 2 (1949), 149.

[194] Williams, L. J. Tunnel vision induced by a foveal load manipulation. *Human factors 27*, 2 (1985), 221–227.

[195] Williamson, D. F. The box plot: A simple visual method to interpret data. *Annals of Internal Medicine 110*, 11 (June 1989), 916.

[196] Woolgar, S., and Suchman, L. A. Plans and Situated Actions: The Problem of Human Machine Communication. *Contemporary Sociology 18*, 3 (1989), 414.

[197] Wrightsman, L. S. Interpersonal trust and attitudes toward human nature. In *Measures of Personality and Social Psychological Attitudes*. Elsevier, 1991, pp. 373–412.

[198] Wu, W., Ekaette, E., and Far, B. H. Uncertainty management framework for multi-agent system. In *Proceedings of ATS* (2003), vol. 2003.

[199] Xu, M., Fralick, D., Zheng, J. Z., Wang, B., Tu, X. M., and Feng, C. The differences and similarities between two-sample t-test and paired t-test. *Shanghai archives of psychiatry 29*, 3 (2017), 184.

[200] Zaphiris, P., and Ang, C. S. HCI issues in computer games. *Interacting with Computers 19*, 2 (03 2007), 135–139.

[201] Zhou, J., Li, Z., Hu, H., Yu, K., Chen, F., Li, Z., and Wang, Y. Effects of influence on user trust in predictive decision making. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (May 2019), ACM.

Part IV

Appendices

Appendix A

Agent Reliability and Predictability Study

A.A Pre-Hoc Survey

Start of Block: General Questions

Q0 Thank you for participating in this study. Please answer the questions below.

Q1
Please enter your participant ID:

____

Q2 What is your age?

____

Q3
What is your gender?

○ Male (1)

○ Female (2)

○ Non-Binary (3)

○ Self-Defined (4)

○ Prefer not to say (5)

Q9 What is your current status?

○ Student (1)

○ Employed (2)

○ Unemployed (3)

○ Other (4)

○ Prefer not to say (5)

Q4 Are you a native English Speaker?

○ Yes (1)

○ No (2)

Q5 How would you rate your English comprehension skills?

	Very Poor (1)	Poor (2)	Average (3)	Good (4)	Very Good (5)	I'm a native speaker (6)
Make your selection: (1)	○	○	○	○	○	○

Q6 How often do you play video-games?

	Never (1)	Once a year (2)	Once in Several Months (3)	Once a Month (4)	Several times a month (5)	Everyday (6)
Make your selection: (1)	○	○	○	○	○	○

Q7 What kind of games do you usually play (example: 3D shooters, 2D plateformers, short mobiles games etc...) if you selected "never" on the last question, you can skip this one.

____

End of Block: General Questions

Start of Block: Please read and rate the following statements

Q0 Please rate the following statements.

Q8 Manually sorting through card catalogues is more reliable than computer-aidedsearches for finding items in a library.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q10 If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because computerized surgery is more reliable and safer than manual surgery.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q11 People save time by using automatic teller machines (ATMs) rather than a bank teller in making transactions.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)

Select an answer: (1)	○	○	○	○	○

Q51 I do not trust automated devices such as ATMs and computerized airline reservations.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q12 People who work frequently with automated devices have lower job satisfaction because they feel less involved in their job than those who work manually.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q13 I feel safer depositing my money at an ATM than with a human teller.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q14 I have to tape an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my recorder rather than manual taping.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)

Select an answer: (1)	○	○	○	○	○

Q15 People whose jobs require them to work with automated systems are lonelier than people who do not work with such devices.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q16 Automated systems used in modern aircraft, such as the automatic landing system, have made air journey safer.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q17 ATMs provide safeguard against the inappropriate use of an individual's bank account by dishonest people.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q18 Automated devices used in aviation and banking have made work easier for both employees and customers.

	Strongly	Disagree (2)	Neutral (3)	Agree (4)	Strongly

	Disagree (1)				Agree (5)
Select an answer: (1)	○	○	○	○	○

Q19 I often use automated devices.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q20 People who work with automated devices have greater job satisfaction because they feel more involved than those who work manually.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q21 Automated devices in medicine save time and money in the diagnosis and treatment of diseases.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q22 Even though the automatic cruise control in my car is set at a speed below the speed limit, I worry when I pass a police radar speed-trap in case the automatic control is not working properly.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q23 Bank transactions have become safer with the introduction of computer technology for the transfer of funds.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q24 I would rather purchase an item using a computer than have to deal with a sales representative on the phone because my order is more likely to be correct using the computer.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q25 Work has become more difficult with the increase of automation in aviation and banking.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q27 I do not like to use ATMs because I feel that they are sometimes unreliable.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q28 I think that automated devices used in medicine, such as CAT-scans and ultrasound, provide very reliable medical diagnosis.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

End of Block: Please read and rate the following statements

Start of Block: Block 2

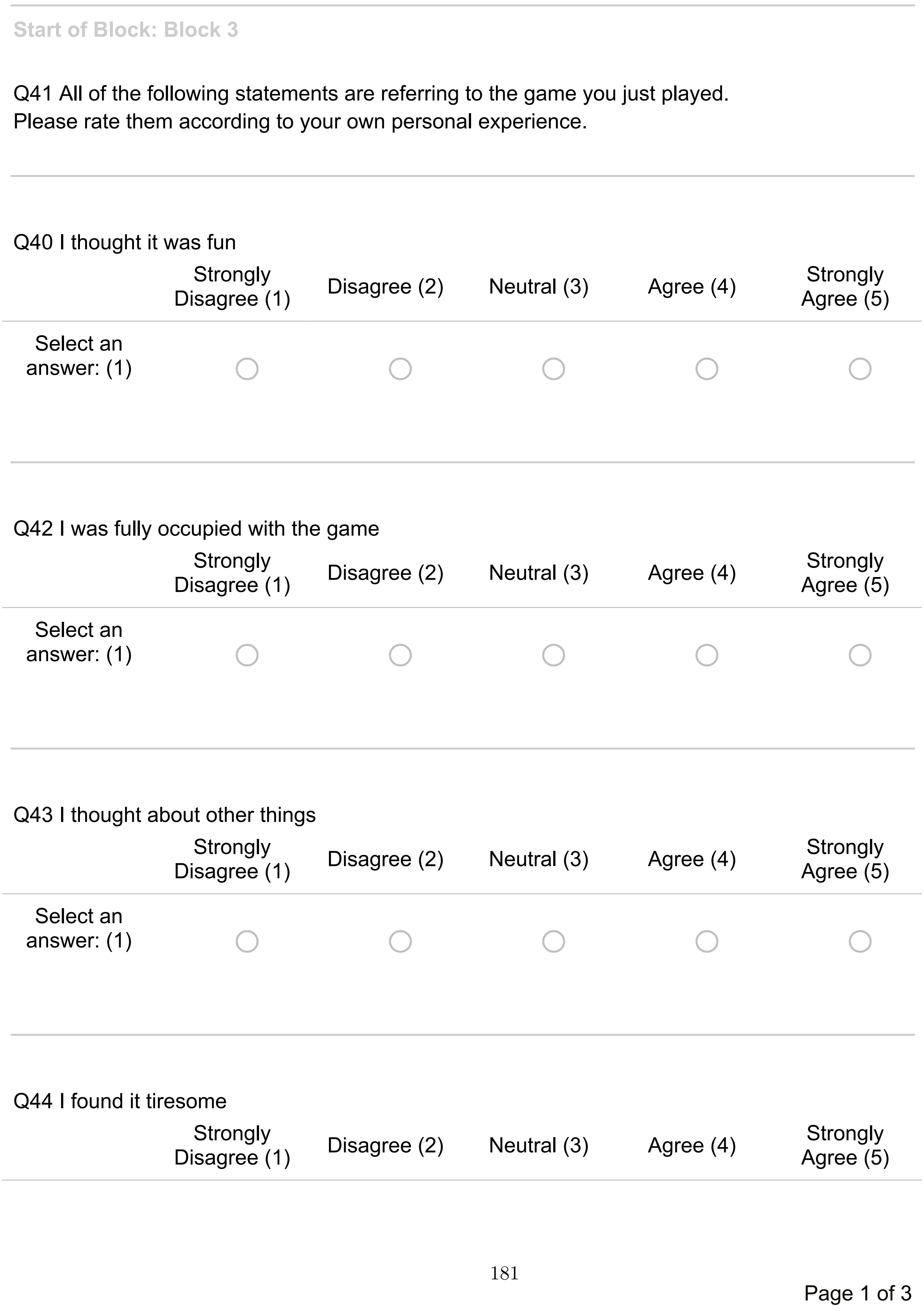
Q37 Thank your for answering this preliminary set of questions.

You can now begin to play the game.

When you're done, please click on the "next" button to answer a final short set of questions.

End of Block: Block 2

A.B Post-Hoc Survey



Start of Block: Block 3

Q41 All of the following statements are referring to the game you just played.
Please rate them according to your own personal experience.

Q40 I thought it was fun

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q42 I was fully occupied with the game

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q43 I thought about other things

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q44 I found it tiresome

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)

Select an answer: (1)	○	○	○	○	○

Q45 I felt competent

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q46 I thought it was hard

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q47 I felt frustrated

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q48 I felt time pressure

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q49 I felt successful

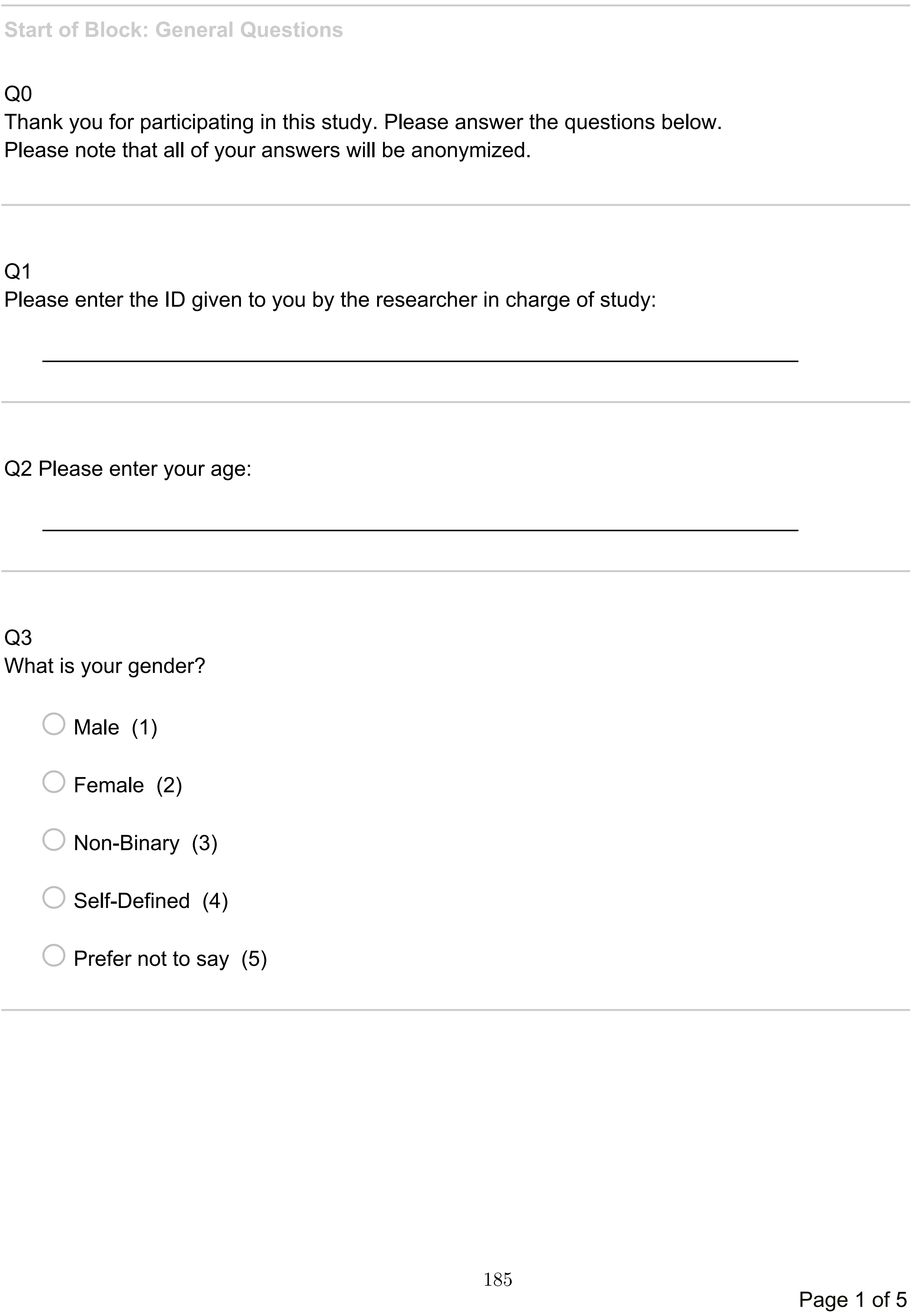
	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

End of Block: Block 3

Appendix B

Agent Errors and Behaviours Study

B.A Pre-Hoc Survey



Start of Block: General Questions

Q0
Thank you for participating in this study. Please answer the questions below.
Please note that all of your answers will be anonymized.

Q1
Please enter the ID given to you by the researcher in charge of study:

____

Q2 Please enter your age:

____

Q3
What is your gender?

○ Male (1)

○ Female (2)

○ Non-Binary (3)

○ Self-Defined (4)

○ Prefer not to say (5)

Q4 Education: What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

○ No schooling completed (1)

○ Nursery school to 8th grade (2)

○ Some high school, no diploma (3)

○ High school graduate, diploma or equivalent (4)

○ Some college credit, no degree (5)

○ Trade/technical/vocational training (6)

○ Associate degree (7)

○ Bachelor’s degree (8)

○ Master’s degree (9)

○ Professional degree (10)

○ Doctorate degree (12)

○ Prefer not to say (13)

Q5 In which field are you currently employed? (e.g "automotive industry", "medical" etc...)
If you are studying or on the lookout for a job, which field is your main focus?

____

End of Block: General Questions

Start of Block: Block 5

Q6 How often do you play video-games?

	Never (1)	Very Rarely (2)	Rarely (7)	Occasionally (8)	Frequently (9)	Very Frequently (10)

Make your selection: (1)	○	○	○	○	○	○

Q7 What kind of games do you usually play (example: 3D first person shooters, 2D plateformers, small mobiles games etc...) if you selected "never" on the last question, you can skip this one.

____

End of Block: Block 5

Start of Block: Block 3

Q8 Please rate the following statements

	Never (1)	Rarely (2)	Sometimes (6)	Very Often (7)	Constantly (8)
When I have a lot to do, it makes sense to delegate a task to automation. (1)	○	○	○	○	○
If life were busy, I would let an automated system handle some tasks for me. (2)	○	○	○	○	○
Automation should be used to ease people’s workload. (3)	○	○	○	○	○
If automation is available to help me with	○	○	○	○	○

something, it makes sense for me to pay more attention to my other tasks. (4)					
Even if an automated aid can help me with a task, I should pay attention to its performance. (5)	○	○	○	○	○
Distractions and interruptions are less of a problem for me when I have an automated system to cover some of the work. (6)	○	○	○	○	○
Constantly monitoring an automated system’s performance is a waste of time. (7)	○	○	○	○	○
Even when I have a lot to do, I am likely to watch automation carefully for errors. (8)	○	○	○	○	○
It’s not usually necessary to pay much	○	○	○	○	○

attention to automation when it is running. (9)					
Carefully watching automation takes time away from more important or interesting things. (10)	○	○	○	○	○

End of Block: Block 3

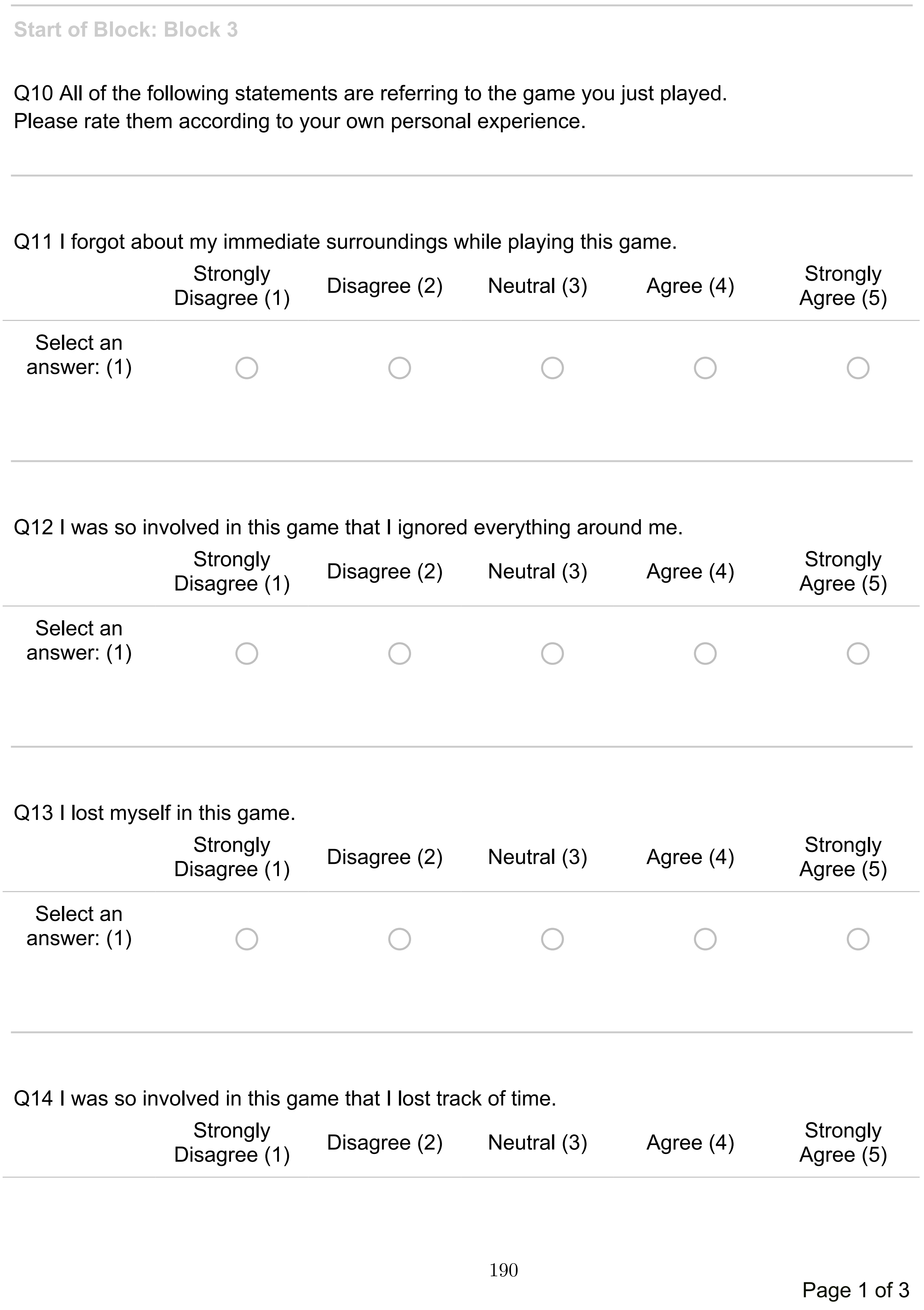
Start of Block: Block 2

Q9 Thank your for answering this preliminary set of questions.

You can now begin to play the game.

After you're done playing, please click on the button below to answer a final short set of questions.

B.B Post-Hoc Survey



Start of Block: Block 3

Q10 All of the following statements are referring to the game you just played.
Please rate them according to your own personal experience.

Q11 I forgot about my immediate surroundings while playing this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q12 I was so involved in this game that I ignored everything around me.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q13 I lost myself in this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q14 I was so involved in this game that I lost track of time.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)

Select an answer: (1)	○	○	○	○	○

Q15 I blocked out things around me when I was playing this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q16 When I was playing this game, I lost track of the world around me.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q18 The time I spent playing this game just slipped away.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q19 I was absorbed in this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q20 During this game I let myself go.

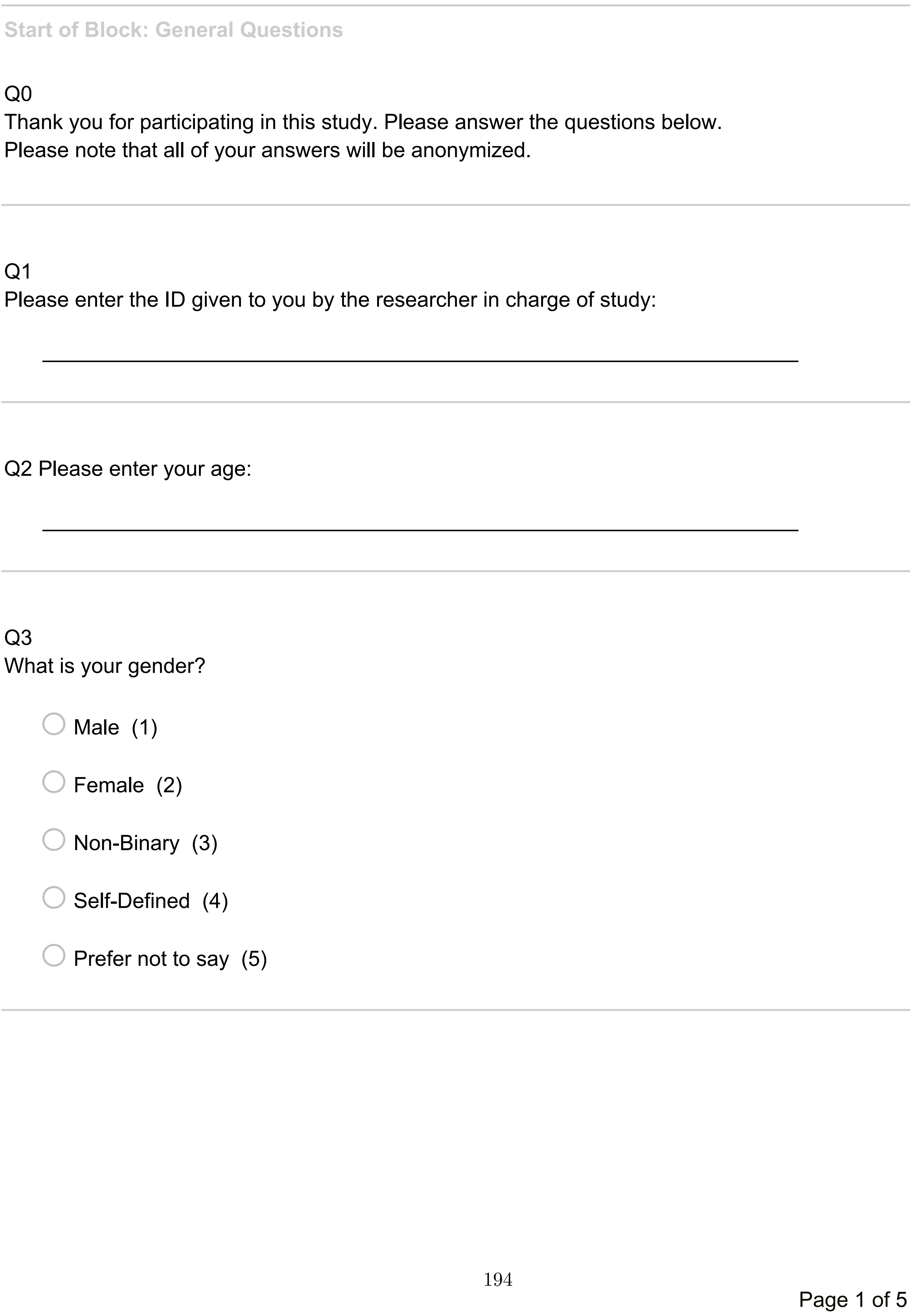
	Strongly agree (1)	Disagree (6)	Neutral (7)	Agree (8)	Strongly Agree (9)
Click to write Statement 1 (1)	○	○	○	○	○

End of Block: Block 3

Appendix C

Visual Uncertainty Study

C.A Pre-Hoc Survey



Start of Block: General Questions

Q0
Thank you for participating in this study. Please answer the questions below.
Please note that all of your answers will be anonymized.

Q1
Please enter the ID given to you by the researcher in charge of study:

____

Q2 Please enter your age:

____

Q3
What is your gender?

○ Male (1)

○ Female (2)

○ Non-Binary (3)

○ Self-Defined (4)

○ Prefer not to say (5)

Page 1 of 5

Q7
What is your gender?

○ Male (1)

○ Female (2)

○ Non-Binary (3)

○ Self-Defined (4)

○ Prefer not to say (5)

Q8 What is your country of origin?

○ Please enter your country of origin below: (1)
____

○ Prefer not to say (2)

Q9 Education: What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

○ No formal education (1)

○ High school diploma (2)

○ College degree (3)

○ Vocational training (14)

○ Bachelor's degree (15)

○ Master's degree (16)

○ Professional degree (17)

○ Doctorate degree (18)

○ Prefer not to say (13)

Q10 In which field are you currently employed? (e.g "automotive industry", "medical" etc...)
If you are studying or on the lookout for a job, which field is your main focus?

____

End of Block: Demographic Survey

Start of Block: Gaming-related Questions

Q11 How often do you play video-games?

	Never (1)	Very Rarely (2)	Rarely (7)	Occasionally (8)	Frequently (9)	Very Frequently (10)
Make your selection: (1)	○	○	○	○	○	○

Q12 What kind of games do you usually play (example: 3D first person shooters, 2D platformers, casual mobiles games etc...). If you selected "never" on the last question, you can skip answering this one.

____

Q13 How old is the computer that you are using for this study?

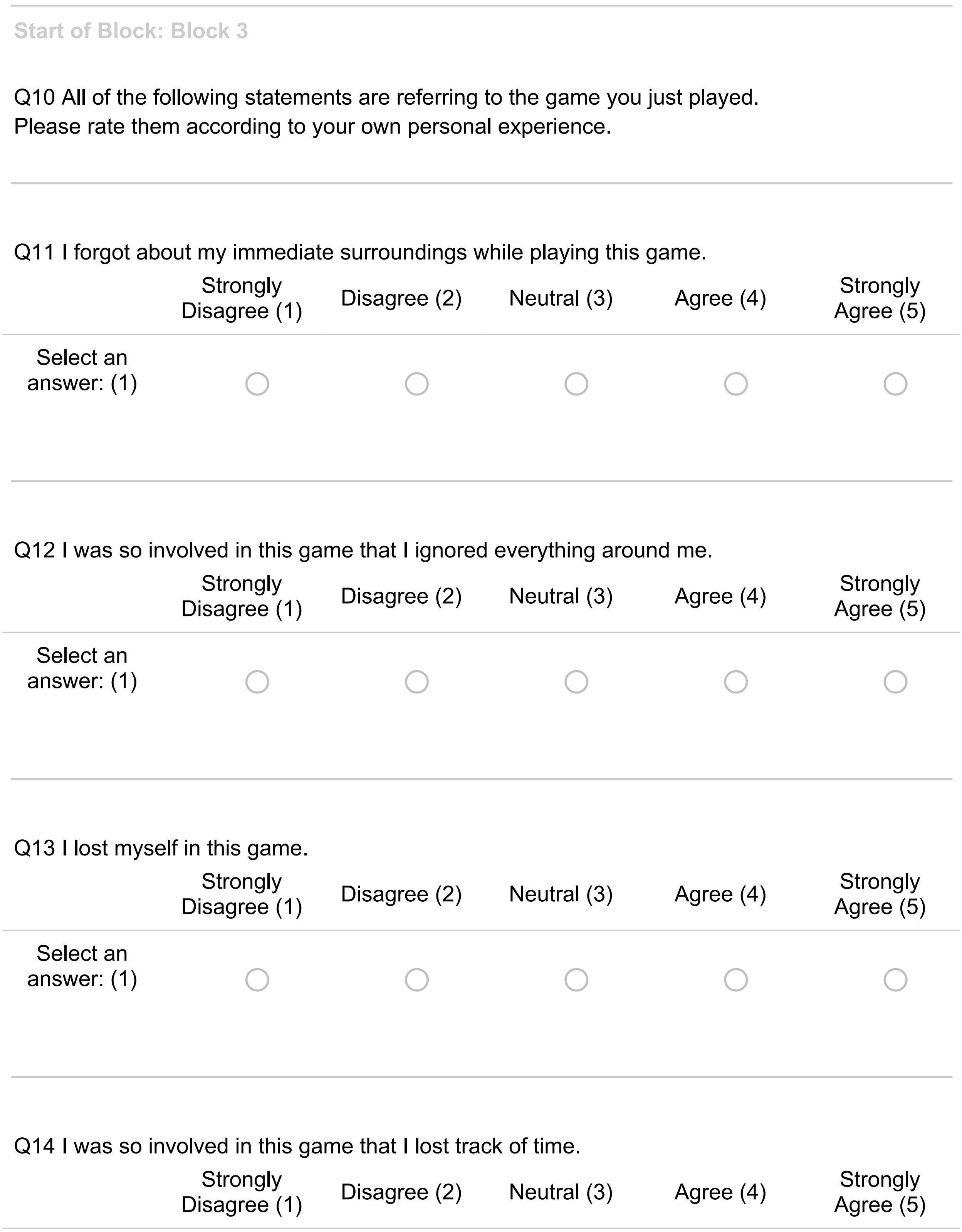
____

Q14
Please indicate below the name of the CPU and/or GPU that your computer is currently using. If you do not know it, simply write the brand/name of your computer (ex: "HP Laptop", "Apple Macbook Air" etc...)

____

End of Block: Gaming-related Questions

C.B Post-Hoc Survey



Start of Block: Block 3

Q10 All of the following statements are referring to the game you just played.
Please rate them according to your own personal experience.

Q11 I forgot about my immediate surroundings while playing this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q12 I was so involved in this game that I ignored everything around me.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q13 I lost myself in this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q14 I was so involved in this game that I lost track of time.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)

Q8 During one of the session, **a "disruption" (static-like effect) hid the TOP of the screen**. How did that affected your ability to play and your reliance on the agent?

____

____

____

____

____

Q10 During one of the session, **a "disruption" (static-like effect) hid** the BOTTOM of the screen. How did that affected your ability to play and reliance on the agent?

____

____

____

____

____

Q13
Please rate the difficulty of each conditions from **0 (very easy)** to **100 (very hard)**:

	Easy Medium Hard
	0 10 20 30 40 50 60 70 80 90 100
Darkness & Spotlight ()	
Clouds ()	
Disruption (top of the screen) ()	
Disruption (bottom of the screen) ()	

Q15 Please give us some feedback or recommendations about the study below:

____

____

____

____

____

End of Block: Default Question Block

Appendix D

Visual Explanation and Agent Transparency Study

D.A Pre-Hoc Survey

Start of Block: General Questions

Q0
Thank you for participating in this study. Please answer the questions below.
Please note that all of your answers will be anonymized.

Q1
Please enter the ID given to you by the researcher in charge of study:

____

Q2 Please enter your age:

____

Q3
What is your gender?

○ Male (1)

○ Female (2)

○ Non-Binary (3)

○ Self-Defined (4)

○ Prefer not to say (5)

Q7
Gender: How do you identify?

○ Female (1)

○ Male (2)

○ Non-Binary (3)

○ Self-Defined: (4) ____

○ Prefer not to say (5)

X→

Q19 Origins: Where are you from?

▼ Afghanistan (1) ... Zimbabwe (1357)

Q9 Education: What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

○ No formal education (1)

○ High school diploma (2)

○ College degree (3)

○ Vocational training (14)

○ Bachelor's degree (15)

○ Master's degree (16)

○ Professional degree (17)

○ Doctorate degree (18)

○ Prefer not to say (13)

End of Block: Demographic Survey

Start of Block: Gaming-related Questions

Q11 Please read and rate the following statements.

	1 Strongly Disagree (12)	2 Somewhat Disagree (13)	3 Neither disagree or agree (14)	4 Somewhat Agree (15)	5 Strongly Agree (16)
I am good at playing computer games. (2)	○	○	○	○	○
Playing computer games is easy for me. (3)	○	○	○	○	○
I understand and play computer games well.	○	○	○	○	○

(4)					
I am skilled at playing computer games (5)	○	○	○	○	○

End of Block: Gaming-related Questions

Start of Block: Block 4

Q18 Please read and rate the following statements.

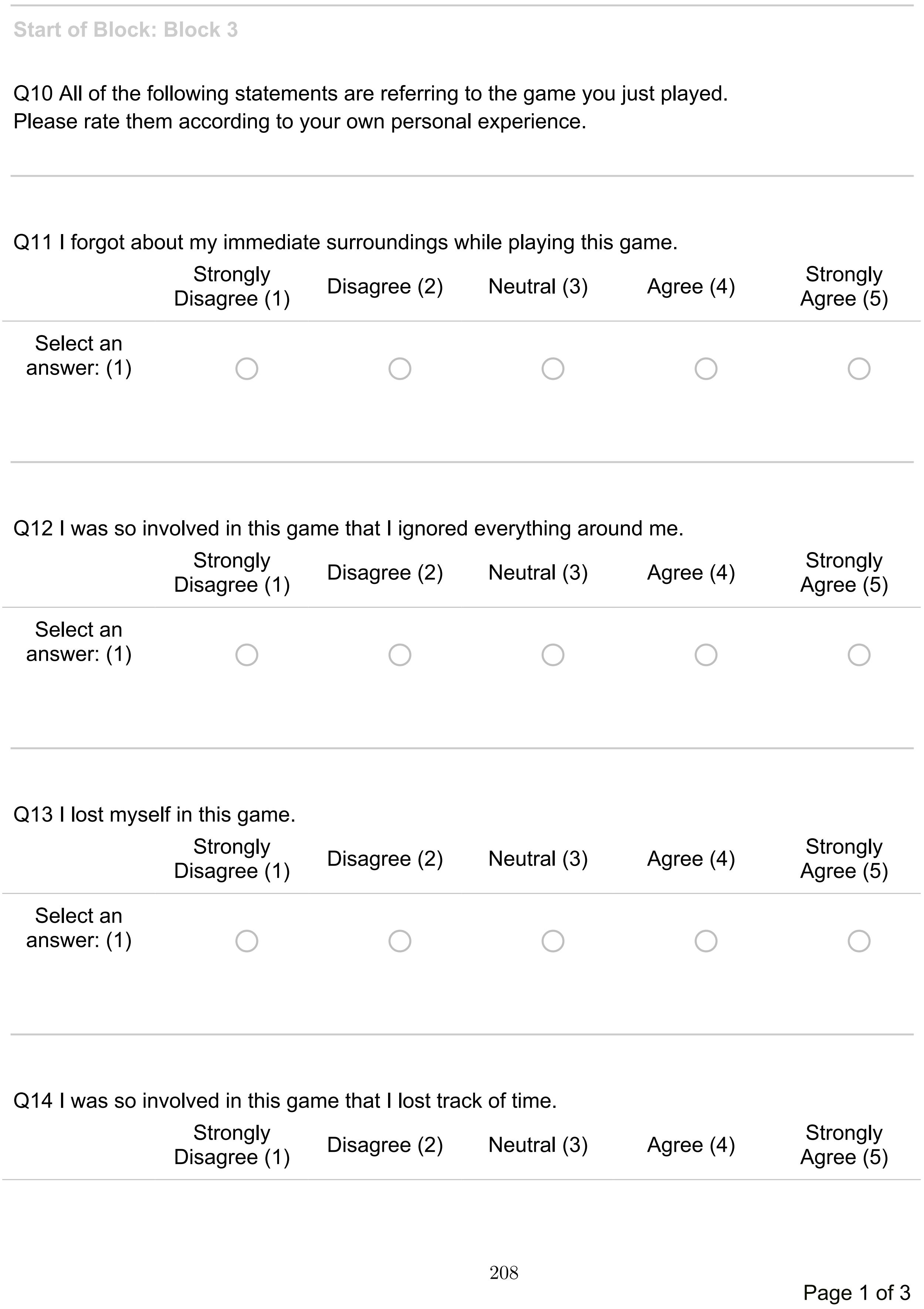
	1 Strongly disagree (18)	2 Somewhat disagree (19)	3 Neither agree nor disagree (20)	4 Somewhat agree (21)	5 Strongly agree (22)
When I have a lot to do, it makes sense to delegate a task to automation. (1)	○	○	○	○	○
If life were busy, I would let an automated system handle some tasks for me. (2)	○	○	○	○	○
Automation should be used to ease people’s workload. (3)	○	○	○	○	○
If automation is available to help me with something, it makes sense for me to pay more attention to my other tasks. (4)	○	○	○	○	○

Even if an automated aid can help me with a task, I should pay attention to its performance. (5)	○	○	○	○	○
Distractions and interruptions are less of a problem for me when I have an automated system to cover some of the work. (6)	○	○	○	○	○
Constantly monitoring an automated system’s performance is a waste of time. (7)	○	○	○	○	○
Even when I have a lot to do, I am likely to watch automation carefully for errors. (8)	○	○	○	○	○
It’s not usually necessary to pay much attention to automation when it is running. (9)	○	○	○	○	○
Carefully watching automation takes time	○	○	○	○	○

away from more important or interesting things. (10)

End of Block: Block 4

D.B Post-Hoc Survey



Start of Block: Block 3

Q10 All of the following statements are referring to the game you just played.
Please rate them according to your own personal experience.

Q11 I forgot about my immediate surroundings while playing this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q12 I was so involved in this game that I ignored everything around me.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q13 I lost myself in this game.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
Select an answer: (1)	○	○	○	○	○

Q14 I was so involved in this game that I lost track of time.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)

Q12 You can use the textbox below to give us some additional feedback about the study:

____

____

____

____

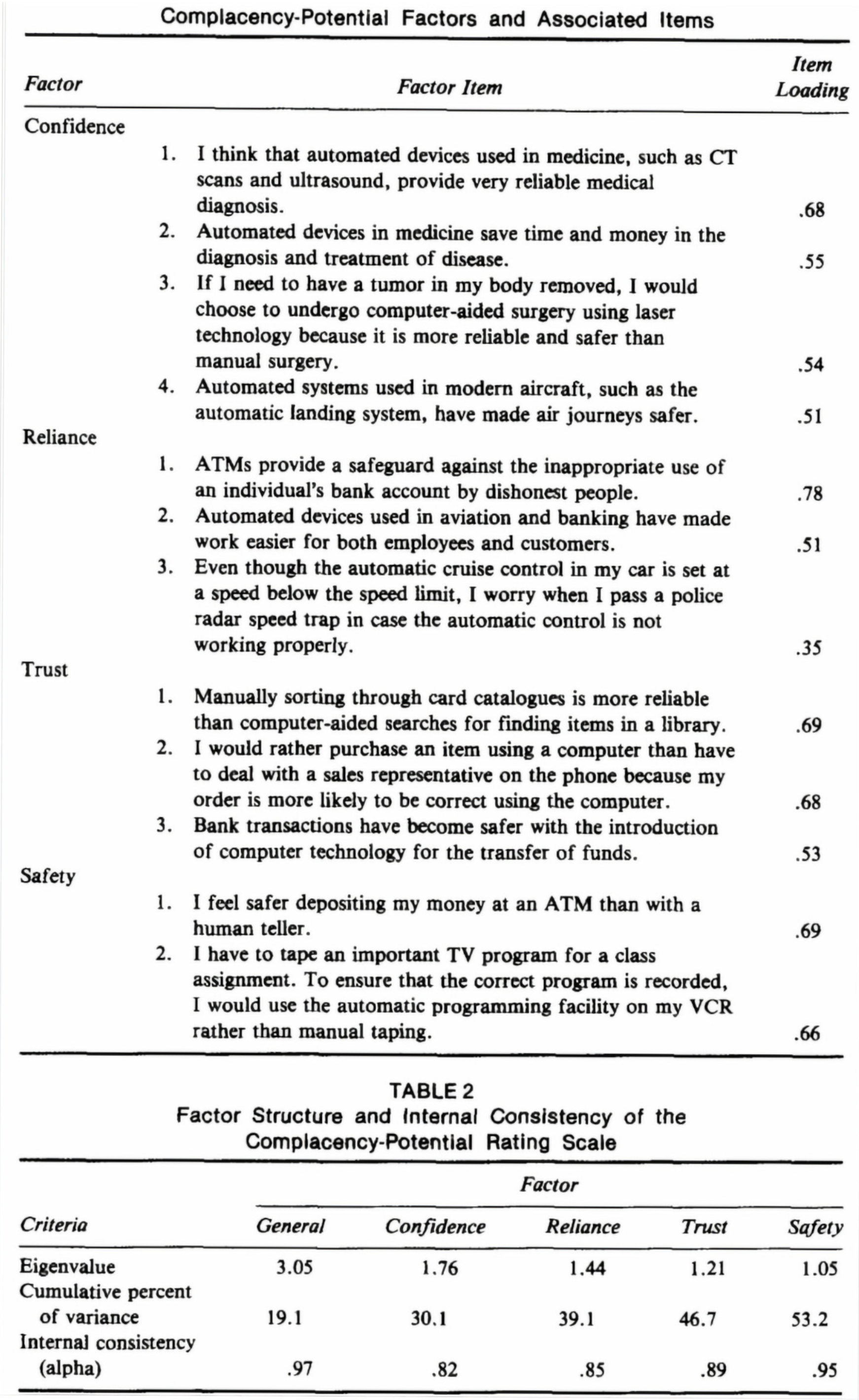
____

End of Block: Last questions

Appendix E

Survey Instruments

E.A Complacency Potential Questionnaire



Complacency-Potential Factors and Associated Items

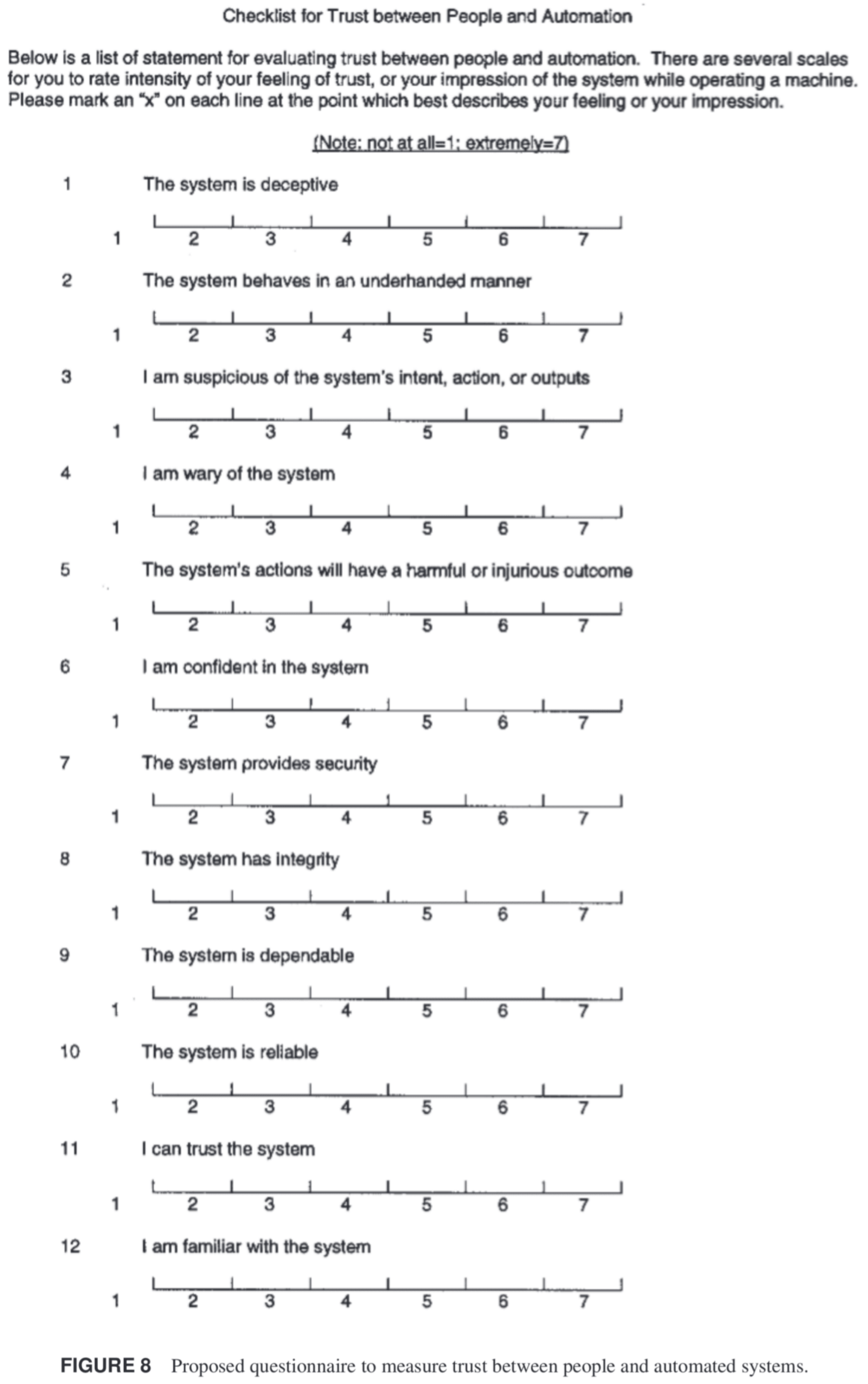
Factor	Factor Item	Item Loading
Confidence		
	1. I think that automated devices used in medicine, such as CT scans and ultrasound, provide very reliable medical diagnosis.	.68
	2. Automated devices in medicine save time and money in the diagnosis and treatment of disease.	.55
	3. If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because it is more reliable and safer than manual surgery.	.54
	4. Automated systems used in modern aircraft, such as the automatic landing system, have made air journeys safer.	.51
Reliance		
	1. ATMs provide a safeguard against the inappropriate use of an individual's bank account by dishonest people.	.78
	2. Automated devices used in aviation and banking have made work easier for both employees and customers.	.51
	3. Even though the automatic cruise control in my car is set at a speed below the speed limit, I worry when I pass a police radar speed trap in case the automatic control is not working properly.	.35
Trust		
	1. Manually sorting through card catalogues is more reliable than computer-aided searches for finding items in a library.	.69
	2. I would rather purchase an item using a computer than have to deal with a sales representative on the phone because my order is more likely to be correct using the computer.	.68
	3. Bank transactions have become safer with the introduction of computer technology for the transfer of funds.	.53
Safety		
	1. I feel safer depositing my money at an ATM than with a human teller.	.69
	2. I have to tape an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my VCR rather than manual taping.	.66

TABLE 2
Factor Structure and Internal Consistency of the Complacency-Potential Rating Scale

	Factor				
Criteria	General	Confidence	Reliance	Trust	Safety
Eigenvalue	3.05	1.76	1.44	1.21	1.05
Cumulative percent of variance	19.1	30.1	39.1	46.7	53.2
Internal consistency (alpha)	.97	.82	.85	.89	.95

Figure E.1: Complacency Potential Questionnaire by Singh et al. [167]

E.B Checklist for Trust between People and Automation



Checklist for Trust between People and Automation

Below is a list of statement for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please mark an "x" on each line at the point which best describes your feeling or your impression.

(Note: not at all=1; extremely=7)

1 The system is deceptive

1 2 3 4 5 6 7

2 The system behaves in an underhanded manner

1 2 3 4 5 6 7

3 I am suspicious of the system's intent, action, or outputs

1 2 3 4 5 6 7

4 I am wary of the system

1 2 3 4 5 6 7

5 The system's actions will have a harmful or injurious outcome

1 2 3 4 5 6 7

6 I am confident in the system

1 2 3 4 5 6 7

7 The system provides security

1 2 3 4 5 6 7

8 The system has integrity

1 2 3 4 5 6 7

9 The system is dependable

1 2 3 4 5 6 7

10 The system is reliable

1 2 3 4 5 6 7

11 I can trust the system

1 2 3 4 5 6 7

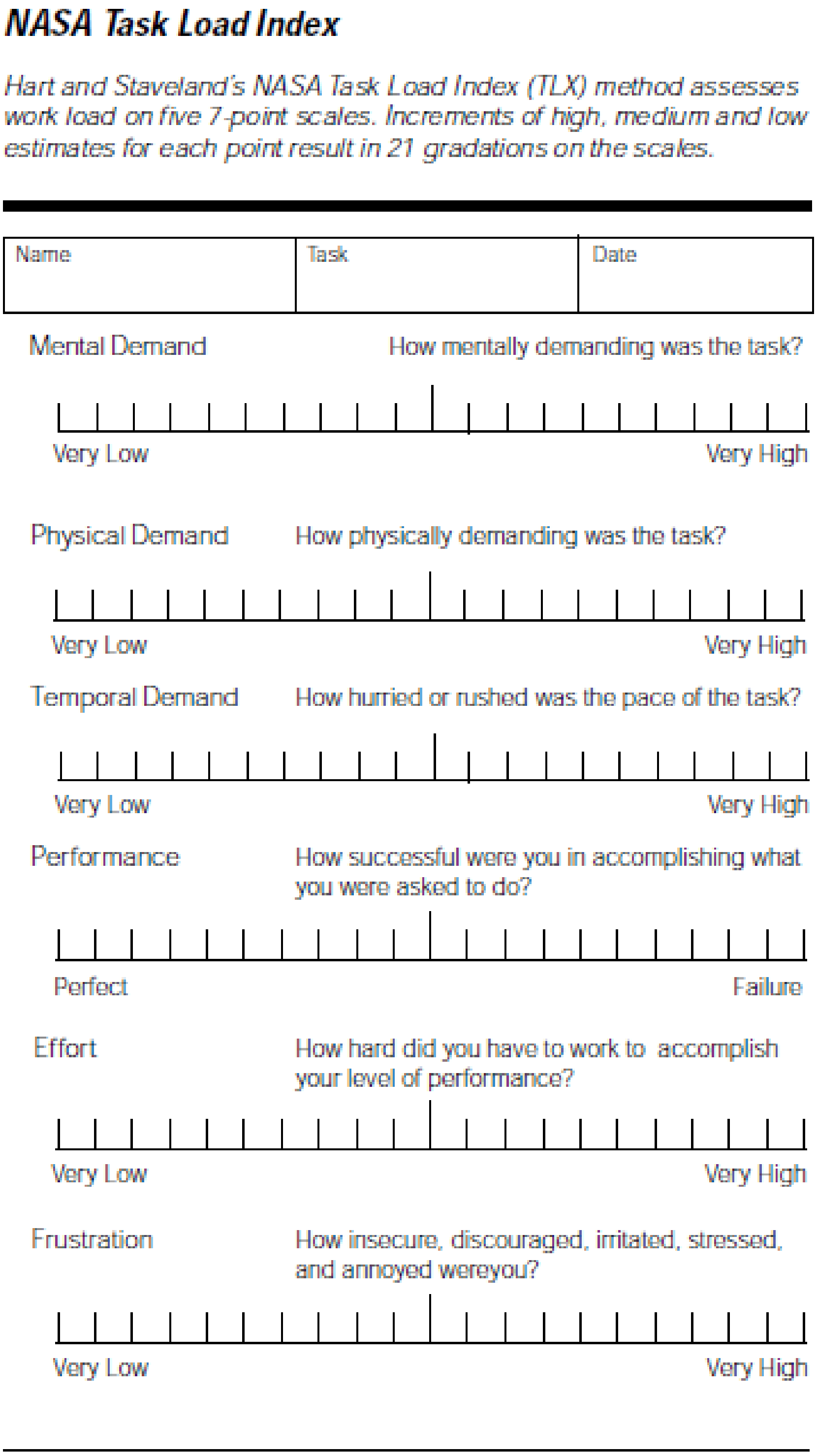
12 I am familiar with the system

1 2 3 4 5 6 7

FIGURE 8 Proposed questionnaire to measure trust between people and automated systems.

Figure E.2: Checklist for Trust Between People and Automation by Jian et al. [90].

E.C NASA Task Load Index (TLX)



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

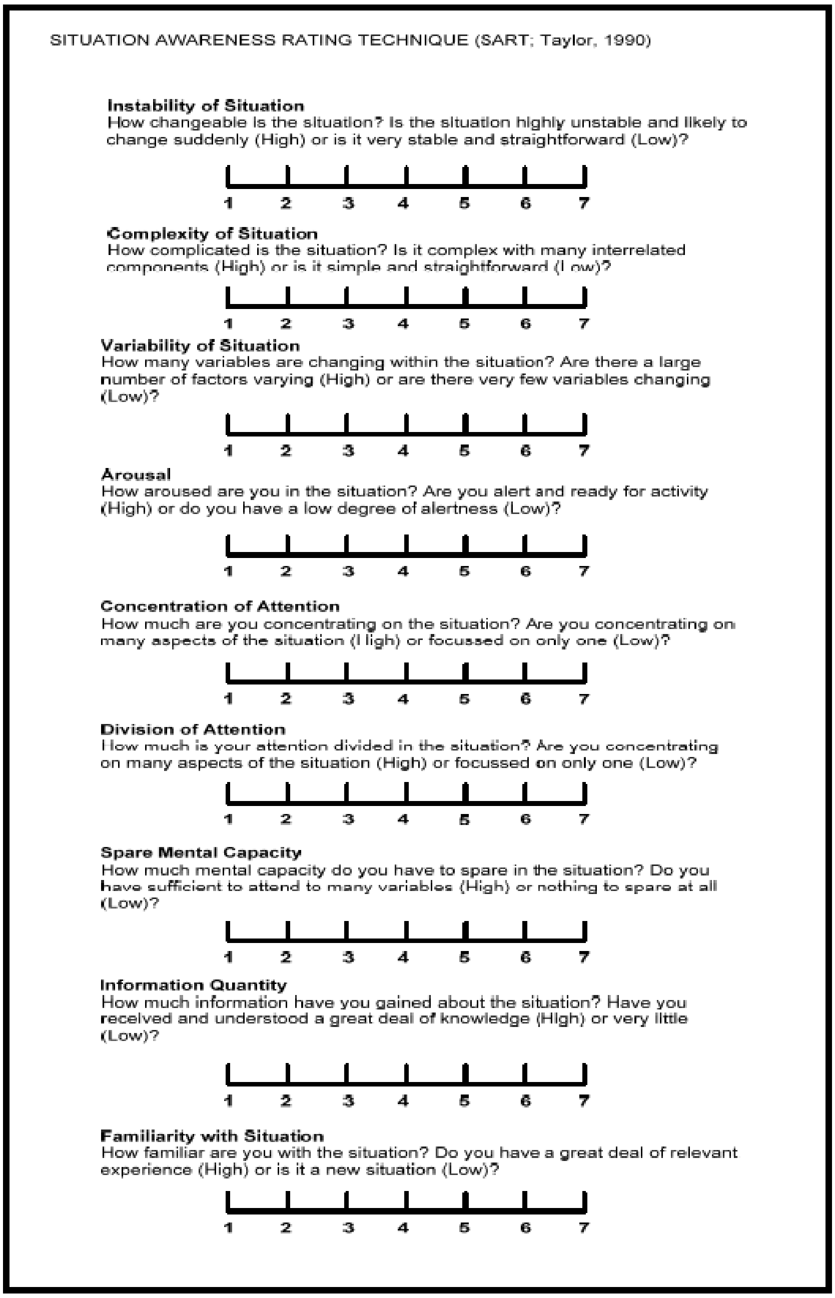
Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

Figure E.3: NASA TLX Cognitive Workload survey designed by Hart et al. [76]

E.D Situational Awareness Rating Technique (SART)



SITUATION AWARENESS RATING TECHNIQUE (SART; Taylor, 1990)

Instability of Situation
How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?

1 2 3 4 5 6 7

Complexity of Situation
How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?

1 2 3 4 5 6 7

Variability of Situation
How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)?

1 2 3 4 5 6 7

Arousal
How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?

1 2 3 4 5 6 7

Concentration of Attention
How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (I ligh) or focussed on only one (Low)?

1 2 3 4 5 6 7

Division of Attention
I low much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

1 2 3 4 5 6 7

Spare Mental Capacity
How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?

1 2 3 4 5 6 7

Information Quantity
How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?

1 2 3 4 5 6 7

Familiarity with Situation
How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?

1 2 3 4 5 6 7

Figure E.4: Situational Awareness Rating Technique designed by Taylor et al. [173]