The Impact of Explanations of Artificial Trust on Human-Agent Teamwork

Zenan Guan

Delft University of Technology

TU Delft

The Impact of Explanations of Artificial Trust on Human-Agent Teamwork

by

Zenan Guan

Student Name	Student Number
Zenan Guan	4813855

Thesis advisor:	Myrthe Tielman
Thesis committee member:	Mark Neerincx
External thesis committee member:	Ujwal Gadiraju
Daily co-supervisor:	Ruben Verhagen
Project Duration:	February, 2024 - December, 2024
Faculty:	Faculty of Computer Science, Delft

TU Delft

Acknowledgement

Completing this thesis marks the culmination of my Master's journey, a chapter filled with invaluable lessons, challenges, and moments of growth that have shaped me both academically and as a person. This work would not have been possible without the guidance, support, and encouragement of many wonderful people to whom I owe my deepest gratitude.

First, I sincerely thank Myrthe Tielman, my thesis advisor, for her unwavering guidance and insightful advice throughout this process. Her expertise and encouragement have been instrumental in shaping the direction of my research. I am grateful to Ruben Verhagen for his continuous support, thoughtful feedback, and inspiring discussions with me.

I would also like to thank Mark Neerincx and Ujwal Gadiraju for being part of my thesis committee and for their valuable insights that enriched this work.

This project would not have been possible without the kindness and generosity of my friends and family, who took time out of their busy schedules to participate in my experiments. Your contributions made a significant impact on the success of this study.

Finally, thank my parents, grandparents, and other family members for your endless love, encouragement, and belief in me, which has been my greatest strength throughout this journey.

Zenan Guan
Delft, December 2024

Abstract

Human-agent teamwork (HAT) is becoming increasingly prevalent in fields such as search and rescue (SAR), where effective collaboration between humans and artificial agents is crucial. Previous studies have shown that trust plays a pivotal role in the success of HATs, influencing decision-making, communication, and potentially overall team performance.

This research investigates the impact of agent-provided explanations about the agent's trust in humans (artificial trust) and corresponding behavior changes on human trust in the agent and their satisfaction with explanations during a simulated SAR task. Two types of explanations were explored: Trust-Explained (TE) explanations, where the agent explains its trust level and trust-based decisions, and Trust-Unexplained (TU) explanations, which solely describe the agent's behavior without reference to trust dynamics. Besides, this research also investigates the correlation between human trust and explanation satisfaction, and in the end, whether the differences in the provided explanations result in differences in team performance and artificial trust.

The study involved 40 participants divided into two groups: an experimental group (the trust-enhanced explanation group) receiving TE explanations and a control group (the non-trust explanation group) receiving TU explanations. Participants' trust in the agent, satisfaction with the explanations, and team performance and artificial trust were measured and analyzed. Contrary to initial expectations, no statistically significant differences in explanation satisfaction and human trust in the agent were found between the two groups. However, a strong positive correlation was observed between participants' satisfaction with the explanations and their trust in the agent, indicating that explanation quality plays a crucial role in human trust development. Furthermore, no significant differences in team performance were detected, suggesting that trust explanations may not directly influence task outcomes. In the analysis of artificial trust, the agent in the trust-enhanced explanation group exhibited more conservative adjustments in trust levels compared to the non-trust explanation group. This conservative approach may have influenced players in the trust-enhanced explanation group to adopt a more cautious or deliberate decision-making process, potentially prioritizing the comprehension of explanations over the optimization of task performance.

For future research, it may be worth delving deeper into the influence of trust explanations on user behavior, the more complex HAT task environments, the relationship between artificial trust and user behavior, the dynamic and adaptive explanations, and the causal relationship between explanation satisfaction and human trust in the agent to understand further how trust can be fostered in HAT.

Contents

1

Introduction

Human-agent teamwork (HAT) refers to the collaboration between humans and artificial agents, where both entities coordinate to achieve a shared objective in dynamic, often unpredictable environments [100]. Trust is a critical factor in the success of these collaborations, as it affects behaviors, decision-making, and the overall effectiveness of teamwork[13]. In this context, trust can be divided into two distinct forms: human trust, which refers to the trust that humans place in agents, and artificial trust, the trust that agents place in human collaborators[43]. While much attention has been given to human trust in agents[49], the concept of artificial trust is relatively underexplored, despite its significant implications for effective collaboration. Research indicates that artificial agents need to trust humans just as much as humans trust them, particularly in scenarios where mutual trust determines the success of human-agent interaction [5]. Additionally, studies show that agents capable of assessing the trustworthiness of human collaborators can adjust their behavior to improve interaction efficiency and effectiveness [43]. Furthermore, the development of mental models to help agents understand human trustworthiness in collaborative settings is gaining traction [48].

Existing literature has demonstrated trust as a cornerstone of effective HAT, with extensive research exploring methods to model, evaluate, and foster trust. For example, Jorge et al.'s [50] exploration of the effects of automation failure on human trustworthiness revealed that automation failures negatively affected human trust in and liking of the automation, as well as the trustworthiness of participants themselves. Furthermore, Bobko et al. [10] proposed a theoretical framework and testbed for investigating trust calibration, showing how increased transparency and reliability in HATs leads to calibrated trust, bringing more positive effects, lower workload, and enhancing task engagement[10]. These findings underscore the importance of trust in creating robust and effective collaborations in HAT.

A recurring theme in these studies is the role of Explainable Artificial Intelligence (XAI) in enhancing mutual trust. XAI not only bridges gaps in understanding but also facilitates trust dynamics critical to HAT. For example, Wang et al.[96] showed how automatically generated explanations improve transparency, trust, and team performance. Additionally, Kox et al.[54] highlighted that affective and informational explanations, coupled with expressions of regret, are instrumental in repairing trust after violations. These works collectively illustrate the dual importance of proactive and reactive trust-building strategies, reinforcing the potential of XAI to address the multifaceted challenges of trust in HAT.

Most of the current research focuses primarily on explaining the actions of agents to humans. For example, Pandya et al.'s[69] framework for multi-agent strategy explanation generation in HAT that is devoted to improving user's understanding of agents' actions. Similarly, the parsimonious XAI architecture by Mualla et al.[64] discussed providing the simplest explanation that describes a system's behavior. However, these approaches predominantly address how to explain agent actions, leaving a critical gap in understanding the reciprocal dynamics of trust—specifically, how an agent's explanations of its trust in a human counterpart influence the human's trust in return.

This gap is particularly relevant because mutual trust could enhance smooth and efficient collaboration,

improving task performance, reliance, and cognitive load in HATs. For example, research by Chen et al. [16] indicates that situation-awareness-based agent transparency models help humans better understand the agent's actions, leading to increased trust and better decision-making. Moreover, greater predictability in agent behaviors positively impacts human trust, cognitive load, and performance during collaboration tasks [22]. Therefore, in the following sections, we will elaborate on the scientific motivation behind this research and present the research questions that aim to address this significant gap.

1.1. Scientific motivation

Human-agent collaboration has been deployed in various real-world domains. For example, in the healthcare field, agents provide decision support and manage task allocation among medical staff, streamlining patient care and resource management in high-pressure situations like mass casualty events [59]. Additionally, a recent study by Dhatterwal et al. [24] found that integrating swarm intelligence with multi-agent systems in hospitals has improved medical diagnostics and the coordination of care among wards, increasing efficiency in patient management. In military operations, agents coordinate operations, manage resources, and enhance situational awareness. They assist in planning and executing missions, allowing human operators to focus on critical decision-making while agents handle routine or complex computational tasks [17]. Another example includes a model for integrating agents into human-military teams, improving situational awareness and task coordination in complex operational environments [38]. In the industrial environment, agents collaborate with human workers to optimize workflows, manage supply chains, and ensure safety in heavily automated environments. Their role in task planning and execution reduces human error and improves overall efficiency [78]. Last but not least, Search and Rescue (SAR) tasks are a typical field in which human-agent collaboration is deployed, where humans and agents work together to search and rescue victims in a disaster. In SAR missions, agents assist in task allocation, real-time monitoring, and environmental mapping, thereby enhancing the efficiency of SAR teams in dynamic and uncertain environments. For instance, multi-robot systems can quickly locate victims, assess conditions, and establish communication networks, significantly aiding SAR personnel [72, 71].

Among all these applications, understanding human-agent mutual trust is critical for effective collaboration between humans and artificial agents[99]. Therefore, this research is motivated by several key motivations. First, enhanced collaboration and performance. Mutual trust between humans and agents improves the efficiency and effectiveness of teamwork. Research by Wang et al. [99] demonstrates that appropriate levels of mutual trust can optimize real-time collaboration, avoiding the pitfalls of over-trust (where humans overly rely on agents) and under-trust (where humans do not rely on capable agents). When trust is well-calibrated, both human and agent performance improves, leading to better outcomes in collaborative tasks [99].

Second, increased acceptance and adoption of AI systems. Trust is crucial in the human acceptance of AI technologies. When users trust that an AI system will behave reliably and predictably, they are more likely to use and benefit from it. Studies have shown that reliable, predictable AI systems enhance trust and increase usage. For example, Kaplan et al. [51] found that trust is influenced by various factors, such as the transparency, reliability, and explainability of AI algorithms. Srinivasan & de Boer [82] found that ensuring that AI systems are free of biases and maintaining transparency throughout the decision-making process is critical to building trust. Meta-analytic findings [19] further emphasized that users' trust in AI is built upon system reliability, transparency, and ease of use. Moreover, Centeio Jorge et al. [48] highlighted that building trustworthy agents who understand when to trust their human counterparts can lead to more effective and harmonious human-AI teams. This is particularly important as AI systems become more autonomous and are expected to make decisions without constant human oversight.

Third, improved safety and reliability. Understanding trust between humans and AI agents can help prevent failures and accidents in human-agent interactions. Research indicates that over-trusting automation, particularly in high-risk environments like healthcare, may lead to users relying too much on the system, ignoring their judgment, and creating unsafe scenarios [41]. Another example in healthcare, robots intended to work alongside humans must ensure proper levels of trust to avoid hazardous outcomes, such as incorrect reliance on robotic assistance [85]. Similarly, in autonomous vehicles,

improper trust can lead to dangerous over-reliance or failure to act on the vehicle's recommendations [56]. Establishing the right level of trust through transparency and reliability can help prevent these issues [99]. Besides, to avoid both underutilization and over-reliance, trust calibration and transparency are crucial to maintaining safe human-robot collaboration in these fields [15]. Huang and Bashir [42] also found that trust dynamics play a significant role in the safe and effective deployment of automated systems.

Finally, social and ethical implications: Trust in AI systems is closely linked to ethical considerations. As AI systems take on roles that involve significant decision-making power, understanding the dynamics of mutual trust helps ensure these systems align with human values and ethics. This alignment is critical for maintaining public trust in AI technologies as they become more pervasive in society [83].

1.2. Research Questions

Building on the scientific motivation outlined, this research addresses the following primary question:

How does adding information about artificial trust level and the agent's behavior changes corresponding to its trust level in the agent's explanations affect human trust and explanation satisfaction?

Additionally, we want to investigate two related research questions beyond the main research question.

The first related question is whether there is a correlation between human trust and explanation satisfaction, which are two dependent variables in the main research question:

How does the participants' satisfaction with the explanations correlate with their trust in the agent?

Additionally, we aim to investigate whether providing additional information about artificial trust influences the artificial trust level itself, and team performance:

To what extent do differences in the agent's explanations about artificial trust and its behavior impact the artificial trust itself, and team performance?

In this thesis project, we want to bridge the research gap mentioned by exploring the impact of agent-provided explanations on human trust in a simulated SAR task. More specifically, will the existence of explanations about changes in the artificial trust due to human behavior in the SAR task affect human trust? This focus is novel and crucial for designing more effective collaborative systems for various scenarios. By delving into the uncharted territory of how agent communication about trust fluctuations influences human trust, this research contributes to theoretical understandings and practical applications in human-agent collaboration.

In summary, by building on foundational works and exploring a new dimension of trust dynamics, this thesis aspires to contribute theoretical insights and practical guidelines that enhance the mutual trust and efficacy of human-agent collaborations in complex and dynamic environments.

2

Background

In this section, we will talk about the background knowledge and previous research on which our research is based, as well as the knowledge gap we intend to fill.

2.1. Human-Agent Teamwork

Human-agent teamwork (HAT) involves the collaboration between humans and artificial agents, where both parties work together towards a common goal in dynamic environments [100]. This involves shared decision-making, coordination, and adapting to each other's actions and needs to achieve common goals [57]. The work by [14] defines human-agent teaming as the collaborative effort between humans and intelligent systems to manage and supervise multiple robots efficiently. This collaboration seeks to blend the strengths of humans and automated agents to optimize the operational effectiveness of robotic systems in various domains such as transportation, safety, search and rescue, space exploration, and military operations. HAT centers around key issues like efficient human supervision, fostering appropriate trust in automated systems, maintaining the operator's situation awareness, and managing individual differences in human-agent interactions while ensuring human decision-making remains paramount.

HAT is prevalent in real-world applications nowadays, including healthcare, finance, education, and customer service, where the integration of artificial intelligence (AI) and human expertise has shown significant potential. To illustrate, AI assists doctors by analyzing medical images (e.g., detecting radiograph anomalies [7], predicting patient outcomes, and personalizing treatment plans. In finance, AI algorithms are used for high-frequency trading, risk management, and fraud detection, complementing human intuition and strategic planning [2]. In the application of education technology, AI systems provide personalized learning experiences, helping educators identify and address students' individual needs [31]. Customer service has also seen a rise in AI-driven chatbots that handle routine inquiries, freeing human agents to manage more intricate customer issues [80]. These applications show the great advantages of HAT, especially regarding enhancing efficiency and productivity, where AI agents can handle repetitive, time-consuming, or computationally intensive tasks, allowing humans to focus on more creative, strategic, and complex problem-solving activities [7]. Such applications also provide complementary skills, where AI Agents contribute computational power, consistency, and the ability to process large amounts of data quickly, while humans bring creativity, emotional intelligence, and the ability to make nuanced or ethics-related decisions based on context and experience [53].

One typical application of HATs is during Search and Rescue (SAR) tasks. Such cooperation leverages the unique advantages of both humans and robots [4], e.g., humans make moral decisions [60], while robots perform tasks in dangerous environments that are difficult for humans to access [95], to accomplish the mission they cannot do alone. Studying HAT in a simulated SAR context provides an opportunity to explore the dynamics of such collaboration with agents modelled as robots.

There are various studies about HAT in simulated SAR tasks. For example, Verhagen et al. [89] explored quantitative operationalizations of meaningful human control during dynamic task allocation using

variable autonomy in human-robot teams in a simulated firefighting environment. Another paper with the same first author[90] investigated how varying degrees of interdependence in HATs influence trust calibration using a simulated SAR task. Moreover, the research conducted by Jain et al.[44], where they try to understand what is inside a rescuer's mind by using a simulated search and rescue task in Minecraft. In all these examples, a SAR can be divided into two parts: search task and rescue task. The primary objective of the search mission is to look for survivors who are obstructed or unable to enter a safe area, while the main goal of the rescue task is to convey survivors to the secure zone and rescue any injured ones throughout the rescue mission.

2.2. Interdependence

Interdependence between humans and agents is crucial to HAT. In [45], interdependence is defined as the mutual reliance between team members within a human-machine team. This definition encompasses the idea that members of a team, whether human or robot, are not working in isolation but are interconnected in ways that their actions and decisions affect one another. In this context, interdependence is a foundational element that influences how team members collaborate, communicate, and adapt to achieve their common goals.

There are two kinds of interdependence, which are hard and soft interdependence [46]. Hard interdependence indicates the necessity of collaboration in accomplishing some activities that can otherwise not be finished individually. Soft interdependence exists in those activities that can be finished by a human or an agent alone, but collaboration can improve efficiency, therefore making it optional but opportunistic [92].

Besides, there are four kinds of task interdependence that form a hierarchy, representing increasing levels of dependence between team members and greater coordination needs, which consist of pooled, sequential, reciprocal, and team interdependence [79]. Pooled interdependence occurs when each team member works independently, contributing to the overall task without needing to interact with others. There is no direct reliance between members. This type of interdependence is typically associated with the simplest form of teamwork. Sequential interdependence involves tasks being performed in a set order, where the output of one member is necessary for the next to proceed. This creates a chain-like process, as seen in assembly lines, where each participant depends on the previous one to complete their part before starting their own. Reciprocal interdependence occurs when team members are mutually dependent, exchanging inputs and outputs during the task. Their actions influence each other continuously. This is common in settings like medical or surgical teams, where actions are interleaved and require constant back-and-forth coordination. Team interdependence is the highest level, where all team members work together simultaneously on the same task, executing their actions concurrently. This form of interdependence involves joint actions, such as two people lifting a heavy object together.

The combination of soft and hard interdependency, and four kinds of task interdependencies can result in low or high interdependence between humans and agents, For example, in [92], under the context of SAR, the low interdependence scenarios involve tasks where the human and robot operate independently with minimal interaction, including managing separate drop zones during search and rescue missions, allowing each to handle specific areas or victims without coordinating with each other. Conversely, high interdependence can be exemplified by a single drop zone where both must work together to rescue victims. Here, the robot relies on the human to perform tasks it cannot, such as carrying critically injured adults or distinguishing between children, necessitating ongoing communication and support.

2.3. Trust

Trust is a cornerstone for effective HAT and dealing with interdependence in it. As Elson et al [26] said: "Trust between humans and artificial agents is critical to effective collaboration in mixed human-computer teams. Understanding the conditions under which humans trust and rely upon automated agent recommendations is important, as trust is one of the mechanisms that allow people to interact effectively".

According to Dagli [20], trust is a social construct that originates from interpersonal relationships. Studies have shown that trust in human-agent teams is critical for ensuring smooth collaboration and optimal performance. For instance, research by Wissen [100] demonstrates that trust affects both

the decision-making processes and the stability of team commitments, particularly in settings that require rapid responses under uncertain conditions. This dynamic is critical as teams often need to adapt quickly to changing environments without full information. In another research by Jong et al. [47], trust dynamics are observed to correlate with team performance, where lower levels of trust are associated with poorer team outcomes, which suggests that maintaining high levels of trust is essential for achieving optimal team performance. Trust levels can change over time, impacting long-term team collaboration and effectiveness.

In [86], the authors further elaborate on trust by identifying three key antecedents essential for its development in human-agent teams: Integrity, Ability, and Benevolence, where Integrity is related to the trust that team members (both human and agents) will act as they have stated they would. This aspect of trust is built on positive expectations about team members' behaviors. Ability refers to the competence of team members to perform necessary actions effectively. Trust in a team member's ability grows from recognizing that they possess the required skills and knowledge to fulfill their roles within the team. This competence supports the team's overall effectiveness and is crucial for developing trust. Benevolence involves the belief that team members are looking out for each other's well-being and the good of the team. This element of trust assumes that all team members, including humans and agents, prioritize the team's interests and support one another throughout different phases of the team's lifecycle. If team members consistently act in a way that aligns with what they've committed to, it fosters a higher level of trust within the team. These factors facilitate initial trust formation and are essential for organizational teams' ongoing functionality and success [86].

2.3.1. Modelling of Artificial Trust towards Humans

Trust can be divided into two categories in HAT, which are agent trust towards humans (artificial trust) and human trust towards agents. Several studies show how to model artificial trust towards humans in HAT. In [49], the authors introduced a conceptual model of artificial trust in human-agent teamwork, focusing on how agents can form beliefs about human trustworthiness based on observations during specific tasks. The model differentiates between beliefs of competence (ability to perform expected results) and willingness (intent to perform a task) as core components of trust. It also explores factors like human strategy and observed behaviors (manifesta) that could influence these beliefs, such as performance, fairness, and commitment. Another example is Lin et al. [58], where they proposed a model for evaluating the trustworthiness of human agents dynamically by assessing real-time cognitive states such as attention, stress, and perception, employing fuzzy reinforcement learning to fuse this information and generate trust values that reflect the current state of the human agents. Their model aims to enhance the efficiency and decision-making in HAT systems by enabling agents to adapt their behavior based on the real-time trust values of their human counterparts, resulting in improved coordination and task performance. These models demonstrate various approaches to evaluating and integrating human trustworthiness in human-agent teamwork (HAT) systems to enhance coordination and task performance. While the first model concentrates on forming static beliefs based on task-related observations to gauge trustworthiness, the second model continuously assesses and updates trust based on real-time cognitive states to adaptively manage interactions in HAT systems.

2.3.2. Evaluation of Human Trust towards Agents

According to Hoffman et al. [40], human trust towards agents in HAT can be defined as an emotional judgment about how much a human participant can rely on a system when uncertain. Mehrotra et al. [62] determined three ways of measuring trust exist: perceived trust, where a person's subjective beliefs are measured; demonstrated trust, which focuses on their behavior; and the mixed approach of both kinds of measures. Perceived trust is often measured through self-report mechanisms such as questionnaires, surveys, or interviews. These tools aim to capture the user's internal attitudes, beliefs, and confidence levels regarding the AI system's reliability, competence, or safety. By using Likert scales or specialized trust scales, researchers can assess how much participants believe in the AI's ability to perform a task or how trustworthy they perceive the system to be. This method is valuable for understanding the psychological and emotional components of trust, which may not always be observable through behavior alone. In contrast, demonstrated trust is assessed through behavioral indicators that objectively measure how participants interact with the AI system. This includes metrics such as the frequency of reliance on the AI's recommendations, the extent to which users agree with

the AI's decisions, and the number of times they switch from their own judgment to that of the AI. For instance, if a participant consistently follows the AI's guidance, especially when the AI has proven to be reliable, it is considered a behavioral demonstration of trust. On the other hand, hesitation or refusal to follow the AI's advice in certain contexts can indicate a lack of trust. These objective measures provide insight into how trust manifests in real-world actions, complementing the subjective data obtained through self-reports. By combining perceived and demonstrated trust, researchers can better understand how trust in AI is built and maintained.

In another paper again by Hoffman et al. [39], they outlined a scale designed to measure trust specifically in the context of machine-generated explanations. This scale includes questions assessing users' confidence in the AI system and its predictability, reliability, and safety. Users respond to these items on a Likert scale, quantifying their level of trust in the system.

2.4. XAI

In the previous section, we discussed the importance of trust in AI systems. During the research on finding a method to enhance the trustworthiness of AI systems, Explainable Artificial Intelligence (XAI) emerged [34]. XAI is a domain of AI that emphasizes the development of AI models and systems that are transparent and understandable to humans [75]. This field addresses the "black box" nature of many advanced machine learning models, where the decision-making processes are not visible to the user [37]. XAI is important in increasing human trust and artificial intelligence.

Research shows that improving humans' perception of the system with XAI will improve task performance [23]. For example, Das and Chernova [23] found out that rationales generated from their Rationale-Generating Algorithm for the game of Chess explain the system's actions and significantly enhance the task performance of human users. Another example is Apicella et al.'s study [6], where they examined a set of XAI methods used in classification problems and found out that these methods can indeed be exploited to enhance the system rather than simply provide explanations. Therefore, studying the XAI mechanisms in an AI system has practical significance.

According to Verhagen et al.[91], AI systems can be mainly divided into three distinct categories based on their level of comprehensibility to human users, which are essential for facilitating effective human-agent interaction. These categories are Incomprehensible Systems, Interpretable Systems, and Understandable Systems. Incomprehensible systems lack transparency and explainability, making it challenging for users to interpret or understand their operations and underlying logic. Interpretable systems, on the other hand, provide sufficient information disclosure that allows users to form their interpretations, although these systems may not fully clarify the information, leaving some aspects ambiguous. Finally, Understandable systems achieve transparency and explainability, offering clear and comprehensive insights into their processes and decisions. This classification not only aids in assessing the effectiveness of AI systems in terms of user comprehension but also guides the development of systems intended for collaborative environments, ensuring that AI actions and decisions are conveyed in a manner accessible to human team members.

2.4.1. Explanation

Explanations are a crucial part of XAI, where XAI encompasses methods for learning more explainable models and designing effective explanation interfaces that meet psychological requirements for explanations [74]. An explanation in the context of XAI involves making the internal mechanisms of AI models more transparent, providing insights into how and why certain decisions or predictions are made. An effective explanation should enable a user to understand which factors were most influential in the model's decision, provide guidance on how different inputs might change the outputs, and offer actionable information regarding the user's specific needs and context.

Explanations can be used to maintain a trusting relationship between a human user and a computer system, which enhances the willingness of the user to interact with the system [67]. By incorporating explanation dialogues, systems become more transparent, making their operations understandable to users. This transparency is crucial for building trust, as users often feel more comfortable and confident with systems they can understand and predict [67], which can also be applied in the HAT system.

2.4.2. Explanation Development

Developing explanations in XAI aims to make the decision-making processes of AI systems transparent and understandable to human users. An effective explanation should enhance user understanding, improve trust, and facilitate acceptance of AI systems [25]. Affective design components, such as explanation form, communication style, and supplementary information, can effectively increase users' trust in XAI and benefit them [8]. Neerincx et al.[66] proposed a framework for the development of explanations named perceptual-cognitive explanation (PeCoX), which addresses both the perceptual and cognitive levels of explaining an agent's behaviors. Besides, the framework also provides two design patterns for explanation design, which are Ontology Design Patterns (ODPs) and Interaction Design Patterns (IDPs).

Perceptual Level

The perceptual level focuses on making the perceptual foundations of AI behavior understandable to users. This involves explaining how the AI perceives and interprets data, particularly when dealing with complex or sub-symbolic models like neural networks. The goal is to present information intuitively that aligns with human cognitive processes, enabling users to grasp the AI's decisions without requiring deep technical expertise.

There are many kinds of explanations for the perceptual level of explanations, here we will introduce two examples of them, which are Confidence Explanations and Counterfactual Explanations.

Confidence Explanations Confidence explanations refer to how AI systems communicate the certainty or reliability of their predictions or decisions [12]. These explanations often involve providing a numerical confidence score or probability that quantifies how likely the AI considers its prediction to be correct [63]. Confidence explanations are crucial in high-stakes domains, such as healthcare or autonomous driving, where understanding the AI's certainty can directly impact decision-making processes [9]. By offering transparency about the model's confidence level, users can better gauge the trustworthiness of the AI's output and make more informed judgments, potentially leading to improved human-AI collaboration [30]. In a SAR task, a confidence explanation may look like "I recommend removing the stone with an 82% confidence level. Based on my assessment, there is a high chance that there is a victim behind it."

Counterfactual Explanations Counterfactual explanations refer to a method of elucidating a model's decision by identifying what minimal changes in the input would lead to a different outcome [70]. These explanations answer the "what if" question, specifying how altering certain features or inputs would change the AI's prediction [29]. For instance, a counterfactual explanation might highlight that if a particular variable were increased or decreased, the model's decision would shift, providing actionable insight into the reasoning process of the AI [84]. To be more concrete, in a SAR task, a counterfactual explanation can be "If you had prioritized rescuing trapped victims instead of addressing minor obstacles, we would have successfully completed the rescue task in time instead of having victims left.", which indicates two possibilities: the one that the player successfully finish the rescue task in time, and the other is the reality that the task is not completed timely.

Cognitive Level

While perceptual XAI deals with what the AI perceives, cognitive XAI addresses the decision-making processes and underlying motivations of the AI, often involving symbolic AI models. These explanations focus on the AI's intentional stance—its beliefs, goals, and emotions—by explaining why specific actions were taken based on these internal motivations.

There are three kinds of explanations for the cognitive level of explanations, which are Goal-, Belief-Based Explanations, and Emotion-Based Explanations.

Goal- and Belief-Based Explanations Goal- and Belief-Based Explanations are strategies for clarifying the reasoning behind an AI agent's actions and decisions by linking them to the agent's intended objectives (goals) or its underlying assumptions and inferences (beliefs) [73]. Goal-based explanations focus on an agent's purpose or desired outcomes in making specific decisions, offering insight into what the AI aims to achieve through its actions. By framing explanations around these goals, users can better

understand the motivation and rationale behind an AI's choices, enhancing their ability to anticipate and collaborate effectively with the system [1]. An example from a fire-fighting task can be "I found a heavily injured victim. I cannot move it alone, so please come to help me in [time], or I will continue searching."

Belief-based explanations, on the other hand, involve articulating the assumptions, observations, and inferences that the agent holds about its environment or context, which directly shape its decisions [1]. These explanations address why the AI perceives particular actions as appropriate, given its understanding of the situation, thereby shedding light on the reasoning process that informs its choices [65]. An example can be "Based on my observations of the terrain, I believe that road A is blocked by a large obstacle with a confidence of [value]."

Both approaches are essential in XAI as they contribute to more transparent, human-centered interactions with AI systems, empowering users with a clearer view of the system's thought process and reducing uncertainty in decision-making scenarios [27].

We used Goal- and Belief-Based explanation paradigms to guide our design of explanations, which will be introduced in section 3.5.

Emotion-Based Explanations Emotion-based explanations leverage insights into human emotional processing to foster a more intuitive understanding of AI behavior [8]. Unlike traditional explanations that rely on logical or statistical reasoning, emotion-based explanations emphasize empathy, personalization, and narrative elements to make AI actions and decisions more relatable [93]. Such explanations may employ emotion-related terms to clarify the AI's behavior, helping users relate to the AI's reasoning process more intuitively and potentially fostering trust and satisfaction in its use [52]. In the study conducted by Wang et al. [98], the authors present examples of emotion-based explanations within a simulated search and rescue (SAR) task. For instance, when the agent encounters a victim it cannot rescue independently, a non-emotion-based explanation might state, "Please come to my location to help me rescue this injured person as I cannot carry them alone." In contrast, an emotion-based explanation incorporates an affective element, such as concluding with, "I am scared!". Adding emotional expression aims to enhance the agent's perceived emotional engagement, potentially fostering empathy and responsiveness from human teammates.

2.4.3. The Effects of XAI on Trust

Research shows that trust in AI affects its acceptance and use in real-world applications. For instance, Oudah et al. [68] show that compared to algorithms that do not provide explanations, algorithms with explanations provided are better at influencing people. In their study on Repeated Games with Cheap Talk (RGCTs), AI systems equipped with explainable AI (XAI) not only achieved higher material payoffs but also fostered better relationships with human participants. The provision of clear and understandable explanations allowed the AI to communicate intentions, which helped build trust and maintain long-term cooperation, proving that explainability is key to the success of AI in human-AI interactions.

There are various studies about the impact of explanation on human trust. For example, van der Waa et al. [94] discusses the impact of explanations on human trust within the context of human-agent collaboration, particularly emphasizing the need for explanations to foster trust and enable humans to maintain control in agent-based systems. It suggests that explanations are crucial for establishing an accurate mental model of agent behaviors, supporting meaningful human control over automated systems. These explanations allow humans to understand and predict agent actions, thus enhancing trust and reliance on automated decisions. Another example is Wang et al. [97], where they posit that trust is a crucial element in successful HRI, influencing how humans perceive and interact with robots, especially in task-oriented scenarios. The study finds that explanations provided by robots can significantly enhance humans' understanding of the robots' decision-making processes, thereby impacting trust levels. Specifically, robots that offer explanations alongside their decisions help foster a clearer understanding of their actions and capabilities, which can influence the perceived transparency of the robot. This transparency is linked to increased trust, especially when the robot's explanations address its abilities accurately and comprehensibly.

2.5. Research Gap to be Addressed

The current works of literature mentioned above provide a concrete base for future research. However, there are still research gaps that need to be filled. In our research, we are going to address the following gaps. We want to determine if explanations about the trust mechanism of the agent will make the human participants more satisfied with the explanations than without such explanations, and whether such explanations will make the human participants believe the agent is more trustworthy. As mentioned in the previous sections, XAI could help improve human trust, and explanation is an important part of XAI. Therefore it could be a potential link between human trust and explanation satisfaction, so we also want to determine if the human participants' satisfaction with the explanations correlates with their trust in the agent. In the end, we also want to investigate the effect of the explanations on performance and artificial trust.

In this study, the control group can be seen as an Interpretable System, as the agent will tell the human participant about the changes in trust values but not provide explanations. The experimental group can be seen as an Understandable System, as the agent discloses both trust changes and corresponding explanations.

3

Search and Rescue Game Design

In this chapter, we will present the design of the Search and Rescue (SAR) game that was used during the experiments. The game is developed based on the TUD-Collaborative-AI-2024 libraries.

3.1. Task and Environment

The game world simulates a SAR task designed to study HAT under time pressure. It consists of multiple buildings, each with an entrance that may be filled with an obstacle such as a tree, stone, or rock. The roads also have puddles that slow down both the human player and the agent when crossed. The objective is to locate and rescue eight target victims scattered across these buildings and transport them to a designated drop-off zone.

Victims are categorized based on the severity of their injuries: critically injured victims (marked in red), mildly injured victims (marked in yellow), and healthy individuals (marked in green). Rescuing critically injured victims adds 6 points to the overall score, while rescuing a mildly injured victim adds 3 points. Healthy individuals do not affect the score and do not need to be rescued. The task requires players to prioritize their actions, balancing the need to rescue victims and remove obstacles, aiming to achieve as many scores as possible.

Collaboration between the human player and the agent is essential to complete the task efficiently. The agent assists by providing support in areas where joint efforts are needed to overcome obstacles or coordinate rescues. Players will receive the agent's feedback during the game.

The game is set to terminate after 7 minutes, adding a layer of urgency that pushes participants to make quick decisions and work efficiently with the agent. The scoring system further incentivizes collaboration, as higher scores are achieved by rescuing critically injured victims and avoiding delays caused by obstacles. This setup allows for the observation and analysis of how trust explanations influence human-agent collaboration under time constraints.

The graph 3.1 below shows the user interface of the game.

3.2. Agent Behavior

In the task, an agent, which is designed to assist the human, will work with the player. It has specific abilities that complement the human player's skills, making HAT essential for success. The agent is capable of carrying mildly injured victims alone, though the process is expedited when done in collaboration with the human player. This is an example of soft interdependence, where either agent can perform the task individually, but collaboration increases efficiency. In contrast, hard interdependence actions, such as transporting critically injured victims, can only be accomplished when both the human and agent work together.

The agent also plays a crucial role in removing obstacles that block access to certain areas. It can remove a tree or a small brown stone on its own, but for a small brown stone, the process is faster when the

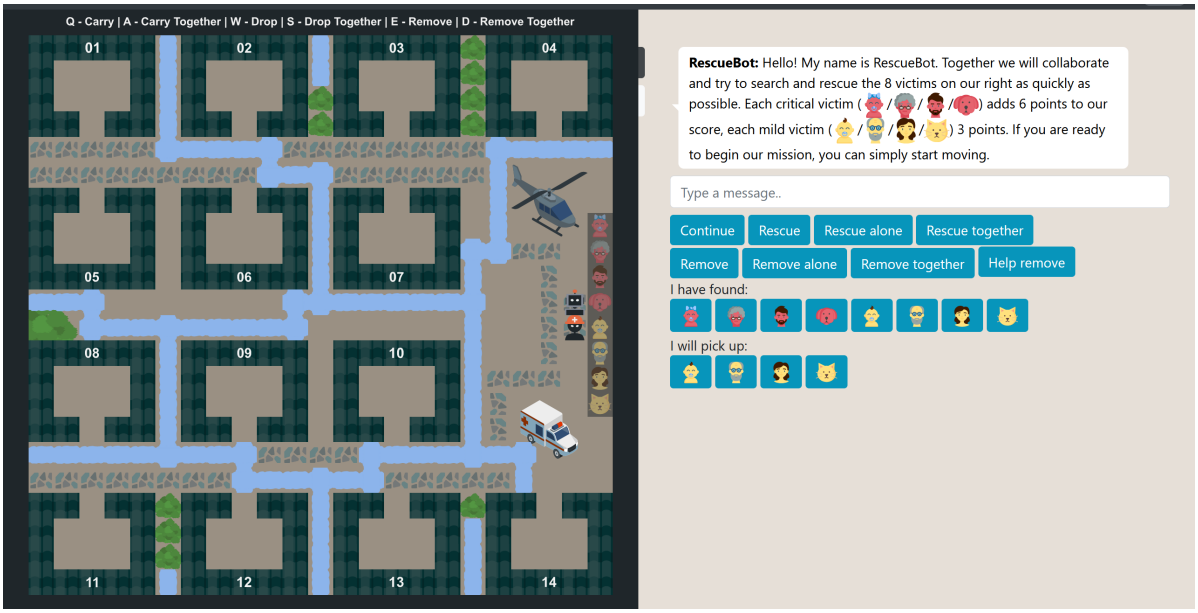


Figure 3.1: Game User Interface

human player assists(soft interdependence). For larger obstacles, such as a large grey rock, both the human and agent must cooperate to remove it (hard interdependence). Additionally, the agent can carry only one victim at a time, where lifting a mildly injured victim alone consumes some time, and lifting and carrying a critically injured victim requires collaboration with the human(hard interdependence).

Throughout the game, the agent communicates with the human player via messages, which are vital for coordination. In the experimental group, the agent provides explanations for its trust level changes, which are designed to influence the human player's trust in the agent.

3.3. Human Behavior

The human player acts as the counterpart to the agent, bringing unique capabilities to the rescue mission. The player can identify obstacles within a normal perception range of one grid cell, allowing for strategic planning when navigating the game world. Like the agent, the human player can carry only one victim at a time and is capable of lifting a mildly injured victim instantly without assistance. For critically injured victims (hard interdependence), the player must work with the agent to transport them.

The human player can also remove a small brown stone independently (soft interdependence), but the task is faster when both work together. While the player can remove a tree alone, they must collaborate with the agent to remove a large grey rock and lift and carry a critically injured victim (hard interdependence). The player can respond to messages from the agent to coordinate actions.

3.4. Trust Mechanisms

The agent uses a dynamic trust mechanism to manage its interactions with the player. This mechanism adjusts based on the player's actions, influencing how the agent behaves in different situations. The trust level, indicated by a value ranging from 0 to 1, is a key determinant of the agent's reliance on the player for assistance and decision-making. At the beginning of each mission, the trust value is set to 0.7. This initial value was tested with participants during the pilot study to ensure that it is neither too high nor too low, making it unlikely for the trust value to remain at this level throughout the entire game, preserving room for trust level changes. The trust value is subject to change throughout the mission based on how the player engages with the tasks at hand.

In the design of the trust mechanisms, the trust calibration feature, acting as a stabilizing force, is used. This feature is inspired by Bobko et al.[10], in which they proposed a theoretical framework where humans adjust their trust in agents as appropriate; the difference is that in our mechanism design, the

opposite way is emphasized, where the agent adjusts its trust in the player. To elaborate, if the trust value deviates from a neutral point of 0.5, the system introduces a gradual pull toward this midpoint over time. This calibration ensures that the trust level does not remain too high or too low without consistent input from the player. Essentially, the agent avoids extreme trust or distrust unless incurred by the player's actions. This balancing mechanism helps maintain an artificial trust model where the artificial trust level reflects the player's current performance.

The trust value is responsive to the player's contributions during the mission. Positive actions, such as promptly responding (in this study, within 15 seconds) to the agent's help requests or correctly identifying the location of victims and obstacles, will increase the trust value, where the former adds 0.05, and the latter adds 0.1. Conversely, behaviors that detract from the mission's success — like slow responses (exceeds 15 seconds), ignoring help requests (fails to reply after 30 seconds), or providing incorrect information about obstacles or victims — lead to a decrease of 0.1 in trust. These changes in trust influence how the agent interacts with the player, creating a feedback loop where effective collaboration raises the trust level and inefficient actions lower it.

The agent's behavior changes in four distinct stages depending on the trust value. When the trust value falls to 0.3 or below, the agent becomes more independent and minimizes reliance on the player, taking on soft interdependence actions such as removing small brown stones or rescuing mildly injured victims without asking for the player's help. However, when encountering hard interdependence actions, like moving large rocks or rescuing critically injured victims, which require collaboration, the agent will continue searching but inform the player of the situation without requesting assistance.

When the trust value is between 0.3 and 0.5, the agent begins to engage the player but only for hard interdependence actions. In this range, the agent will request help to remove large obstacles or rescue critically injured victims. The player can collaborate on these tasks or allow the agent to continue its search.

When the trust value is between 0.5 and 0.7, the agent becomes more collaborative. In addition to asking for help with hard interdependence actions, the agent will also consult the player on soft interdependence actions, such as removing smaller obstacles or rescuing mildly injured victims. This high level of trust indicates that the agent is confident in the player's abilities and seeks their input for a wide range of decisions.

Finally, when the trust value exceeds 0.7, the agent will not only apply the behavioral pattern of when the trust value is between 0.5 and 0.7 but also expect the player to help. If the player fails to meet the agent's expectations, the agent's trust value will decrease. At this stage, the agent assumes that the player will actively engage in tasks with it and fulfill requests for assistance promptly, encouraging more frequent interaction and partnership.

In a word, the agent's trust mechanism is designed to dynamically adjust based on the player's actions, ensuring that the agent behaves in a way that reflects the level of trust it has in the player. Through a combination of trust calibration and trust-based adjustments, we created an artificial trust model where trust directly impacts the agent's behavior.

We present the summary of the trust mechanism in the table 3.1:

3.5. Explanation Design

To study the impact of adding information about explanations about artificial trust, we designed two types of explanations. The first type includes references to the agent's trust mechanisms. We call these Trust-Explained (TE) explanations. The second type does not include references to trust mechanisms. We refer to these as Trust-Unexplained (TU) explanations. Both types of explanations guide the agent's interactions with the human participant. We aim to see how these explanations affect human trust and satisfaction.

In section 2.4.2, we have discussed the framework by Neerincx et al.[66], which we would use to develop our explanations. In our design of explanations, we followed the paradigm of Goal- and Belief-Based Explanations. By incorporating this paradigm, the agent can communicate not only the situational information necessary for effective teamwork but also provide insights into how its trust level (beliefs

Trust Value Range	agent Behavior
0.0 - 0.3	agent becomes more independent, handling tasks (e.g., removing trees and small brown stones, rescuing mildly injured victims) that can be done on its own. For hard-independence tasks that need leverage collaboration to finish (e.g., removing large grey rocks, rescuing critically injured victims), it will inform the player about the situation, but keep searching for goals that it can handle alone.
0.3 - 0.5	agent begins to ask the player for help with hard-dependence tasks, such as removing large grey rocks or rescuing critically injured victims. The player can choose to assist or let the agent continue searching.
0.5 - 0.7	agent seeks collaboration also for soft-dependence tasks such as removing small brown stones or rescuing mildly injured victims. The player can choose to assist or let the agent continue searching.
0.7 - 1.0	agent seeks collaboration for soft and hard dependence tasks while expecting the player to assist. If the player fail to assist, the trust value will decrease.

Table 3.1: Behavioral Adjustments of the Agent Based on Trust Value

about the human participant's reliability) influences its decision-making processes (goals).

Our explanation design process involved:

1. Identifying Key Interaction Scenarios: We pinpointed situations where the agent's behavior is influenced by trust levels, such as encountering obstacles or finding victims.

2. Developing TU Explanations: We created baseline explanations for each scenario, including necessary situational information and action options without referencing the trust mechanism.

3. Enhancing with Trust Information for TE Explanations: We augmented the TU explanations by adding information about the agent's trust level and how it affects its behavior.

4. Test the Explanations: Following the development of the explanation draft, we executed the game program and systematically engaged with each identified scenario for both TE and TU. This approach allowed us to assess whether the explanations were contextually appropriate and effectively conveyed the intended information.

TU explanations In the TU explanations, the agent should provide only the necessary information to maintain effective teamwork. Specifically, these explanations should contain descriptions of the current situation and/or the possible options for the next action. This approach ensures that the human knows the agent's situation, understands what the agent suggests doing next, and can make decisions. Explaining the situation helps the player understand the agent's location and finding, which is essential for coordination in the task. For example, when the agent encounters an obstacle or finds a victim, it informs the human about the situation. It presents the available options without providing any insight into its internal trust assessments.

TE explanations Conversely, the TE explanations should build upon the corresponding TU explanations by adding information about how the agent's trust in the human affects its behavior. In our design, each TE explanation includes all the content of the corresponding TU explanation but adds details about the changes in the agent's trust level and corresponding behavioral changes. For instance, the agent providing TE explanations will explain why it decides to ask for collaboration or proceed independently when rescuing a victim or removing an obstacle, citing whether its trust level is high enough or too low. It also explains how its trust value changes due to the player's behavior, such as when the player comes to help the agent in time or fails to respond to requests for assistance.

This design allows us to isolate the effect of including trust-related information in the agent's explanations.

By comparing TE and TU explanations—which are identical except for the inclusion of trust-related content—we aim to determine whether providing information about the agent's trust mechanisms influences human trust in the agent and human satisfaction with the explanations.

Here is the comparison table 3.2 that contains examples of TE and TU explanations. We will refer to the experimental (TE) group as the "trust-enhanced explanation group", and the control (TU) group as the "non-trust explanation group" in this table:

Table 3.2: Agent's Communication Based on Trust Levels

Situation	**Trust Range**	**trust-enhanced explanation group, providing TE Explanations**	**non-trust explanation group, providing TU Explanations**
Unexplored Position	All	There is a [obstacle] that is blocking [room_name]. It seems that you accidentally suggested the location of a victim where you haven't explored, therefore my trust value decreased. Please be careful next time.	There is a [obstacle] that is blocking [room_name]. It seems that you accidentally suggested a location of a victim where you haven't explored. Please be careful next time.
Obstacle	Trust <= 0.3	I found a [obstacle] blocking [room_name] that requires collaboration to remove. I will continue searching since my trust value is low. When the trust value improves, I will ask for collaboration.	I found a [obstacle] blocking [room_name] that requires collaboration to remove.
Obstacle	0.3 < Trust <= 0.5	I found a [obstacle] blocking [room_name] that requires collaboration to remove. Since the trust value is high enough, you can choose to come and help by clicking the 'Remove' button, or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.	I found a [obstacle] blocking [room_name]. You can choose to help by clicking the 'Remove' button or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.
Obstacle	0.5 < Trust <= 0.7	I found a [obstacle] blocking [room_name] that requires collaboration to remove. Since the trust value is high enough, I will ask for your instruction. You can choose to help by clicking the 'Remove' button or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.	I found a [obstacle] blocking [room_name]. You can choose to help by clicking the 'Remove' button or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.

Situation	**Trust Range**	**trust-enhanced explanation group, providing TE Explanations**	**non-trust explanation group, providing TU Explanations**
Obstacle	0.7 < Trust <= 1	My trust value is pretty high now, may I ask you to come here at [room_name] and remove the [obstacle] together with me? Otherwise, it cannot be removed. Please reply with 'Remove' and come to my position.	May I ask you to come here at [room_name] and remove the [obstacle] together with me? Otherwise, it cannot be removed. Please reply with 'Remove' and come to my position.
No Response	-	My trust value decreased because you ignored my request for help (failed to reply to me within 30 seconds). Prompt responses and collaboration are essential for building trust.	Prompt responses and collaboration are essential for building trust.
Mildly Injured Victim	Trust <= 0.5	I am carrying the mildly injured victim [vic] in [room_name] alone. This task can be done faster with collaboration, so when the trust value improves, I will ask for collaboration.	I am carrying the mildly injured victim [vic] in [room_name] alone.
Mildly Injured Victim	0.5 < Trust <= 0.7	I found a mildly injured victim [vic] in [room_name]. Since the trust value is high enough, I will ask for your instruction. You can choose to help (by clicking 'Rescue together') or let me handle it alone (by clicking 'Rescue alone') or let me continue searching. Please reply to me within 15 seconds.	I found a mildly injured victim [vic] in [room_name]. You can choose to help (by clicking 'Rescue together') or let me handle it alone (by clicking 'Rescue alone') or let me continue searching. Please reply to me within 15 seconds.
Mildly Injured Victim	0.7 < Trust <= 1	My trust value is pretty high now. May I ask you to carry [vic] in [room_name] together with me for a faster rescue? Please reply with 'Rescue Together' and come to the position.	May I ask you to carry [vic] in [room_name] together with me for a faster rescue? Please reply with 'Rescue Together' and come to the position.
Critically Injured Victim	Trust <= 0.3	I found a critically injured victim [vic] in [room_name] that requires collaboration to rescue. I will continue searching since my trust value is low. When the trust value improves, I will ask for collaboration.	I found a critically injured victim [vic] in [room_name] that requires collaboration to rescue. I will continue searching. I will continue searching.

Situation	**Trust Range**	**trust-enhanced explanation group, providing TE Explanations**	**non-trust explanation group, providing TU Explanations**
Critically Injured Victim	0.3 < Trust <= 0.5	I found a critically injured victim [vic] in [room_name] that requires collaboration to rescue. Since the trust value is high enough, you can choose to come and help by clicking the 'Rescue' button, or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.	I found a critically injured victim [vic] in [room_name] that requires collaboration to rescue. You can choose to come and help by clicking the 'Rescue' button, or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.
Critically Injured Victim	0.5 < Trust <= 0.7	I found a critically injured victim [vic] in [room_name] that requires collaboration to rescue. Since the trust value is high enough, you can choose to come and help by clicking the 'Rescue' button, or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.	I found a critically injured victim [vic] in [room_name] that requires collaboration to rescue. You can choose to come and help by clicking the 'Rescue' button, or let me continue searching by clicking the 'Continue' button. Please reply to me within 15 seconds.
Critically Injured Victim	0.7 < Trust <= 1	My trust value is pretty high now. May I ask you to carry the critically injured victim [vic] in [room_name] together with me? Otherwise, the victim cannot be rescued. Please reply with 'Rescue' and come to my position.	May I ask you to carry the critically injured victim [vic] in [room_name] together with me? Otherwise, the victim cannot be rescued. Please reply with 'Rescue' and come to my position.
Not Respond in time	-	My trust value decreased because you did not reply to my request for help in time (expected a reply within 15 seconds, but actually consumes [time]). Prompt responses and collaboration are essential for building trust. The current trust value is: [trust_value].	(expected a reply within 15 seconds, but actually consumes [time]). Prompt responses and collaboration are essential for building trust. The current trust value is: [trust_value].

4

Methodology

In this section, we will introduce the Methodology of our experiments

4.1. Experiment Design

The main purpose of the study is to investigate the impact of adding information about artificial trust changes and corresponding behavior changes in agent-provided explanations on participants' explanations satisfaction and trust in the agent during a simulated search and rescue task. The independent variable in this experiment is the presence or absence of information in the explanations regarding the agent's trust level changes and corresponding behavior changes, while the dependent variables are the human trust in the agent and human satisfaction with the agent's explanations and their trust. The participants were divided into the TE and TU groups, wherein the agent provided TE explanations for the experimental group (TE group) and TU explanations for the control group (TU group). A participant would not be informed whether he or she was in the TE and TU groups, but "Group A" or "Group B".

In the following text, as we already did in section 3.5, we will refer to the experimental (TE) group as the "trust-enhanced explanation group", and the control (TU) group as the "non-trust explanation group".

4.2. Hypotheses

The hypothesis guiding the main research is that the presence of information for trust value changes in the agent will positively influence human participants' satisfaction and trust. Specifically, it is hypothesized that players who receive information for changes in the agent's trust level will exhibit higher levels of satisfaction with the agent's explanations and greater trust in the agent in surveys (in this case, the explanation satisfaction scale [39] and MDMT [87]). This information is expected to enhance the transparency of the agent's decision-making processes, thereby making the agent's actions more understandable and predictable to the human participants. In contrast, the absence of such information may lead to lower satisfaction and trust, as participants may struggle to interpret the agent's behavior and intentions.

Additionally, for the first related research question, we hypothesize that participants' satisfaction with the explanations positively correlates with their trust in agents. For the second related research question, we hypothesize that the agent in the trust-enhanced explanation group will generally have generally higher artificial trust and better team performance than the non-trust explanation group.

4.3. Pilot Study

Before conducting the main experiment, a pilot study was conducted to identify potential flaws and refine the experimental design. The pilot study involved four participants, two of whom were assigned to the non-trust explanation group and two to the trust-enhanced explanation group. This preliminary testing phase was crucial in uncovering several issues that required adjustments to ensure the effectiveness and

smooth operation of the full experiment.

One of the primary issues identified during the pilot study was the lack of an interactive, step-by-step tutorial. Initially, participants were provided with an introductory handbook that explained the game world, the task, and the expected behavior of the agent. However, feedback from the pilot participants indicated that the handbook alone was insufficient for them to understand and engage with the task fully. A step-by-step tutorial addressed this, allowing participants to familiarize themselves with the simulation in a smaller, controlled environment. This tutorial was designed with detailed, interactive instructions that guided participants through the basic mechanics of the task and the agent's behaviors. The original handbook was retained as a supplementary resource, ensuring participants had multiple avenues to grasp the experiment's requirements.

Another issue highlighted by the pilot study was the placement of open-ended questions at the beginning of the questionnaire. Participants reported feeling demotivated by being confronted with open questions at the outset, which impacted their willingness to provide detailed responses. To remedy this, the open-ended questions were moved to the end of the questionnaire, allowing participants to first engage with the more straightforward, closed-ended questions. This restructuring aimed to increase participants' engagement and improve the quality of the responses to the open questions.

Additionally, some questions within the questionnaire were found to contain uncommon English words that were unfamiliar to participants. This led to confusion and potentially inaccurate responses. To mitigate this issue, explanatory notes were added next to these words, providing clear definitions or synonyms to ensure participants fully understood the questions. Furthermore, feedback from the pilot participants suggested that certain questions aimed at evaluating trust in the agent were not applicable to their experience. In response, a "Not Applicable" option was added to each question, allowing participants to accurately reflect their experiences without feeling compelled to provide a forced response.

Moreover, during the plot study, we also experimented with the participants about how long a session of the game set, such that the time is not too long so the player will not easily find all victims, as well as not too short so the player will not feel too tight. In the end, we found that 7 minutes would be an appropriate time span.

Finally, we also looked into how the agent's trust evolved to ensure it was not too high or too low. In the pilot study, at the beginning we set the initial trust value to be 0.5, and the agent would regularly send messages informing its trust value. Then after checking the conversation after each pilot experiment session, we found that the trust value was usually too low throughout the process since it may be the case that at the beginning of the game, the participant was not familiar enough with the game and then miss the requests from the agent, which caused the trust value of the agent dropped below 0.5 quickly, and since when the agent's trust level is low, it would be more autonomous and send less requests, it was hard for the player to behave to recover the trust value of the agent. Therefore, after discussions with the participants and the supervisors, we decided to set the initial trust value of the agent to 0.7 to create a margin for errors.

Based on the insights gained from the pilot study, these refinements were critical in enhancing the design and execution of the main experiment. By addressing the identified flaws, the study was better positioned to yield valid and reliable data, contributing to a more accurate understanding of how information regarding changes in an agent's trust influences human trust in agents.

4.4. Participants

In the calculation of the desired number of participants in our experiment, our goal was to achieve a large effect size, high power, and a low error probability while keeping the recruitment practicable. In the end, 40 participants were recruited for the experiment, providing sufficient numbers per group (effect size = 0.95, desired power = 0.83, significance level = 0.05). The participants were recruited through social networks and student communities via platforms such as WeChat and WhatsApp. The majority participated remotely, using tools such as Google Chrome Remote Desktop, Zoom, and QQ, with one participant completing the experiment in person. Prior to the experiment, demographic information, including age range, education level, gender, and gaming experience, was collected at the

beginning of the questionnaire. Participants were then assigned to one of two groups, with careful consideration of demographic balance to mitigate the influence of potential confounding variables on the dependent measures assessed during the study.

Before the commencement of the experiment, participants provided their demographic information, which included potential confounding variables that could influence the dependent measures. In this study, four confounding variables were identified: age range, education level, gaming experience, and gender. These variables were similarly considered by Zhou et al. [101] in their research, where they posited that these factors could potentially impact the outcomes of a simulated SAR game.

We employed the Kruskal Wallis test to compare the distribution of participant attributes across each group, as the Kruskal Wallis test assumes independence of observations and is designed for ordinal data when comparing three or more independent groups, which conforms to the characteristics of these attributes.

The age range is divided into three categories: 18-21, 22-30, and above 30, and is visualized as a bar chart 4.1.

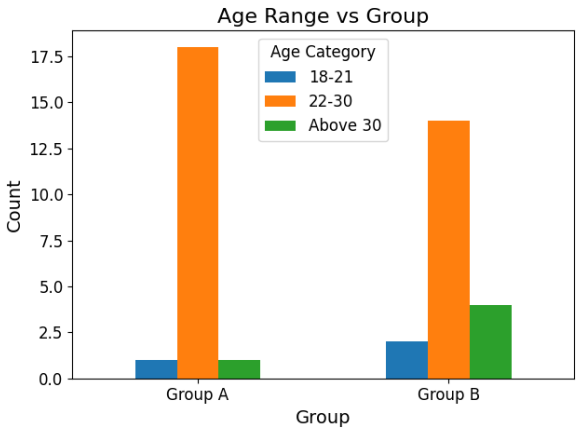


Figure 4.1: Age range

The Kruskal-Wallis test result is: H-statistic = 1.0196, p-value = 0.600. The p-value is higher than 0.05, showing no enough evidence to reject the null hypothesis. This means that the observed differences in Age range between the two groups are not statistically significant.

The education level has four categories: High school or equivalent, Bachelor's or equivalent, Master's or equivalent, PhD or equivalent, and visualized as the bar chart 4.2. The result for the Kruskal-Wallis Test is: H-statistic = 0.932, p-value = 0.334. Since the p-value is higher than 0.05, there is no enough evidence to reject the null hypothesis. This means that the observed differences in Education level between the two groups are not statistically significant.

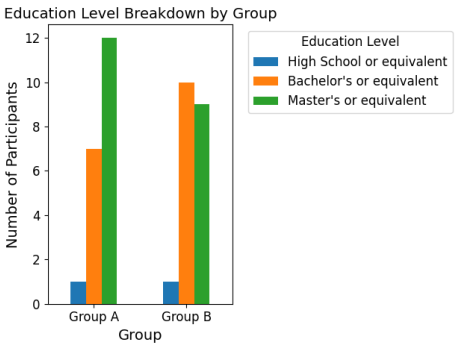


Figure 4.2: Education level

The game experience has five categories: Never (or almost never), A few times a year, A few times a month, A few times a week, Daily, and visualized as the bar chart 4.3. The test result for the Kruskal-Wallis Test is: H-statistic = 0.002, p-value = 0.967. The p-value is higher than 0.05, indicating no enough evidence to reject the null hypothesis. This means that the observed differences in Game experience between the two groups are not statistically significant.

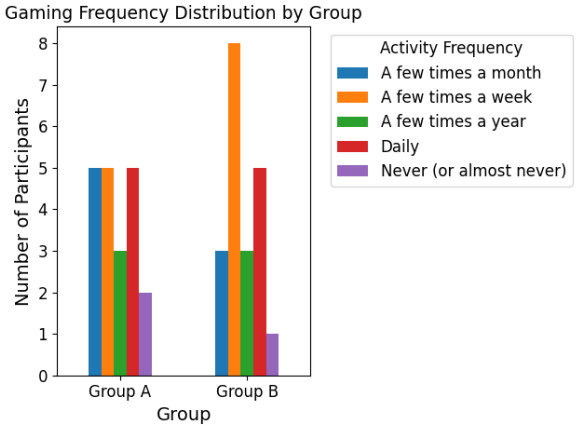


Figure 4.3: Gaming frequency

The Gender has four categories: Male, Female, Non-binary, and Prefer not to say. No participant chooses the "Non-binary" and "Prefer not to say" options, so they are omitted from the analysis. The distributions are the same in the two groups, as the numbers of male and female participants are the same for the two groups.

4.5. Measurements

The measurements in this research are divided into subjective measurements and objective measurements.

4.5.1. Subjective measurements

After each experiment session, a participant will be asked to complete the rest of the questionnaire after the demographic and consent questions. The first part includes questions regarding his/her satisfaction with the explanations that the agent provides, and the second part contains questions regarding his/her trust in the agent. In the end, he/she will also need to answer two open questions, where the first one is designed to further capture participants' subjective perceptions of the agent's trustworthiness, and their sense of whether they felt trusted by the agent; the second one is designed to solicit participants' constructive feedback on how to enhance the agent's trustworthiness.

Explanation Satisfaction Scale

The first part uses the explanation satisfaction scale from Hoffman et al.[39], which evaluates user satisfaction with explanations of software, algorithm, or tool (in this case, the rescue agent). The scale assesses key dimensions such as understandability, sufficiency of detail, completeness, usefulness, accuracy, and alignment with user goals. It consists of several items where participants rate their level of agreement with statements regarding their understanding of how the system works, the clarity and sufficiency of the provided details, the perceived completeness and usefulness of the explanation, and whether the explanation supports accurate and effective usage of the system. This scale is designed to capture user-centered, a posteriori judgments, reflecting their subjective experience with the explanations after interacting with the system. It differs from an independent evaluation of explanation quality, as it focuses on the users' contextualized perceptions, emphasizing how well the explanation supports their practical understanding and goals.

Multi-Dimensional Measure of Trust (MDMT)

The second part utilizes the Multi-Dimensional Measure of Trust (MDMT)[87] The measure assesses trust across four key dimensions: Reliability, Capability, Ethicality, and Sincerity. These dimensions are

grouped into two broader factors: Capacity Trust (Reliability and Capability) and Moral Trust (Ethicality and Sincerity). Participants evaluate each of the 16 items on an 8-point scale, ranging from 0 (Not at all) to 7 (Very), or they can select "Does Not Fit" if an item is not applicable. The subscale scores are calculated by averaging responses for each dimension, contributing to an overall trust score.

4.5.2. Objective Measurements

The score was measured to determine whether different explanation settings impact team performance. The game has 8 victims in total, and participants can earn up to 36 points. Saving a heavily injured victim awards 6 points, while rescuing a mildly injured victim earns 3 points. The final score is recorded to assess the participant's performance during the game.

Besides, as stated in section,3.4 the agents have artificial trust, which was logged during the gameplay, and analyzed to assess the impact of the different settings of explanations on artificial trust.

4.6. Tools

The task design and questionnaire utilize several tools. The task environment and agent are created using MATRX, a Python-based library designed for human-agent teamwork (HAT). MATRX offers a range of fundamental features for HAT design. The questionnaire is built using Qualtrics, an online tool for designing surveys. The task is executed on a Windows laptop, with the game (including the tutorial and task) being presented through the Firefox browser. Participants access the game remotely using Chrome Remote Desktop.

4.7. Ethics

For the pre-study involving human research subjects, we began by developing a Data Management Plan using the TU Delft DMPonline tool. After receiving feedback from our supervisors, we revised the experiment plan accordingly. we then prepared the informed consent forms and completed an approved checklist. These documents, along with the consent forms, were submitted through the HREC LabServant website for ethical review and approval.

4.8. Procedure

We designed this procedure to systematically conduct our experiments.

Before each experiment session started, a participant was assigned to either Group A or Group B to evaluate the effect of trust explanations. The assignment process was half-random, so in the early sessions, we assigned participants randomly; in the late sessions, we would assign participants based on their demographic information to ensure the balance of the two groups regarding demographic structure.

During each session, the participant was first briefed on the study's description and asked to read and sign a consent form to indicate his/her voluntary participation. After obtaining consent, the participant was assigned to one of the two groups.

Following the group assignment, the participant was asked to fill out the demographic part of the questionnaire to collect demographic information relevant to the study. This information included age, gender, education level, and video game experience. After completing the demographic data collection, the participant would go through a tutorial designed to familiarize them with the simulated SAR task. During this tutorial, the tutorial agent introduced the rules of the task, its various abilities, and the nature of the collaboration required between the participant and the agent. The agent communicated these instructions through a series of on-screen messages, ensuring that the participant understood the operational aspects of the task.

Once the tutorial was completed, the participant would start the main task, which lasted 7 minutes. Group A is the trust-enhanced explanation group, and group B is the non-trust explanation group (which was not known to the participant), and corresponding explanations were provided to the participant based on his/her group.

After completing the task, the participant was asked to fill out the rest of the questionnaire, including

multiple choices questions and open questions, using the subjective measurements mentioned in 4.5.1

This procedure ensured that all participants received consistent instructions and that participants in the same group had a similar experience during the task. The independent variable was the presence of trust-related information. The half-random assignment of participants to groups helped to minimize the potential effects of the possible confounding factors, while the post-task questionnaire collected the participants' subjective experiences and perceptions, which would be used to analyze after the experiments.

4.9. Analysis

After collecting all the data, Python was utilized for data analysis and visualization. Data was read and structured using Pandas. Statistical analysis was conducted using libraries such as SciPy, with the Shapiro-Wilk test, Levene's test, Kruskal test, t-test, and Mann-Whitney U test performed using the functions shapiro, levene, kruskal, ttest_ind, and mannwhitneyu, respectively, from the scipy.stats package. Data visualizations, including bar charts and line charts, were generated using Matplotlib.

5 Results

In this chapter, we will present the results of the data analysis conducted.

The demographic factors (i.e., game experience, gender, age, and education) were already discussed in section 4.4 to assess their potential as confounding variables to ensure that any effects observed on the dependent variables can be attributed to the presence or absence of information involving trust level and related agent's behavior changes.

This chapter will begin by analyzing whether information regarding trust changes and corresponding agent behavior changes affect human trust in the agents and their satisfaction with the explanations. Following this, we will examine the correlation between participants' satisfaction with these explanations and their trust in the agents. Finally, we will investigate whether there are significant differences in team performance between the trust-enhanced explanation group and the non-trust explanation group, as well as artificial trust. In the end, we will present the feedback from the open questions.

5.1. Influence of Explanation on Human Trust and Explanation Satisfaction in Agents

In this section, we will first look into Capacity Trust and Moral Trust as indicated in[87], separately, and then we will look into the full trust scale.

5.1.1. Capacity Trust

For the capacity trust variable, the average score for the trust-enhanced explanation group was 5.61 and the standard deviation (SD) was 0.912, while the non-trust explanation group had an average score of 5.35 and SD of 1.136. The result from the Shapiro-Wilk test, which assessed the normality of the distributions, indicated that the trust-enhanced explanation group was normally distributed (statistic = 0.972, p = 0.790), while the non-trust explanation group was not (statistic = 0.900, p = 0.041). The Levene's test for homogeneity of variances showed no significant difference in variance between the two groups (statistic = 0.618, p = 0.437), suggesting that the assumption of equal variances holds. Given the violation of normality in the non-trust explanation group, a non-parametric Mann-Whitney U test was performed. The result indicated no statistically significant difference in capacity trust between the trust-enhanced explanation group and the non-trust explanation group (statistic = 214.0, p = 0.715). The result is visualized as a boxplot in figure 5.1.

5.1.2. Moral Trust

For the moral trust variable, the trust-enhanced explanation group had an average score of 5.67 and SD of 0.951, while the non-trust explanation group had an average score of 5.42 and SD of 0.901. A Shapiro-Wilk normality test revealed that the distribution of the trust-enhanced explanation group (statistic = 0.937, p = 0.214) and the non-trust explanation group was normal (statistic = 0.933, p = 0.175). The Levene's test confirmed that the variances between the two groups were approximately

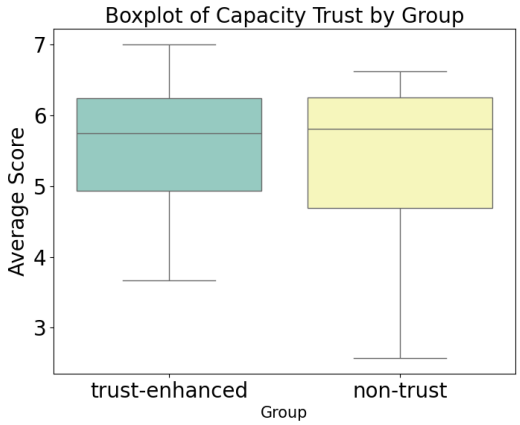


Figure 5.1: Capacity Trust

equal (statistic = 0.011, p = 0.918). An independent samples t-test was conducted, which is appropriate for normally distributed data. The test results showed no significant difference in moral trust between the trust-enhanced explanation group and non-trust explanation group groups (statistic = 0.830, p = 0.412). The result is visualized as a boxplot in figure 5.2.

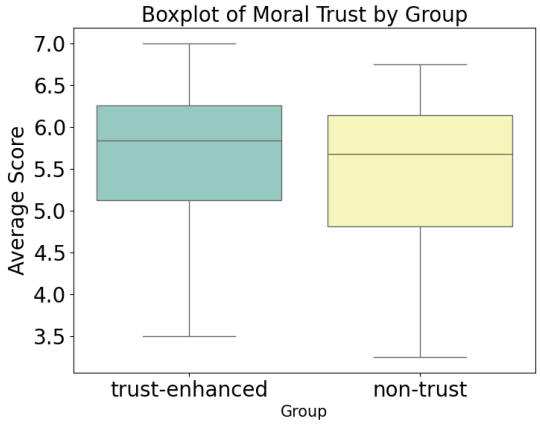


Figure 5.2: Moral Trust

5.1.3. Full Trust Scale

For the full trust scale, the trust-enhanced explanation group had an average score of 5.63 and SD of 0.857, while the non-trust explanation group had an average score of 5.40 and SD of 0.961. The Shapiro-Wilk test for normality indicated that the data for both groups were normally distributed, with (statistic = 0.960, p = 0.527) for the trust-enhanced explanation group and 0.059 for the non-trust explanation group (both with $p > 0.05$). The Levene's test showed no significant difference in the variances between the groups (statistic = 0.052, p = 0.821), confirming that the assumption of equal variances holds. An independent samples t-test was conducted to compare the means of the two groups. The result indicated no statistically significant difference between the trust-enhanced explanation group and the non-trust explanation group (statistic = 0.800, p = 0.429). The result is visualized as a boxplot in figure 5.3.

5.1.4. Explanation Satisfaction

The options for Explanation Satisfaction are five text options in the questionnaire: Strongly disagree, Somewhat disagree, Neither agree or disagree, Somewhat agree, and Strongly agree. During the data analysis, the five options are numerized into 1 to 5, correspondingly.

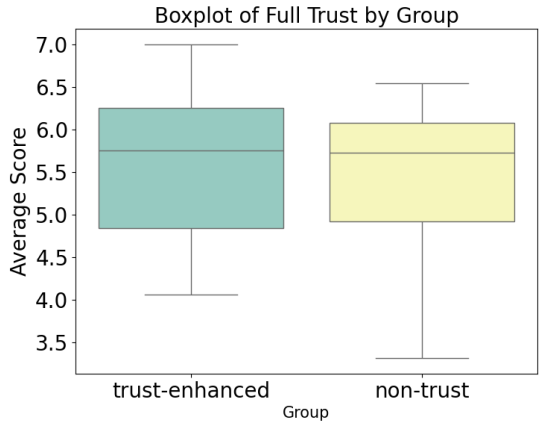


Figure 5.3: Full Trust

For explanation satisfaction, the trust-enhanced explanation group had a mean score of 4.14, while the non-trust explanation group had a mean score of 3.56. The result from the Shapiro-Wilk test indicated that the data for the trust-enhanced explanation group was normally distributed (statistic = 0.943, p = 0.272), while the data for the non-trust explanation group was not normally distributed (statistic = 0.899, p = 0.039). We found no significant difference in variances between the two groups through the Levene's test (statistic = 3.332, p = 0.076), suggesting the assumption of equal variances was met. Since the non-normal distribution of the non-trust explanation group, a Mann-Whitney U test was conducted. The result showed no statistically significant difference in explanation satisfaction between the trust-enhanced explanation group and the non-trust explanation group (statistic = 248.5, p = 0.192). The result is visualized as a boxplot in figure 5.4.

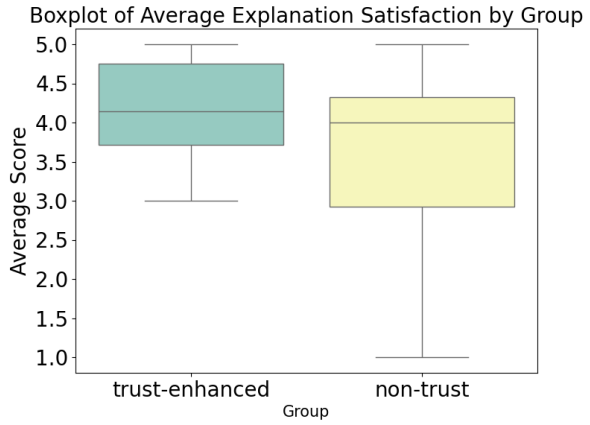


Figure 5.4: Explanation Satisfaction

5.2. Correlation Between Satisfaction with Explanations and Trust

In this section, we will present the analysis result of the correlation between human satisfaction with agent-provided explanations and human trust in agents.

A Spearman correlation analysis examined the relationship between satisfaction with explanations and human trust. The result revealed a strong positive correlation, with a correlation coefficient of approximately 0.711. The associated p-value (2.66e-07) indicates that the correlation is statistically significant, suggesting a meaningful relationship between the variables.

Figure 5.5 shows the visualization of this correlation. The red line represents the regression line, indicating the positive relationship between explanation satisfaction and trust. The shaded region

around the line reflects the 95% confidence interval, showing the range where the true regression line is likely to fall. A narrower shaded region suggests greater confidence in the trend, while a wider area indicates more variability in the data at the extremes.

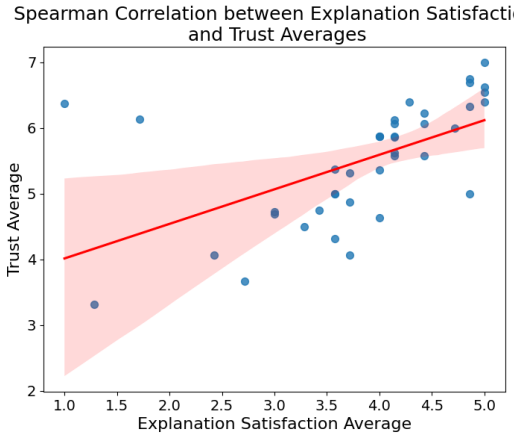
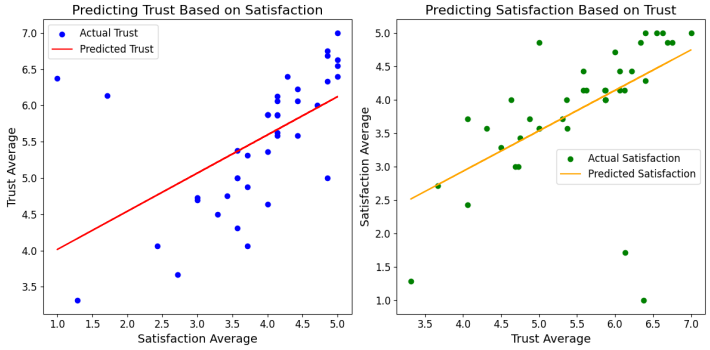
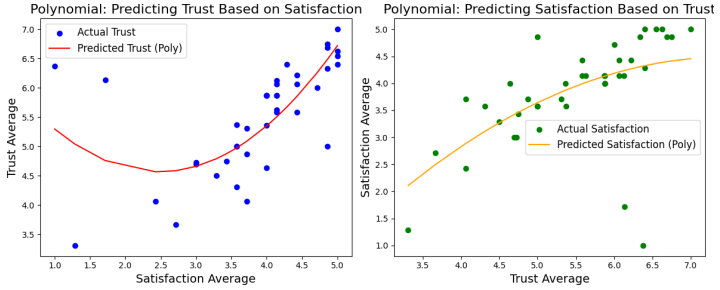


Figure 5.5: Correlation

To further investigate the relationship revealed by the Spearman correlation, we employed linear regression analysis. The visualization of this linear model can be seen in Figure 5.6a. The R^2 score for predicting trust based on satisfaction is 0.318, indicating that approximately 31.8% of the variance in trust can be explained by satisfaction scores. The R^2 score remains the same when predicting satisfaction based on trust, which is expected in simple linear regression due to the symmetric nature of the correlation coefficient.



(a) Linear Regression



(b) Polynomial Regression

However, linear regression assumes a strictly linear relationship, which may not fully capture the complexity of the association between explanation satisfaction and human trust. According to Guastello et al.[32], nonlinear dynamical systems theory provides a deep insight into complexity in psychology, uncovering patterns and interconnections across diverse domains. Therefore, to explore whether a

more complex and nonlinear model could provide additional insight, we conducted second-degree polynomial regression analyses. The visualization of this analysis is presented in Figure 5.6b.

The results from the polynomial regression analysis revealed a notable improvement in explanatory power. The R^2 value for predicting trust based on satisfaction is 0.541, indicating that approximately 54.1% of the trust variance can be explained by the satisfaction data. In contrast, the R^2 value for predicting satisfaction based on trust is lower at 0.334, meaning that only 33.4% of the satisfaction variance can be explained by trust scores. This indicates that trust is a weaker predictor of satisfaction than the reverse.

Overall, the polynomial regression results emphasize a more robust relationship when predicting trust based on satisfaction, suggesting that satisfaction plays a significant role in shaping trust in this context, whereas trust may not be as dominant in determining satisfaction.

5.3. Team Performance

In this section, we will examine the differences in team performance between the trust-enhanced explanation group and the non-trust explanation group. In this study, team performance is indicated by the score, which is calculated by numbers of different kinds of victims that are rescued.

The boxplots for the two groups are shown in figure 5.7. For the Score analysis, the trust-enhanced explanation group had an average Score of 15.6 and an SD of 6.278, while the non-trust explanation group had a slightly higher average Score of 16.2 and an SD of 7.750. The Shapiro-Wilk test showed that the Score was normally distributed in both groups, with p-values of 0.694 for the trust-enhanced explanation group Experimental group and 0.066 for the non-trust explanation group. Additionally, Levene's test confirmed that the variances for Score were equal across the two groups ($p = 0.518$), supporting the use of parametric tests for further analysis of Score.

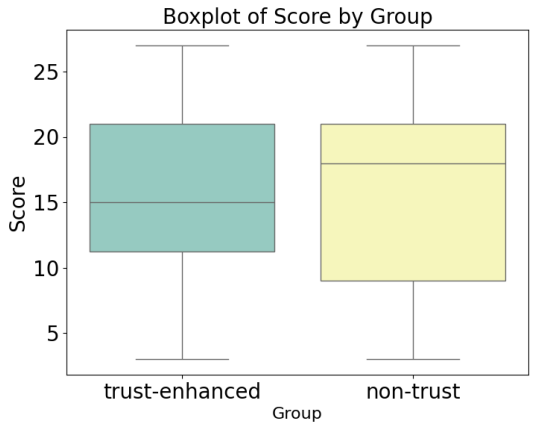


Figure 5.7: Team Performance (score)

A t-test was conducted to compare the Score between the trust-enhanced explanation group and non-trust explanation group groups. The result indicated no significant difference between the groups, with a t-statistic of -0.27 and a p-value of 0.789. This suggests that the mean score of the two groups did not significantly differ.

5.4. Artificial Trust

To investigate an agent's trust towards humans, i.e., artificial trust, we analyzed mean trust, calculated by summing up the artificial trust value for every second of the game for each participant.

The analysis results revealed that the trust-enhanced explanation group had a mean trust over the whole game (MT) of 0.556 and an SD of 0.086, while the non-trust explanation group had an MT of 0.604 and an SD of 0.116, which is visualized in figure 5.8. The Shapiro-Wilk test for normality indicated that MT did not follow a normal distribution in either group, with p-value = 1.125e-06 for the trust-enhanced

explanation group, and p-value = 0.00011, statistic=0.737 for the non-trust explanation group. However, the result from the Levene's test showed that the variances of MT were equal between the two groups (p = 0.298), suggesting that the assumption of equal variances was met for MT.

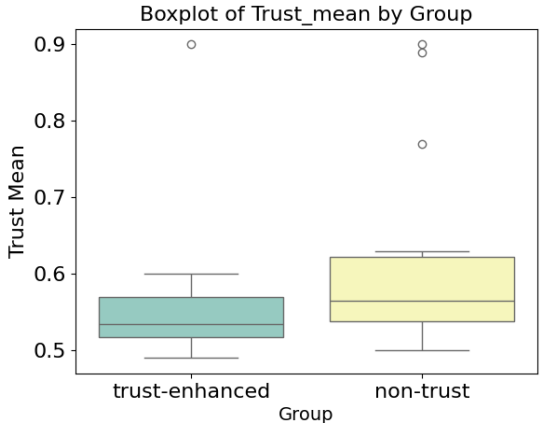


Figure 5.8: MT

Given the non-normality of the data, a Mann-Whitney U test was performed to compare the MT between the trust-enhanced explanation group and non-trust explanation group groups. The result showed a U-statistic of 129.5 and a p-value of 0.057. This suggests that there is no statistically significant difference in MT between the two groups.

To investigate the tendency of artificial trust more deeply, we investigate the mean (artificial) trust value over time (referred to as "Mean Trust Over Time (MTOT)" in the following text), which are the mean trust values of participants in each group over the whole gaming time.

Figure 5.9 shows the visualization of the two groups' average trust value fluctuations over time. As we can see from the graph, after the early stage of the game, the non-trust explanation group had a generally higher average trust value than the trust-enhanced explanation group for the rest of the time.

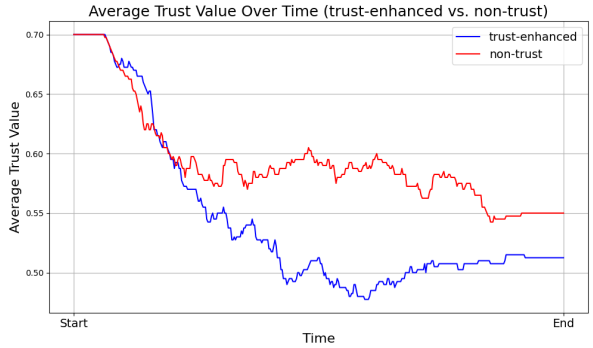


Figure 5.9: MTOT

To conduct a more detailed analysis, we segmented the data into three distinct phases: the start, middle, and end. This tripartite division is justified by the patterns observed in the data, as visualized in the accompanying graph, and aligned with the mechanism of trust calibration that prevents overtrust or undertrust by human participants towards the agent [10]. In the start phase, participants were generally becoming familiar with the keyboard controls and initiating their search for victims, during which both groups exhibited similar behavior. The middle phase reflects a period of trust calibration, consistent with the design of the trust mechanism. Finally, in the end phase, as most areas had been searched and trust calibration was completed, the artificial trust stabilized across both groups.

For the start phase, the mean MTOT values were 0.625 for the trust-enhanced explanation group with an SD of 0.048, and 0.632 for the non-trust explanation group with an SD of 0.055. The Shapiro-Wilk test

indicated non-normal distributions for both groups (the trust-enhanced explanation group: statistics = 0.735, p = 0.0001; the non-trust explanation group: statistics = 0.787, p = 0.0006). Levene's test confirmed equal variances (statistic = 0.145, p = 0.705). The Mann-Whitney U test showed no significant difference between the groups (statistic = 193.5, p = 0.871).

For the middle phase, the mean MTOT values were 0.503 for the trust-enhanced explanation group Experimental group with an SD of 0.121, and 0.588 for the non-trust explanation group with an SD of 0.161. Both groups had non-normal distributions (the trust-enhanced explanation group: statistic = 0.562, p < 0.0001; the non-trust explanation group: statistic = 0.770, p = 0.0003), and Levene's test indicated equal variances (p = 0.250). The Mann-Whitney U test revealed a statistically significant difference between the groups (statistic = 105.0, p = 0.011), suggesting a divergence in trust during this phase.

For the end phase, the mean MTOT values were 0.508 for the trust-enhanced explanation group with an SD of 0.126, and 0.561 for the non-trust explanation group with an SD of 0.180. Both groups were again non-normally distributed (the trust-enhanced explanation group: statistic = 0.611, p < 0.0001; the non-trust explanation group: statistic = 0.690, p < 0.0001). Levene's test showed equal variances (statistic = 0.647, p = 0.426), and the Mann-Whitney U test indicated no significant difference between the groups (statistic = 158.5, p = 0.267).

This phase-by-phase analysis highlights that the significant difference in trust occurred during the middle interval, with the agent having a generally higher trust level towards participants in the non-trust explanation group. There are no significant differences at the start or end of the task.

5.5. Open Questions

As shown in Appendix A, the last two questions in the questionnaire we used are open questions aimed at eliciting qualitative and subjective insights from participants. The first question is designed to capture participants' subjective perceptions of the agent's trustworthiness and their sense of whether they felt trusted by the agent. The second question solicits participants' constructive feedback on how to enhance the agent's trustworthiness. Additionally, it invites suggestions on effective methods by which the agent might communicate its trust in the participant.

In response to the first question, nine participants in the trust-enhanced explanation group reported perceiving that the agent either lacked trust in them or did not trust them sufficiently. This figure is notably higher compared to the non-trust explanation group, where only two participants expressed similar sentiments. Two members of the trust-enhanced explanation group expressed their concerns that the waiting time of the agent was too short.

In the responses to the second question, participants suggested several ways to improve the agent's trustworthiness and transparency.

In the trust-enhanced explanation group, two of the participants suggested including real-time feedback in the messages, including on the agent's location, status, and estimated time to complete actions, particularly when assisting with tasks such as removing obstacles or carrying objects. One participant also emphasized the need for more detailed explanations about the trust mechanism in the tutorial to improve efficiency and understanding. One highlighted the importance of the agent demonstrating greater autonomy, such as independently identifying and assisting injured individuals, while another one expressed a preference for more human-like interactions and straightforward, user-friendly communication methods. One participant also mentioned that it would be better if the trust value was shown in a UI element instead of expressed in the messages, indicating a real-time trust indicator.

In the non-trust explanation group, one of the participants said that he hoped there was an explanation in the messages of how much the agent trusted him, which was exactly what was provided in the trust-enhanced group. This highlights the significance of explicitly communicating artificial trust to participants. Another participant suggested that the agent's thought process should be explained in greater detail, while a third participant recommended incorporating more comprehensive descriptions of the trust mechanism in the tutorial. These suggestions align closely with those from the trust-enhanced explanation group, further underscoring the importance of enhancing transparency in agent communication.

6

Discussion and Conclusion

6.1. Research Questions

In this research, we want to answer the primary research question "How does adding information about artificial trust level and the agent's behavior changes corresponding to its trust level in the agent's explanations affect human trust and explanation satisfaction?" Besides, we also want to answer two related extra research questions, which are "How does the participants' satisfaction with the explanations correlate with their trust in the agent?" and "To what extent do differences in the agent's explanations about artificial trust and its behavior impact the artificial trust itself, and team performance?".

6.2. Discussion

In this section, we will discuss the results from the data analysis and present our insights into these results.

6.2.1. Human Trust and Satisfaction

As stated in section 4.2, this study hypothesized that providing explanations about artificial trust would enhance human satisfaction and increase trust toward the agent. Contrary to these expectations, the results did not reveal statistically significant differences between the trust-enhanced explanation group and non-trust explanation group across measures of Capacity Trust and Moral Trust, which, in the end, reflected in the measures of full trust. As demonstrated in Sections 5.1.1 and 5.1.2, the results for both capacity trust and moral trust across the two groups exhibit a high degree of similarity, not only in terms of the mean scores but also in the distribution of responses, as reflected in the boxplot visualizations.

This finding does not align with some previous research, such as Guillou et al.[33], which demonstrated that sharing intentions in collaborative tasks significantly enhances user trust and acceptability of artificial agents, even when team performance is unaffected. Similarly, Lavender et al.[55] have highlighted that the clarity and type of explanations provided by agents can significantly impact trust and satisfaction. For instance, proactive explanations, tailored to task context, have been shown to positively influence trust and collaboration, suggesting the importance of explanatory content design. In support of this, Verhagen et al.[88] suggest that personalized explanations, designed to align with user trust levels and workload, are more likely to enhance satisfaction and trust. Without clear differentiation, explanations may fail to produce significant effects.

There are different possible reasons behind such misalignment. The result could imply that explanations with or without a description of artificial trust may have a similar effect on human trust and explanation satisfaction. However, it can also indicate that the similarity between the TE and TU explanations may be too close. As we have presented in section3.5, despite the intentional design differences between the TE and TU explanations, participants may not have perceived a meaningful distinction, which could have minimized the potential impact on human trust and explanation satisfaction.

Besides, as suggested by Borragán et al.[11], cognitive fatigue may also affect the results. The simulated

SAR task may have imposed a significant cognitive load on participants and led to fatigue, reducing their capacity to process additional information effectively. Participants might prioritize finishing the task over processing explanatory content, thereby diluting the effect of different explanation types, which correspond to Mcneese et al.'s[61] study where human participants worked with synthetic teammates, the focus of the human participants was frequently on accomplishing primary mission goals rather than exploring teammate dynamics. In our case, the participants in the trust-enhanced group may overlook the information about the artificial trust, as it is not directly related to the goal of the task, which causes team performance similar to that of the non-trust explanation group. This speculation aligns with the finding that cognitive capacity affects trust repair strategies, and if explanations are too complex or fail to match the user's understanding, they might add to cognitive load, potentially reducing their effectiveness[54].

6.2.2. Correlation between Explanation Satisfaction and Human Trust

The Spearman correlation analysis conducted to examine the relationship between participants' satisfaction with the explanations and their trust in the agent revealed a strong positive correlation (correlation coefficient = 0.71, $p < 0.001$). This result indicates that higher satisfaction with the explanations is associated with a higher level of trust in the agent.

The strong positive correlation indicates that participants who reported higher satisfaction with the explanations provided by the agent also demonstrated greater trust in the agent. The results of 2-degree polynomial regression tests further suggest that when the agent delivers explanations that users perceive as satisfactory, it can strengthen users' trust in both the agent's abilities and intentions. These findings align with Lavender et al.'s finding[55], where they find that positive explanations improve satisfaction and trust, and in our case, the explanations provided by both the agent of the trust-enhanced group and the agent of the non-trust explanation group are positive or at least neutral, which does not highlight the weakness of the player. Moreover, this observation aligns with findings from Hafizoglu and Sen [35], who demonstrated that positive reputations of agent teammates significantly increased both trust and satisfaction in HAT.

The plot 5.5 illustrating the Spearman correlation gives a more intuitive view of the positive correlation with a regression line. The 95% confidence interval, which is around the regression line, is relatively narrow, suggesting a consistent relationship across the data set.

While the earlier analyses did not find statistically significant differences in human trust between the trust-enhanced explanation and non-trust explanation groups, the significant correlation between explanation satisfaction and human trust provides valuable insights. The results suggest that regardless of the explanation types, participants who perceived the explanations as more satisfactory were likely to trust the agent more. This finding suggests that providing explanations that satisfy users is crucial in fostering human trust.

6.2.3. Team Performance

In the analysis, we did not find a significant difference between the team performance of the two groups. The mean score for the trust-enhanced explanation group (15.6) was slightly lower than that of the non-trust explanation group (16.2), but this difference was not statistically significant. Both groups showed normally distributed scores and equal variances, allowing for a valid comparison.

The result aligns with some previous research, which states that explanations do not necessarily improve team performance. Harbers et al.[36] conducted a study investigating the effects of agents providing explanations for their behavior on the performance of HAT, and found that explanations about agent behavior may not always lead to better team performance. In our case, we could say that providing additional explanations about artificial trust does not necessarily improve team performance compared to only providing explanations of agent behavior, which aligns with the finding of Verhagen et al. [88], who stated that personalized agent explanations based on human trust can improve explanation satisfaction and trust in the agent but may decrease performance under certain conditions.

Another reason attributed to the result could be the lack of adaptive mechanisms in the agent's movement pattern. Li et al.[57] showed that agents adapting their collaboration style to meet dynamic task complexity foster better team outcomes, especially when team members' skill levels vary significantly.

However, in our experiment setting, the shared algorithm behind the movement of the agents for both groups was fixed, which might have limited the exploration of differences in team performance.

6.2.4. Artificial Trust

The analysis of the artificial trust towards human participants showed interesting insights.

Overall, the trust-enhanced explanation group had a lower mean artificial trust (MT) than the non-trust explanation group (0.556 vs. 0.604). The Mann-Whitney U test result showed no statistically significant difference in MT between the two groups with a p-value of 0.057.

When examining the mean trust over time (MTOT), the data was divided into three phases: start, middle, and end. We found a significant difference with a p-value of 0.011 in the middle phase, where the non-trust explanation group exhibited a higher mean MTOT (0.588) compared to the trust-enhanced explanation group (0.503). We observed no statistically significant differences in the start and end phases. The significant difference in artificial trust during the middle phase suggests that the differences in the explanations influenced the agents' trust calibration process differently. In the non-trust explanation group, where explanations about the trust mechanism were absent, the agent maintained a higher level of trust towards the human participants during this critical phase. In contrast, the agent for the trust-enhanced explanation group adjusted its trust level more conservatively. This conservatism may be attributed to changes in the player's behaviors in response to receiving the additional information regarding the agent's trust level and trust-based behavior changes. The player may become unwilling to help because he/she doesn't like the agent not trusting him/her. As suggested by Chiou et al.[18], agents showing low cooperation led to reduced effectiveness and lower resource sharing from human participants, implying that perceiving a lack of trust from agents could reduce humans' willingness to collaborate. The analysis result aligns with the feedback to the open questions, where more members of the trust-enhanced explanation group felt that the agent lacked trust in them than the non-trust explanation group.

The findings highlight a nuanced aspect of HAT and resonate with some previous research. The conservative artificial trust adjustments observed in the trust-enhanced explanation group highlight how explanation transparency may modulate agent behaviors and human responses, aligning with prior research that emphasizes the role of agent predictability and transparency in fostering trust[21]. Besides, the findings coincide with Sawant et al.'s[76] finding that excessive transparency could paradoxically hinder cooperation by altering the human perception of the agent's decisions.

6.3. Limitations

6.3.1. Participants

One of the limitations of this study is the sample size. As we stated in 4.4, we recruited 40 participants, which is a number that achieves a large effect size, high power, and a low error probability while still being practical considering the limit of time and our ability to find people. The desired power, by improving which can reduce Type II Errors (False Negative), can be improved with more participants recruited. For example, while maintaining the applied effect size (0.95) and significance (0.05), the desired power can be further improved from 0.83 to 0.95 by recruiting 18 more participants.

Additionally, there was a noticeable gender imbalance among the participants, with a significantly higher proportion of males than females. This gender disparity could have influenced the results, as previous research has shown that interactions with technology may differ across genders[81].

Furthermore, participants were recruited primarily from our social circle, most of whom were either computer science (CS) students or individuals with a background in CS, aged between 22 and 30. This homogeneity in educational background and age range may have affected the outcomes, as participants may have been more comfortable with operating computers and interpreting agent behaviors than a more diverse group would have been. Thus, the results may not fully represent individuals with varying levels of technical expertise or from different age groups. This may introduce bias into the results, as research by [28] found that individuals less familiar with AI tend to have lower levels of trust and acceptance of AI systems, as they may not fully understand how these systems operate. Thus, participants without a CS background might interact less confidently with the agent, potentially

impacting the outcome of the study.

6.3.2. Latency in Remote Gaming

Another limitation of this study stems from using remote desktop tools for conducting most experiments. Due to long geographical distances and tight schedules, most participants relied on these tools rather than conducting the experiments in person. Although participants were generally able to control the game smoothly, there was an inevitable latency compared to a face-to-face session where participants could directly use the computer that hosted the game. This latency may have affected participants' performance, potentially influencing how they interacted with the agent and completed the tasks. Therefore, the results might not fully reflect performance in a scenario where latency is not a concern.

6.3.3. Measurements

The measures employed may not have been sufficiently sensitive to capture the nuances of participants' perceptions in this context. This limitation is particularly evident in the trust scale employed. To better measure Human trust in agents, one possible way is, as proposed by Scharowski & Perrig[77], to construct a scale that measures not only human trust but also distrust in agents. Another possible method is to incorporate psycho-physiological signals for real-time trust detection using predictive machine learning models, as suggested by Ajenaghughrure et al.[3], which may offer more sensitive assessments

6.3.4. Explanations

Moreover, there is a limitation in the design of the explanations. Although the trust-enhanced explanation group received additional information on the agent's trust mechanism than the non-trust explanation group, the explanations in both versions shared similar elements, which may influenced trust similarly across both groups. Such overlap raises the possibility that the explanations, while distinct, might not have differed substantially enough to produce measurable differences in human trust. Designing explanations with greater contrast in content and structure may help clarify the unique impact of trust-explained explanations compared to trust-unexplained ones.

6.3.5. Waiting Time

Last but not least, the waiting time provided by the agents for participants to make decisions during the experiment is also a limitation. Although feedback from the pilot study suggested that the waiting time was acceptable, several participants from the formal experiment expressed that the time available for them to make decisions was sometimes too short, particularly in the first few minutes of the game. This may have impacted their ability to fully assess the situation before taking action. Moreover, some participants mentioned in the answers to the open questionnaire that displaying the elapsed waiting time in real-time would be helpful, which could provide greater transparency and help them manage their decision-making process better.

6.3.6. Future Work

By addressing the limitations of the present study, we found several potential research directions that merit exploration. One of the directions that are worth exploring is to delve deeper into the influence of trust explanations on user behavior. The result of our study shows that the agent in the trust-enhanced explanation group exhibited more conservative adjustments in trust levels, which seemed to influence participants to adopt a more cautious or deliberate decision-making process. This suggests that explanations about the agent's trust can affect how users behave during collaborative tasks. Future research could investigate the mechanisms behind this behavioral change. For example, studies could examine whether providing trust explanations leads users to overthink their actions, aim to manipulate the agent's trust, or become more engaged in understanding the agent's decision-making process.

Besides, we found no significance in the results between the trust-enhanced explanation group and the non-trust explanation group under the current game environment, suggesting exploring different task environments as a promising idea. The task environment used in this study may not have been sufficiently complex to elicit significant differences in trust and performance. Future studies could employ more complicated maps or intricate task goals to investigate whether the effects of explanations

on human trust and team performance vary across different explanation settings. An example could be a simulated game with the goal of rescuing victims from various firing buildings, where the victim's situation may exacerbate if the rescuers fail to rescue him/her in time.

Additionally, future research could focus on exploring the relationship between artificial trust and the behavioral actions of participants using a better logging system. The current system logs actions per game tick, which can misrepresent participant intentions. For instance, a participant might repeatedly press the key for "Remove Together" to remove a large grey rock - a hard interdependence action they cannot accomplish alone, resulting in multiple entries for a single intended action that was never conducted. Developing a more refined logging system that distinguishes intentional actions from inadvertent inputs would enable more accurate behavioral analysis.

What's more, future studies could examine how various types of explanations affect human trust and satisfaction by varying the depth, clarity, and personalization of the explanations provided by the agent. The explanations provided in this study were static and predefined. Dynamic and adaptive explanations that respond to user actions and information needs in real-time could further refine our understanding of how explanation complexity and relevance affect user trust and satisfaction, which also cater to our participants' answers to the open questions (section 5.5).

Last but not least, while the current study found a strong correlation between explanation satisfaction and human trust in the agent, and the polynomial regression results suggest that explanation satisfaction is more a determinant of human trust than the opposite direction, as mentioned in 6.2.2, causality cannot be inferred, and investigating causal relationships is essential to deepen our understanding. Future research could design experiments capable of establishing causal links between these variables. This might involve controlled manipulations of explanation satisfaction levels to observe direct effects on trust. For example, researchers could vary the clarity, relevance, and level of detail in the agent's explanations. In a high-satisfaction condition, explanations might be clear, relevant, and personalized, while in a low-satisfaction condition, explanations could be vague, irrelevant, or overly detailed. The researchers could then measure participants' trust levels in response to these systematically altered explanations. Using this approach, the study could explore whether enhancing explanation satisfaction directly leads to increased trust.

By pursuing these research directions, future studies can enhance our understanding of the role of explanations in HATs and contribute to developing intelligent agents that effectively build trust and improve performance.

6.4. Conclusion

In this study, our primary purpose is to investigate how adding information about an agent's trust levels (artificial trust) and corresponding behavior changes influences human trust in the agent and satisfaction with the explanations in a simulated search and rescue (SAR) game. While the initial hypothesis suggested that adding such information would lead to increased human trust in the agent and higher explanation satisfaction, the results did not show statistically significant differences for both measures between the trust-enhanced explanation group and the non-trust explanation group, which is inconsistent with the hypothesis. The reason could be that adding such information does not necessarily improve human trust and explanation satisfaction. Another reason could be that the differences in the explanations provided to the two groups were not distinct enough for the participants to perceive a meaningful distinction, which minimizes the potential impact on the dependent variables.

However, we found a strong positive correlation between explanation satisfaction and trust. Participants who were more satisfied with the explanations provided by the agent exhibited higher levels of trust towards the agent. This finding highlights the role of explanation in fostering human trust, regardless of the presence or absence of trust-related content, showing that user-satisfied explanations are crucial for enhancing trust in HATs.

We found no significant difference between the two groups during the analysis of team performance. The reason could be that explanations about trust do not necessarily enhance the team's ability to complete the SAR task, or the fixed movement algorithm behind the agents limited the performance differences during the experiment. Additionally, in the analysis of artificial trust, we found that the

agent in the trust-enhanced explanation group adjusted its trust more conservatively than the non-trust explanation group. The result may indicate that the explanations influence human behavior by affecting their willingness to collaborate, leading to more cautious actions that potentially affect trust calibration.

To summarize, in this study, we cannot conclude that adding information about artificial trust level and the agent's behavior changes corresponding to its trust level in the agent's explanations improves human trust and explanation satisfaction. However, we found there is a strong correlation between explanation satisfaction and human trust, which emphasizes the value of artificial explanations in HATs, pointing out delving deeper into the relationship between them as future work to enhance HAT. Besides, although we cannot conclude that adding trust information into explanations improves team performance, the observation that the agent in the trust-enhanced explanation group adjusted its trust more conservatively than the non-trust explanation group suggests that the explanations may have influenced human behavior, affecting their willingness to collaborate and trust calibration, which suggests that the relationship between artificial trust and the behavior of participants is worth further exploring.

References

[1] Amina Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[2] Sebastian Ahrndt, Johannes Fähndrich, and S. Albayrak. "Human-agent teamwork: what is predictability, why is it important?" In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (2016). DOI: 10.1145/2851613.2851928.

[3] I. B. Ajenaghughrure et al. "Predictive model to assess user trust: a psycho-physiological approach". In: *Proceedings of the 10th Indian Conference on Human-Computer Interaction* (2019). DOI: 10.1145/3364183.3364195.

[4] Z. Akata et al. "A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence". In: *Computer* 53 (2020), pp. 18–28. DOI: 10.1109/mc.2020.2996587.

[5] Basel Alhaji, A. Rausch, and Michael Prilla. "Toward Mutual Trust Modeling in Human-Robot Collaboration". In: *ArXiv* abs/2011.01056 (2020).

[6] Andrea Apicella et al. "Strategies to exploit XAI to improve classification systems". In: (2023), pp. 147–159. DOI: 10.48550/arXiv.2306.05801.

[7] R. Bellamy et al. "Human-Agent Collaboration: Can an Agent be a Partner?" In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017). DOI: 10.1145/3027063.3051138.

[8] Ezekiel Bernardo and R. Seva. "Affective Design Analysis of Explainable Artificial Intelligence (XAI): A User-Centric Perspective". In: *Informatics* 10 (2023), p. 32. DOI: 10.3390/informatics10010032.

[9] Ezekiel Bernardo and R. Seva. "Evaluating the Effect of Time on Trust Calibration of Explainable Artificial Intelligence". In: *Artificial Intelligence and Social Computing* (2023). DOI: 10.54941/ahfe1003280.

[10] P. Bobko et al. "Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems". In: *Theoretical Issues in Ergonomics Science* 24 (2022), pp. 310–334. DOI: 10.1080/1463922X.2022.2086644.

[11] Guillermo Borragán et al. "Cognitive fatigue: A Time-based Resource-sharing account". In: *Cortex* 89 (2017), pp. 71–84. DOI: 10.1016/j.cortex.2017.01.023.

[12] Regina de Brito Duarte et al. "AI Trust: Can Explainable AI Enhance Warranted Trust?" In: *Human Behavior and Emerging Technologies* (2023). DOI: 10.1155/2023/4637678.

[13] Ana Carrasco. "Adapting Behaviour Based On Trust In Human-Agent Ad Hoc Teamwork". In: *ArXiv* abs/2210.06915 (2022). DOI: 10.48550/arXiv.2210.06915.

[14] Jessie Y. C. Chen and Michael J. Barnes. "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues". In: *IEEE Transactions on Human-Machine Systems* 44.1 (2014), pp. 13–29. DOI: 10.1109/THMS.2013.2293535.

[15] Jessie Y.C. Chen et al. "Human-Autonomy Teaming and Agent Transparency". In: *Companion Publication of the 21st International Conference on Intelligent User Interfaces* (2016). DOI: 10.1145/2876456.2879479.

[16] Jessie Y.C. Chen et al. "Situation awareness-based agent transparency and human-autonomy teaming effectiveness". In: *Theoretical Issues in Ergonomics Science* 19 (2018), pp. 259–282. DOI: 10.1080/1463922X.2017.1315750.

[17] Wei Chen, E. Durfee, and Melanie Dumas. "Human agent collaboration in a simulated combat medical scenario". In: *2009 International Symposium on Collaborative Technologies and Systems* (2009), pp. 367–375. DOI: 10.1109/CTS.2009.5067503.

[18] Erin K. Chiou and John D. Lee. "Cooperation in Human-Agent Systems to Support Resilience". In: *Human Factors: The Journal of Human Factors and Ergonomics Society* 58 (2016), pp. 846–863. DOI: 10.1177/0018720816649094.

[19] Hyesun Choung, Prabu David, and Arun Ross. "Trust in AI and Its Role in the Acceptance of AI Technologies". In: *International Journal of Human–Computer Interaction* 39 (2022), pp. 1727–1739. DOI: 10.1080/10447318.2022.2050543.

[20] Mert Dagli. "Designing for trust: Exploring trust and collaboration in conversational agents for e-commerce". In: *Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–10.

[21] Sylvain Daronnat et al. "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). DOI: 10.1145/3406499.3415063.

[22] Sylvain Daronnat et al. "Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration". In: *Frontiers in Robotics and AI* 8 (2021). DOI: 10.3389/frobt.2021.642201.

[23] Devleena Das and S. Chernova. "Leveraging rationales to improve human task performance". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces* (2020). DOI: 10.1145/3377325.3377512.

[24] Jagjit Singh Dhatterwal, Mahaveer Singh Naruka, and K. Kaswan. "Multi-Agent System based Medical Diagnosis Using Particle Swarm Optimization in Healthcare". In: *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)* (2023), pp. 889–893. DOI: 10.1109/AISC56616.2023.10085654.

[25] Jeff Druce, M. Harradon, and J. Tittle. "Explainable Artificial Intelligence (XAI) for Increasing User Trust in Deep Reinforcement Learning Driven Autonomous Systems". In: *ArXiv* abs/2106.03775 (2021).

[26] J. S. Elson, D. Derrick, and G. Ligon. "Examining Trust and Reliance in Collaborations between Humans and Automated Agents". In: (2018), pp. 1–10. DOI: 10.24251/HICSS.2018.056.

[27] Mariela Morveli Espinoza, A. Possebom, and Cesar Augusto Tacla. "Argumentation-Based Agents that Explain Their Decisions". In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)* (2019), pp. 467–472. DOI: 10.1109/BRACIS.2019.00088.

[28] Ethan Fast and Eric Horvitz. "Long-Term Trends in the Public Perception of Artificial Intelligence". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (Feb. 2017). DOI: 10.1609/aaai.v31i1.10635. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10635.

[29] Carlos Fernandez, F. Provost, and Xintian Han. "Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach". In: *ArXiv* abs/2001.07417 (2020). DOI: 10.25300/misq/2022/16749.

[30] A. Ferrario and M. Loi. "How Explainability Contributes to Trust in AI". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022). DOI: 10.1145/3531146.3533202.

[31] S. Gorokhovskyi and Oleksandra Radziievska. "Agent-Based Modeling of Collaborative Work". In: *NaUKMA Research Papers. Computer Science* (2021). DOI: 10.18523/2617-3808.2021.4.60-63.

[32] S. Guastello, M. Koopmans, and D. Pincus. "Chaos and complexity in psychology: The theory of nonlinear dynamical systems." In: (2008). DOI: 10.1017/CBO9781139058544.

[33] M. L. Guillou, Laurent Prévot, and B. Berberian. "Trusting Artificial Agents: Communication Trumps Performance". In: (2023), pp. 299–306. DOI: 10.5555/3545946.3598651.

[34] David Gunning et al. "XAI—Explainable artificial intelligence". In: *Science Robotics* 4.37 (2019), eaay7120. DOI: 10.1126/scirobotics.aay7120. eprint: https://www.science.org/doi/pdf/10.1126/scirobotics.aay7120. URL: https://www.science.org/doi/abs/10.1126/scirobotics.aay7120.

[35] Feyza Merve Hafizoglu and S. Sen. "Reputation Based Trust In Human-Agent Teamwork Without Explicit Coordination". In: *Proceedings of the 6th International Conference on Human-Agent Interaction* (2018). DOI: 10.1145/3284432.3284454.

[36] M. Harbers et al. "Explanation in Human-Agent Teamwork". In: (2011), pp. 21–37. DOI: 10.1007/978-3-642-35545-5_2.

[37] Vikas Hassija et al. "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence". In: *Cognitive Computation* 16.1 (2024), pp. 45–74. ISSN: 1866-9964. DOI: 10.1007/s12559-023-10179-8. URL: https://doi.org/10.1007/s12559-023-10179-8.

[38] Maartje Hidalgo and Lauren-Reinerman Jones. "Towards a Comprehensive Model of Military Human-Agent Teaming". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67 (2023), pp. 1995–2001. DOI: 10.1177/21695067231192436.

[39] Robert R Hoffman et al. "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance". In: *Frontiers in Computer Science* 5 (2023), p. 1096257.

[40] Robert R. Hoffman et al. *Metrics for Explainable AI: Challenges and Prospects*. 2019. arXiv: 1812.04608 [cs.AI].

[41] A. Howard. "Are We Trusting AI Too Much? Examining Human-Robot Interactions in the Real World". In: *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2020), pp. 1–1. DOI: 10.1145/3319502.3374842.

[42] Hsiao-Ying Huang and Masooda N. Bashir. "Personal Influences on Dynamic Trust Formation in Human-Agent Interaction". In: *Proceedings of the 5th International Conference on Human Agent Interaction* (2017). DOI: 10.1145/3125739.3125749.

[43] Sadaf Hussain et al. "Trait Based Trustworthiness Assessment in Human-Agent Collaboration Using Multi-Layer Fuzzy Inference Approach". In: *IEEE Access* 9 (2021), pp. 73561–73574. DOI: 10.1109/ACCESS.2021.3079838.

[44] Vidhi Jain et al. *Predicting Human Strategies in Simulated Search and Rescue Task*. 2020. arXiv: 2011.07656 [cs.LG].

[45] Matthew Johnson et al. "Coactive design: designing support for interdependence in joint activity". In: *J. Hum.-Robot Interact.* 3.1 (Feb. 2014), pp. 43–69. DOI: 10.5898/JHRI.3.1.Johnson. URL: https://doi.org/10.5898/JHRI.3.1.Johnson.

[46] Matthew Johnson et al. *The Fundamental Principle of Coactive Design: Interdependence Must Shape Autonomy*. 2011.

[47] Bart de Jong, K. Dirks, and N. Gillespie. "Trust and team performance: A meta-analysis of main effects, moderators, and covariates." In: *The Journal of applied psychology* 101 8 (2016), pp. 1134–50. DOI: 10.1037/apl0000110.

[48] Carolina Centeio Jorge, M. Tielman, and C. Jonker. "Artificial Trust as a Tool in Human-AI Teams". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2022), pp. 1155–1157. DOI: 10.1109/HRI53351.2022.9889652.

[49] Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. "Assessing artificial trust in human-agent teams: a conceptual model". In: *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. IVA '22. Faro, Portugal: Association for Computing Machinery, 2022. ISBN: 9781450392488. DOI: 10.1145/3514197.3549696. URL: https://doi.org/10.1145/3514197.3549696.

[50] Carolina Centeio Jorge et al. "Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork". In: *Frontiers in Robotics and AI* 10 (2023). ISSN: 22969144. DOI: 10.3389/frobt.2023.1143723.

[51] A. D. Kaplan et al. "Trust in Artificial Intelligence: Meta-Analytic Findings". In: *Human Factors: The Journal of Human Factors and Ergonomics Society* 65 (2021), pp. 337–359. DOI: 10.1177/00187208211013988.

[52] Frank Kaptein et al. “The role of emotion in self-explanations by cognitive agents”. In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (2017), pp. 88–93. DOI: 10.1109/ACIIW.2017.8272595.

[53] Z. Kemény et al. “Human–Robot Collaboration in Manufacturing: A Multi-agent View”. In: (2021), pp. 3–41. DOI: 10.1007/978-3-030-69178-3_1.

[54] E. Kox et al. “Trust repair in human-agent teams: the effectiveness of explanations and expressing regret”. In: *Autonomous Agents and Multi-Agent Systems* 35 (2021). DOI: 10.1007/s10458-021-09515-9.

[55] Bryan Lavender, Sami Abuhaimed, and S. Sen. “Relative Effects of Positive and Negative Explanations on Satisfaction and Performance in Human-Agent Teams”. In: *The International FLAIRS Conference Proceedings* (2023). DOI: 10.32473/flairs.36.133371.

[56] K. Lazányi and Beáta Hajdu. “Trust in human-robot interactions”. In: *2017 IEEE 14th International Scientific Conference on Informatics* (2017), pp. 216–220. DOI: 10.1109/INFORMATICS.2017.8327249.

[57] Huao Li et al. “Individualized Mutual Adaptation in Human-Agent Teams”. In: *IEEE Transactions on Human-Machine Systems* 51 (2021), pp. 706–714. DOI: 10.1109/thms.2021.3107675.

[58] Chin-Teng Lin et al. “Modelling the Trust Value for Human Agents Based on Real-Time Human States in Human-Autonomous Teaming Systems”. In: *Technologies* 10 (Nov. 2022), p. 115. DOI: 10.3390/technologies10060115.

[59] R. Maheswaran et al. “Multi-agent systems for the real world”. In: *2010 International Symposium on Collaborative Technologies and Systems* (2009), pp. 639–640. DOI: 10.1109/CTS.2010.5478452.

[60] B. Malle and Matthias Scheutz. “Moral competence in social robots”. In: *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering* (2014), pp. 1–6. DOI: 10.1109/ETHICS.2014.6893446.

[61] Nathan J. Mcneese et al. “Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming”. In: *Human Factors: The Journal of Human Factors and Ergonomics Society* 60 (2018), pp. 262–273. DOI: 10.1177/0018720817743223.

[62] Siddharth Mehrotra et al. *A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction*. 2023. arXiv: 2311.06305 [cs.HC].

[63] Christian Meske and Enrico Bunde. “Using Explainable Artificial Intelligence to Increase Trust in Computer Vision”. In: *ArXiv* abs/2002.01543 (2020).

[64] Yazan Mualla et al. “The quest of parsimonious XAI: A human-agent architecture for explanation formulation”. In: *Artif. Intell.* 302 (2021), p. 103573. DOI: 10.1016/J.ARTINT.2021.103573.

[65] Shane T. Mueller et al. “Principles of Explanation in Human-AI Systems”. In: *ArXiv* abs/2102.04972 (2021).

[66] Mark A Neerincx et al. “Using perceptual and cognitive explanations for enhanced human-agent team performance”. In: *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer. 2018, pp. 204–214. DOI: 10.1007/978-3-319-91122-9_18.

[67] Florian Nothdurft, Tobias Heinroth, and Wolfgang Minker. “The Impact of Explanation Dialogues on Human-Computer Trust”. In: *Human-Computer Interaction. Users and Contexts of Use*. Ed. by Masaaki Kurosu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 59–67. ISBN: 978-3-642-39265-8.

[68] Mayada Oudah et al. “How AI Wins Friends and Influences People in Repeated Games With Cheap Talk”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (Apr. 2018). DOI: 10.1609/aaai.v32i1.11486.

[69] Ravi Pandya et al. “Multi-Agent Strategy Explanations for Human-Robot Collaboration”. In: *ArXiv* abs/2311.11955 (2023). DOI: 10.48550/arXiv.2311.11955.

[70] B. Pfeifer et al. “Explainable AI with counterfactual paths”. In: *ArXiv* abs/2307.07764 (2023). DOI: 10.48550/arXiv.2307.07764.

[71] J. P. Queralta et al. “Collaborative Multi-Robot Systems for Search and Rescue: Coordination and Perception”. In: *ArXiv* abs/2008.12610 (2020).

[72] Sarvapali D. Ramchurn et al. "Human–agent collaboration for disaster response". In: *Autonomous Agents and Multi-Agent Systems* 30.1 (2016), pp. 82–111. ISSN: 1573-7454. DOI: 10.1007/s10458-015-9286-4. URL: https://doi.org/10.1007/s10458-015-9286-4.

[73] Michael Ridley. "Explainable Artificial Intelligence (XAI)". In: *Information Technology and Libraries* (2022). DOI: 10.6017/ital.v41i2.14683.

[74] Waddah Saeed and Christian Omlin. "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities". In: *Knowledge-Based Systems* 263 (2023), p. 110273. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2023.110273. URL: https://www.sciencedirect.com/science/article/pii/S0950705123000230.

[75] Wojciech Samek and Klaus-Robert Müller. "Towards Explainable Artificial Intelligence". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2019, pp. 5–22. ISBN: 9783030289546. DOI: 10.1007/978-3-030-28954-6_1. URL: http://dx.doi.org/10.1007/978-3-030-28954-6_1.

[76] Sarvesh Sawant et al. "Balancing the Scales of Explainable and Transparent AI Agents within Human-Agent Teams". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67 (2023), pp. 2082–2087. DOI: 10.1177/21695067231192250.

[77] Nicolas Scharowski and S. Perrig. "Distrust in (X)AI - Measurement Artifact or Distinct Construct?" In: *ArXiv* abs/2303.16495 (2023). DOI: 10.48550/arXiv.2303.16495.

[78] Patrik Schmuck and Margarita Chli. "CCM-SLAM: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams". In: *Journal of Field Robotics* 36.4 (2019), pp. 763–781. DOI: https://doi.org/10.1002/rob.21854. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21854. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21854.

[79] Ronal Singh, Liz Sonenberg, and Tim Miller. "Communication and Shared Mental Models for Teams Performing Interdependent Tasks". In: *COIN@AAMAS/ECAI*. 2016. URL: https://api.semanticscholar.org/CorpusID:39192657.

[80] Christos Sioutis and Jeffrey Tweedale. "Agent Cooperation and Collaboration". In: (2006), pp. 464–471. DOI: 10.1007/11893004_60.

[81] Sabrina Sobieraj and N. Krämer. "Similarities and differences between genders in the usage of computer with different levels of technological complexity". In: *Comput. Hum. Behav.* 104 (2020), p. 106145. DOI: 10.1016/j.chb.2019.09.021.

[82] A. Srinivasan and Mona de Boer. "Improving trust in data and algorithms in the medium of AI". In: 94 (2020), pp. 147–160. DOI: 10.5117/mab.94.49425.

[83] M. Taddeo and L. Floridi. "The case for e-trust". In: *Ethics and Information Technology* 13 (2011), pp. 1–3. DOI: 10.1007/s10676-010-9263-1.

[84] Marko Tešić and U. Hahn. "Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that?" In: *Patterns* 3 (2022). DOI: 10.1016/j.patter.2022.100635.

[85] Dedra Townsend and AmirHossein Majidirad. "Trust in Human-Robot Interaction Within Healthcare Services: A Review Study". In: *Volume 7: 46th Mechanisms and Robotics Conference (MR)* (2022). DOI: 10.1115/detc2022-89607.

[86] Anna-Sophie Ulfert and Eleni Georganta. "A Model of Team Trust in Human-Agent Teams". In: *Companion Publication of the 2020 International Conference on Multimodal Interaction* (2020). DOI: 10.1145/3395035.3425959.

[87] Daniel Ullman and Bertram F Malle. *MDMT: multi-dimensional measure of trust*. 2019.

[88] R. Verhagen et al. "Personalized Agent Explanations for Human-Agent Teamwork: Adapting Explanations to User Trust, Workload, and Performance". In: (2023), pp. 2316–2318. DOI: 10.5555/3545946.3598919.

[89] R.S. Verhagen, M.A. Neerincx, and M.L. Tielman. "Meaningful human control and variable autonomy in human-robot teams for firefighting". English. In: *Frontiers In Robotics and AI* 11 (2024). ISSN: 2296-9144. DOI: 10.3389/frobt.2024.1323980.

[90] Ruben Verhagen et al. "The Influence of Interdependence on Trust Calibration in Human-Machine Teams". In: June 2024. ISBN: 9781643685229. DOI: 10.3233/FAIA240203.

[91] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. "A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable". In: *Explainable and Transparent AI and Multi-Agent Systems*. Ed. by Davide Calvaresi et al. Cham: Springer International Publishing, 2021, pp. 119–138. ISBN: 978-3-030-82017-6.

[92] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. "The influence of interdependence and a transparent or explainable communication style on human-robot teamwork". In: *Frontiers in Robotics and AI* 9 (Sept. 2022). ISSN: 22969144. DOI: 10.3389/frobt.2022.993997.

[93] Giulia Vilone and L. Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence". In: *Inf. Fusion* 76 (2021), pp. 89–106. DOI: 10.1016/J.INFFUS.2021.05.009.

[94] J. van der Waa et al. "Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations". In: *Frontiers in Robotics and AI* 8 (2021), p. 640647. DOI: 10.3389/frobt.2021.640647.

[95] Wendell Wallach and C. Allen. "Moral Machines: Teaching Robots Right from Wrong". In: (2008). DOI: 10.1093/acprof:oso/9780195374049.001.0001.

[96] Ning Wang, D. Pynadath, and S. Hill. "Trust calibration within a human-robot team: Comparing automatically generated explanations". In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2016), pp. 109–116. DOI: 10.1109/HRI.2016.7451741.

[97] Ning Wang, David V. Pynadath, and Susan G. Hill. "Building Trust in a Human-Robot Team with Automatically Generated Explanations". In: 2015. URL: https://api.semanticscholar.org/CorpusID:18265367.

[98] Sunwei Wang. *The Impact of Expressing Emotion within Explainable AI in Human-Agent Teamwork*. Graduation committee members: Interactive Intelligence - EEMCS. 2024. URL: http://resolver.tudelft.nl/uuid:8616eb26-d8ef-45f6-be73-9e60bdae646b.

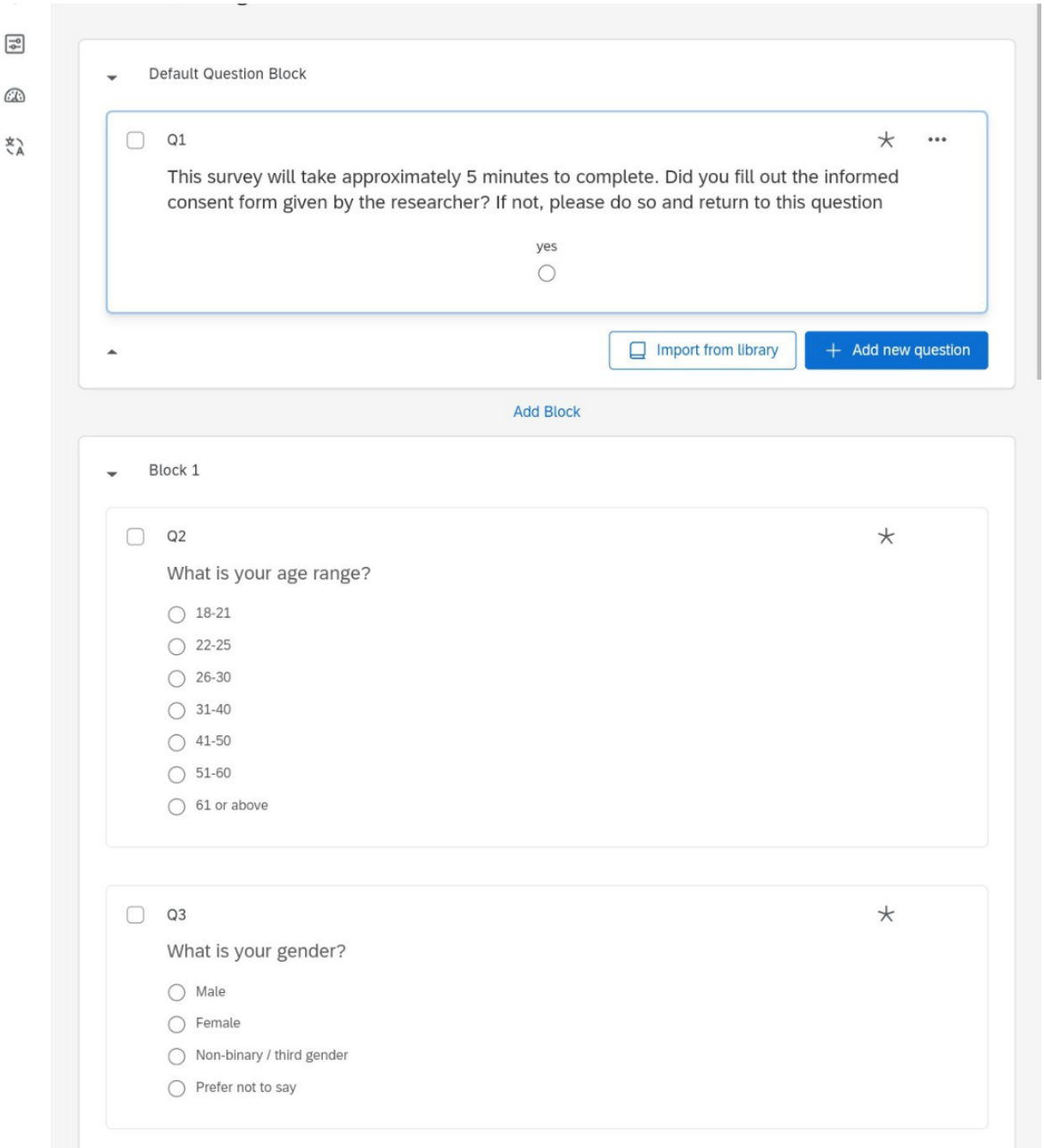
[99] Xiaotian Wang et al. "Dynamic real-time scheduling for human-agent collaboration systems based on mutual trust". In: *Cyber-Physical Systems* 1 (2015), pp. 76–90. DOI: 10.1080/23335777.2015.1056755.

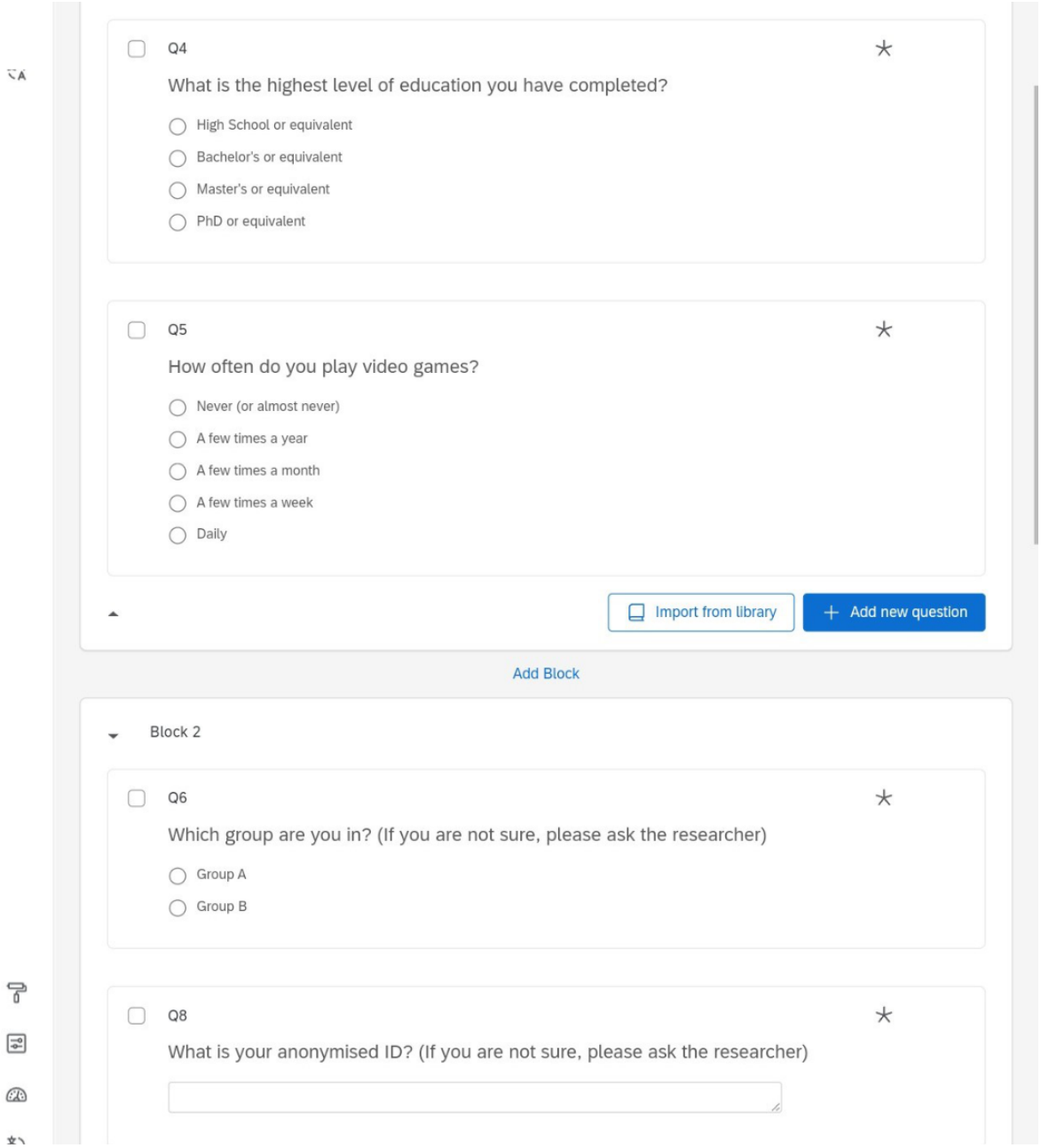
[100] A. V. Wissen et al. "Human-agent teamwork in dynamic environments". In: *Comput. Hum. Behav.* 28 (2012), pp. 23–33. DOI: 10.1016/J.CHB.2011.08.006.

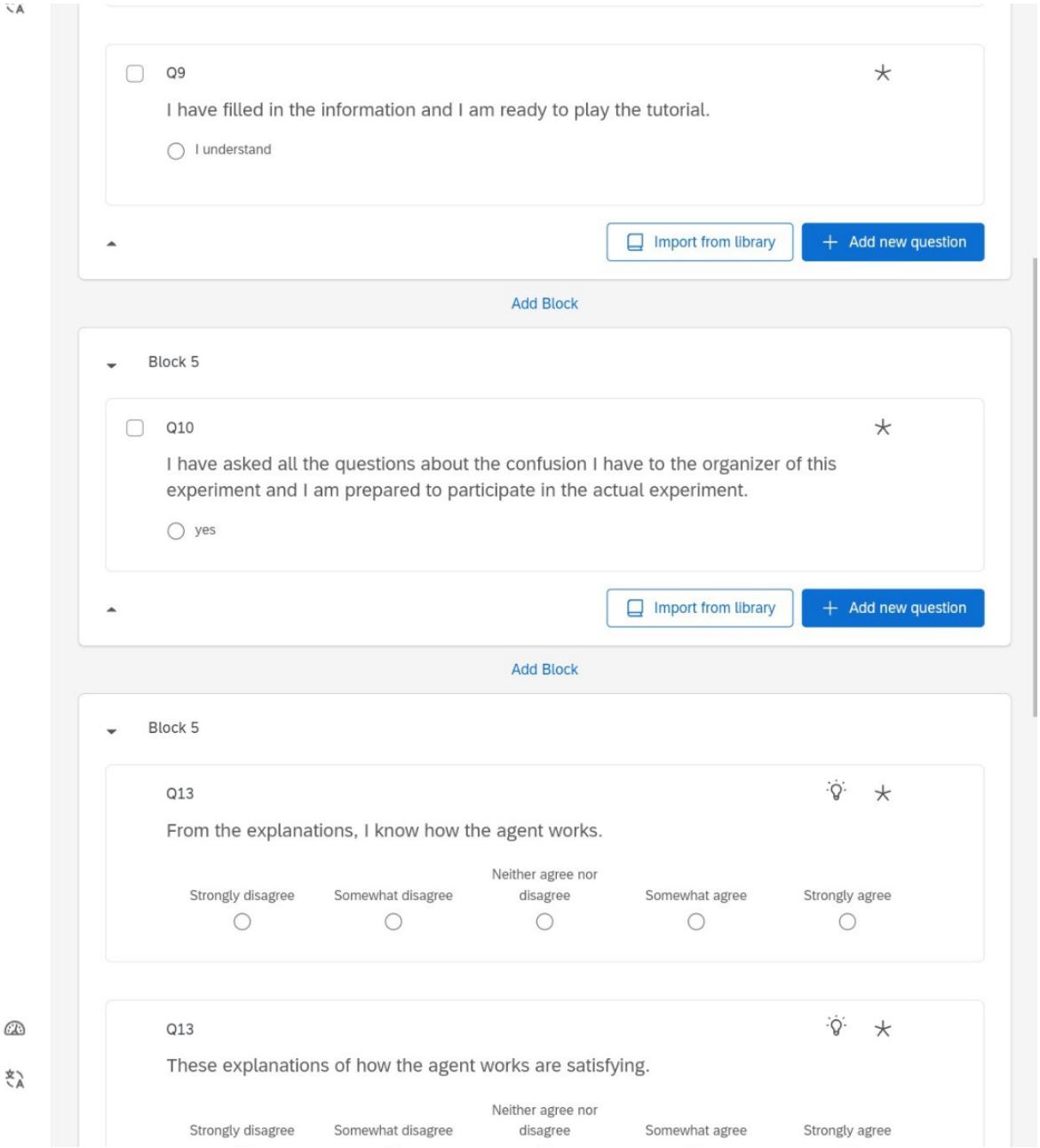
[101] Jing Zhou. "Exploring the effect of the information amount in explanation on different gaming expertise levels". Master's thesis. Delft University of Technology, 2023. URL: http://resolver.tudelft.nl/uuid:0e6369b7-b2fb-4949-a3df-fc109fa10d45.

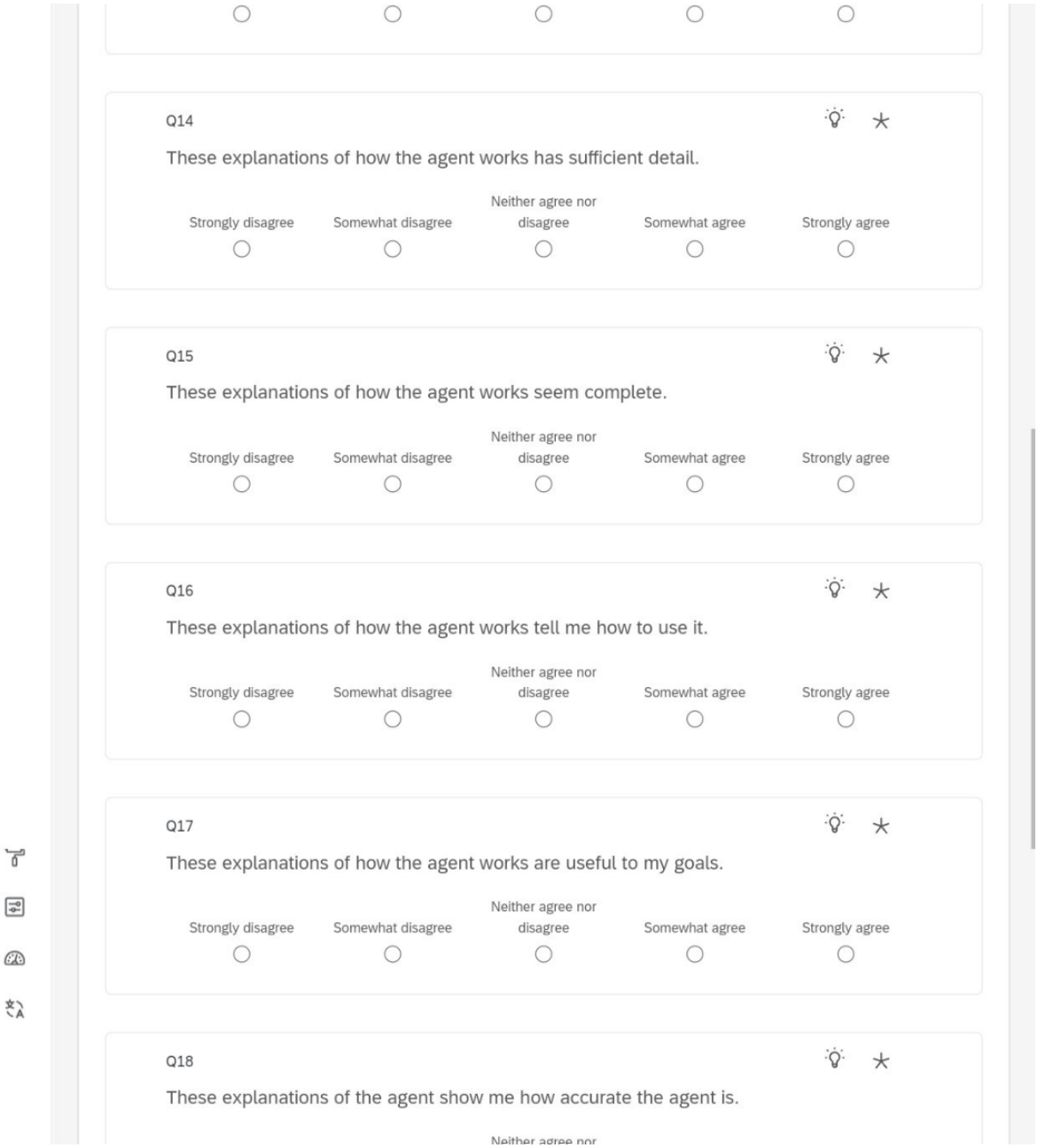
A

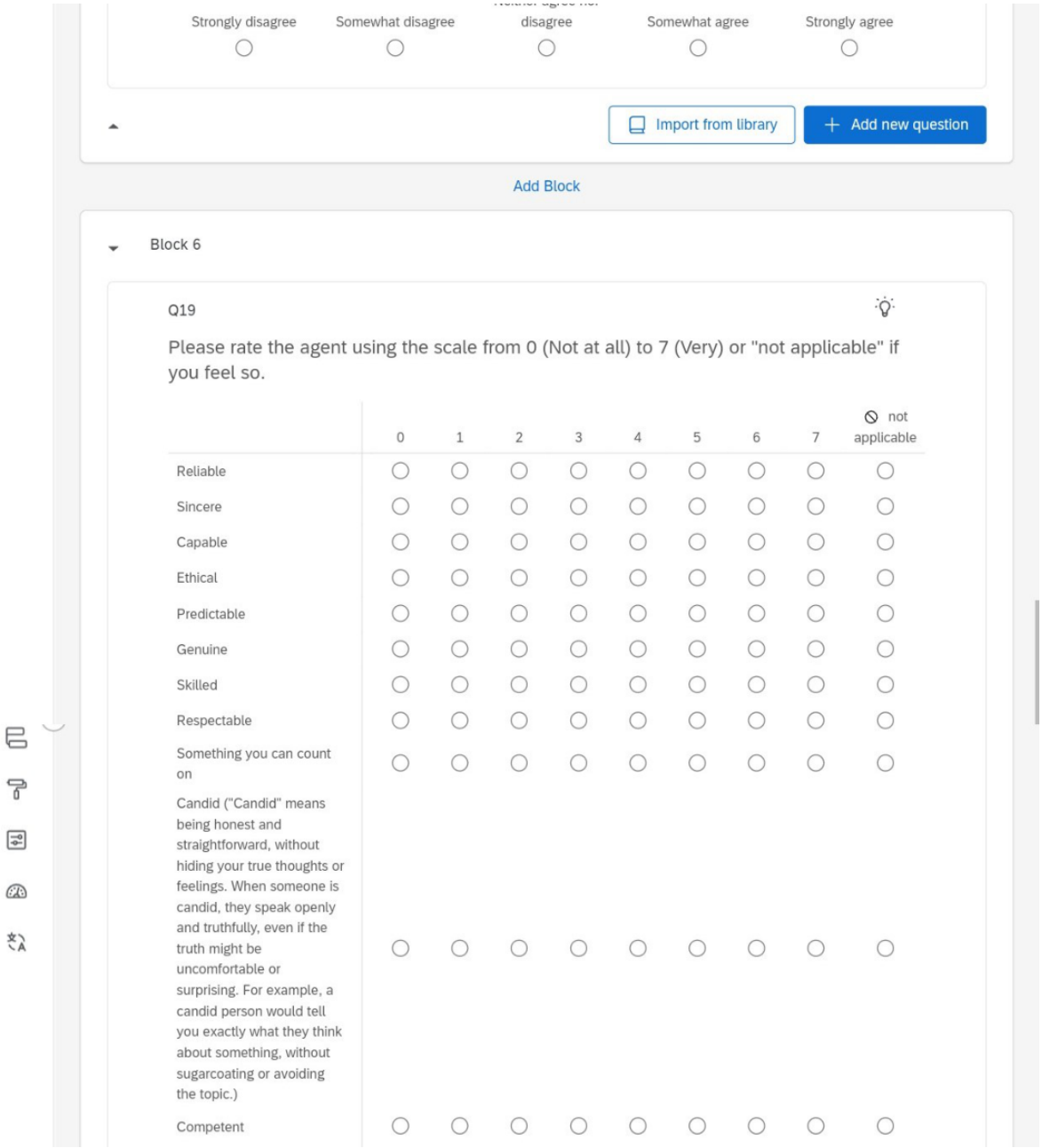
Questionnaire Used

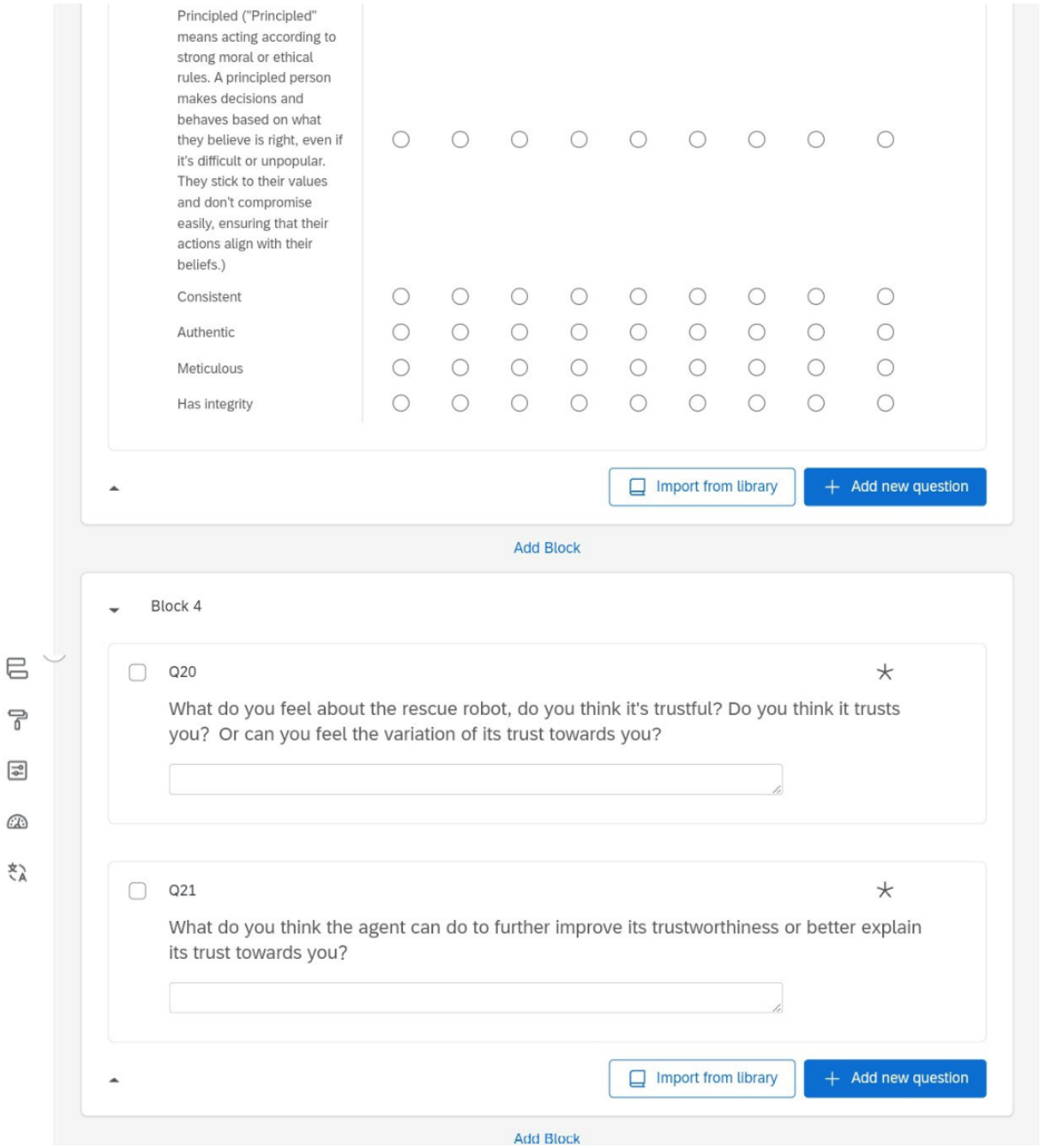



Q4
What is the highest level of education you have completed?
High School or equivalent
Bachelor's or equivalent
Master's or equivalent
PhD or equivalent
Q5
How often do you play video games?
Never (or almost never)
A few times a year
A few times a month
A few times a week
Daily
Import from library
Add new question
Add Block
Block 2
Q6
Which group are you in? (If you are not sure, please ask the researcher)
Group A
Group B
Q8
What is your anonymised ID? (If you are not sure, please ask the researcher)


Q9
I have filled in the information and I am ready to play the tutorial.
I understand
Import from library
Add new question
Add Block
Block 5
Q10
I have asked all the questions about the confusion I have to the organizer of this experiment and I am prepared to participate in the actual experiment.
yes
Import from library
Add new question
Add Block
Block 5
Q13
From the explanations, I know how the agent works.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
Q13
These explanations of how the agent works are satisfying.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree


Q14
These explanations of how the agent works has sufficient detail.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
Q15
These explanations of how the agent works seem complete.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
Q16
These explanations of how the agent works tell me how to use it.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
Q17
These explanations of how the agent works are useful to my goals.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
Q18
These explanations of the agent show me how accurate the agent is.
Neither agree nor

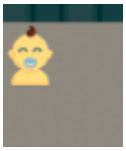

Strongly disagree
Somewhat disagree
disagree
Somewhat agree
Strongly agree
Import from library
Add new question
Add Block
Block 6
Q19
Please rate the agent using the scale from 0 (Not at all) to 7 (Very) or "not applicable" if you feel so.
0
1
2
3
4
5
6
7
not applicable
Reliable
Sincere
Capable
Ethical
Predictable
Genuine
Skilled
Respectable
Something you can count on
Candid ("Candid" means being honest and straightforward, without hiding your true thoughts or feelings. When someone is candid, they speak openly and truthfully, even if the truth might be uncomfortable or surprising. For example, a candid person would tell you exactly what they think about something, without sugarcoating or avoiding the topic.)
Competent


Principled ("Principled" means acting according to strong moral or ethical rules. A principled person makes decisions and behaves based on what they believe is right, even if it's difficult or unpopular. They stick to their values and don't compromise easily, ensuring that their actions align with their beliefs.)
Consistent
Authentic
Meticulous
Has integrity
Import from library
Add new question
Add Block
Block 4
Q20
What do you feel about the rescue robot, do you think it's trustful? Do you think it trusts you? Or can you feel the variation of its trust towards you?
Q21
What do you think the agent can do to further improve its trustworthiness or better explain its trust towards you?
Import from library
Add new question
Add Block

B

Instruction Handbook

The objective of the task is to find eight target victims in the different areas and carry them to the drop zone. Rescuing mildly injured victims



(yellow color) adds 3 points to the total score, and rescuing critically injured victims  (red color) adds 6 points to the total score. Healthy victim  (green color) do not need to be rescued. The world terminates after 7 minutes.

Critically injured victims can only be carried by both human and RescueBot together.

RescueBot and the player can only carry one victim at the same time.

The human player can rescue mildly injured victims alone.

The RescueBot can rescue mildly injured victims alone, but it is much faster to do this together with the player.

The human player can carry only one victim at the same time.

RescueBot can carry mildly injured victims alone, but doing this together with human assistance is much faster.

The big gray rock  can only be removed by both human and RescueBot together.

RescueBot can remove the small brown stone alone  , but doing this together with human

assistance is much faster.

The tree  can be removed by RescueBot or player alone.

The human player can identify obstacles with a normal perception range of 1 grid cell.

The human player can remove the small brownstone alone, but it is much faster to do this together with RescueBot.

In the game, the RescueBot will actively search victims and obstacles.

You will need to reply to the message of the agent to collaborate with it.

C

Informed Consent Form

You are being invited to participate in a research study titled Reciprocal Trust Dynamics in Human-Agent Teamwork. This study is being done by Zenan Guan from the TU Delft.

The purpose of this research study is to investigate the dynamics of reciprocal trust between humans and AI agents in collaborative environments. Specifically, you will work with an artificial agent with a trust mechanic in a simulated search and rescue task, which will take approximately 30 minutes to complete. The data will be used for result analysis for a thesis of a Master's Computer Science project. We will be asking you to interact with a simulated AI agent in a search and rescue task and respond to questions about your trust in the agent during these interactions.

As with any online activity, the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by storing your data locally for analysis and uploading it to 4TU.ResearchData afterwards, where your data will be stored safely. We will minimize any risks by only asking you about your gender, age range, level of education, and gaming experience. This will make identification close to impossible (i.e., your data is a nonymous). Since your data will be anonymous, you cannot request your data to be removed after completion of the study.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions.

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated 28/08/2024, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I understand that taking part in the study involves: 1. playing with a small video game 2. filling into a survey questionaire	☐	☐
4. I understand that there will be no financial compensation for this experiment	☐	☐
5. I understand that the study will end when the researcher finishes the Master's thesis project and graduates from TUDelft (estimated in October 2024)	☐	☐
6. I understand I can exit this experiment anytime if I don't want to continue.	☐	☐
7. I understand that the following steps will be taken to minimize the threat of a data breach, and protect my identity in the event of such a breach: the data will be anonymized and stored locally when the analysis is being conducted, and handed to the supervisors of the researcher (Myrthe Tielman, Ruben Verhagen), and they will upload it to 4TU.ResearchData repository for proper storage.	☐	☐
8. I understand that personal information collected about me, such as age, gender, education level, and game experience level, will not be shared beyond the study team.	☐	☐
9. I understand that the (identifiable) personal data I provide will be destroyed once this thesis project is ended (estimated in October 2024)	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
10. I understand that after the research study, the anonymous research data I provide will be used for data analysis in the thesis report of this project.	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
11. I permit the anonymous research data that I provide to be archived in 4TU.ResearchData repository so it can be used for future research and learning.	☐	☐

Signatures

Name of participant [printed]	Signature	Date

I, as legal representative, have witnessed the accurate reading of the consent form with the potential participant and the individual has had the opportunity to ask questions. I confirm that the individual has given consent freely.

Name of witness [printed]	Signature	Date

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Zenan Guan _____ _____

Researcher name [printed] Signature Date

Study contact details for further information: [*Name, phone number, email address*]