Communicating Trust Beliefs and Decisions in Human-AI Teams

Tamer Sahin[1]

Supervisor(s): Myrthe Tielman[1]**, Carolina Ferreira Gomes Centeio Jorge**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Tamer Sahin
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Carolina Ferreira Gomes Centeio Jorge, Ujwal Gadiraju

Abstract

As technological capabilities progress, there is a growing imperative to enhance collaborative dynamics within human-agent teams. Artificial agents and humans possess capabilities that compensate for each other's limitations. This paper outlines the effect of communication using real-time textual explanations of the artificial agents' mental model of its trust in the human teammate. An experiment (n = 40) was conducted by examining the impact of real-time textual explanations of the artificial agents' trust beliefs. The participants collaborated with an artificial agent during a search and rescue mission in a 2D grid world, where they had to rescue six victims in a 10 minute time frame. The results show that real-time textual explanations did affect the natural trust and satisfaction positively compared to the baseline method.

1 Introduction

In recent years, Artificial Intelligence (AI) has become a significant factor in various domains such as healthcare, business, and industry [1, 2]. While AI is still growing as technology, humans have high expectations for AI [3]. AI is becoming more capable of working alongside humans on a wide range of tasks, including data processing to decision making [3, 4, 5]. While agents are effective in processing large amounts of data quickly and accurately, humans have emotional intelligence and are flexible [6]. Together, they effectively cover each others shortcomings.

AI joining together with humans as teammates introduces a new type of team known as Human-AI teams (HAT). In HATs, there is at least one human and one artificial agent teammate [7]. The artificial agent and human collaborate, leveraging their complementary strengths to achieve better outcomes than either could alone. Effective communication in human-agent teams ensures that both human and AI team members can understand each other, which is crucial for optimal performance and decision-making.

Trust plays an important role in HATs. McAllister (1995) defines trust as "the extent to which a person is confident in and willing to act on the basis of the words, actions, and decisions of another" [8]. Mutual trust in human-agent teams means that both humans and artificial agents rely on each other to do their tasks well, making teamwork more effective [9]. Since trust itself is a very wide definition and can be interpreted globally, it is important to distinguish between two kinds of trust. One of them is natural trust, which is the trust from the human towards the agent. The other is artificial trust. Artificial trust is the trust from the agent towards the human. [10].

Trust develops over time as teammates interact in a dynamic task environment, by collaborating and exchanging information to each other [11]. The artificial agent is able to decide its actions based on the behavior of the human. In other words, the actions of the human teammate impact the artificial trust of the agent. The agent may model the human capabilities and characteristics based on trust beliefs, e.g. competence and willingness [12]. These trust beliefs form the mental model of the artificial agent. The mental model decides how the agent interacts with the human, impacting the natural trust of the human [13].

The artificial agent communicating to the human impacts the natural trust. The communication of the mental model could be through textual explanations, visual explanations or verbally. There are already existing studies that have researched communication methods and strategies in Human-AI teams, such as Zhang et al. (2023) [14]. However, the depth of investigation into how these methods specifically influence natural trust remains limited, leading to a knowledge gap. This research is significant because enhancing trust and communication in human-AI teams lead to more effective collaboration [11, 14]. In this way, it is aimed to develop interactive agent technology that enables individuals and groups to face challenges.

This paper will extend this by explicitly focusing on communication through real-time textual explanations whereby the following research question will be answered:

> How does a real-time textual explanation of the agent's mental model of trust in the human teammate affect the human teammate's trust in the agent and overall satisfaction?.

The research question will be answered with the help of the following sub-questions:

- How does a real-time textual explanation of the agent's mental model of trust in the human teammate affect the teammate's trust in the agent?
- How does a real-time textual explanation of the agent's mental model of trust in the human teammate affect the teammate's satisfaction in the agent?
- What does the mental model of the artificial agent look like?

Although it does not directly answer the research question, this study will also examine whether the player's performance has decreased or increased by comparing the trust values with the baseline.

To answer the research question, a user study was conducted in a 2D grid environment, where the human and agent communicate with each other during a search and rescue game.

This research paper is structured as follows. Section 2 introduces the background of the experiment and explains concepts like Explainable AI, trust and Human-Agent Interaction. In Section 3, the mental model used in this study is explained, including the setup of the environment. Section 4 outlines the methodology and the involvement of the participants. Section 5 reports the results obtained after conducting the user study. As in Section 6, insights are provided into the responsible research. Section 7 discusses the results of the user study. Finally, the conclusion of the paper is provided in Section 8.

2 Background

2.1 Explainable AI

As AI systems become more essential to decision making across various domains, the need for transparency and interpretability has become crucial [15]. Transparency is the process of understanding how the AI systems make decisions [16]. E.g. understanding what kind of results they produce and the specific data that is used. The degree to which a cause and effect may be seen in a system is known as interpretability [17]. Explainable AI (XAI) describes strategies and tactics that help artificial intelligence systems decision-making processes become visible and intelligible to people [15, 18].

Artificial agents can effectively communicate with humans through real-time explanations. Neerincx et al. (2018) emphasize that real-time perceptual and cognitive explanations enhance collaboration by enabling quick adjustments during task execution [19].

Another way of communicating to humans is through textual explanations. Research showed that when artificial agents give clear reasons for why it makes decisions right away, people better understand how it works. This makes them more confident in the AI system and improves how well teams work together and perform [20].

2.2 Trust

Understanding trust in human-agent teamwork is vital for the successful integration of AI systems into various fields. As AI technologies become increasingly powerful, they are often designed to collaborate with humans, creating a need to understand and increase trust dynamics in these interactions. Trust forms the basis of effective collaboration, which influences decision-making, communication, and overall performance in human-agent teams [21]. By exploring trust more deeply, researchers and practitioners can create effective strategies to build trust in AI systems. This, in turn, will enhance their successful integration and positive impact across various fields.

According to Jorge et al. (2022) trust involves two parties in a dyadic relation. One of them is the trustor, who places trust in the other party. The other party is the trustee, who is responsible for upholding that trust. The trust between the trustor and trustee is dynamic and can change based on the actions and behavior of the trustee. Trust can be seen as the perceived trustworthiness, where trustworthiness is a property of the trustee [22].

Ali et al. (2022) mentions that trust in an agent is computed from an artificial trust model, where trust is assessed along a capability dimension [23]. Historical research has explored various trust models. One of them is Falcone et al. (2004), where they deconstruct the artificial trust in two beliefs regarding trustee's trustworthiness, namely competence and willingness [24]. Competence refers to the belief that an individual possesses the necessary skills and abilities to successfully perform a specific task or role, while willingness refers to the probability that an individual will execute a certain task [24]. This mental model using the competence and willingness beliefs decides how the robot behaves to the human in HATs.

2.3 Human-Agent Interaction

Human-agent interaction is the interdisciplinary study and design of the interaction between humans and agents, e.g. the interaction with a virtual assistant at an e-shop [25]. An important aspect of human-agent interaction is that there are tasks which humans are unable to perform, such as processing big data [26]. On the other hand, recognizing emotion remains a challenge for agents. Human-agent interaction provides an opportunity to bridge this gap and collaborate on tasks together [26].

Ulfert et al. (2020) mentions three factors influencing the trust in Human-agent interaction [27]. One of them are individual factors, such as trust in technology. Research shows that prior experience with technology has influences on individual factors. Gaining more experience with technology prevents the members from making system errors resulting in a stronger trust. Another factor is the team factor. When team members are perceived as similar and there is high interdependence, trust is likely to be higher, whereas low interdependence may hinder trust development. Lastly, system factors, including agent autonomy and reliability, impact team trust in human-agent teams. High autonomy and reliability favor trust, whereas system failures, especially with new agents, can decrease trust.

A study by Bussone et al. (2015) explores the role of explanations in human-agent interaction. They found that clear and contextually relevant explanations from the agent improved user trust. Additionally, feedback from users helped improve the artificial agents' performance, indicating a bidirectional trust-building process [28] (see Figure 1).

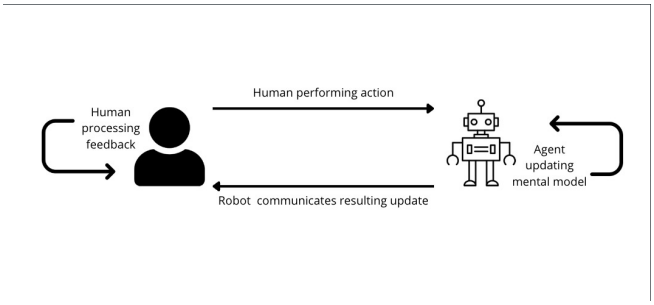

Figure 1: Communication loop between the human and artificial agent.

3 Trust Mechanism

3.1 Environment

The experiment was conducted in a 2D world search and rescue environment. The environment is implemented with the help of MATRX. The world consists of 10 areas, each area has its own unique scenario. The world consists of 6 victims and 7 obstacles (see Figure 2). The mission is to search and rescue the victims and drop them to the drop zone. During the game, two types of victims are distinguished: one being the mildly injured victim (yellow ones) and the critically injured victims (red ones). The mildly injured victims can be rescued by the human and RescueBot alone, so there is no

Figure 2: The environment of the map, with the obstacles and victims.

help needed. However, rescuing the mildly injured victims together is more time efficient since it costs less time to rescue. For the critically injured victims, both human and RescueBot need assistance from each other to carry the victim to the drop zone. During the mission, the human will encounter various obstacles such as rocks, stones and trees obstructing access to rooms housing potential victims. Trees (as in area 3) cannot be removed by humans. So the RescueBot has to remove it on his own. Rocks (as in area 1) cannot be removed alone. It can only be removed with the RescueBot and human together. Stones (as in area 5) can be removed by the human alone, by the RescueBot alone or by the human and the RescueBot together. Removing the stones together costs less removal time. Another environmental factor is the flooded area. Half of the map is a flooded area, which makes it less preferable for the player to go there, as the speed of the user decreases.

3.2 Mental Model

In this research, the core focus lies in the development and application of a mental model. This mental model is essential for guiding the actions and decisions of artificial agents in a dynamic and uncertain environment. This mental model is underpinned by a nuanced understanding of human beliefs, particularly centered around the concepts of competence and willingness [24].

Within the scope of the mission, which encompasses three primary tasks: search, rescue, and navigating or removing obstacles, our agent tracks both the competence and willingness associated with each of these tasks. The competence and willingness for these values are tracked in range of [-1, 1], where -1 denotes that the artificial agent distrust the human maximally, while 1 refers to the fact that the artificial agent trusts the human maximally for a certain task. The mental model used in this study updates the value based on a certain task executed by the human. This update could be a small, medium or a large update based on the importance of the task.

- A small update increases or decreases the trust value with 0.1
- A medium update increases or decreases the trust value with 0.2
- A large update increases or decreases the trust value with 0.4

In essence, the trust value dynamically adjusts according to the significance of the task completed by the human agent.

3.3 Behaviour Adaptation

The artificial agent adjusts its trust in the human's willingness and competence based on direct experiences and observed actions. This trust influences how the artificial agent responds to tasks, deciding whether to rely on the human or act independently. For instance, if the artificial agent finds a stone blocking its path, the artificial agent's trust determines if it will ask for help or just remove it alone.

Besides the trust values of competence and willingness the artificial agent tracks a confidence score, ranging from -1 to 1, which reflects its certainty in its trust beliefs. This score is updated based on recent trends in trust values: increasing with consistent trends and decreasing with inconsistent changes. The confidence score affects the likelihood of the agent trusting the human's declarations. A random value is sampled, and if it is less than the confidence score, the artificial agent further checks trust thresholds before deciding to trust the human. If the sample is higher, the artificial agent defaults to trusting the human. Lower confidence makes the artificial agent more likely to trust the human by default. In essence, instead of simply checking if willingness or competence meet a threshold, the agent first ensures it has confidence in its own trust beliefs.

3.4 Preference Modelling

The mental model used in this study includes a preference integration mechanism that adjusts the update of the willingness for the artificial agents based on environmental factors, distance and type of victims. For instance, when a task is located far away a difficulty factor is applied, ranging from 0 to 1. This factor mitigates the negative impact on the willingness value when the task is challenging due to the distance. Thus, the further the task, the less severe the penalty on these values, reflecting an understanding of the increased difficulty involved.

Similarly, the model differentiates between flooded and non-flooded areas in the environment. When the agent navigates through flooded regions, a difficulty factor is applied to account for the added challenges. This factor reduces the negative impact on the willingness, recognizing the higher difficulty of operating in such conditions. In contrast, in non-flooded areas, the agent perceives the tasks as easier and therefore expects the human to assist more readily.

Lastly, the model takes into consideration the type of victim involved. If it is a difficult victim, the model applies a factor that adjusts the agent's willingness, ensuring that the difficulty level is appropriately reflected in the agent's decision-making process.

Overall, this approach allows the artificial agents to make more informed and balanced decisions, taking into account the various complexities of the environment and the task.

4 Methodology

4.1 Design

To understand whether the real-time textual explanations do improve the overall trust and satisfaction of the human, an user study was conducted. In this experiment the communication model with the real-time textual explanations was compared with the baseline, which did not contain the communication of the mental model. No changes were made in the environment, mental model and time.

4.2 Participants

To conduct the user study, a total of 40 participants were recruited. All participants provided informed consent before taking part in the study, ensuring they were fully aware of the study's purpose, procedures, and any potential risks or benefits. No specific demographic, age, gender, or other personal details were targeted during recruitment, so the participants were selected randomly. All of the participants reside in Europe. 7 of the participants were woman, while the other 33 were men. 23 of the participants have a background related to a computer science field, while the other 17 had a different background. 19 participants have a lot of gaming experience, 12 reported some experience, 7 have little experience, while 2 had no experience in gaming. 26 participants aged between 18-24, 12 participants aged between 25-34, while 2 participants aged between 35-44. The participants were evenly split into two groups for the experiment. This distribution allowed for a comparison between the new model and the baseline, providing insights into the natural trust and satisfaction.

4.3 Agents

There are two agent added in the search and rescue mission. One of the agents is an autonomous rule-based virtual agent named RescueBot. During the game the RescueBot tracks which victims are rescued and which areas has been searched. If the agent finds an obstacle, it also mentions its removal time. Also for the mildly victims it mentions the extra rescue time if the human wants the RescueBot to rescue the victim alone. The other agent is a human agent which is controlled by the participant. The participants are able to control the human agent by using their keyboard. The participants were able to share their actions with the RescueBot through a chat by using buttons (see Figure 3). RescueBot incorporated the shared information into its memory and modified its behavior correspondingly, e.g. if the human mentions it already visited area 2, the RescueBot won't search area 2 even if it is the closest area. The agents could identify each other within a range of two grid cells, detect and clear obstacles or pick up victims within a single grid cell, and recognize walls and doors from any location. Both agents were able to see the flooded area from any location.

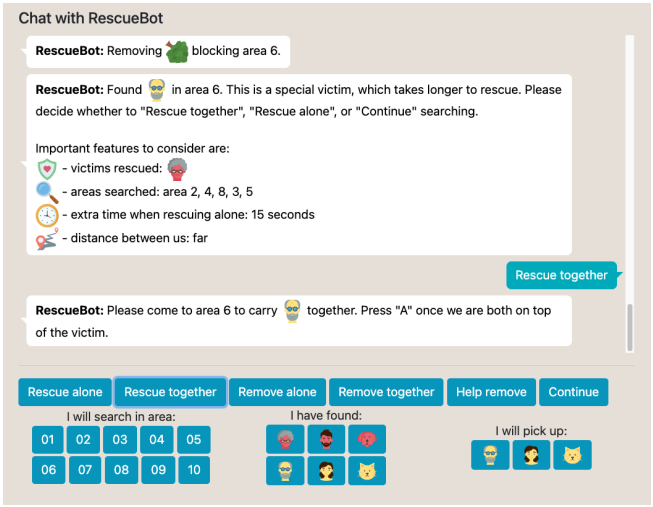

Figure 3: The chat box to communicate with RescueBot

4.4 Real-time Textual Explanations

This paper aims to answer the research question: How does a real-time textual explanation of the mental model of the agent's trust in the human teammate affect the human teammate's trust in the agent and overall satisfaction? In the baseline version the participants were playing the search and rescue mission without having the explanations of the agent's trust in the human teammate. In the communication model with real-time textual explanations, the RescueBot explained how the actions of the human impacted the artificial trust for tasks and provided reasons for its behavior.

The communication of the preference modeling mentioned in Section 3.4 is not taken into account when implementing the real-time textual explanations, even though it affects the artificial trust. Mayer et al. (1996) conducted some experiments based on communication methods, including textual, and came to a conclusion that information overload in texts should be avoided [29].

Since the mental model tracks the artificial trust of rescuing, searching and victims, the RescueBot specifically mentioned for which task the trust did increase or decrease. If the artificial agent asks the human to come over to remove an obstacle and the human comes over, instead of directly removing it together the RescueBot communicates that the trust for removing obstacles has increased. If the human lets the RescueBot know that it will search an area that the RescueBot did search before, the RescueBot will mention that the area has been searched by him before and his trust to the human regarding searching decreased. For victims the same principle is applied. Imagine that the human mentions it found a victim in e.g. area 4 and the victim is not there then the RescueBot will mention that the trust regarding victims decreased. These are just 3 examples of possible scenarios for searching, rescuing and removing obstacles. Table 1 provides an overview of the real-time textual explanations for various scenarios.

Action	Message of the RescueBot
Human mentions it searched an area, but the robot found a stone blocking the path	You mentioned that you had searched in area 1, but I found a stone blocking the way. I will now check the other area you claimed to have searched, as my trust in your searching has significantly decreased.
Human mentions that it found a victim in a area, without mentioning it searched for that area	Thanks for mentioning, however, since you did not mention that you were going to search that area, my trust regarding your searching decreased.
Human communicates that it is going to search a area	Thanks for mentioning, now I can search for other unsearched areas. My trust in you regarding searching has increased.
Human said it will come to remove an obstacle together, but did not come.	I decided to remove the stones blocking area 1, since you did not respond to me and it is important to remove obstacles blocking areas. My trust in regarding obstacles decreased.
Human asks for help removing an obstacle in area 1	Thanks for asking for help in area 1. My trust regarding obstacles increased.
Human responses to the RescueBot that it will come to remove an obstacle	It seems that you are willing to help removing obstacles.
RescueBot sees a stone or tree, while it does not trust the human regarding obstacles	Since I do not trust you with removing obstacles based on your previous actions, I decided to remove alone stones blocking area 1.
RescueBot finds a victim the human said it collected	Found the victim in area 1 although you said you collected it. My trust in you regarding victims decreased heavily.
Humans mentions it found a critical victim, while the victim is not there	Since you lied about finding a critical victim, my trust regarding victims decreased significantly. I consider critical victims as very important.
Human comes to rescue a victim, because the robot asked.	Thanks for coming over. My trust in you regarding victims increased.

Table 1: Overview of real-time textual explanations of the RescueBot

4.5 Tools and software

The experiments were conducted on a Apple Macbook Pro M1 to ensure that participants could effectively engage with the search and rescue environment. The environment is created with the help of Human-Agent Teaming Rapid Experimentation (MATRX) [1], which is a Python package designed for simulating and experimenting with environments where humans and artificial agents interact and collaborate. The main objective of MATRX is to facilitate the study of complex interactions and dynamics in such mixed teams to better understand and improve teamwork and decision-making processes.

4.6 Procedure

Before starting the experiment, participants were asked to read and sign the informed consent form. Information about the user was also collected before the game, including age, gender, gaming experience, expertise in computer science, region, and highest level of education. After the participants filled the form, they were randomly assigned to the baseline or the communication model containing the real-time textual explanations. To ensure they get familiar with the environment, they followed a tutorial to understand the game mechanism, chat box and controls. This tutorial environment was different from the actual environment, and the tutorial results were not considered in the results. The search and rescue mission was the same for both parties. However, the communication model with the real-time explanation contained information about how the agent thinks about the human and whether he can trust the human with certain tasks (search/rescue/victims). The participants had 10 minutes to rescue the six victims. The mission ends when the 10 minutes are over, or the 6 participants were rescued.

After the mission, the participants were asked to fill in a questionnaire regarding satisfaction and trust in the agent (natural trust). The questionnaire contained 15 questions (7 for satisfaction and 8 for trust). These results were used to analyze the data and compare the baseline results with the real-time textual explanations communication model.

4.7 Measures

To analyse the (natural) trust and satisfaction of the human towards the agent, objective and subjective data is collected. The subjective measures were collected throughout a questionnaire using Microsoft Forms, while the objective measures were logged after the search and rescue mission using MATRX.

Subjective Measures

For the measurement of trust and satisfaction, a 5-point Likert scale was used. The questionnaire was adopted from a previous study by Hoffman et al. (2023), as it provides assurance that trust and satisfaction are measured accurately [30]. The trust in RescueBot was measured using 8 questions. These questions assessed confidence in the RescueBot's performance, reliability, predictability, and efficiency. Participants evaluated their sense of safety and correctness when relying on RescueBot, as well as its ability to perform tasks better than a novice human user. Additionally, they consid-

[1] MATRX Software: https://matrx-software.com/

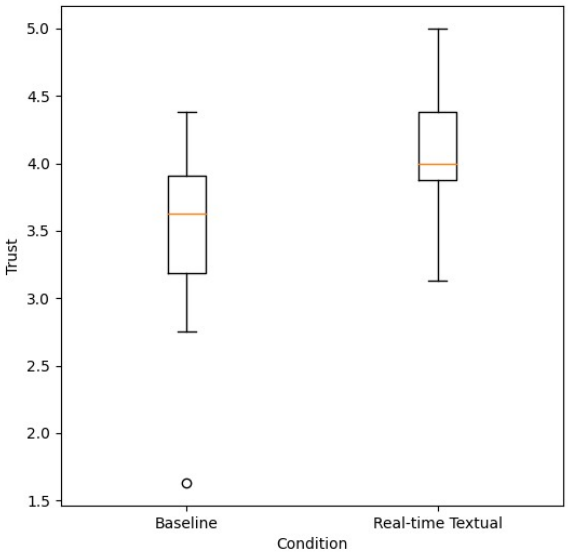

Figure 4: Box-plots for comparing natural trust between baseline and real-time textual explanation model.

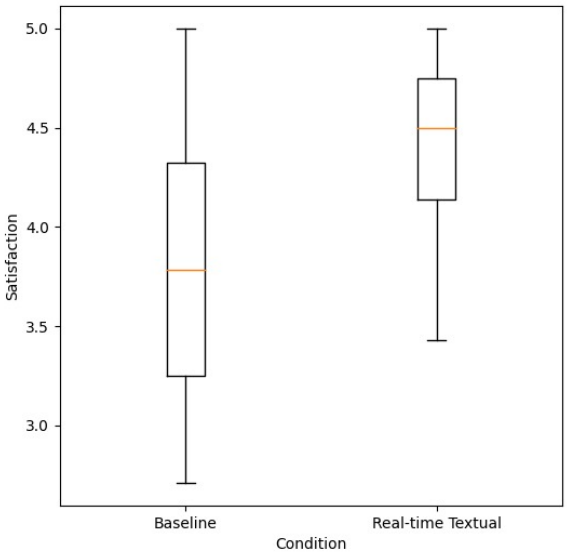

Figure 5: Box-plots for comparing satisfaction between baseline and real-time textual explanation model.

ered their overall comfort and preference for using RescueBot's guidance in decision-making.

The satisfaction with RescueBot was measured using 7 questions. Like the questions of measuring trust, the questions for satisfaction were also taken from the study of Hoffman et al. (2023) [30]. These questions focused on the quality of RescueBot's explanations regarding its functionality. Participants evaluated how well the explanations conveyed operational details, completeness, and usability. Additionally, the questions assessed whether the explanations were useful for achieving user goals and demonstrated the system's accuracy.

Additional open-ended questions were asked to the participants to gather their opinions on aspects they liked or disliked about the interaction. Participants were also asked what information they felt was missing from RescueBot, what they would like to see improved, and how they perceive RescueBot's opinion of them, including how that perception affects their feelings. This addition makes it clear that participants were given an opportunity to provide feedback on specific aspects related to their interaction with RescueBot.

Objective Measures

Objective data was collected to assess human performance in the game. In Human-Agent Teams (HATs), The presence of a continuous feedback loop significantly impacts player performance. This feedback loop ensures that players receive ongoing information and adapt their strategies based on interactions with the system, leading to diverse responses from the players.

Throughout the game, several performance indicators were logged. These were the time taken to complete the game, whether the game was successfully completed, the number of victims rescued and the trust values regarding obstacles, victims and searching. These measurements provided a comprehensive view of the player's effectiveness and efficiency during gameplay.

5 Results

5.1 Natural Trust

The average score for the natural trust of the baseline condition was 3.49 with a standard deviation (SD) of 1.082, while the average natural trust score for the real-time textual explanations condition was 4.09 with an SD of 0.907 (see Figure 4). We maintain assumptions on normality and equal variance by Shapiro-Wilk test on the baseline ($p = 0.078$, test statistic = 0.91) and experimental data ($p = 0.521$, test statistic = 0.96) and Levene's test ($p = 0.329$, test statistic = 0.98). The standard t-test indicated a significant difference in trust scores between the two conditions ($p = 0.001$, test statistic = -3.54)

5.2 Satisfaction

The average score for the satisfaction of the baseline condition was 3.79 with a standard deviation (SD) of 0.810, while the average natural trust score for the real-time textual explanations condition was 4.41 with an SD of 0.298 (see Figure 5). We maintain assumptions on normality and equal variance by Shapiro-Wilk test on the baseline ($p = 0.500$, test statistic = 0.96) and experimental data ($p = 0.184$, test statistic = 0.93) and Levene's test ($p = 0.092$, test statistic = 2.99). The standard t-test indicated a significant difference in trust scores between the two conditions ($p = 0.002$, test statistics = -3.306)

5.3 Artificial Trust

The average score for the artificial trust of the baseline condition was 0.75 with a standard deviation (SD) of 0.170, while the average natural trust score for the real-time textual explanations condition was 0.82 with an SD of 0.105. We maintain assumptions on normality and equal variance by Shapiro-Wilk test on the baseline ($p = 0.348$, test statistic = 0.95) and experimental data ($p = 0.461$, test statistic= 0.96) and Levene test ($p = 0.126$, test statistic = 2.45). The standard t-test did not indicate a significant difference in trust scores between the two conditions ($p = 0.188$, test statistic = -1.34)

6 Responsible Research

The user study described in this paper received approval from the Human Research Ethics Committee (HREC) at TU Delft. The approval ensures that the rights of the participants are protected. Before the experiments were carried out, potential risks were identified and measures to mitigate these risks were implemented. The participants were asked to fill in an informed consent form before the experiment. The data collected from the participants will be used for the purposes of this study, which includes analyzing, drawing conclusions, and contributing to the field. The personal data that is collected from the participant will not be shared beyond the study team. The participants were free to refuse to answer questions and withdraw from the study at any time without the obligation to give a reason.

Regarding reproducible research, this study adheres to the principles of openness and transparency. The methodologies and data are documented thoroughly to allow other researchers to replicate the study. Ensuring reproducibility is essential for validating results and advancing the field.

7 Discussion

7.1 Natural Trust

The results indicate a statistically significant difference in natural trust scores between conditions, with the mean trust score for the real-time textual explanation model (4.09) being higher than that of the baseline condition (3.49). This suggests that providing real-time textual explanations of the agent's trust in the human teammate significantly enhances the human's trust in the agent (natural trust).

These findings are consistent with existing research that emphasizes the importance of reliability, predictability, and transparency in building and maintaining trust in artificial agents [31, 32]. Studies have shown that when agents provide clear and predictable explanations for their actions, users are more likely to understand the agent's behavior, which enhances trust [31].

7.2 Satisfaction

The analysis of satisfaction scores from the study reveals a statistically significant difference between the baseline condition (mean = 3.79) and the real-time textual explanations (mean = 4.41) condition, demonstrating that the introduction of real-time explanations significantly enhances user satisfaction.

A reason for this difference could arise from the fact that the participants were not indented to lie towards the agent. Since the agent mentions that the artificial trust for a specific task increased in the real-time textual explanations, this makes the player of the game more satisfied. A study from Lavender et al. (2024) indicates that explanations provided by agents, whether positive or negative, can significantly affect user satisfaction [33].

By providing explanations, agents can engage users on an emotional level. Emotional engagement is crucial for satisfaction as it can make interactions feel more personal and less mechanical [34]. As adding the explanations of the mental model into the communication added more emotion to the output of RescueBot, this increased the satisfaction for the real-time textual explanation model.

7.3 Artificial Trust

The statistical analysis conducted on the artificial trust scores between the baseline condition and the real-time textual explanations condition revealed no significant difference. Therefore, this experiment does not provide evidence to support the findings of previous studies.

7.4 Limitations and Future Work

Due to the nature of the project, there are some limitations. One of the limitations is the small number of participants conducting the user study. The current experiment had 40 participants in total, where twenty of them were experimenting with the baseline and the other twenty the communication model with the real-time explanations. Having more participants in a user study improves the reliability, validity, and generalizability of the findings.

Another limitation of this study pertains to its demographic scope, primarily focused on participants residing in Europe. This geographical constraint may restrict the generalizability of the findings to populations outside of Europe, potentially limiting the study's applicability to diverse cultural contexts. As a result, the conclusions drawn from this research may not fully represent the experiences or perspectives of individuals from other regions of the world. Future research endeavors could benefit from including participants from a more geographically diverse range of locations to enhance the external validity of the findings. Henrich et al. (2010) discuss the limitations of research that predominantly samples from western societies, where many findings are based on samples that are not representative of the world global population, meaning that the results may not be generalizable to other cultural or geographical contexts [35].

In future research, it would be beneficial to explore alternative methods of presenting the agent's mental model of artificial trust through real-time textual explanations. The communication method mentioned in section 4.3 is not the only way of presenting textual explanation real-time. For instance, there could also be a third explanation model, which shows the updated trust values directly after an event or action, e.g. after the human removing an obstacle the RescueBot mentions the changes as competence_obstacles = 0.5 and willingness_obstacles = 0.5. Investigating multiple real-time textual explanation models, rather than relying on a single

approach, can provide better insights into their effectiveness and help identify the most effective strategies for enhancing natural trust and overall satisfaction.

8 Conclusion

The aim of this research was to investigate how real-time textual explanations of the agent's mental model of trust in the human teammate affect the human teammate's trust in the agent and overall satisfaction. The study focused on explaining how the mental model of trust in the human teammate is modeled, namely with competence and willingness values for the 3 main tasks: searching, rescuing and (removing) obstacles. The artificial agent also contained a preference integration, taking into account the difficulty of certain tasks. To answer the research question, a user study was conducted in a search and rescue mission, simulated in a 2D real-time environment. Every action that caused a change in the mental model of the artificial agent was communicated to the human through the chat box. This real-time explanation communication model was compared to a baseline model, where the changes were not communicated.

The findings of the study show that the participants were more satisfied and trusted the artificial agent more than without the real-time textual explanation. A possible reason for this is because the players intentions were good and they didn't try to lie causing positive outputs from the agent communicating the trust in the human.

In summary, this research is relevant to the field as it addresses critical aspects of transparency, trust, user satisfaction, and the design of a communicating method in AI systems. These contributions are essential for advancing the integration of AI into diverse areas of human-agent teams, ensuring that these systems are not only technically proficient but also trusted and accepted by their human counterparts.

References

[1] A. Bohr and K. Memarzadeh, "The rise of artificial intelligence in healthcare applications," pp. 25–60, 2020.

[2] D. E. O'Leary, "Artificial intelligence and big data," *IEEE intelligent systems*, vol. 28, no. 2, pp. 96–99, 2013.

[3] R. Zhang, N. J. McNeese, G. Freeman, and G. Musick, "" an ideal human" expectations of ai teammates in human-ai teaming," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1–25, 2021.

[4] D. Wang, J. D. Weisz, M. Muller, P. Ram, W. Geyer, C. Dugan, Y. Tausczik, H. Samulowitz, and A. Gray, "Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–24, 2019.

[5] Y. Hayashi and K. Wakabayashi, "Can ai become reliable source to support human decision making in a court scene?," pp. 195–198, 2017.

[6] J. H. Korteling, G. C. van de Boer-Visschedijk, R. A. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, "Human-versus artificial intelligence," *Frontiers in artificial intelligence*, vol. 4, p. 622364, 2021.

[7] C. Flathmann, B. G. Schelble, R. Zhang, and N. J. McNeese, "Modeling and guiding the creation of ethical human-ai teams," pp. 469–479, 2021.

[8] D. J. McAllister, "Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations," *Academy of management journal*, vol. 38, no. 1, pp. 24–59, 1995.

[9] A. C. Costa, R. A. Roe, and T. Taillieu, "Trust within teams: The relation with performance effectiveness," *European journal of work and organizational psychology*, vol. 10, no. 3, pp. 225–244, 2001.

[10] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human–robot collaboration," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 5913–5920, 2021.

[11] M. Demir, N. J. McNeese, J. C. Gorman, N. J. Cooke, C. W. Myers, and D. A. Grimm, "Exploration of teammate trust and interaction dynamics in human-autonomy teaming," *IEEE transactions on human-machine systems*, vol. 51, no. 6, pp. 696–705, 2021.

[12] M. Heyns and S. Rothmann, "Dimensionality of trust: An analysis of the relations between propensity, trustworthiness and trust," *SA Journal of Industrial Psychology*, vol. 41, no. 1, pp. 1–12, 2015.

[13] K. I. Gero, Z. Ashktorab, C. Dugan, Q. Pan, J. Johnson, W. Geyer, M. Ruiz, S. Miller, D. R. Millen, M. Campbell, *et al.*, "Mental models of ai agents in a cooperative game setting," pp. 1–12, 2020.

[14] R. Zhang, W. Duan, C. Flathmann, N. McNeese, G. Freeman, and A. Williams, "Investigating ai teammate communication strategies and their impact in human-ai teams for effective teamwork," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 1–31, 2023.

[15] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," pp. 1–19, 2021.

[16] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards transparency by design for artificial intelligence," *Science and engineering ethics*, vol. 26, no. 6, pp. 3333–3361, 2020.

[17] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[18] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," pp. 563–574, 2019.

[19] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, "Using perceptual and cognitive explanations

for enhanced human-agent team performance," pp. 204–214, 2018.

[20] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," pp. 263–274, 2019.

[21] H. J. Wilson and P. R. Daugherty, "Collaborative intelligence: Humans and ai are joining forces," *Harvard Business Review*, vol. 96, no. 4, pp. 114–123, 2018.

[22] C. C. Jorge, M. L. Tielman, and C. M. Jonker, "Artificial trust as a tool in human-ai teams," pp. 1155–1157, 2022.

[23] A. Ali, H. Azevedo-Sa, D. M. Tilbury, and L. P. Robert Jr, "Heterogeneous human–robot task allocation based on artificial trust," *Scientific Reports*, vol. 12, no. 1, p. 15304, 2022.

[24] R. Falcone and C. Castelfranchi, "A belief-based model of trust," pp. 306–343, 2004.

[25] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, p. 92, 2020.

[26] J. M. Bradshaw, P. J. Feltovich, and M. Johnson, "Human–agent interaction," pp. 283–300, 2017.

[27] A.-S. Ulfert and E. Georganta, "A model of team trust in human-agent teams," pp. 171–176, 2020.

[28] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," pp. 160–169, 2015.

[29] R. E. Mayer, W. Bove, A. Bryman, R. Mars, and L. Tapangco, "When less is more: Meaningful learning from visual and verbal summaries of science textbook lessons.," *Journal of educational psychology*, vol. 88, no. 1, p. 64, 1996.

[30] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance," *Frontiers in Computer Science*, vol. 5, p. 1096257, 2023.

[31] S. Daronnat, "Factors influencing trust, reliance, performance and cognitive workload in human-agent collaboration," 2021.

[32] S. Daronnat, L. Azzopardi, M. Halvey, and M. Dubiel, "Inferring trust from users' behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration," *Frontiers in Robotics and AI*, vol. 8, p. 642201, 2021.

[33] B. Lavender, S. Abuhaimed, and S. Sen, "Positive and negative explanation effects in human–agent teams," *AI and Ethics*, pp. 1–10, 2024.

[34] B. Biancardi, S. Dermouche, and C. Pelachaud, "Adaptation mechanisms in human–agent interaction: Effects on user's impressions and engagement," *Frontiers in Computer Science*, vol. 3, p. 696682, 2021.

[35] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?," *Behavioral and brain sciences*, vol. 33, no. 2-3, pp. 61–83, 2010.