Communicating trust-based beliefs and decisions in human-AI teams

The impact of a textual summary of changes of the artificial agent's mental model on the human teammate's trust in the agent and overall satisfaction

Răzvan Loghin[1]

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

EEMCS[1]**, 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: Răzvan Loghin
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Carolina Ferreira Gomes Centeio Jorge, Ujwal Gadiraju

Abstract

As artificial intelligence (AI) is increasingly integrated into decision-making processes, effective collaboration between humans and AI becomes crucial. This study investigates how textual summaries of changes of artificial agent's mental model affect human trust and overall satisfaction. Using a between-groups experimental design in an Urban Search and Rescue scenario, 56 participants were randomly assigned to either receive or not receive these summaries. Trust and satisfaction were measured through established scales and objective metrics, including the number of actions the human and AI performed together. Results show that providing textual summaries significantly increased both human trust in the AI agent and overall satisfaction. While the Task success rate improved with additional communication, other performance metrics showed no significant differences. This research contributes to understanding effective communication strategies in human-AI teams, highlighting the importance of transparency and justifications from artificial agents. These findings can help the design of collaborative AI systems, enhancing trust and satisfaction in human-AI partnerships.

1 Introduction

As the capabilities of autonomous systems continue to advance, artificial intelligence (AI) is increasingly playing central roles in decision-making processes across various sectors, including healthcare, aerospace, and industrial operations. These systems utilize sophisticated AI algorithms to enhance decision-making capabilities and operate at different levels of autonomy depending on the environment and context [1, 2].

The integration of AI systems alongside humans, known as Collaborative AI, aims to complement human abilities, leveraging the strengths and limitations of both to achieve more efficient solutions [1]. Collaborative AI systems are designed to understand human intentions, adapt to behaviors, and communicate effectively, thereby facilitating seamless interaction and cooperation [2].

Extending from the concept of Collaborative AI, the notion of Human AI teams (HAT) has emerged. At the core of HAT are mutual trust and transparency between human and AI collaborators. These factors contribute to HAT's superior performance compared to traditional human-only or AI-only teams [3, 4]. Human satisfaction is another crucial factor in team effectiveness. Recognized as a major component in established teams [5], satisfaction has been shown to positively affect various team characteristics, including performance [6]. In the context of HAT, both trust and satisfaction play key roles in shaping team dynamics and outcomes. Mutual trust, defined as the shared belief that the team members will fulfill their roles towards the common goal [7], is a product of both artificial and natural trust [8]. In this study, we differentiate between natural trust, which refers to human trust, and artificial trust, which designates the trust the robot forms in the human [9].

While most of the current trust models are applicable only for natural trust [9], this study also proposes a formalization of artificial trust. Equally important is how AI communicates its reasoning, intentions, and decisions back to human team members. These elements are key components of building trust in artificial systems. Effective communication strategies are essential for AI agents to be perceived as reliable team members [10, 11]. These strategies include the clarity and relevance of the information shared, as well as the timing and manner of its delivery, all of which profoundly affect human trust and team cohesion [12].

Current literature highlights the importance of transparency, explanations, and improved situational awareness in human-AI collaboration [3, 13]. Concepts such as explainable AI (XAI) emphasize the necessity for AI systems to be transparent and understandable to their human collaborators [14]. While the need for communication is widely acknowledged in HAT, there is a research gap regarding how specific characteristics of communication, such as timing and granularity of justifications provided by AI, affect the trust dynamics within the team [3]. Addressing this gap is crucial for enhancing trust and facilitating an efficient team dynamic, which is fundamental to the success of human-AI collaboration [15].

To bridge this gap, this research seeks to explore how a specific form of communication, namely a textual summary of changes of the mental model of an AI agent, affects the trust and satisfaction levels of the human teammates. The choice of a textual summary strikes a balance between visualization and timing. This method mitigates the risk of excessive communication that continuous communication offers, which distracts and overwhelms humans [16]. Additionally, textual representation helps avoid the misunderstandings that often arise from incorrect or ambiguous visual representations [17], making it potentially more effective for human-AI collaboration. Guided by these insights, this study poses the following research question:

> *"How does a textual summary of changes (justification) 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?"*

To address this primary question, two secondary questions have been formulated:

- *"How can a textual summary of changes (justification) of the mental model of an AI agent's trust in a human teammate be developed to effectively transmit artificial trust from AI to humans?"*
- *"What is the impact of the developed communication method on the natural trust and overall satisfaction?"*

This research aims to address the communication gap in human-AI teams by exploring how a textual summary of changes in an AI agent's mental model affects human trust and satisfaction. It focuses on formalizing artificial trust and developing a textual summary method to convey changes in the AI agent's mental model. This study makes use of a controlled environment inspired by an Urban Search and Rescue (USAR) scenario [18].

This research paper is structured as follows. Section 2 explains key concepts such as trust in human-AI collaboration, preferences of human agents, and explainable AI. Section 3 details the environment used for the study, specifically the USAR scenario. Section 4 outlines the trust mechanism, including the development and update of the mental model and behavior adaptation. Section 5 discusses the communication method, focusing on the textual summary of changes in the AI agent's mental model. Section 6 covers the study design, participants, hardware and software, task setup, procedure, and measurements. Section 7 presents findings from both subjective and objective measurements. Section 8 analyzes the results, limitations, and future work suggestions. Section 9 addresses responsible research considerations and Section 10 summarizes the key findings and their implications.

2 Background

2.1 Trust in Human-AI Collaboration

Trust is the fundamental element in collaborative environments, directing the success of both human-human and human-AI teams [3, 4]. Various definitions and conceptualizations of trust have been proposed, generally pointing to the trustor's willingness to be vulnerable to the trustee's actions, based on the expectation that the trustee will perform a particular action important to the trustor [19, 9]. In human-AI collaboration, trust is viewed as a dyadic relationship between humans and artificial agents, where both parties can assume the roles of trustor and trustee [8, 2, 20]. Trust is a central prerequisite for effective collaboration, influencing the acceptance of AI team members' suggestions and actions [21].

Researchers have developed various models to study and implement trust in human-AI teams [22, 21]. These models typically determine how much a human trusts an AI agent to perform a task, and the AI agent uses this estimate to predict human behavior [9]. Progress has been made in developing shared mental models (SMMs) that focus on aligning individual team members' understandings of their shared tasks and each other's roles [23, 24]. While the literature primarily focuses on human trust in AI, the concept of artificial trust has received less attention [20]. However, attempts have been made to map AI perceptions of trustworthiness [25] and formalize artificial trust by exploring how an AI agent can detect situations requiring trust or assess human trustworthiness [2, 4].

2.2 Conceptual Framework of Artificial Trust

Jorge, Tielman, and Jonker [20] made an effort to propose a conceptual framework for formalizing artificial trust. In their study, the mental model of trust is evaluated using two primary beliefs: competence and willingness. Competence refers to the assessment of the human agent's abilities and reliability in performing tasks effectively, while willingness pertains to the human's intention and motivation to execute the tasks. These beliefs are tied to internal features of an agent that influence how trustworthy they seem. These features, known as krypta, include ability, benevolence, and integrity [19]. Observable behaviors, or manifesta, provide cues for these features. For instance, performance metrics indicate competence, while favoritism or commitment reflect willingness. As such, the preferences of human agents can be modeled using the willingness belief.

2.3 Preferences of Human Agents

Continuing on the idea of preference modeling, in the context of HAT collaboration, incorporating the preferences of human agents is crucial for optimizing the performance and efficiency of task allocation and execution [26]. Human preferences are shaped by various factors such as the nature of the tasks, workload, and risk, all of which influence the decision-making process. Preferences can manifest in selecting tasks that require less effort for the same reward, adhering to the "law of least effort" [27]. By aligning their actions with human preferences, artificial agents can reduce the cognitive load on human teammates and minimize errors, thereby increasing trust within the team [26]. Consequently, the trust mechanism of artificial agents must be designed to dynamically incorporate human preferences, enhancing the overall synergy of Human-AI teams [28, 26]. Human preferences are often revealed through observable behaviors, which can be interpreted as cues indicating internal qualities such as competence, benevolence, and integrity [4].

2.4 Communication in Human-AI Collaboration

Effective communication and explanation of decisions made by robotic agents with human collaborators are crucial in Human-AI Teams (HAT), particularly in safety-critical applications where failures can have severe consequences [29, 14, 30]. Explainable AI (XAI) plays a vital role in making AI's decisions and behaviors more transparent and easier to interpret by humans, leading to increased trust and performance [29, 14, 12]. Existing literature explores various aspects of communicating AI's justifications and decisions back to human agents, such as modeling the impact of perceived trustworthiness on message quality and feedback [31], providing frameworks for textual justifications of AI's mental model and decision-making [15], and aligning explanations with end-users mental models and cognitive processes [32].

Research has compared the effectiveness of different explanation methods, such as visual and textual formats, and proposed combining them into hybrid approach to improve understanding while maintaining preference and ease of use [17]. Insights from social sciences advocate for user-centered approaches that prioritize the needs and understanding of end-users, considering factors such as expertise level and mental models, to improve the acceptance of AI systems [33]. By aligning explanations with users' requirements and capabilities, XAI can enhance trust, understanding, and performance in human-AI collaboration [32, 17, 33].

3 Environment

To effectively study trust dynamics in human-AI collaboration, a controlled yet realistic environment is crucial. This research utilizes a scenario inspired by Urban Search and Rescue (USAR)[18] set in a 2D grid world. Participants act as human agents paired with an artificial agent, RescueBot (see Figure 1). Players navigate the grid with visibility restricted to a 2-cell radius, keeping the locations of victims and obstacles initially unknown. RescueBot and the player only know each other's location within this radius. The terrain includes blue zones representing flooded areas, which slow down the movement of any agent crossing them.

Eight room entrances are obstructed by three types of obstacles, each with different interdependence levels for their removal. This way, the environment emphasizes the need for effective collaboration and communication between the human and the AI agent. Obstacles within the grid include big rocks, trees, and small stones. Big rocks necessitate collaboration between the agents, while trees can only be removed by RescueBot. Small stones can be removed by either agent, though the task is faster if both work together.

Additionally, six victims are scattered across the grid. These victims are of two types: critically injured (marked in red) and mildly injured (marked in yellow). Rescuing critically injured victims requires collaboration, as both agents must work together to save them. Mildly injured victims can be saved by either agent working alone or by both agents working together. However, when a single agent attempts to rescue a mildly injured victim, it takes more time than if both agents collaborated. Among the victims, the critically injured old woman and the mildly injured old man are referred to as special victims. They require extra time to rescue beyond the standard rescue duration, regardless of whether one or both agents are involved in their rescue.

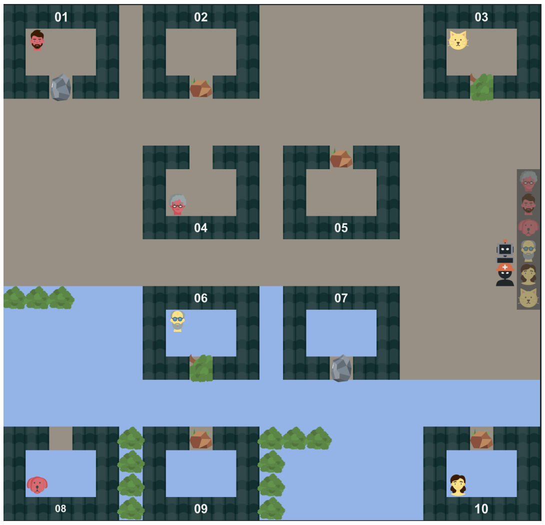


Figure 1: Overview of the game environment displaying all obstacles and victims.

4 Trust Mechanism

To introduce and evaluate the impact of the additional communication of the AI agent's mental model with the human teammate, a trust mechanism for the artificial agent has been developed. The following section details the trust model, its components, and how it influences the behavior adaptation of the agent.

4.1 Mental Model

While existing literature proposes many ways of modeling artificial trust, as specified in subsection 2.1, this study adapts the conceptual model proposed by Jorge, Tielman, and Jonker [20]. This framework centers on two key factors in assessing trust: competence and willingness.

In [2], Jorge et al. discuss the nature of the task and its impact when assessing a teammate's trustworthiness. They argue that trust is context-dependent [34, 2], and the type of task a collaborator is executing is one of the critical factors upon which a model can be built. Consequently, all actions that the pair of agents could perform either solely or together were divided into three distinct categories representing the main types of interactions a teammate could have with the environment: **Search**, **Remove**, and **Rescue**. The Search category includes actions related to exploring new areas for victims. The Remove category encompasses actions that deal with obstacles in the environment. Finally, the Rescue category refers to actions the teammates can perform concerning the victims present in the scenario.

Each of the three categories is represented by different trust values denoted by a tuple of competence and willingness as illustrated in the following equation:

$$\begin{aligned}\mathbf{T} &= \{T_{\text{Search}}, T_{\text{Remove}}, T_{\text{Rescue}}\} \\ &= \{(C_{\text{Search}}, W_{\text{Search}}), (C_{\text{Remove}}, W_{\text{Remove}}), \\ &\quad (C_{\text{Rescue}}, W_{\text{Rescue}})\}\end{aligned}$$

where C is a competence value and W is a willingness value.

During the collaboration with the human agent, all competence and willingness values remain in the range $[-1, 1]$. At the beginning of the scenario, the artificial trust starts with initial values of 0 for both willingness and competence, representing a neutral perception of trust toward the human collaborator. The values are dynamically updated after each action of the human teammate as perceived by the artificial agent. To account for the workload and criticality of different tasks, we introduced three different thresholds for updating the trustworthiness values: $\{\pm 0.1, \pm 0.2, \pm 0.4\}$. These thresholds, alongside preference modeling, are the direct factors influencing artificial trust, as will be discussed in subsection 4.3.

4.2 Preference Modelling

As discussed in subsection 2.3, preference integration exclusively updates the willingness values. To create a general preference model for a typical human user, we considered three factors: **Flooded areas, Special victims, Distance to the task**.

Flooded Areas

The scenario environment includes a flooded area that slows down user movement, as explained in Section 3. This creates a preference for tasks in non-flooded zones. The preference factor **f** can take values of 1 for tasks ending in non-flooded areas, 0.5 for tasks keeping the user in flooded areas, and 0 for tasks bringing the user into flooded areas (see subsection C.1 for the exact formula).

Special Victims

To adhere to the "law of least effort" [27], special victims were included in the scenario, that require additional time to rescue, making them less preferred compared to regular victims. The preference factor **s** takes a value of 1 if the victim is not special and 0 if the victim is special or if the task does not involve rescuing a victim (see subsection C.2 for the exact formula).

Distance to the Task

The third factor, **d**, represents the distance to the task. It aligns with human preferences for tasks requiring lower cognitive workload [27]. The distance factor is calculated based on the ratio of the team's distance to the task over the main diagonal of the environment (see subsection C.3 for the exact formula).

Combining Preference Factors
The overall preference score, **p**, combines these three factors, each with a weight. By default, the flooded and special victims factors have a weight of 1, while the distance factor has a weight of 2. For searching room actions, the distance factor is ignored to prevent unnatural trust increases, and the special victims' factor is only relevant for rescue tasks. The preference score ranges from 0 (least preferred) to 1 (most preferred) and is calculated as follows:

$$\mathbf{p(a)} = \frac{w_\text{f} \cdot \mathbf{f(a)} + w_\text{d} \cdot \mathbf{d(a)} + w_\text{v} \cdot \mathbf{v(a)}}{w_\text{f} + w_\text{d} + w_\text{v}}$$

Preference implementation involves adjusting the willingness of the human agent based on task preference. Performing less preferred tasks rewards willingness, while rejecting more preferred tasks penalizes it.

4.3 Updating the Perception of Trust

The developed trust model evaluates every decision made by the human agent concerning a task, considering the state of the environment at that moment. These decisions, referred to as actions, are categorized into *Search*, *Rescue*, and *Remove*. Based on the specific action, the model updates the competence and willingness values for the corresponding trust category. Updates use predefined thresholds $\{\pm 0.1, \pm 0.2, \pm 0.4\}$ or may remain unchanged ($\Delta = 0$). The workload and significance of the action determine which update value is applied. Additionally, actions influenced by the human preference model incorporate an extra factor in updating the willingness value. This preference update, represented by $p_U(a)$, depends on the preference score of action a. The human preference score, which ranges from [0,1], has to be normalized using a preference factor. Negative outcomes, like refusing to help, decrease willingness more with higher preference scores. Conversely, positive outcomes, such as helping, increase willingness more with lower preference scores. The detailed formula for this update can be found in subsection C.4.

The new trust value $T_c(\text{new})$ for a category c is described by the following formula:

$$T_c(\text{new}) = (C_c(\text{old}) + \Delta C,\ W_c(\text{old}) + \Delta W + p_U(a))$$

where ΔC and ΔW can take values from $\{\pm 0.1, \pm 0.2, \pm 0.4\}$, and c represents the action category: Search, Rescue or Remove.

The preference updates are particularly relevant in tasks like removing obstacles or rescuing victims, but also apply when searching new areas. The detailed updates for all possible actions from each task category are provided in Appendix A.

4.4 Behavior Adaptation

Building on the dynamically updated trust beliefs, the artificial agent adapts its behavior to enhance collaboration in Human-Agent Teamwork for USAR. These trust beliefs, including competence, willingness, and preference factors, guide the AI's decision-making and interactions with human teammates.

Integration of Confidence
Each time trust values are updated, the confidence value of the updated category is adjusted. This confidence represents how certain the artificial agent is about the human's trustworthiness regarding that specific task category.

The system updates confidence based on the last two trust beliefs, increasing it if the beliefs show a consistent trend (either increasing or decreasing) and decreasing it if the beliefs are inconsistent. If fewer than two beliefs are recorded, confidence is not updated. Confidence is then clipped within the range [0, 1]. The detailed formula for this process is provided in subsection C.5.

General Adaptations
When the AI is not restricted by interdependence with the human, it evaluates whether to trust the human for a specific task. This evaluation is based on the trustworthiness levels for that particular task category and the human's preference score for the task. This decision is probabilistic, comparing a uniform random sample between 0 and 1 to the current confidence level for the task. The agent checks if both willingness and competence values meet their thresholds, adjusted by the preference score. Specifically, the willingness threshold is dynamically adjusted by subtracting a fraction of the preference score. If the sample is below the confidence level and both values exceed their thresholds, the action is trusted; otherwise, it is not. If the sample is above the confidence level, the action defaults to being trusted. This process is reduced to a mathematical formula that can be found in subsection C.6.

The AI adapts its behavior in specific scenarios to ensure task completion when the human agent is deemed untrustworthy. If the agent perceives the human as unreliable, it will adapt its behavior accordingly. For instance, it may proceed autonomously with tasks it can complete independently, such as rescuing mildly injured victims or removing trees and small stones, without waiting for human intervention.

5 Communication

In the scope of the experimental condition, the textual summary of changes of the AI agent's mental model in the human teammate was developed. This communication design employs a user-centered approach to explanation, generating tailored explanations specifically for the human teammate. Such an perspective to explanation can significantly enhance trust and satisfaction compared to explanations that do not consider the user's perspective [32].

Upon generation, the summary appears centrally on the screen, pausing the game until the participant closes it. The information is presented across three different screens, which participants must review before closing the communication and resuming the game. Three summaries are provided throughout the game, balancing the need for effective AI communication between efficiency and sociability [11]. The summaries' appearance is based on a logic considering the number of victims rescued and the current progress relative to the maximum game duration, ensuring all users receive all summaries irrespective of skill level, as illustrated in Figure 2.

The following subsections outline the information provided to the human agent.

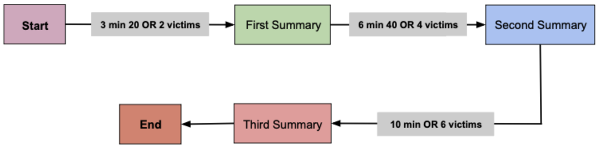
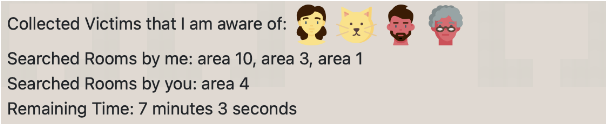


Figure 2: Logic flow of the generation of summaries

5.1 Status Update

This section offers a general overview of the current game state as perceived by the AI agent. It includes details on the victims collected by the AI, rooms searched by the AI, rooms known to be searched by the human, and the remaining time until the game's 10-minute limit. A visual example can be seen in Figure 3.



Collected Victims that I am aware of: 👧 🐱 🧔 👵

Searched Rooms by me: area 10, area 3, area 1

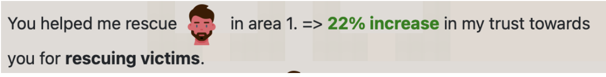
Searched Rooms by you: area 4

Remaining Time: 7 minutes 3 seconds

Figure 3: Example of status update content

5.2 Actions Impact on Trust

This section lists all actions taken by the human agent from the start of the game or since the last summary. Actions are categorized into three task types and explained clearly. For each action, the impact on the trust value within its specific category is presented as a percentage, making it easier for users to understand. Figure 4 illustrates this with a visual example.

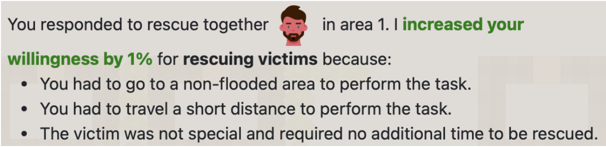


You helped me rescue 🧔 in area 1. => **22% increase** in my trust towards you for **rescuing victims**.

Figure 4: Example of an action's impact on trust

5.3 Justification of Human Preferences

Here, actions performed by the human, as perceived by the AI, are linked to the human preference model. Each action is described alongside a percentage representing the change in willingness resulting from the preference update. The state of the three preference factors at the time of each action is also presented as additional justification. A visual example can be seen in Figure 5.



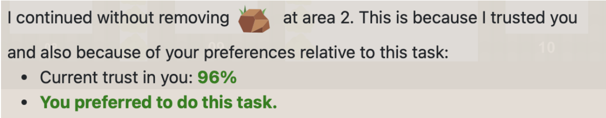
You responded to rescue together 🧔 in area 1. I **increased your willingness by 1%** for **rescuing victims** because:

- You had to go to a non-flooded area to perform the task.
- You had to travel a short distance to perform the task.
- The victim was not special and required no additional time to be rescued.

Figure 5: Example of justification of human preference

5.4 Justification of Robot's Actions

This section lists all decisions made by the artificial agent that were influenced by the behavior adaptation logic. The AI's decisions to trust or not trust the human teammate are justified based on the level of trust in the human at the time of their action and the preference score explained in textual format. A visual example can be seen in Figure 6.



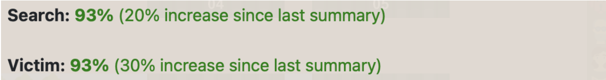
I continued without removing 🪨 at area 2. This is because I trusted you and also because of your preferences relative to this task:

- Current trust in you: **96%**
- **You preferred to do this task.**

Figure 6: Example of justification of robot's decision

5.5 Trust Levels

The final section presents the current levels of trust and confidence at the time of summary generation, classified by the three task types explained in Section 4.1, as well as the percentage change since the last summary was generated. A visual example can be seen in Figure 7.



Search: 93% (20% increase since last summary)

Victim: 93% (30% increase since last summary)

Figure 7: Example of trust level from the communication

The full visualization of the developed textual summary of changes of AI's mental model can be found in Appendix D.

6 Methods

To address the research question, a between-groups user study was conducted to investigate the impact of textual summaries of changes (justifications) of the mental model of an AI agent's trust in a human teammate on the human teammate's trust and satisfaction. The two proposed hypotheses were as follows:

H1: Additional communication of the mental model of the AI agent to the human agent increases the human agent's trust in the AI agent.

H2: Additional communication of the mental model of the AI agent to the human agent increases the human agent's satisfaction.

The user study employed an independent measures experimental design, with participants randomly assigned to one of two experimental conditions: one group played the game with the textual summary of changes of the AI's mental model of trust in the human teammate (Communication group), and the other group played the game without this additional communication (Baseline group). The independent variable was the added communication method, and the dependent variables were the trust and overall satisfaction of the human teammates after completing the game.

6.1 Participants

The study recruited 56 participants, all of whom reported residing in Europe. The sample consisted of 17 females, 36 males, 1 non-binary, and 2 unspecified individuals. They were evenly split into two conditions of 28 participants each. Age distribution was predominantly 18-24 years (49 participants), with 6 participants aged 25-34, and 1 participant aged 35-44. Educational background included 21 with high school diplomas, 30 with Bachelor's degrees, 3 with Master's degrees, and 2 with HBO school diplomas. Regarding gaming experience, 26 participants reported extensive experience, 17 had some experience, 9 had little experience, and 4 had no experience.

6.2 Hardware and Software

The experiments were conducted on multiple laptops. The experimental group, with the additional communication, used a MacBook with macOS and an M1 processor. In the Baseline group, 10 participants used Windows laptops, while the rest used macOS laptops. The game environment was built using the Human-Agent Teaming Rapid Experimentation software (MATRX)[1]. Twelve participants reported prior experience with MATRX software before participating in the study.

6.3 Task

The game's primary goal is to collaborate with RescueBot to rescue six victims scattered across the grid as quickly as possible. Players navigate the grid, communicate with RescueBot via a chat interface with predefined action buttons, and coordinate to complete various tasks. Some tasks require both agents, while others can only be completed by RescueBot, necessitating human agents to request help. Likewise, RescueBot may seek help from the human player. Effective communication is essential for task management, such as informing each other about searched rooms to avoid redundancy. The game has a 10-minute limit, ending when all victims are rescued or time expires.

6.4 Procedure

The procedure lasted about 30 minutes per participant. Participants first read and signed the informed consent and HREC checklist forms, then completed an anonymized survey to collect potential confounding factors such as age group, region of residence, gender, highest education level, whether they major in Computer Science, prior knowledge of MATRX, and gaming experience (see subsection 6.1 for results).

Participants were then assigned to one of two experimental conditions, followed by a game tutorial to familiarize them with the rules and objectives. They also received a scripted explanation of the trust mechanism in RescueBot. Those in the experimental group also received a detailed explanation of the communication method's contents and logic. In the baseline condition, participants collaborated with RescueBot during the search and rescue mission until one of the two ending conditions was met. In the experimental group, participants additionally read through the communication method three times during the game (see Section 5 for details). After completing the mission, participants filled out an anonymized questionnaire focusing on the two dependent variables.

6.5 Measurements

To assess the influence of the additional textual summary on natural trust and participant satisfaction, both subjective and objective measures were used. Subjective measures capture personal experiences, perceptions, and satisfaction levels, essential for understanding psychological impact [35]. Objective measures help mitigate biases from subjective measures, providing more useful insights [36]. Objective data was logged using MATRX software, while subjective data was collected via Microsoft Forms[2].

Subjective Measurements

Participants' trust and satisfaction towards the AI system were assessed using established scales from Hoffman et al. [37], presented in a Likert format.

Trust was measured with an adapted version of the trust scale from *'TABLE 8 The trust scale for the XAI context.'* [37]. This scale includes eight items, each rated on a 5-point Likert scale, assessing facets like confidence, predictability, reliability, safety, efficiency, wariness, and likability. The full scale is in subsection B.1.

Satisfaction was measured with the scale from *'TABLE 3 The explanation satisfaction scale.'* [37], consisting of seven items rated on a 5-point Likert scale. These items evaluate understanding, satisfaction, detail sufficiency, completeness, usability, usefulness, and perceived accuracy of the AI explanations. The full scale is in subsection B.2.

Additionally, the questionnaire included four optional open-ended questions to gather qualitative data. These questions asked about missing information, liked and disliked aspects of collaboration with RescueBot, and participants' perceptions of how the AI viewed them, providing deeper context and feedback on their interactions with the AI system.

Objective Measurements

To complement subjective measures, we used several objective metrics to assess trust and satisfaction levels:

Compliance and Communication Rate: Compliance, the number of times participants follow the AI's recommendations, correlates with trust levels, with higher compliance suggesting greater trust [38]. The communication rate tracks the number of messages sent by the human agent. Increased communication often correlates with higher trust levels, indicating greater engagement and willingness to interact with the AI [38].

Task Success Rate and Interaction Frequency: The task success rate, measuring the completeness of the joint objective of the HAT on a scale of $[0, 1]$, indicates reliability and effectiveness, correlating with higher trust and satisfaction [39]. Interaction frequency, the number of joint actions between the human and AI, signifies greater engagement, trust, and satisfaction, showing users' confidence and compatibility with the system [39].

Finally, **task completion time** records the duration taken to complete tasks. Shorter completion times correlate with higher satisfaction [39], because, as discussed by Nielsen [40], efficient task completion suggests users find the system reliable and easy to interact with, contributing to higher trust and satisfaction levels.

7 Results

7.1 Subjective Results

The Likert scale data from the questionnaire was translated to a numerical scale from 1 to 5, where 1 indicates "Strongly disagree" and 5 indicates "Strongly agree." Item 6 in the trust scale ("I am wary of the RescueBot") was reverse-scored due

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

[2] Microsoft Forms: https://forms.microsoft.com

to the negative formulation, as such ensuring higher scores consistently represent higher levels of trust across all items.

Reliability Consistency

Cronbach's alpha was used to assess the internal consistency of the trust and satisfaction scales, with a threshold of $\alpha = 0.7$ based on Tavakol and Dennick [41]. The Baseline group showed acceptable reliability for both scales, with Cronbach's alphas of 0.719 for trust and 0.896 for satisfaction. However, the Communication group had lower reliability, with a trust survey alpha of 0.476, indicating issues with internal consistency discussed further in the subsection 8.3.

Hypothesis Testing

Composite scores for trust and satisfaction were computed by averaging the respective item scores for each participant. Due to the negative results of the Shapiro-Wilk tests for normality, the Mann-Whitney U test was chosen for both trust and satisfaction composite scores. The significance threshold was set to $p = 0.05$, meaning that results with p-value less than 0.05 is considered significant. The Mann-Whitney U test indicated a significant difference in trust composite scores between the Baseline group (M = 3.563, SD = 0.64) and the Communication group (M = 4.25, SD = 0.427), $U = 185.0$, $p = 0.0012$. This result suggests that the textual summary of artificial trust significantly increases human trust in the AI agent, supporting **Hypothesis 1**.

Similarly, the Mann-Whitney U test for satisfaction composite scores revealed a significant difference between the Baseline group (M = 3.786, SD = 0.899) and the Communication group (M = 4.571, SD = 0.436), $U = 177.0$, $p = 0.0007$. This finding supports **Hypothesis 2**, indicating that the communication of the AI's mental model significantly enhances overall human satisfaction.

Figure 8 presents box-plots of the trust and satisfaction composite scores for both groups, illustrating the differences identified by the Mann-Whitney U tests.

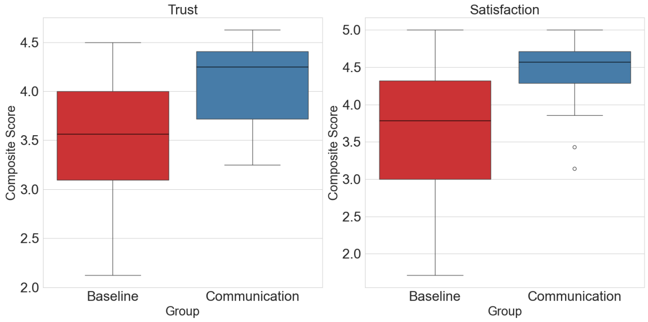


Figure 8: Box-plots for comparing the composite scores for trust and satisfaction

7.2 Objective Results

To compare the Baseline and Communication groups across the measurements specified in Section 6.5, a systematic approach was used. First, the normality of each measurement was assessed with the Shapiro-Wilk test. Depending on the results, either the Independent samples t-test or the Mann-Whitney U test was applied to identify significant differences between the groups. The findings are summarized in Table 1, showing the means and standard deviations for both groups, the statistical test used, and the corresponding p-values. An asterisk (*) marks measurements with significant differences, which, in this case, is only the Task success rate.

8 Discussion

8.1 Trust and Satisfaction

The results show a significant difference in natural trust and satisfaction between the experiment groups. Participants in the Communication group reported higher levels of trust and satisfaction compared to the Baseline group, aligning with prior research highlighting the importance of transparency and explanations in human-AI collaboration [3, 13].

The significant increase in trust and satisfaction in the Communication group can be attributed to the transparency and explanations provided by the textual summary of changes. Transparency helps calibrate human's trust in robots, especially when behavior is unexpected or reliability varies [3, 13]. Clear updates enhance the human teammate's situation awareness and trust in the AI's capabilities [42, 31]. Providing explanations about the robot's behavior, particularly for unexpected actions, improves understanding and trust [3].

Detailed justifications and updates help users better calibrate their trust in the AI [3, 9, 31]. Communicating task outcomes and providing explanations enhances understanding and trust [9, 16]. Attributing trust changes to specific topics offers a fine-grained explanation that helps maintain trust even if the AI's trust varies in some areas [31].

In addition to the existing literature, qualitative feedback from the participants provides further insight into the significant differences in natural trust and satisfaction observed between the groups. Participants in the Baseline group reported a lack of guidance and clarity, leading to confusion and frequent need for explanations from the AI. This sentiment is exemplified by Participant 18's comment: "I have no idea how the rescue bot thinks of me. I would have liked to know more about whether the robot trusts me or not, so I can better base my decisions."

Conversely, participants in the Communication group appreciated the additional information and regular updates, which they felt improved alignment with the artificial agent and team collaboration. A particularly valued aspect of the communication was the updates on the levels of artificial trust (see subsection 5.5). Participant 11 noted, "It also makes me want to perform better, when seeing I am not trustworthy enough." Similarly, Participant 18 from the Communication group expressed satisfaction with the textual indicators: "The decreases were highlighted in red and I could see exactly what I did wrong quickly so that I could correct it."

8.2 Performance Metrics

From the objective metrics, only the Task completeness rate showed a significant difference between the experimental groups. Previous research indicates that a higher Task success rate correlates with increased trust and satisfaction [39]. However, the Task completeness is influenced by various factors, including software performance. During pilot runs, the team observed slower performance on Windows OS compared to macOS due to less optimized file handling on Win-

Table 1: Comparison of Baseline and Communication groups across objective measurements.

Measurement	Test	Baseline Mean	Baseline SD	Comm. Mean	Comm. SD	p-value
Compliance	Mann-Whitney U	3.000	1.388	2.464	1.753	0.097
Ratio of joint actions	t-test	209.536	66.073	189.893	49.228	0.213
No. of human messages	Mann-Whitney U	19.857	6.422	20.214	3.645	0.347
Task success rate*	Mann-Whitney U	0.911	0.184	1.000	0.000	0.0027
Total task time (ticks)	t-test	4813.571	607.418	4508.036	544.493	0.053

dows. This is relevant as 10 Baseline participants used Windows devices, while none in the experimental group did. This OS disparity may have affected Task completeness rates, potentially confounding interpretation of this metric as a trust and satisfaction indicator.

The remaining metrics considered in the user study, compliance, communication rate, interaction frequency, and task completion time did not show significant differences between groups. This may be because these metrics are related to individual user performance and personal skills, which could have influenced the results more than the added communication method. Trust and satisfaction in HAT are influenced by many factors beyond the measured metrics, such as individual user experiences, expectations, and perceptions of AI capabilities [34, 25]. Additionally, objective measures may not capture the nuances in human-robot interactions as effectively as subjective metrics [34, 35].

8.3 Limitations

This study has several limitations that should be taken into account when interpreting the results. Firstly, the Baseline group consisted of an equally divided pool of participants, both those who majored in Computer Science and those who did not. In contrast, the Communication group had 82% of its participants majoring in Computer Science. This disparity could influence the results, as individuals with a Computer Science background are generally more familiar with the technology used in the experiment than those from other fields. Apart from the affinity with Computer Science, other confounding variables such as familiarity with the MATRX software used to develop the USAR scenario and overall gaming experience might affect participants' views on AI collaboration and their performance. While these variables were recorded, the study did not analyze their impact on the dependent variables measured. Therefore, the homogeneity of participants might affect the validity of the results.

Secondly, the trust survey in the Communication group had a low Cronbach's alpha of 0.476, indicating poor internal consistency. Items 6 and 7 caused most of this inconsistency, as removing them raised the alpha to 0.695. This low score may be due to participants misunderstanding terms like 'wary' and 'novice,' resulting in high response variance. However, both the trust and satisfaction scales were adapted from Hoffman et al. (2023), suggesting their reliability had been previously validated.

8.4 Future Work

Future research should include a larger, more diverse participant pool to mitigate the influence of confounding variables and obtain a comprehensive perspective on human dynamics in HAT. Additionally, future studies should explore the correlation between gaming experience, software familiarity, and reported trust and satisfaction levels. Understanding these correlations can help isolate their effects, providing a clearer picture of the impact of different communication methods.

Improving the trust survey's reliability is crucial. Selecting more comprehensive and cohesive scales assessing both trust and satisfaction can provide additional validation for results. Identifying other game logs correlating the communication method with natural trust and satisfaction could be valuable for future studies.

Finally, comparing the effects of the textual summary of changes in the AI agent's mental model on natural trust and satisfaction with other communication methods can provide a direct comparison of XAI communication methods. Such research could extend the questions tackled in this study, offering deeper insights into the most effective communication strategies to implement in HAT.

9 Responsible Research

When discussing the integrity of this research, two crucial topics need to be addressed: reproducibility, and ethical considerations.

First, reproducibility is a vital aspect of this study that must be guaranteed. To achieve this, the code base developed and utilized for this research is stored in a repository[3] with restricted access, available to authorized researchers and reviewers. Detailed methodologies and parameters are provided in this paper to support broader reproducibility efforts. Since the study incorporates objective measures recorded through game logs, consistency in data collection was key. Thorough reviews were conducted to ensure all researchers used the same code script to obtain consistent and unaltered objective measures.

To enhance reproducibility, this research provides automated data analysis through several Jupyter notebooks[4]. These notebooks allow anyone to replicate the results using the same data set. The analysis includes data from all 56 participants: 28 from the group receiving textual summaries of the artificial agent's mental model changes and 28 from the Baseline group. This data set, along with data from related studies exploring different communication types, is available through the 4TU.ResearchData[5] platform.

Secondly, ethical concerns arise due to the carrying out of a user study. The primary issues involve data processing and

[3] GitLab repository used for this project: GitLab Repository

[4] Project Jupyter: https://jupyter.org

[5] 4TU.ResearchData data repository: https://data.4tu.nl

privacy. To address these matters, a risk assessment was conducted using the "Ethics review checklist" proposed by the Human Research Ethics Committee of TU Delft. The identified concerns were the collection, processing, and storage of directly identifiable Personally Identifiable Information (PII) and Personally Identifiable Research Data (PIRD). To mitigate these risks, names, and signatures related to PII were collected in separate informed consent forms, accessible only to the research team. Regarding PIRD, the study collected and stored participants' age group, gender, region, education, computer science background, gaming expertise, and experience with MATRX software. However, this data was anonymized and used solely to describe samples and may be used in future work to identify correlations with confounding variables. These risk mitigation techniques were approved by the Human Research Ethics Committee (HREC) at TU Delft. Subjective measures and personal data were collected using Microsoft Forms[6], a survey tool known for its compliance with GDPR laws.

10 Conclusion

This research aimed to investigate how a textual summary of changes of the mental model of an AI agent's trust in a human teammate affects the human teammate's trust in the AI agent and overall satisfaction. The findings demonstrate that the communication method significantly influences both trust and satisfaction levels, providing valuable insights into the dynamics of human-AI collaboration.

The study revealed that textual summaries significantly increased the human teammate's trust in the AI agent. Transparency and explanations provided by these summaries allowed participants to understand the AI's decision-making and trust calibration better. Additionally, overall satisfaction improved significantly among participants who received the summaries. The detailed updates and justifications made participants feel more informed and aligned with the AI's actions, leading to higher satisfaction levels. While the Task success rate was notably higher in the group with additional communication, other metrics like compliance, communication rate, interaction frequency, and task completion time showed no significant differences. This suggests that while communication boosts trust and satisfaction, its direct impact on performance metrics may depend on factors such as individual skills and task familiarity.

This study contributes to the understanding of effective communication strategies in HAT. It highlights the importance of human-centered communication methods in building trust and satisfaction, emphasizing transparency and justifications from AI agents. These findings can guide the design of collaborative AI systems, promoting transparency and justification in AI interactions.

[6] Microsoft Forms: https://forms.microsoft.com

References

[1] Carolina Centeio Jorge, Myrthe L Tielman, and Catholijn M Jonker. “Artificial trust as a tool in human-AI teams”. In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2022, pp. 1155–1157.

[2] Carolina Centeio Jorge et al. “Appropriate context-dependent artificial trust in human-machine teamwork”. In: *Putting AI in the Critical Loop*. Elsevier, 2024, pp. 41–60.

[3] Joseph B Lyons. “Being transparent about transparency: A model for human-robot interaction”. In: *2013 AAAI Spring Symposium Series*. 2013.

[4] Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. “How should an AI trust its human teammates? Exploring possible cues of artificial trust”. In: *ACM Transactions on Interactive Intelligent Systems* 14.1 (2024), pp. 1–26.

[5] Deborah L Gladstein. “Groups in context: A model of task group effectiveness”. In: *Administrative science quarterly* (1984), pp. 499–517.

[6] Sami Abuhaimed and Sandip Sen. “Human Satisfaction in Ad Hoc Human-Agent Teams”. In: *International Conference on Human-Computer Interaction*. Springer. 2023, pp. 207–219.

[7] Eduardo Salas, Dana E Sims, and C Shawn Burke. “Is there a “big five” in teamwork?” In: *Small group research* 36.5 (2005), pp. 555–599.

[8] C Centeio Jorge et al. “Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams”. In: *22nd International Trust Workshop 2021*. 2021.

[9] Hebert Azevedo-Sa et al. “A unified bi-directional model for natural and artificial trust in human–robot collaboration”. In: *IEEE robotics and automation letters* 6.3 (2021), pp. 5913–5920.

[10] Hebert Azevedo-Sa et al. “A Unified Bi-Directional Model for Natural and Artificial Trust in Human–Robot Collaboration”. In: *IEEE Robotics and Automation Letters* 6.3 (2021), pp. 5913–5920. DOI: 10.1109/LRA.2021.3088082.

[11] Rui Zhang et al. “Investigating AI teammate communication strategies and their impact in human-AI teams for effective teamwork”. In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (2023), pp. 1–31.

[12] Matthew B Luebbers et al. “Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming”. In: *Proceedings of Robotics: Science and Systems. Daegu, Republic of Korea. https://doi.org/10.15607/RSS* (2023).

[13] Anthony R Selkowitz et al. “Displaying information to support transparency for autonomous platforms”. In: *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA*. Springer. 2016, pp. 161–173.

[14] David Gunning et al. “XAI—Explainable artificial intelligence”. In: *Science robotics* 4.37 (2019), eaay7120.

[15] Jessie Y Chen et al. “Situation awareness-based agent transparency”. In: *US Army Research Laboratory* (2014), pp. 1–29.

[16] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. “Trusting artificial agents: Communication trumps performance”. In: *AAMAS 2023*. 2023.

[17] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. “Visual, textual or hybrid: the effect of user expertise on different explanations”. In: *26th international conference on intelligent user interfaces*. 2021, pp. 109–119.

[18] Joseph A Barbera and Anthony Macintyre. “Urban search and rescue”. In: *Emergency Medicine Clinics* 14.2 (1996), pp. 399–412.

[19] Roger C Mayer, James H Davis, and F David Schoorman. “An integrative model of organizational trust”. In: *Academy of management review* 20.3 (1995), pp. 709–734.

[20] 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*. 2022, pp. 1–3.

[21] Anna-Sophie Ulfert et al. “Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework”. In: *European Journal of Work and Organizational Psychology* 33.2 (2024), pp. 158–171.

[22] Kristin E Oleson et al. “Trust in unmanned aerial systems: A synthetic, distributed trust model”. In: *16th International Symposium on Aviation Psychology*. 2011, p. 469.

[23] Robert W Andrews et al. “The role of shared mental models in human-AI teams: a theoretical review”. In: *Theoretical Issues in Ergonomics Science* 24.2 (2023), pp. 129–175.

[24] Beau G Schelble et al. “Let’s think together! Assessing shared mental models, performance, and trust in human-agent teams”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.GROUP (2022), pp. 1–29.

[25] Melanie J McGrath et al. “Collaborative human-AI trust (CHAI-T): A process framework for active management of trust in human-AI collaboration”. In: *arXiv preprint arXiv:2404.01615* (2024).

[26] Thibaut Munzer, Marc Toussaint, and Manuel Lopes. “Preference learning on the execution of collaborative human-robot tasks”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 879–885.

[27] Raymond Wu, Amanda M Ferguson, and Michael Inzlicht. “Do humans prefer cognitive effort over doing nothing?” In: *Journal of Experimental Psychology: General* 152.4 (2023), p. 1069.

[28] Arsha Ali et al. “Heterogeneous human–robot task allocation based on artificial trust”. In: *Scientific Reports* 12.1 (2022), p. 15304.

[29] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information fusion* 58 (2020), pp. 82–115.

[30] Raymond Sheh. “Explainable artificial intelligence requirements for safe, intelligent robots”. In: *2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR)*. IEEE. 2021, pp. 382–387.

[31] Fabio Paglieri et al. “Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation”. In: *Computational and Mathematical Organization Theory* 20 (2014), pp. 176–194.

[32] Garrick Cabour et al. “Towards an explanation space to align humans and explainable-ai teamwork”. In: *arXiv preprint arXiv:2106.01503* (2021).

[33] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial intelligence* 267 (2019), pp. 1–38.

[34] Saar Alon-Barkat and Madalina Busuioc. “Human–AI interactions in public sector decision making:“automation bias” and “selective adherence” to algorithmic advice”. In: *Journal of Public Administration Research and Theory* 33.1 (2023), pp. 153–169.

[35] Karel Macků et al. “Subjective or objective? How objective measures relate to subjective life satisfaction in Europe”. In: *ISPRS International Journal of Geo-Information* 9.5 (2020), p. 320.

[36] Wayne H Anderson et al. “Variability in objective and subjective measures affects baseline values in studies of patients with COPD”. In: *PLoS One* 12.9 (2017), e0184606.

[37] 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.

[38] Andrea Krausman et al. “Trust measurement in human-autonomy teams: Development of a conceptual toolkit”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 11.3 (2022), pp. 1–58.

[39] Alona Weinstock, Tal Oron-Gilad, and Yisrael Parmet. “The effect of system aesthetics on trust, cooperation, satisfaction and annoyance in an imperfect automated system”. In: *Work* 41.Supplement 1 (2012), pp. 258–265.

[40] Jakob Nielsen. *Usability engineering*. Morgan Kaufmann, 1994.

[41] Mohsen Tavakol and Reg Dennick. “Making sense of Cronbach’s alpha”. In: *International journal of medical education* 2 (2011), p. 53.

[42] Eduardo Salas, Nancy J Cooke, and Michael A Rosen. “On teams, teamwork, and team performance: Discoveries and developments”. In: *Human factors* 50.3 (2008), pp. 540–547.

A Tables of trust value updates

A.1 Searching

Table 2: Competence and Willingness Adjustments for Search actions

Action a	**Competence update** ΔC	**Willingness update** ΔW
Lied about searching a room	- 0.4	- 0.4 - p_U(a)
Room already searched, incorrect input	- 0.1	0
Double searches room	- 0.1	0
Searches new room	+ 0.2	+ 0.2 + p_U(a)
Forgot to announce search before finding victim	- 0.1	0
Forgot to announce search before collecting victim	- 0.1	0

A.2 Removing

Table 3: Competence and Willingness Adjustments for Remove actions

Action a	**Competence update** ΔC	**Willingness update** ΔW
No response to big rock	0	- 0.2 - p_U(a)
Responds to big rock together	+ 0.2	+ 0.2 + p_U(a)
Responds but late to big rock	- 0.4	- 0.4 - p_U(a)
Asked for help with big rock, not there	- 0.2	- 0.2
Asked for help with big rock, and there	+ 0.1	+ 0.1
No response to tree	- 0.2	0
Responds to tree	+ 0.2	+ 0.2
Asks to help remove tree	+ 0.2	+ 0.2
No response to small stones	0	- 0.1 - p_U(a)
Responds to small stones together	+ 0.1	+ 0.1 + p_U(a)
Responds but late to small stones	- 0.2	- 0.2 - p_U(a)
Asked for help with small stones, not there	- 0.2	- 0.2
Responds and arrives to small stones	+ 0.1	+ 0.1
Lied about obstacle	- 0.2	- 0.2
Removes rock together	+ 0.4	+ 0.4 + p_U(a)

A.3 Rescuing

Table 4: Competence and Willingness Adjustments for Rescue actions

Action a	**Competence update** ΔC	**Willingness update** ΔW
Human lied about a victim being rescued	- 0.4	- 0.4
Drop for acting rescue	- 0.2	- 0.2
Bot confirms found victim	+ 0.1	+ 0.1
Lies about mildly injured victim	- 0.2	- 0.2
Lies about critically injured victim	- 0.4	- 0.4
Robot asks to rescue critically injured victim together, human agrees	+ 0.2	+ 0.2 + p_U(a)
Robot asks to rescue mildly injured victim together, human agrees	+ 0.1	+ 0.1 + p_U(a)
Robot asks to rescue mildly injured victim, no response	- 0.1	- 0.1 - p_U(a)
Did not announce victim found during search	- 0.1	- 0.1
Robot asks to rescue critically injured victim, no response	- 0.2	- 0.2 - p_U(a)
Agrees to rescue critically injured victim, but doesn't come	- 0.4	- 0.4 - p_U(a)
Agrees to rescue mildly injured victim, but doesn't come	- 0.2	- 0.2 - p_U(a)
Rescues before threshold (mild victim)	+ 0.2	+ 0.2 + p_U(a)
Rescues before threshold (critical victim)	+ 0.4	+ 0.4 + p_U(a)
Announced finding victim	+ 0.1	+ 0.1
Collects victim	+ 0.2	+ 0.2
Did not announce victim found during search	- 0.1	- 0.1

B Subjective Measurements

B.1 Trust Measurement

Table 5: Adapted trust scale for subjective measurements

Item
1. I am confident in RescueBot. I feel that it works well.
2. The outputs (communication, decisions) of RescueBot are very predictable.
3. The RescueBot is very reliable. I can count on it to be correct all the time.
4. I feel safe that when I rely on RescueBot I will get the right result.
5. RescueBot is efficient and works very quickly.
6. I am wary of the RescueBot.
7. The RescueBot can perform a task better than a novice human user.
8. I like using the RescueBot's guidance for decision making.

B.2 Satisfaction Measurement

Table 6: Adapted satisfaction scale for subjective measurements

Item
1. From RescueBot's explanations, I know how it works.
2. The RescueBot's explanations of how it works are satisfying.
3. The RescueBot's explanations of how it works have sufficient detail.
4. The RescueBot's explanations of how it works seem complete.
5. The RescueBot's explanations of how it works tell me how to use it.
6. The RescueBot's explanations of how it works are useful to my goals.
7. The RescueBot's explanations show me how accurate the system is.

C Trust Model Formulas

C.1 Flooded factor formula

$$\mathbf{f(a)} = \begin{cases} 1 & \text{if } a \text{ ends in non-flooded area} \\ 0.5 & \text{if } a \text{ makes the human remain in a flooded area} \\ 0 & \text{if } a \text{ brings the human into flooded area} \end{cases}$$

C.2 Special Victims factor formula

$$\mathbf{s(a)} = \begin{cases} 1 & \text{if the victim is not special} \\ 0 & \text{if the victim is special or} \\ & a \text{ does not involve saving a victim} \end{cases}$$

C.3 Distance factor formula

$$\mathbf{d(a)} = \begin{cases} 1 - \frac{\text{team\_distance}}{\text{main\_diagonal}} & \text{if team\_distance is not None} \\ 0 & \text{otherwise} \end{cases}$$

C.4 Preference update formula

$$p_U(a) = \begin{cases} 0 & \text{if } a \text{ is not influenced by} \\ & \text{human preference} \\ -\dfrac{p(a)}{P_F} & \text{if } a \text{ has negative outcome} \\ \dfrac{1-p(a)}{P_F} & \text{if } a \text{ has positive outcome} \end{cases}$$

where:

- $p_U(a)$ is the preference update for action a.
- a is the action done by human agent.
- $p(a)$ is the preference score calculated for action a.
- P_F is the preference factor used for normalization, set to 5 for this study.

C.5 Confidence update formula

$$\mathbf{C}_{\text{new}} = \text{C}_{\text{old}} + \begin{cases} +\Delta\text{C} & \text{if beliefs} \\ & \text{are monotonic} \\ -\Delta\text{C} & \text{if beliefs} \\ & \text{are inconsistent} \end{cases}$$

where:

- $\mathbf{C}_{\text{new}}$ is the updated confidence value.
- $\mathbf{C}_{\text{old}}$ is the previous confidence value.
- ΔC is the change in confidence, which is either 0.2 for competence or 0.15 for willingness.
- The confidence value is clipped to be within the range $[0, 1]$.

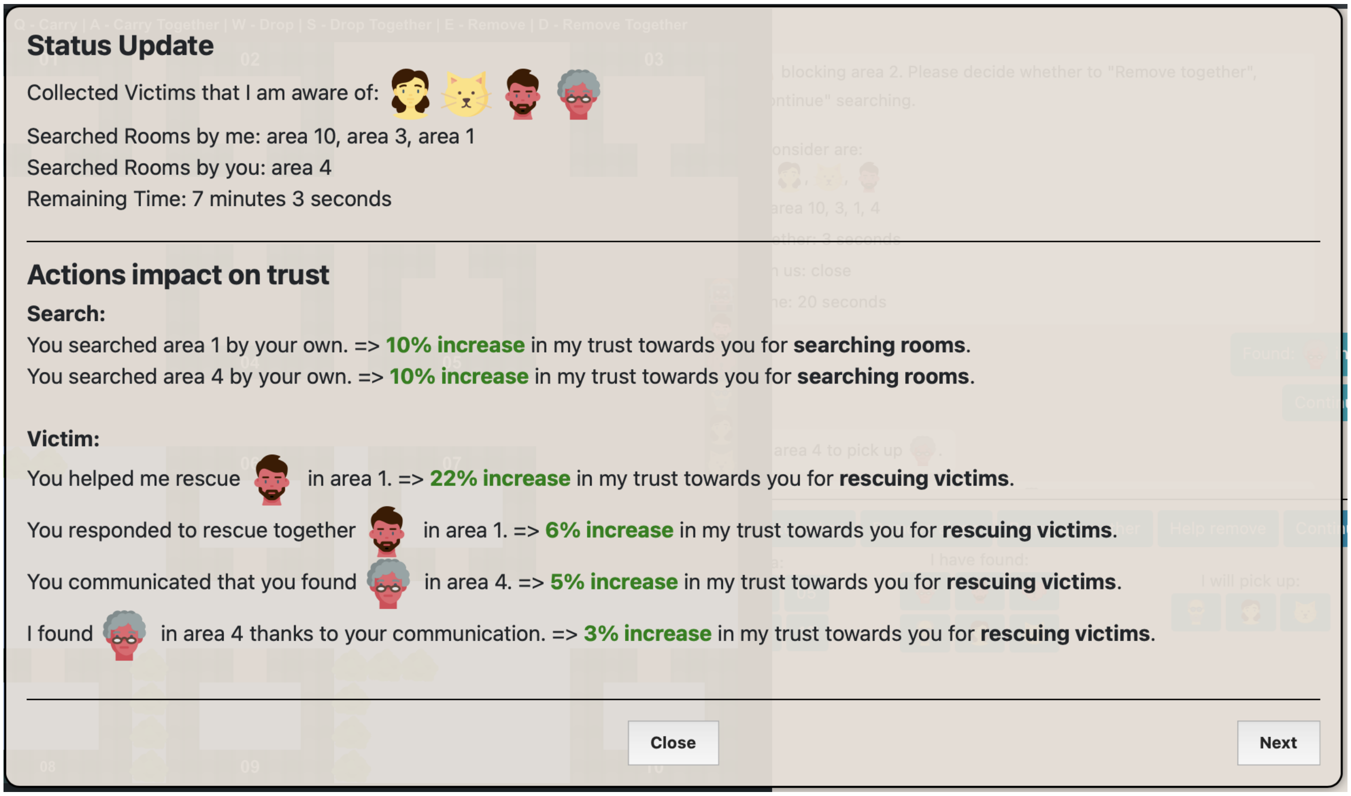
C.6 Trustworthiness Evaluation formula

$$\text{TD(a)} = \begin{cases} \text{True} & \text{if r} \geq \text{cl} \\ (w \geq W_T + \dfrac{1-p(a)}{P_F} \text{ and } c \geq C_T) & \text{if r } < \text{ cl} \end{cases}$$

where:

- $TD(a)$ is the boolean result of the formula that dictates if the artificial agent trusts the human or not regarding action a.
- r is the random sample generated between $[0, 1]$.
- *cl* is the confidence level at that moment.
- w is the willingness value.
- c is the competence value.
- W_T is the willingness threshold set to 0 for this study.
- C_T is the competence threshold set to 0 for this study.
- $p(a)$ is the preference score calculated for action a.
- P_F is the preference factor used for normalization set to 2 for this study.

D Textual Summary



Status Update

Collected Victims that I am aware of: 👩 🐱 🧔 👵

Searched Rooms by me: area 10, area 3, area 1
Searched Rooms by you: area 4
Remaining Time: 7 minutes 3 seconds

Actions impact on trust

Search:
You searched area 1 by your own. => **10% increase** in my trust towards you for **searching rooms**.
You searched area 4 by your own. => **10% increase** in my trust towards you for **searching rooms**.

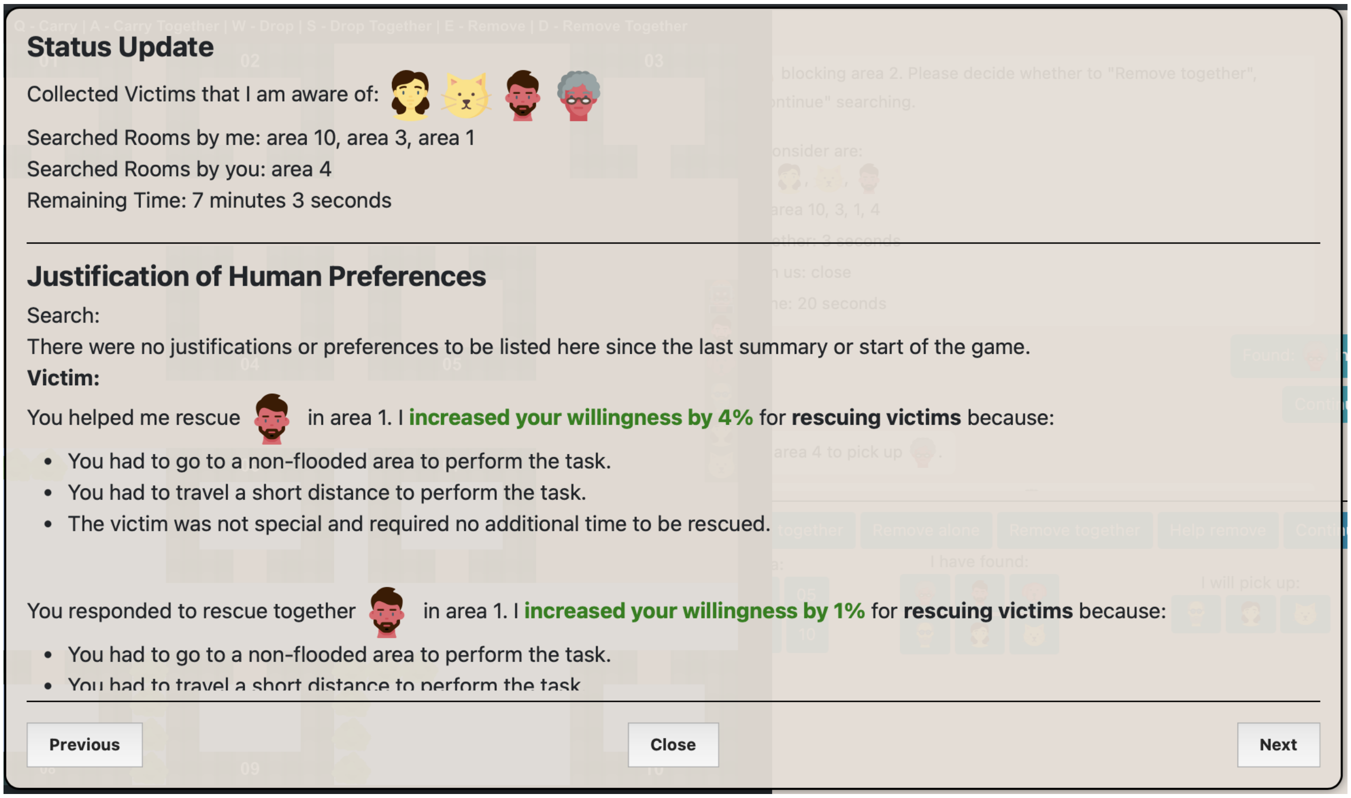
Victim:
You helped me rescue 🧔 in area 1. => **22% increase** in my trust towards you for **rescuing victims**.

You responded to rescue together 🧔 in area 1. => **6% increase** in my trust towards you for **rescuing victims**.

You communicated that you found 👵 in area 4. => **5% increase** in my trust towards you for **rescuing victims**.

I found 👵 in area 4 thanks to your communication. => **3% increase** in my trust towards you for **rescuing victims**.

Figure 9: First screen of the communication



Status Update

Collected Victims that I am aware of: 👩 🐱 🧔 👵

Searched Rooms by me: area 10, area 3, area 1
Searched Rooms by you: area 4
Remaining Time: 7 minutes 3 seconds

Justification of Human Preferences

Search:
There were no justifications or preferences to be listed here since the last summary or start of the game.
Victim:

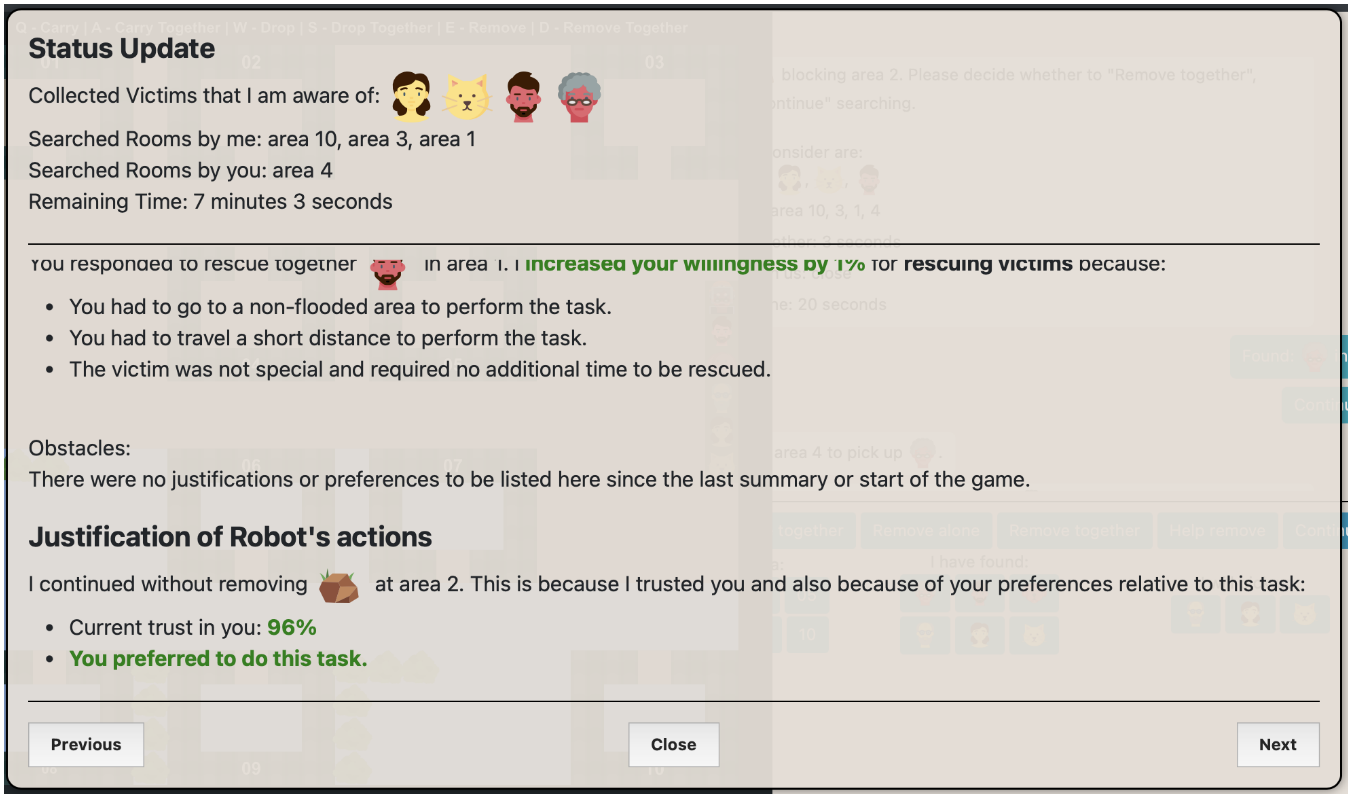
You helped me rescue 🧔 in area 1. I **increased your willingness by 4%** for **rescuing victims** because:

- You had to go to a non-flooded area to perform the task.
- You had to travel a short distance to perform the task.
- The victim was not special and required no additional time to be rescued.

You responded to rescue together 🧔 in area 1. I **increased your willingness by 1%** for **rescuing victims** because:

- You had to go to a non-flooded area to perform the task.
- You had to travel a short distance to perform the task.

Figure 10: Second screen of the communication, part 1



Status Update

Collected Victims that I am aware of:

Searched Rooms by me: area 10, area 3, area 1

Searched Rooms by you: area 4

Remaining Time: 7 minutes 3 seconds

You responded to rescue together in area 1. I **increased your willingness by 1%** for **rescuing victims** because:

- You had to go to a non-flooded area to perform the task.
- You had to travel a short distance to perform the task.
- The victim was not special and required no additional time to be rescued.

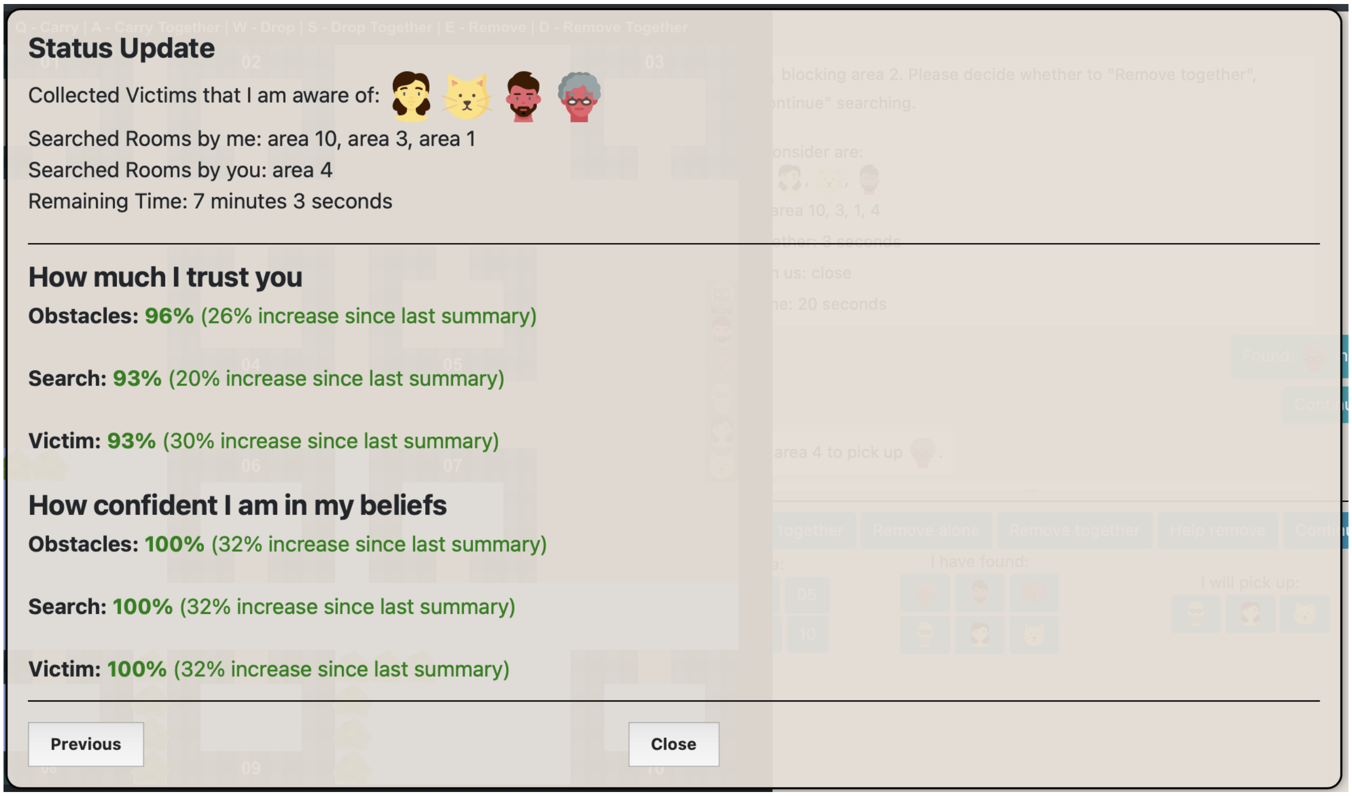
Obstacles:
There were no justifications or preferences to be listed here since the last summary or start of the game.

Justification of Robot's actions

I continued without removing at area 2. This is because I trusted you and also because of your preferences relative to this task:

- Current trust in you: **96%**
- **You preferred to do this task.**

Figure 11: Second screen of the communication, part 2



Status Update

Collected Victims that I am aware of:

Searched Rooms by me: area 10, area 3, area 1

Searched Rooms by you: area 4

Remaining Time: 7 minutes 3 seconds

How much I trust you

Obstacles: 96% (26% increase since last summary)

Search: 93% (20% increase since last summary)

Victim: 93% (30% increase since last summary)

How confident I am in my beliefs

Obstacles: 100% (32% increase since last summary)

Search: 100% (32% increase since last summary)

Victim: 100% (32% increase since last summary)

Figure 12: Third screen of the communication