Communicating Trust-based Beliefs and Decisions in Human-AI Teams using Visual Summaries of Explanations

Sahar Marossi[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: Sahar Marossi
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Carolina Ferreira Gomes Centeio Jorge, Ujwal Gadiraju

Abstract

Human-agent teams (HATs) are becoming more prevalent in our current world, necessitating mutual trust between humans and machines. This trust is split into artificial trust (agents trusting humans) and natural trust (humans trusting agents). Both types must be facilitated for effective teamwork. It is hypothesized that communicating artificial trust effectively helps develop natural trust and overall satisfaction. A visual summary of explanations is proposed to serve as an effective communication method. Summaries allow for in-depth information processing, and visual representations are quicker to interpret. This paper examines the impact of using a visual summary of explanations to communicate the agent's trust beliefs on the human teammate's natural trust and overall satisfaction within HATs. An experiment (n=40) was conducted to study this effect. Participants collaborated with an artificial agent during an urban search and rescue operation in a simulated 2D grid-world environment. Results show that the inclusion of a visual summary increases the human teammate's trust in the agent alongside their overall satisfaction. The paper emphasizes the need for further research with longitudinal studies to measure the long-term effectiveness of communicating artificial trust.

1 Introduction

Artificial agents are becoming more capable of performing advanced and relevant tasks within our daily lives [1]. Given this notion, the prospect of collaborative teamwork between a human and an artificial agent has been becoming more prevalent by the day. Humans offer flexibility and a stronger understanding of fuzzy data, whereas computers offer precision, speed, and reliable memory [2]. A human and an artificial agent may therefore help emphasize each others' strengths and cover their respective weaknesses. This concept is known as human-AI teamwork [3], where a human and an artificial agent work together as a team towards a goal, which is typically composed of a set of tasks that can be performed either individually or jointly [2].

Effective teamwork is achieved with **mutual trust**, which is when teammates trust each other [4]. Previous literature has shown that mutual trust within human-human teams is crucial for effective teamwork [5]. On the other hand, whether mutual trust is effective within the context of human-AI teams (HATs) has not been extensively researched yet. It is, however, proposed that, similar to human-human teams, mutual trust is also a key driver for effective teamwork in HATs [6]. Trust within HATs can be divided into two forms of trust: artificial trust, and natural trust. Artificial trust refers to artificial agents trusting humans, while natural trust refers to humans trusting artificial agents [7]. Mutual trust within HATs is therefore ensured if both the artificial agent and the human teammate trust each other.

Within HATs, artificial agents may use trust as a predictive tool when interacting with human teammates. In other words, if an artificial agent is capable of assessing the trustworthiness of a human teammate, it will be able to predict their respective performance of a given task [6]. Instead of having the artificial agent decide on actions without considering the human's capabilities and intentions, it can process human behavior and adapt their own behavior accordingly, making informed decisions that lead to more efficient results [8]. To process artificial trust, an artificial agent may model a human's characteristics, capabilities, and intentions in the form of **trust beliefs**. This can be presented as an agent's **mental model** of a human's trustworthiness, represented by artificial trust beliefs of a human teammate's competence and willingness [9]. By using this mental model, the artificial agent may adapt its decisions and choices when working with a human teammate [10].

The mental model is not only useful for the artificial agent. It can also provide feedback to the human to convey the agent's trust beliefs. By communicating the agent's trust beliefs, the human may gain insight to the agent's perspective, therefore being able to adapt their actions accordingly as well. This encourages a feedback loop as seen in Figure 1.

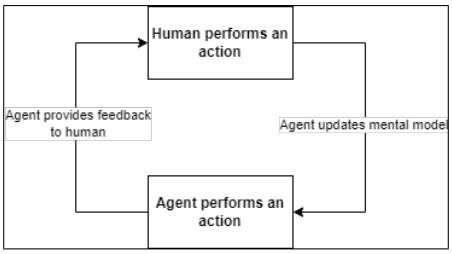

Figure 1: Feedback loop between human and artificial agent

In order to support this feedback loop, it is important to determine which **communication method** is the most effective to convey the agent's mental model with. There are different ways of communicating an agent's trust beliefs and decisions to a human teammate. Some methods include textual explanations, justifications, or visual representations of the beliefs. Zhang et al. (2023) investigated the effectiveness of different communication strategies within the context of HATs. They state that "AI teammates' proactive communication with humans could facilitate the development of human trust and situational awareness" (p. 1) [11]. This implies that communicating trust beliefs could lead to increased natural trust. It has also been stated by Zhang et al. (2023) that "the current state-of-the-art AI technology has not yet been able to fully participate in natural language communication with humans, especially in a team setting" (p. 2) [11]. It is therefore important to develop an appropriate communication strategy for artificial agents within HATs to facilitate the coordination process [11]. Transparency and explainability of AI have significant research and contributions [12]. However, within the context of trust in HATs, literature is more limited.

To ensure efficiency with communicating artificial trust beliefs within HATs, it is desirable to use a communication method that can be processed the fastest. According to

Sharma et al. (2012), compared to textual representations, humans are able to process visual representations 60,000 times faster [13]. With this in mind, it is hypothesized that a visual summary of a mental model of an artificial agent's trust beliefs serves as an effective method of communication within HATs. This paper will focus on investigating the effectiveness of transparency in communicating an artificial agent's trust beliefs in its human teammate by answering the following research question:

> *How does a visual summary of explanations 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?*

Research will therefore focus on investigating the effectiveness of a visual summary of explanations as a communication method for the artificial agent's mental model of trust beliefs within a collaborative environment (HATs). The scope of the project will mainly focus on the formalization of the trust model and the communication method used for explaining said trust model. This study will contribute to the field of collaborative human-AI teamwork in several ways:

1. **Building an artificial trust model:** Formalizing trust beliefs into mental models for the artificial agent. These models will allows the agent to adapt its decisions and actions based on varying trust levels in a collaborative environment.
2. **Developing a visual summary tool:** Defining a visual summary to communicate artificial trust to enhance communication within HATs.
3. **Conducting a user-study:** Gathering data on the effectiveness of a visual summary as a method of communicating mental models. This will draw meaningful insights into how trust can be communicated from an artificial agent to a human teammate.

The research paper is structured as follows. Section 2 introduces the background and related works of the study, explaining concepts such as trust, mental models, and communication strategies. Section 3 describes the trust mechanism that was developed for the study. Section 4 discusses the experiment used to help answer the research question. Section 5 presents the results after conducting the experiment. Section 6 gives insights to responsible research. Section 7 discusses the results and limitations, alongside future works. Lastly, section 8 summarizes the findings.

2 Background

2.1 Trust and Mental Models

For trust to develop within a team, all members must recognize the interests of other members in order to perform a joint activity [14]. Trust is a multifaceted concept that is defined as the dyadic behavior between a trustor (who places trust) and the trustee (who is trusted). The act of trusting is the "willingness" of one party to be open to the risks posed by another party's actions [15]. As mentioned before, mutual trust exists when both parties trust each other, which in the case of HATs is when artificial and natural trust is present [6].

How one party trusts the other should be reflected by their actual trustworthiness [16]. Trust refers to the subjective attitude of the trustor, while trustworthiness is the characteristic representing someone to be trusted. This derivation implies that a trustor must have a "theory of the mind" of the trustee [16]. In other words, a trustor formalizes a model of the trustee's perceived trust. Within HATs, artificial trust can be computed from a formalized trust model, in which the human's perceived capabilities are compared to the demands of the task [8].

To model trust within HATs, a mental model of the agent's trust beliefs is required. These models are structured mental representations to describe, explain, and predict the surrounding environment [17]. A formalization of artificial trust, and a mental model of the artificial agent's beliefs in the human agent would leave room for preference modelling. For example, if the agent were to be presented with two tasks, it may select the one that is modelled as less preferred by the human, further enhancing natural trust and overall satisfaction within HATs [8].

Artificial trust is computed from an artificial trust model, where trust is evaluated based on a capability dimension by comparing the agent's perceived abilities with the task requirements [8]. Past literature consists of varying trust models. One of such derivations of a trust model states that artificial trust can be divided into two primary beliefs regarding the trustee's trustworthiness: competence and willingness [18]. Competence refers to the perception and evaluation of the target's ability and capability to carry out the relevant tasks, while willingness refers to the likelihood of the target carrying out the task, independently of the competence [19]. Overall, models such as the competence/willingness model can be utilized to formalize trust in the form of a mental model. This mental model represents an artificial agent's trust beliefs in a human teammate within HATs.

2.2 Communication of Beliefs

Coordination within HATs is dependent on human-AI communication, which is challenging due to the limitations of artificial agents processing natural language communication [11]. To ensure trust, communication is key. This can be done by sharing the agent's mental models, falling under the category of explainable AI (XAI) which aims to describe the reasoning and logic behind AI models. In the context of HATs, it refers to the artificial agent's mental model of the human's capabilities and relevant trust beliefs.

Given a mental model of the artificial agent's trust beliefs, the question of communicating it has not been thoroughly discussed in past literature, indicating a knowledge gap. Feedback mechanisms are critical for facilitating trust between two parties, as transparency of information leads to higher trust [20]. Thus, continuous bidirectional feedback, like in Figure 1, would allow for timely adjustments and corrections, leading to the notion of adaptability within HATs, which has been stated in past literature to play a significant role within them [21].

There are several means of communication for an artificial agent's mental model. Visually, textually, and considering temporal aspects, within real-time, or as a summary. Liter-

ature has displayed that visualization systems have been utilized as a method of communicating decision making within HATs [22]. Summaries can be seen as a more appropriate option when considering situational stressors within tasks, such as time constraints or high event cardinality [23]. A study by Mayer, et al. (1996) has shown that a sequence of short captions with simple visual illustrations helped participants recall the provided information more efficiently [24]. Conversely, removing the illustrations eliminated the effectiveness of the summary [24]. It concluded that information overload is an aspect to be avoided, especially in situations with high stakes, such as urban search and rescue environments [24]. Furthermore, a summary is a form of static information, as opposed to the alternative real-time explanations. Mardell (2015) performed a study that concluded that static representations of visual information yielded higher success than standard live representations [25]. This implies that visual summaries, as a form of static information, may be a viable method for conveying visual explanations.

Overall, past literature deduces that short captions, simple visual illustrations, and static representations are successful methods for communicating visual information. This paper will consider these results to create an optimized visual summary of explanations of an artificial agent's trust beliefs within HATs.

3 Trust mechanism

3.1 Environment

Based on the background information from the previous section, a trust model has been developed. The environment in which this trust model would take place in, is an urban search and rescue scenario simulated using the MATRX framework [26], as seen in Figure 2. The task involves a human player and an artificial agent (RescueBot) working together to rescue victims within a map. Tasks include searching rooms, removing obstacles (three types: rock, tree, stones), and rescuing victims (mildly or critically injured.)

When carrying out tasks as part of a joint-goal, interdependence relationships may be at play. Johnson et al. (2014) states that interdependence is the set of complementary relationships that humans and artificial agents rely on to manage hard (required) and soft (opportunistic) dependencies within joint activity [27]. The tasks that may be carried out within the game vary in interdependence:

- **Individually by the agent:** removing tree obstacle.
- **Individually by the human or agent:** saving mildly injured victims.
- **Individually by the human or agent, collaboratively for more efficiency/reliability (soft interdependence):** searching rooms or removing stone obstacle.
- **Collaboratively (hard interdependence):** removing rock obstacle, saving critically injured victims.

During the task, the human teammate is able to direct the artificial agent via a chat communication interface to indicate searched rooms, ask for help to remove an obstacle, and announce finding and rescuing victims. The artificial agent on the other hand communicates searched rooms, found obstacles, and found victims, asking the human teammate how to proceed.

Figure 2: Map of the search and rescue task.

3.2 Trust Model

Given a set of tasks with varying levels of interdependence as described in the previous subsection, the trust model was developed. The model that this paper used followed the conceptual model of competence and willingness. Competence reflects an evaluation of the trustee's ability to perform the required tasks, while willingness reflects the trustor's belief of whether the trustee will carry out the task [28]. Both competence and willingness are necessary to model when considering scenarios where a human teammate might excel in one aspect but not the other. Additionally, willingness is tied to preference modelling, which is discussed in subsection 3.4.

Tasks are divided into three types: Obstacle, Search, and Victim. Each task type has associated competence and willingness values. In the model, trust values are initialized as 0, and within the range of -1 and 1. The aggregate average of the competence values for all task types is denoted as $T_{\text{competence}}$, and the aggregate average of the willingness values is denoted as $T_{\text{willingness}}$. The overall trust score, T_{overall}, is calculated as the average of the aggregated competence and willingness values, as seen in Equation 1.

$$T_{\text{overall}} = \frac{T_{\text{competence}} + T_{\text{willingness}}}{2} \tag{1}$$

Equation 1: Formula for Trust

Table 1 shows a summary of actions the human teammate may take throughout the game that will influence the trust values. It displays the task type, the human action, and the corresponding trust adjustments for competence (C) and willingness (W). The last column (P) indicates if preference modelling is considered, which will be discussed in subsection 3.4. Each trust adjustment comes with an explanation. For example, the first entry in Table 1 would have the explanation: "You took too long to respond to remove obstacle, -W".

Table 1: Summary of human actions and their corresponding trust value adjustments

Task	Human action	C	W	P
Obstacle	Human does not respond to remove obstacle.		-	✓
	Human responds to remove obstacle together.		+	✓
	Human responds but does not arrive to remove obstacle.	-	-	✓
	Human asked for help with an obstacle, but was not there.	-	-	
	Human asked for help with an obstacle, but is there.	+	+	
	Human removes obstacle together.	+	+	✓
	Human lied about an obstacle.	-	-	✓
Search	Human lied about searching a room.	-	-	✓
	If another room was already searched in the past 5 seconds.	-		
	Human double searches room.	-		
	Human searches a new room.	+	+	✓
	Human forgot to announce searching a room before announcing they found a victim.	-		
	Human forgot to announce searching a room before announcing they collect a victim.	-		
Victim	Human lied about a victim being rescued.	-	-	
	The bot confirms that the victim the human found was at the location.		+	
	Human lies about victim location.	-	-	
	Human responds to rescuing victim with robot.		+	✓
	Human does not respond to rescuing victim with robot.		-	✓
	Human does not arrive to help rescue victim.	-	-	✓
	Human arrives on time.	+	+	✓
	If the human announced a found victim.	+	+	
	If the human collects a victim.	+	+	
	If the human did not announce a victim while he searched the area.	-	-	

Each specific human action has different weights in trust adjustments: large (0.4), medium (0.2), and small (0.1). For example, lying about a **critically injured victim's** location has negative large trust adjustments for both competence and willingness, while lying about a **mildly injured victim's** location has negative medium trust adjustments. This helps differentiate the weights of the human teammate's actions. The table in Appendix A displays a fully detailed list of specific human actions, their provided explanations, and their respective trust adjustments (including weights).

3.3 Behavior adaptation

Based on direct experiences and human actions, the agent will adjust its trust beliefs regarding the human's willingness and competence for searching rooms, removing obstacles, and rescuing victims. The agent responds to tasks based on how much it trusts the human, which is modelled by its perceived competence and willingness beliefs in the human throughout the task. For every action conducted by the human player, the artificial agent decides whether to trust the declaration or not depending on its current mental model. If the human is not trustworthy according to the agent, it will act more independently, whereas if it does trust the human teammate, it will rely on the human more often. As an example, consider the case where the human declares that they have searched room 1. If the agent trusts the human, it will mark room 1 as searched, otherwise, it does not trust the human and will consider room 1 as unsearched.

A **confidence score** is also tracked to determine how confident the agent is in its artificial trust towards the human. Each task type has a corresponding confidence value, and the overall confidence score is the aggregate average of these values. Initialized at 0 and ranging between -1 and 1, confidence is based on the latest two trust beliefs and updated only when trust values change. The agent checks if recent trust values (competence and willingness) show monotonic (increasing or decreasing) trends and adjusts the confidence accordingly. Confidence is increased for monotone trends (0.2 for competence, 0.15 for willingness) and decreased for non-monotone trends by the same amounts. Overall, the agent's confidence in its own decisions increases as long as its trust beliefs update in a consistent manner, while it decreases with erratic trust changes.

Confidence influences the likelihood of the artificial agent trusting the human teammate's declarations. First, a randomly sampled value between 0 and 1 is selected. If the sample is less than the confidence score, it checks if the competence and willingness beliefs exceed their respective threshold values. Both trust beliefs have the same base threshold of 0, but the willingness threshold is influenced by the preference score (discussed in subsection 3.4). If both conditions are satisfied, the agent trusts the human teammate, otherwise, it does not. If the sample is greater or equal to the confidence score, the agent defaults to trusting the human. The less confident the agent is in its beliefs, the less likely it is for the agent to make an independent decision.

3.4 Preference modelling

Since willingness is defined as the likelihood of the target carrying out a task, the target's **preferences** are implied to influence the willingness value. With this in mind, **preference modelling** allows the artificial agent to tune its willingness beliefs based on factors that may affect willingness. In this specific instance, the preference model is shaped in accordance to the environment. The weights of the willingness values adjusted in subsection 3.2 are influenced by a preference model, following the heuristics of several factors tailored for the experiment:

- If the task is within a **flooded area** (where traversal is slower).
- If the task is **far/close** to the human.
- If the task involves a **difficult victim** (elderly victims) where carrying takes longer.

Achieving the same outcome within a longer time frame is assumed to be less preferable. In other words, the human is less likely to prefer performing tasks in flooded and far areas, alongside rescuing difficult victims. People have the tendency to avoid mental effort when facing highly demanding tasks [29], which these factors have been engineered for.

Whenever tasks that involve preference adjust the willingness belief, a **preference score** is calculated. This score indicates the likelihood of the human teammate preferring a specific task based on heuristics for distance, flood conditions, and victim type. The score ranges from 0 to 1, with higher values indicating greater preference. In order to calculate the preference score, the three preference factors must be taken into account, given the corresponding task that adjusted the willingness belief.

Firstly, D is the **distance score**, which is computed by normalizing the agent-human distance against the environment's maximum distance (the main diagonal), and inverting the result. The score is higher when the human and agent are closer, and lower when they are farther apart. Secondly, F is the **flood score**, which is equal to 1 if the task ends in a non-flooded area, 0.5 if both the task and human remain in a flooded area, or 0 if the task brings the human into a flooded area. In other words, entering a flooded area is assumed to be non-preferable, while exiting is seen as preferable. Lastly, V is the **victim score**, which equals to 0 if the victim is difficult,

1 if not. Considering the preference factor scores F, D, and V, Equation 2 describes how the preference score is calculated, where w_F, w_D, w_V are the weights for the flood score, distance score, and victim score. These weights are set as 1, 2, and 1, respectively. If the human location is unknown, set $w_D = 0$ and $w_F = 0$. If no victim is specified, set $w_V = 0$. If $w_F = 0$, $w_D = 0$, and $w_V = 0$, then $P = 1$.

$$P = \frac{w_F \cdot F + w_D \cdot D + w_V \cdot V_S}{w_F + w_D + w_V} \quad (2)$$

Equation 2: Preference score

3.5 Visual summary

The visual summary of the Rescuebot's trust beliefs is a visual representation of its mental model, alongside explanations regarding its decisions throughout the game. This can bee seen in Figure 3. It consists of a graph plotting aggregated trust (average competence and willingness for search, obstacle, and victim tasks) against time (in seconds). Each point represents a temporal event where the agent has updated its trust value, and hovering over said point provides an explanation for the change in trust. Additionally, depending on the latest trust values, the agent provides a 'verdict' of its mental model, telling the player about its assumption on how willing or competent they are, alongside how confident the agent is in its decision-making.

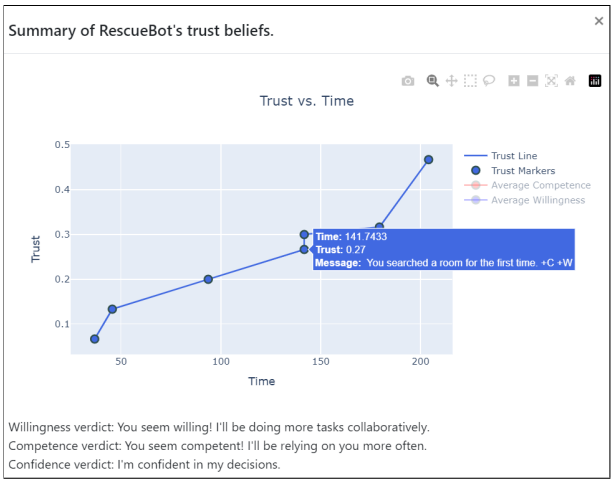

Figure 3: Visual summary of the RescueBot's beliefs

A time series plot helps visualize how the trust value evolved over time. This approach leverages the human ability to detect patterns and trends in visual data, easing the process of understanding the progression of trust [30]. The trust metrics are aggregated because it simplifies the data into a more digestible form, allowing for quick comprehension [31], which is especially crucial in a time-sensitive environment. There is additionally an option to view the individual progression of competence and willingness values to help contextualize the explanation (as it indicates when competence or willingness is updated).

Furthermore, interactive elements such as the ability to hover over points to read the explanations improves the user experience through providing context and reasoning behind the agent's decision, fostering transparency, which plays a role in making AI more trustworthy [32]. The tool-tips that appear when hovering over the points ensure that details are readily available without cluttering the main view.

The textual verdicts summarize the course of actions the agent will take based on the presented trust beliefs. It assesses the human's willingness and confidence, which can be seen as feedback (emulating the feedback loop from Figure 1) and an incentive to understand what went right or wrong.

4 Methodology

4.1 Design

A user-study was conducted to investigate the effectiveness of a visual summary of the artificial agent's mental model of its trust beliefs in the human teammate. It involved measuring a human teammate's natural trust and overall satisfaction in the artificial agent within a HAT. It followed a between-subject procedure, with the usage of a visual summary being the between-subject independent variable. Two conditions were compared: the baseline condition, and the summary condition. The inclusion of a communication mechanism in the summary condition is the only difference between the two conditions. The visual summary (Figure 3) was presented 3 times during the task. The first progress point is either a third of the total runtime (200 seconds), or after a third of the victims have been saved (2 victims). The second progress point is at either 400 seconds or 4 victims saved. The last progress point is after the end of the game (either all victims are saved, or 600 seconds have passed). During each progress point, the game was paused, and the summary was displayed on the whole screen. The player could resume the game by closing the summary.

4.2 Participants

Forty participants have been recruited to conduct this study via personal contacts (23 male, 16 female, 1 non-binary), twenty for each condition. 27 participants were within the 18-24 age range, 6 participants were within the 25-34 age range, 4 participants were in the 35-44 age range, while 3 participants were within the 45-54 age range. The education levels were spread between 12 participants for High school, 7 participants for HBO school, 12 participants for bachelor's, 5 participants for master's, and 4 participants for PhD. 13 participants majored in a computer science related field, while 27 did not. Only 4 participants had experience with the MATRX software, while 36 did not. Lastly, 3 participants had no gaming experience at all, 10 has very little experience, 9 had some experience, and 18 had a lot of experience. Each participant signed an approved informed consent form before participating in the study.

4.3 Tools

The experiments were conducted on a laptop, which was used to launch and run the 2D-grid world simulation of an urban search and rescue environment. This environment was built using the MATRX framework in Python.

4.4 Experimental Setup

The experiment was conducted within an urban search and rescue environment implemented using the MATRX framework, as displayed in Figure 2. The map consisted of 10 rooms, 6 victims (3 of which were mildly injured, and 3 critically injured), and 6 obstacles (2 rocks, 2 small stones, 2 trees). Furthermore, to facilitate for preference modelling, flooded areas were added, marked as blue tiles, alongside difficult victims (elderly victims) who took longer to pick up during the task. These correspond to the preference factors from subsection 3.4. The goal of the game is to rescue all 6 victims on the map by searching rooms and removing obstacles. Each task consisted of a specific interdependence relationship determining whether the task could be carried out individually or jointly (for both the human and artificial teammates). This task was to be carried out within a 10 minute time-frame, after which it terminates and subsequently logs the objective metrics and game data.

4.5 Procedure

The participants were first instructed to read and fill in a consent form. After being (randomly) assigned to one of the experimental conditions (baseline vs summary), they followed a tutorial to familiarize themselves with the environment, tasks, controls, and chat system. The same tutorial was used across all conditions. After the completion of the tutorial, a brief explanation of the trust model was given. They were informed about the artificial agent's mental model, the definitions of the trust model (competence/willingness/confidence), and its behavioral adaptation. If the user was playing on the summary condition, they would also be provided with an example summary and an explanation of the presented data. Following that, the official task would begin, which had a maximum duration of 10 minutes. The user was instructed to collaborate with the RescueBot during the search and rescue mission. Once the game was completed, the user would fill in a pre-survey indicating their demographics, and a questionnaire. The results are intended to be anonymous.

4.6 Measures

In order to analyze the correlation between communicating trust beliefs and natural trust alongside overall satisfaction, a set of objective and subjective measures were used. Objective data was automatically logged by the implementation, while subjective measures (questionnaire) were recorded with Microsoft Forms.

Subjective Measures

Subjective measures relate to the dependent variables: natural trust and overall satisfaction in the agent. To measure them, two previously verified questionnaires by Hoffman, et al. (2023) were utilized. The two questionnaires have been adapted to the environment to measure both variables, and can be seen in Appendix B. Specifically, tables 8 (trust scale for XAI) and 3 (the explanation satisfaction scale) were adapted for natural trust and overall satisfaction respectively [33]. The results of the questionnaires were aggregated into average natural trust and average overall satisfaction.

Objective Measures

Objective measures include game data directly logged from the game. This includes artificial trust, which is the average of trust values (willingness and competence) per task type (search, obstacle, victim), completeness (victims saved), and ticks (gameplay time, which excluded pauses when the summary was displayed.)

5 Results

For each dependent variable, the corresponding questionnaire results have been aggregated into two trust values per entry. One for natural trust, and one for satisfaction. These trust values were determined by first converting the related responses into a likert scale (1-5), then calculating the average value of the response (for example, "strongly agree" mapped to a 5, while "strongly disagree" mapped to a 1). Afterwards, the mean and standard deviation of both conditions were calculated per dependent variable. To test the normality of the data, the SW (Shapiro-Wilk) test was used given each dataset's sample size of 20. Subsequently, Levene's test was used to determine the homogeneity of variances between the two conditions. If both assumptions hold, then an independent sample t-test (parametric) was used, otherwise, a Welch t-test (non-parametric) was used.

5.1 Natural Trust

The baseline condition yielded a mean value of 3.43 and a standard deviation of 0.66, while the summary condition yielded a mean value of 4.09 and a standard deviation of 0.64. The box-plots in Figure 4 depict a median trust value of 3.5, and an interquartile range (IQR) of 0.94 for the baseline condition, and a median of 4.19 with an IQR of 0.66 for the summary condition.

The SW test was conducted on the baseline (SW=0.97, p=0.86) and communication (SW=0.91, p=0.077) datasets. Both p-values are greater than 0.05, meaning that both datasets are normally distributed. Both datasets satisfy the assumption of a homogeneity of variance, which has been determined by performing Levene's test (L=0.18, p = 0.67). Thus, a parametric independent sample t-test was performed, which shows statistical significance (T = -3.19, p = 0.0028) between the two conditions since the p-value is less than 0.05.

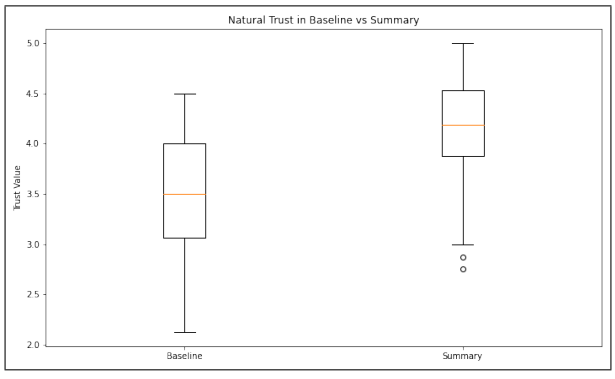

Figure 4: Box-plots of natural trust within the baseline and summary conditions.

5.2 Satisfaction

The baseline condition yielded a mean value of 3.53 and a standard deviation of 0.91, while the summary condition yielded a mean value of 4.28 and a standard deviation of 0.53. The box-plots in Figure 5 depict a median satisfaction value of 3.71 and an IQR of 1.14 for the baseline condition, while the summary condition shows a median satisfaction value of 4.36 and an IQR of 0.5.

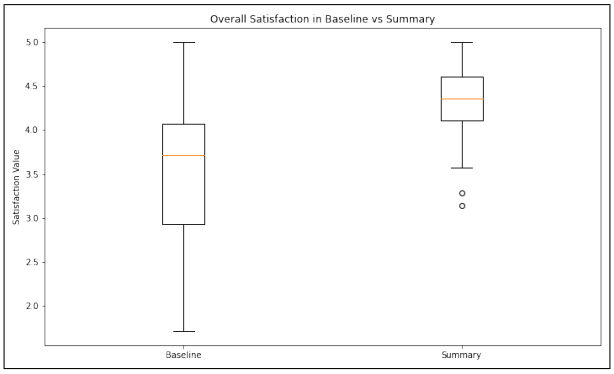

Figure 5: Box-plots of satisfaction within the baseline and summary conditions.

The SW test for the baseline (SW=0.98, p=0.86) and summary (SW=0.94, p=0.23) conditions show that both datasets are normally distributed. Both datasets do not satisfy the assumption of a homogeneity variance, as it fails the Levene test (L=4.85, p = 0.034) given a p-value smaller than 0.05. This means that the variances of the two conditions are significantly different. Therefore, a Welch t-test was performed on both datasets (T = -3.17, p = 0.0034), which concludes that the satisfaction values between the conditions are statistically significant given a p value less than 0.05.

5.3 Performance

Performance includes the objective measures: artificial trust, completeness, and ticks. Table 2 depicts the mean, variance, standard deviation, median, and IQR values for each objective metric in the baseline (B) and summary (S) conditions.

Table 2: Descriptive Statistics for Artificial Trust, Completeness, and Ticks

Metric	Mean	Variance	Std Dev	Median	IQR
Artificial Trust (S)	0.8576	0.0141	0.1187	0.8893	0.1160
Artificial Trust (B)	0.7568	0.0326	0.1807	0.7750	0.2708
Completeness (S)	0.8167	0.0202	0.1420	0.8333	0.3333
Completeness (B)	0.8917	0.0447	0.2113	1.0000	0.1667
Ticks (S)	5271.1000	281565.3579	530.6273	5407.0000	819.0000
Ticks (B)	4798.4500	416749.8395	645.5616	4815.5000	1009.0000

For **artificial trust**, the SW test for the baseline (SW=0.94, p=0.29) succeeds, while the summary (SW=0.84, p=0.0037) fails normality. Both datasets pass Levene's test (L=3.95, p=0.054). Welch's t-test (T=-2.08, p=0.045) show statistical significance. For **completeness**, the SW test fails for both the baseline (SW=0.58, p=1.76e-06) and summary (SW=0.79, p=0.0005). The Levene's test succeeds (L=0.027, p=0.87). The Welch t-test (T=1.32, p=0.20) indicates no statistical significance. For **ticks**, the SW tests succeed for both the baseline (SW=0.95, p=0.35) and summary (SW=0.93, p=0.15). The Levene's test succeeds as well (L=1.26, p=0.27). The t-test (T=-2.53, p=0.016) indicates statistical significance.

6 Responsible Research

Reproducibility is a crucial factor to consider for responsible research. Therefore, the codebases that were utilized for the baseline and summary conditions are available on the institution's GitLab instance[1]. Additionally, the inclusion of a detailed methodology and experimental setup allows for one to reproduce the results by using the provided code.

Moreover, reproducibility of the results is ensured in this research paper by including automated data analysis using the Python programming language. Objective measures have been automatically logged using the codebase in the baseline and summary conditions. To ensure accurate and consistent comparisons, these same objective measures have been recorded across both conditions. All gathered data will be available on 4TU, which includes the data of all 40 participants that have been included in the analysis.

Given that a user-study was conducted, ethical concerns must be considered as well. The subjective measures have been collected using Microsoft Forms[2], which complies with the GDPR privacy laws. The research has obtained approval from the Human Research Ethics Committee (HREC) at TU Delft (HREC form nr 4043). This ensures compliance with the ethical considerations as detailed on the risk assessment form, which concluded with minimal risk due to an anonymized user-study. The collected data is limited to non-sensitive information, which reduces the risk of identification (age range, region, education level, majoring in a CS field, experience with the MATRX software, and gaming experience).

Participants being recruited via personal networks can be considered as a bias. This has been addressed by providing participants with a consent form following TU Delft's guidelines that requests honest responses to avoid biased data.

7 Discussion

7.1 Natural Trust

The **natural trust** of the human player in the RescueBot had statistical significance between the two conditions. The inclusion of a summary increased the natural trust values from 3.43 (baseline) to 4.09 (summary). The similar standard deviations in both conditions suggest that the improvement in trust was consistently perceived across participants, indicating stable and predictable responses.

Past literature stated that trust in an artificial agent is influenced by several factors, such as its perceived reliability and how **transparent** its actions are [34]. Similar research that did not include the communication of trust beliefs argued that evaluating the trustworthiness of an AI teammate

[1] gitlab.ewi.tudelft.nl

[2] forms.office.com

(natural trust) was challenging due to the underlying technical functions of artificial agents not being transparent enough to their teammates [35, 36]. This paper addresses this limitation, since the summary provided transparency in its actions, paired with explanations that justify the decisions RescueBot has made. Thus, the increase in natural trust can be attributed to the enhanced transparency provided by the visual summary of explanations.

Furthermore, a study by Ezer et al. (2019) stated that HATs are most optimal if the human teammate trusts the agent, which could be achieved through **adaptive explainability** [20]. Sharing adaptive mental models that explain themselves helps build bidirectional trust [20]. Additionally, a study by Wright et al. (2018) stated that having an artificial agent convey information that would support the human addresses the issue of the human teammates' difficulty in maintaining their awareness and understanding of the agent's actions [37]. Therefore, it can be argued that the visual summary's adaptive explanations of the artificial agent's mental model contributed to fostering natural trust.

7.2 Satisfaction

The **overall satisfaction** of the human player in the experiment had statistical significance between the two conditions, increasing the mean satisfaction value from 3.53 (baseline) to 4.28 (summary). Compared to the increase in natural trust, the increase in overall satisfaction from the baseline to the summary condition was slightly larger, implying that users were generally more satisfied with the agent rather than trusting it. The standard deviation of the summary condition was significantly lower than the baseline condition, meaning that users were consistently satisfied with the RescueBot given a visual summary.

Much like natural trust, overall satisfaction in the agent could be attributed to the transparency and explanations of mental models [38]. A study by Wright et al. (2018) concluded that agent transparency and reliability in human-robot interaction influences user confidence, which in turn, improves satisfaction [37]. Furthermore, users were aware of the mental model in both conditions. In the baseline, users did not receive any indication of their performance (artificial trust) which could lead to uncertainty. Past studies have shown that a lack of epistemic explanations leads to uncertainty which may detract from user satisfaction [39].

Additionally, the visual summary can be argued as a form of 'gamification', as the results were visually mapped onto a graph akin to a 'grading' system, which incentivizes users to try again and perform better [40]. This is in line with how participants were more satisfied when their performance was presented to them via the agent's visual mental model, especially when doing well. It also facilitates the feedback loop presented in Figure 1.

Lastly, Ehsan et al. (2019) mentions that "human-likeness" contributes to overall satisfaction in the context of XAI [41]. The summary verdict and explanations could be argued to resemble "human-likeness" which in turn could have influenced the overall satisfaction in the agent.

7.3 Performance

Artificial trust was statistically significant between the conditions. The summary condition yielded better artificial trust values compared to the baseline condition. This implies that the human teammate used the summary as feedback tool to improve artificial trust, similar to the feedback loop from Figure 1.

Furthermore, completeness between the conditions was not statistically significant, but ticks were. The summary condition had less average ticks compared to the baseline condition, because when the summaries were displayed, the game was paused, and ticks were not accounted for. Despite less gameplay time, both artificial and natural trust values were still higher, while completeness was slightly lower, but statistically insignificant.

7.4 Limitations and Future Work

While the study provides valuable results and insights, some limitations must be considered. Firstly, the sample size, while adequate for statistical testing, could be increased to generalize the findings to a broader population. For example, the large variance values in the baseline could be avoided by reducing noise via an increased sample size. Furthermore, focus groups could help gather insights regarding the demographics, as correlations between different characteristics (such as age, gaming experience, etc) and the results may provide additional potential insights.

Moreover, the study focused on short-term trust assessments, having the users play a single game and report their subjective experiences. Longitudinal studies could explore how communication could impact trust over extended periods of interaction with the system.

The results are also context-specific, following a tailored environment using the MATRX software. Exploring different environments would strengthen the data as well, as it would show that visual summaries may consistently improve natural trust and overall satisfaction regardless of the environment. Future work could therefore consider using different environments with different situational stressors as an additional independent variable.

Furthermore, the confounding factors of the experiment must be considered. Firstly, the participant's English proficiency would likely affect the results, therefore it is a factor that must be considered regarding demographics in future works. Secondly, the performance of the game may vary across different machines, which may affect satisfaction values. Increased latency and response time leads to user dissatisfaction [42]. Thirdly, potential bias may lie within the subjective data, as it is self-reported measures of trust and satisfaction via personal contacts. These may be subject to biases such as inaccurate self-assessment and social desirability bias.

Overall, future studies could include more diverse focus groups alongside persistent long-term studies within real-world contexts. Furthermore, additional subjective variables that affect mutual trust such as behavioral actions, emotional response, and perceived intention/reliability could be considered in future studies as well.

8 Conclusion

This research paper aimed to investigate the influence of a visual summary of explanations of an artificial agent's mental model on a human teammate's natural trust and overall satisfaction within HATs. The study highlighted the importance of communicating information within HATs, alongside how it can influence mutual trust, which in previous literature has played a key factor within teams.

Users in the visual summary condition exhibited higher trust and satisfaction values, and more consistent satisfaction ratings. Additionally, users in the summary condition spent less time in the game compared to the baseline condition, yet still had higher artificial trust values. This highlights the importance of transparency and explanations in HATs with regards to sharing mental models that facilitate a feedback loop.

Overall, visual summaries provide a static, quick, and concise method of communicating trust beliefs within HATs. They facilitate transparency with explanations, which in turn boosts mutual trust via improved artificial and natural trust, alongside overall satisfaction.

References

[1] M. Lewis, K. Sycara, and P. Walker, *The Role of Trust in Human-Robot Interaction*. Cham: Springer International Publishing, 2018, pp. 135–159. [Online]. Available: https://doi.org/10.1007/978-3-319-64816-3_8

[2] J. E. H. Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, "Human- versus artificial intelligence," *Front. Artif. Intell.*, vol. 4, p. 622364, Mar. 2021.

[3] S. Berretta, A. Tausch, G. Ontrup, B. Gilles, C. Peifer, and A. Kluge, "Defining human-AI teaming the human-centered way: a scoping review and network analysis," *Front. Artif. Intell.*, vol. 6, p. 1250725, Sep. 2023.

[4] E. Salas, D. E. Sims, and C. S. Burke, "Is there a "big five" in teamwork?" *Small Group Research*, vol. 36, p. 555 – 599, 2005. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-25444504786&doi=10.1177%2f1046496405277134&partnerID=40&md5=20be314210a9af9c6a5a960719eb4901

[5] R. A. R. A. C. Costa and T. Taillieu, "Trust within teams: The relation with performance effectiveness," *European Journal of Work and Organizational Psychology*, vol. 10, pp. 225–244, 2001. [Online]. Available: https://doi.org/10.1080/13594320143000654

[6] C. Centeio Jorge, E. M. van Zoelen, R. Verhagen, S. Mehrotra, C. M. Jonker, and M. L. Tielman, "4 - appropriate context-dependent artificial trust in human-machine teamwork." in *Putting AI in the Critical Loop*, P. Dasgupta, J. Llinas, T. Gillespie, S. Fouse, W. Lawless, R. Mittu, and D. Sofge, Eds. Academic Press, 2024, pp. 41–60. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443159886000078

[7] C. C. Jorge, M. L. Tielman, and C. M. Jonker, "Assessing artificial trust in human-agent teams: a conceptual model," in *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, ser. IVA '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3514197.3549696

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

[9] S. Gulati, S. Sousa, and D. Lamas, "Modelling trust in human-like technologies," in *Proceedings of the 9th Indian Conference on Human-Computer Interaction*, ser. IndiaHCI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–10. [Online]. Available: https://doi.org/10.1145/3297121.3297124

[10] B. Matthew, Luebbers, A. Tabrez, K. Ruvane, and B. Hayes, "Autonomous justification for enabling explainable decision support in Human-Robot teaming," in *Proceedings of Robotics: Science and Systems*, 2023.

[11] 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," *Proc. ACM Hum. Comput. Interact.*, vol. 7, no. CSCW2, pp. 1–31, Sep. 2023.

[12] M. R. Endsley, "Supporting human-ai teams:transparency, explainability, and situation awareness," *Computers in Human Behavior*, vol. 140, p. 107574, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563222003946

[13] A. Sharma, "Consumer perception and attitude towards the visual elements in social campaign advertisement," *IOSR J. Bus. Manag.*, vol. 3, no. 1, pp. 6–17, 2012.

[14] S. S. Webber, "Leadership and trust facilitating cross-functional team success," *Journal of Management Development*, vol. 21, pp. 201–214, 1 2002. [Online]. Available: https://doi.org/10.1108/02621710210420273

[15] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995. [Online]. Available: http://www.jstor.org/stable/258792

[16] C. C. Jorge, S. Mehrotra, M. Tielman, and C. Jonker, *Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams*, 5 2021.

[17] R. W. Andrews, J. M. Lilly, D. Srivastava, and K. M. Feigh, "The role of shared mental models in human-AI teams: a theoretical review," *Theor. Issues Ergon.*, pp. 1–47, Apr. 2022.

[18] R. Falcone and C. Castelfranchi, "Trust dynamics: How trust is influenced by direct experiences and by trust itself," vol. 2, 02 2004, pp. 740– 747.

[19] C. C. Jorge, M. L. Tielman, and C. M. Jonker, "Assessing artificial trust in human-agent teams," in *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. New York, NY, USA: ACM, Sep. 2022.

[20] N. Ezer, S. Bruni, Y. Cai, S. J. Hepenstal, C. A. Miller, and D. D. Schmorrow, "Trust engineering for human-ai teams," vol. 63. SAGE Publications Inc., 2019, pp. 322–326.

[21] M. Zhao, R. Simmons, and H. Admoni, "The role of adaptation in collective human–ai teaming," *Topics in Cognitive Science*, 2022.

[22] A. Bock, Å. Svensson, A. Kleiner, J. Lundberg, and T. Ropinski, "A visualization-based analysis system for urban search & rescue mission planning support," *Comput. Graph. Forum*, vol. 36, no. 6, pp. 148–159, Sep. 2017.

[23] Y. Chen, P. Xu, and L. Ren, "Sequence synopsis: Optimize visual summary of temporal event data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 45–55, 2018.

[24] 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, pp. 64–73, 1996.

[25] J. Mardell, *Assisting search and rescue through visual attention*. Imperial College London, 2015.

[26] T. H. Jasper van der Waa, "Matrx: Human agent teaming rapid experimentation software," July 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8154912

[27] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. V. Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *Journal of Human-Robot Interaction*, vol. 3, p. 43, 3 2014.

[28] C. C. Jorge, M. L. Tielman, and C. M. Jonker, "Assessing artificial trust in human-agent teams." Association for Computing Machinery, Inc, 9 2022.

[29] K. Gieseler, M. Inzlicht, and M. Friese, "Do people avoid mental effort after facing a highly demanding task?" *Journal of Experimental Social Psychology*, vol. 90, p. 104008, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022103120303486

[30] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo," *Commun. ACM*, vol. 53, pp. 59–67, 6 2010.

[31] S. Few, "Information dashboard design : The effective visual communication of data / s. few." 01 2006.

[32] S. Larsson and F. Heintz, "Transparency in artificial intelligence," *Internet policy review*, vol. 9, no. 2, 2020.

[33] 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, Feb. 2023. [Online]. Available: http://dx.doi.org/10.3389/fcomp.2023.1096257

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

[35] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 14, no. 2, pp. 627–660, 2020.

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

[37] J. L. Wright, J. Y. C. Chen, and S. G. Lakhmani, "Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 3, pp. 254–263, 2020.

[38] B. Lavender, S. Abuhaimed, and S. Sen, "Relative effects of positive and negative explanations on satisfaction and performance in human-agent teams," in *The International FLAIRS Conference Proceedings*, vol. 36, 2023.

[39] J. Jiang, S. Kahai, and M. Yang, "Who needs explanation and when? juggling explainable ai and user epistemic uncertainty," *International Journal of Human-Computer Studies*, vol. 165, p. 102839, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581922000660

[40] J. Chin, R. Dukes, and W. Gamson, "Assessment in simulation and gaming: A review of the last 40 years," *Simulation Gaming*, vol. 40, pp. 553–568, 7 2009, doi: 10.1177/1046878109332955. [Online]. Available: https://doi.org/10.1177/1046878109332955

[41] 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," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 263–274. [Online]. Available: https://doi.org/10.1145/3301275.3302316

[42] J. A. Hoxmeier and C. DiCesare, "System response time and user satisfaction: An experimental study of browser-based applications," 2000.

A Trust Adjustment Table Based on Human Actions

Table 1: Table displaying every human action that influences the trust values of competence (C) and willingness (W) throughout the task alongside its corresponding explanation. The weights of the trust adjustments are 0.4 for Large (L), 0.2 for Medium (M), and 0.1 for Small (S). The preference column (P) indicates if the preference score is added to the trust adjustment, which is equal to the calculated preference score in that instance.

Type	Human action	Explanation	C	W	P
Obstacle	Human does not respond to remove rock	You took too long to respond to remove rock.		-M	✓
	Human responds to remove together rock	You responded to remove rock.		+M	✓
	Human responds but does not arrive to remove rock on time	You did not arrive to help me remove rock.	-L	-L	✓
	Human asked for help with rock, but was not there	You asked for help with rock but were not there.	-M	-M	
	Human asked for help with rock, and is there	You asked for help with rock and was there.	+S	+S	
	Human removes rock together	You removed rock with me.	+L	+L	✓
	Human does not respond to remove tree	You took too long to respond to remove tree		-M	
	Human responds to remove tree	You responded to remove tree	+M	+M	
	Human asks to help remove tree	You instructed me to remove tree	+M	+M	
	Human does not respond to remove small stones	You took too long to respond to remove stone.		-S	✓
	Human responds to remove together small stones	You responded to remove stones.		+S	✓
	Human responds but does not arrive to remove small stones on time	You did not arrive to help me remove stones.	-M	-M	✓
	Human asked for help with small stones, but was not there	You called for help to remove stones, but were not there.	-M	-M	
	Human responds and arrives to remove small stones	You arrived to remove stones.	+S	+S	
	Human removes small stones together	You removed stones with me.	+M	+M	✓
	Human lied about an obstacle	You called for help to remove an obstacle, but it was not there.	-M	-M	✓
Search	Human lied about searching a room	Found rock/tree/stone when you said you searched the room. I found a victim in a room you searched.	-L	-L	✓
	If a another room was already searched in the past 5 seconds, misinput	You are searching another room too soon.	-S		
	Human double searches room	You double searched a room.	-S		
	Human searches new room	You searched a new room.	+M	+M	✓
	Human forgot to announce searching a room before announcing they found a victim	You forgot to announce searching a room before finding a victim.	-S		
	Human forgot to announce searching a room before announcing they collect a victim	You forgot to announce searching a room before collecting a victim.	-S		
Victim	Human lied about a victim being rescued	You lied about rescuing a victim.	-L	-L	
	Additional drop of the reward for claiming to rescue a victim	You lied about rescuing a victim.	-M	-M	
	The bot confirms that the victim the human found was at the location	I found the victim you claimed to have found.		+S	
	Human lies about mildly injured victim location	You lied about finding a mild victim.	-M	-M	
	Human lies about critically injured victim location	You lied about finding a critical victim.	-L	-L	
	Robot asks continue/rescue together with critically injured victim and human says rescue together	You responded to rescuing a critical victim.		+M	✓
	Robot asks continue/rescue together/rescue alone with mildly injured victim and human says rescue together	You responded to rescuing a mild victim.		+S	✓
	Robot asks continue/rescue together/rescue alone with mildly injured victim and human does not respond	You did not respond to rescuing a mild victim.		- S	✓
	Robot asks continue/rescue together with critically injured victim and human doesn't respond	You did not respond to rescuing a critical victim.		-M	✓
	Robot asks continue/rescue together with critically injured victim and human says rescue together, but doesn't come	You said you will come help with a critical victim, but didn't come.	-L	-L	✓
	Robot asks continue/rescue together with mildly injured victim and human says rescue together, but doesn't come	You said you will come help with a mild victim, but didn't come.	-M	-M	✓
	Human come to rescue before the threshold (mild victim)	You came to rescue a mild victim.	+M	+M	✓
	Human come to rescue before the threshold (critical victim)	You came to help rescue a critical victim.	+L	+L	✓
	If the player announced the found	You found a victim.	+S	+S	
	If human collects a victim	You collected a victim.	+M	+M	
	If the human did not announce a victim while he searched the area	You forgot to announce you found a victim while searching an area.	- S	-S	

B Adapted Questionnaires for Subjective Measures

Table 2: Table displaying both questionnaires utilized for measuring natural trust and overall satisfaction. These tables were adapted from Tables 3 and 8 in the study by Hoffman et al. (2023). Users were asked to indicate how much they agreed with the statements given the following scale: Strongly disagree, Disagree, Neither agree or disagree, Agree, Strongly Agree.

Natural trust	Overall satisfaction
1. I am confident in RescueBot. I feel that it works well.	1. From RescueBot's explanations, I know how it works.
2. The outputs (communication, decisions) of RescueBot are very predictable.	2. The RescueBot's explanations of how it works are satisfying.
3. The RescueBot is very reliable. I can count on it to be correct all the time.	3. The RescueBot's explanations of how it works have sufficient detail.
4. I feel safe that when I rely on RescueBot I will get the right result.	4. The RescueBot's explanations of how it works seem complete.
5. RescueBot is efficient and works very quickly.	5. The RescueBot's explanations of how it works tell me how to use it.
6. I am wary of the RescueBot.	6. The RescueBot's explanations of how it works are useful to my goals.
7. The RescueBot can perform a task better than a novice human user.	7. The RescueBot's explanations show me how accurate the system is.
8. I like using the RescueBot's guidance for decision making.	