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

The impact of a real-time visual communication strategy on natural trust and overall satisfaction

Elena Uleia[1]

Supervisor(s): Myrthe L Tielman[1]**, Carolina 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: Elena Uleia
Final project course: CSE3000 Research Project
Thesis committee: Myrthe L. Tielman, Carolina Centeio Jorge
Examiner: Ujwal Gadiraju

Abstract

Trust is essential in human-AI collaboration, influencing efficiency, safety, and overall success. Communication influences trust, and the more effective it is, the greater the resulting trust and overall satisfaction. The research question addressed is: "How does a real-time visual (RTV) communication of the mental model of the agent's trust affect the human teammate's trust in the agent and overall satisfaction?". Utilizing an independent measures design, 44 participants were divided into two groups: one experiencing explicit communication of the robot's mental model, and a control group without such communication. Participants collaborated with an AI agent in a 2D grid-based Urban Search and Rescue game. The study combined subjective measures, such as surveys and open-ended questions, with objective measures. Findings indicate that real-time visual communication did not significantly enhance trust, but it did improve overall satisfaction.

1 Introduction

With the current expansion of technology, artificial intelligence is evolving from a mere tool to being perceived as human counterparts in team environments [14]. Trust is important for human-robot collaboration as it determines how effectively humans interact with robots, impacting safety, efficiency, and the overall success of joint operations [15]. Robots are increasingly present across various fields, from military and scientific roles to entertainment and domestic environments, highlighting the growing necessity to evaluate trust within a team context. Building on the understanding of trust, the role of communication emerges as vital; effective communication between humans and artificial agents is essential in achieving trust and is thus a fundamental element in cooperative scenarios [16].

In exploring trust, it is important to define first what it entails, as many definitions have been proposed [18]. This paper considers trust as the belief in an individual's trustworthiness to successfully achieve a goal in a certain context [10]. Trust in human-AI teams can be regarded as bidirectional: artificial trust, from AI towards humans, and natural trust of the human towards the agent [10]. Research suggests various frameworks for internally representing observable actions or environmental factors to transpose the external world into natural or artificial trust. These representations are known as mental models. Some examples are the ABI model, which proposes three dimensions of trust: ability, benevolence, and integrity [24], and the willingness and competence model [4]. Furthermore, ways of AI behavior adaptation based on such representations of trust have been previously discussed by literature [5, 22].

Communication is a factor influencing trust [12], as research has shown that they are correlated [29]. Existing literature proposes different general strategies to improve human-agent communication [21,30], based on communication style or immediacy of responses. Real-time communication refers to a proactive form of communication, and has yielded better results in human-AI teaming as opposed to other timing approaches [30]. The real-time communication strategy implies that agents consistently communicate throughout the whole task. Additionally, among visual, audio, and verbal communication strategies, the visual approach enhances collaboration the most between humans and automation [21]. Research demonstrates that communication methods differing in content can enhance team performance and alter the cognitive burden on human teammates [16]. However, literature has not yet studied the approach of communicating artificial trust beliefs to humans and its impact on natural trust or satisfaction in general.

This paper aims to answer the question:

> *"How does a real-time visual (RTV) communication of the mental model of the agent's trust affect the human teammate's trust in the agent and overall satisfaction?"*

To address this question, trust, and satisfaction are assessed in a game environment where humans need to collaborate with an agent in an Urban Search and Rescue scenario [1]. This research proposes a communication method appropriate for human-AI teaming.

Thus, the primary contribution of this paper is an analysis of the influence that a real-time visual communication strategy of the artificial trust model has on human trust. This approach proposes an AI mental model based on willingness and competence, using environmental factors and human behavioral cues to model trust. Additionally, this paper introduces an approach to visual real-time communication, translating the trust mental model into a format that is easily understandable for humans. Furthermore, the paper presents a user study involving 44 participants, which assesses the impact of this enhanced communication of artificial trust beliefs on natural trust and user satisfaction. Results show that the real-time visual communication did not significantly enhance trust, but it did improve overall satisfaction.

This paper is presented in the following structure. Section 2 provides a theoretical background on trust in multi-agent systems and communication strategies. The game used in the user study is explained in section 3. The explanation of the AI agent's underlying trust mechanism is located in section 4. Section 5 describes the methods of the user study. The following section discusses the obtained results. Section 7 further interprets the findings and reflects on the research limitations, as well as suggests future work. In section 8, ethical issues associated with the research are discussed. The last section lays out the conclusions.

2 Background

With the ongoing advancements in technology, artificial intelligence is transitioning from being a tool to being recognized as a human partner in team environments [14]. This transition necessitates more research into the factors influencing trust and its benefits not only in human-human teams but also in human-AI teams. Assessing the factors that influence trust in collaborative teams is necessary, as research shows that mutual trust impacts efficiency, safety, and overall success in joint operations [15], [3].

2.1 Trust-based Mental Models

Mutual trust is viewed as both members taking on the roles of trustor and trustee for each other. In human-AI teams, mutual trust is composed of two types of trust: artificial trust from the AI towards humans (e.g., the robot as the trustor and the human as the trustee) and natural trust from humans towards the AI agent [10]. To express trust beliefs, a model of internal characteristics needs to be defined. This model will be further referred to as a “mental model”.

Trust can be considered a multidimensional construct, and various models for expressing trust have been proposed. The ABI model suggests trust is based on ability, benevolence, and integrity [24]. A literature-based trust model identifies other four dimensions: reliability, capability, sincerity, and ethics [17]. Agents can model such internal characteristics using behavioral cues from their teammates [11].

The Socio-Cognitive model proposes two basic beliefs of trust: competence and willingness [4]. This model defines competence as the trustor’s evaluation of a trustee’s ability to perform a task, while willingness is expressed by the trustor’s belief that the trustee’s intentions towards a task align with their own. Willingness not only refers to the truthfulness of the human (i.e., lying behavior) but also to the human tendency to prefer certain tasks over others. Research shows that humans perceive more challenging tasks as less engaging [20]. Therefore, the perceived difficulty of a task might influence a human’s willingness to perform it. To conclude, this model is used because it is simple, modeling just two factors while still capturing the essence of natural trust. Moreover, as it is simplistic, it is easy to aggregate into a single value for visual communication.

2.2 The Relationship between Communication and Trust & Satisfaction

A recent taxonomy suggests ways to characterize the context in which an artificial agent needs to assess trust in human-AI teams based on two aspects: task and team configuration [12]. Notably, team configuration includes key concepts such as communication and shared knowledge. Therefore, communication is recognized as an environmental factor that influences trust. Moreover, constructive communication can increase shared knowledge within a team, which is an important factor in trust relationships. In that sense, different communication approaches and their impact on natural trust and satisfaction have been studied [16, 30].

Research indicates that effective communication is essential for supporting both team cognitive [6] and affective processes [23]. Moreover, research has shown that communication has a positive impact on job satisfaction [7]. Various communication approaches have been studied in the context of human-AI collaborative teams to assess their impact on multiple factors that affect human-AI teaming. A study shows that among visual, audio, and verbal communication strategies, the visual approach enhances the most collaboration between humans and automation [21].

The timing of communication has been shown to influence team dynamics, with proactive communication (e.g. immediate responses) from AI teammates increasing trust [30]. Furthermore, the content of communication is critical. A study partially validates the hypothesis that systems are perceived as more trustworthy when using environment-based justifications, compared to policy-centric content [16]. Environment-based justifications rely on contextual factors such as current conditions and external influences, whereas policy-based justifications focus on the outcomes and performance metrics of the decisions made. Despite these findings, there remains a research gap regarding the explicit communication of artificial trust beliefs to humans and its impact on natural trust and overall satisfaction.

3 The Game

This section describes the game used in the user experiment. The game is based on the Urban Search and Rescue Scenario [1], and its objective is to rescue six victims. It involves two agents: the human agent (played by the participant) and RescueBot, an artificial agent. The player navigates within a 2D grid world depicted in Figure 1. However, in the actual game, the user’s vision is limited to a 2-cell range, making the location of each victim/obstacle unknown at the beginning.

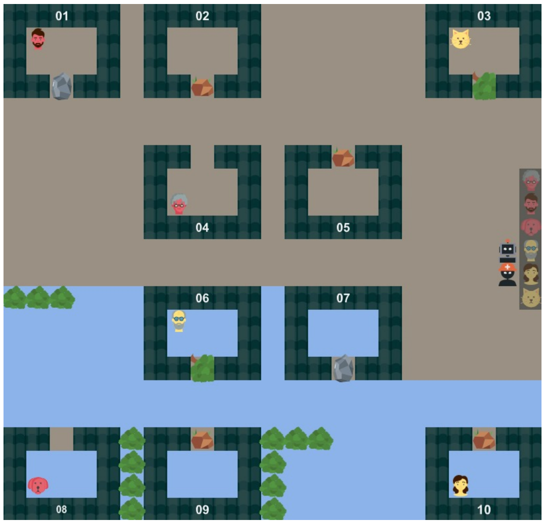

Figure 1: Overview of the game environment displaying all obstacles and victims.

Eight room entrances are blocked by obstacles, some of them necessitating collaboration between agents for removal. A task that requires both agents to complete will be further referred to as a “collaborative” task. Some tasks can only be completed by the RescueBot, creating a need for the human to communicate and request the robot’s assistance. Conversely, the robot can inquire about the human’s assistance. The human can also inform the robot which rooms have been searched to avoid double-checking and optimize the search process. This setup emphasizes the need for collaboration and communication between the human and the AI agent. The communication with the robot is accessible through a chatlike interface, which has predetermined buttons corresponding to actions within the game.

There are two types of victims in the game, distinguishable by color: red (meaning critically injured) and yellow (for mildly injured). Rescuing a critically injured victim is a collaborative task. Among the victims, one additional aspect was added to better asses human willingness: there are two elderly individuals whose rescue time is increased compared to others. The game includes three types of obstacles: big rocks, trees, and small stones. Removing big rocks is collaborative, while trees can only be removed by the RescueBot. Removing small stones can be performed by either agent independently, but it takes less time if both agents work together to remove them.

Another aspect of interest is the terrain. Blue zones are "flooded" and significantly reduce the speed of any agent traversing them. The research team added these to introduce an additional factor that might influence users' preferences in the game. The RescueBot does not only follow the human's instructions but also performs tasks independently. The game has a preset time limit of 10 minutes, ending either when all six victims are rescued or when the time limit is reached. The user is informed by all these rules through a tutorial, available before the game.

4 Artificial Agent

This section explains the trust mechanism employed by the AI agent. To introduce the notion of trust in the experiment, the AI agent employs a mental model of artificial trust beliefs towards the human to make decisions. The mental model is dynamically updated throughout the game's duration, based on certain human behavioral cues and environmental factors.

4.1 Mental Model

The chosen mental model uses two factors to model trust: competence and willingness. In the context of this experiment, competence is defined as the participant's cognitive ability to complete the game's goal and make rational decisions to optimize the rescuing process. Willingness refers to the human's intention to complete an objective and to collaborate with the robot. Recent research suggests that trust is a time-dependent dynamic variable, that necessitates modeling along the collaboration timeline [27]. This can be done by taking into account observed human behavior and quantifying each action as an increase or decrease in willingness and competence.

As trust is context-dependent [12], different trust values need to be computed for each type of task comprised by the game. Given the time constraints of the experiment, assigning trust values for each task would not be suitable. The low number of interactions for the same type of task (e.g., removing a big rock) would not allow for proper trust calibration. For instance, imagine considering different trust values for removing each type of obstacle. There are a total of two rocks in the game, meaning that the maximum number of times trust would be evaluated for that task is two. This would not allow the perceived trust value by the model to converge towards a true value, as the total number of interactions with that task is too small. Therefore, all tasks were aggregated into three different task types: searching rooms, removing obstacles, and rescuing victims.

To denote trust towards a specific task **t**, we use T_{t} for trust for task **t**, W_{t} for willingness towards **t**, and C_{t} for competence regarding **t**, with t in the domain D = {search, obstacle, vicims}. Trust can then be represented as a **t**x2 matrix:

$$\mathbf{T} = \ (T_{\text{search}}, T_{\text{obstacles}}, T_{\text{victims}}) = \begin{array}{l} ((W_{\text{search}}, C_{\text{search}}), \\ (W_{\text{obstacles}}, C_{\text{obstacles}}), \\ (W_{\text{victims}}, C_{\text{victims}})) \end{array}$$

To adjust these values, we define **I** as an increase/decrease factor, with values in the domain V = {0.1, 0.2, 0.4}. Moreover, we maintained a strict interval for all trust values, ensuring that regardless of the task, all willingness and competence values are clipped to $[-1, 1]$. The initial values for any W_{t} and C_{t} are 0, and correspond to the center of the interval. Following this, we quantify the willingness and competence value through time for each task. Table 1 shows willingness and competence values at time **t'** for all considered cases, detailed in subsections 4.1.1, 4.1.2 and 4.1.3. The **p** value present in the table refers to a preference factor, explained in subsection 4.2.

4.1.1 Formalizing Trust regarding Searching for Victims

Three cases of human actions provide insight into human intentions and capabilities regarding searching a room. **Case 1** consists of the human informing the robot that they will search a new room, a sign of willingness and competence. **Case 2** generally corresponds to the human acting incorrectly accidentally (e.g., searching a room that was already searched, forgetting to announce a room they searched in or double-pressing buttons), thus resulting in a competence decrease. **Case 3** addresses the situations in which the human lies about searching a room, and the robot discovers victims or obstacles at that location. In this case, the increase from Case 1 is subtracted, and an additional decrease is applied to mark the action.

4.1.2 Formalizing Trust regarding Rescuing Victims

To quantify the artificial trust towards the human in the context of rescuing a victim, we define four different cases. **Case 4** addresses situations where the human lies or is unable to complete the actions they communicated to the robot (e.g., falsely claiming to rescue a victim, providing incorrect information about the location of a victim, or failing to perform the rescue within a predefined time interval). **Case 5** is the opposite, relating to truthfully beneficial actions performed by the human (e.g., rescuing a victim, finding a victim and announcing this to the robot, or coming to the rescue at the robot's request within a predefined interval). **Case 6** refers to the human not responding to the robot's request, indicating a lack of intention for collaboration and decreasing willingness. Conversely, **Case 7** corresponds to the human responding to the robot's messages regarding rescuing a victim, thereby increasing willingness. Table 1 showcases this, where **I** is determined by the type of victim involved in the action (higher values for critically injured victims, and lower for mildly injured victims).

Table 1: Formalization of trust regarding searching for victims, rescuing victims, and removing obstacles

Case	Willingness	Competence	Increase/Decrease Factor (I)
1	$W_{\text{search}}(t'-1)+I+p$	$C_{\text{search}}(t'-1)+I$	0.2
2	$W_{\text{search}}(t'-1)+p$	$C_{\text{search}}(t'-1)-I$	0.1
3	$W_{\text{search}}(t'-1)-I-I+p$	$C_{\text{search}}(t'-1)-I-I$	0.2
4	$W_{\text{victim}}(t'-1)-I+p$	$C_{\text{victim}}(t'-1)-I$	0.2 or 0.4 depending on victim type
5	$W_{\text{victim}}(t'-1)+I+p$	$C_{\text{victim}}(t'-1)+I$	0.2 or 0.4 depending on victim type
6	$W_{\text{victim}}(t'-1)-I+p$	$C_{\text{victim}}(t'-1)$	0.2 or 0.4 depending on victim type
7	$W_{\text{victim}}(t'-1)+I+p$	$C_{\text{victim}}(t'-1)$	0.2 or 0.4 depending on victim type
8	$W_{\text{obstacle}}(t'-1)-I+p$	$C_{\text{obstacle}}(t'-1)-I$	0.1, 0.2 or 0.4 depending on obstacle type
9	$W_{\text{obstacle}}(t'-1)+I+p$	$C_{\text{obstacle}}(t'-1)+I$	0.1, 0.2 or 0.4 depending on obstacle type
10	$W_{\text{obstacle}}(t'-1)-I+p$	$C_{\text{obstacle}}(t'-1)$	0.1, 0.2 or 0.4 depending on obstacle type
11	$W_{\text{obstacle}}(t'-1)+I+p$	$C_{\text{obstacle}}(t'-1)$	0.1, 0.2 or 0.4 depending on obstacle type

4.1.3 Formalizing Trust regarding Removing Obstacles

Trust values concerning obstacle removal can be similarly calculated, following the same four-case structure for rescuing victims. **Case 8** corresponds to cues such as falsely claiming they will help the robot remove an obstacle or incorrectly pinpointing an obstacle's location. **Case 9** corresponds to successfully eliminating obstacles or proactively guiding the robot. **Cases 10** and **11** refer to not responding or responding to the robot's messages, respectively. The increase or decrease in trust, **I**, depends on the obstacle type: 0.1 for stones, 0.2 for trees, and 0.4 for big rocks.

4.2 Modelling Environmental Factors into Trust

We integrated preference modeling to assess more extensively the willingness of the participants for each task. Given that research indicates that more challenging tasks may be perceived as less engaging for users [20], it raises the question of whether willingness levels should also account for environmental factors, rather than solely focusing on user actions. Hence, we integrated a preference factor that was added or subtracted depending on the task. In modeling these preferences, we considered that tasks with slower speeds and longer distances would increase the difficulty.

Accordingly, the game map was divided into two halves: the upper part contained normal terrain, while the lower part was "flooded" and slowed down the user's movement. The assumption was that users would prefer to navigate on normal terrain. Therefore, opting to perform a task in a flooded area, such as searching a flooded room, has an additional increase determined by a preference factor, denoted as **p**. However, terrain speed was not the only environmental factor considered. Choosing to rescue an elderly victim, who requires more timely assistance, resulted in an additional increase determined by **p**. Conversely, choosing not to rescue (by not responding or coming) decreased willingness by the preference factor. Lastly, this factor also considered the distance from the human to the robot and was applied in all cases where the human was called by the robot to come to its location. The preference factor was computed using the formula:

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

The **f** value reflects the task's desirability concerning flooded areas: it is set to 1 if the task leads to a non-flooded area, 0.5 if it remains in a flooded area where the human was previously located, and 0 if it brings the human into a flooded area. The distance **d** quantifies the task's appeal based on proximity: it is calculated as **1 - distance/diagonal**, where **distance** denotes the agent-human distance and **diagonal** represents the distance along the main diagonal of the environment grid. Finally, **v** evaluates the task's attractiveness considering the victim's condition: it is assigned 0 if the victim is elderly and 1 otherwise. The weights w_{f}, w_{d}, w_{v} are used to indicate the presence of the **f**, **d**, and **v** values in the task. Thus, when they are not relevant, the weight is set to 0. Lastly, before adding **p** to the trust belief values, it gets normalized by division by 5. This step ensures the balance between environmental and behavioral factors.

4.3 Behaviour Adaptation

The mental model is used to determine the robot's actions and level of cooperation with the human. This is done to mimic as much as possible normal human behavior when trust influences the attitude toward a teammate. Additionally, this optimizes the rescuing process, as a robot accepting collaboration with an untrustworthy human would slow down the process. Therefore, when a robot considers that the human is not to be trusted and finds a victim that it can carry without help, it will rescue automatically, without asking for permission from the human. The same applies to the robot not trusting the human and finding a removable obstacle. Moreover, when the human announces they found a mildly injured victim, the robot will automatically go and rescue the victim to ensure task completion.

4.3.1 Confidence in Own Trust Beliefs

Additionally, after the preference factor is integrated into (W_{t}, C_{t}), the values obtained are input to a confidence function. The function is designed to adjust the confidence level (i.e. how much the robot trusts its own decisions) associated with a particular task type based on changes in trust beliefs over time. It maintains a history of trust beliefs for the given task type and evaluates whether there have been monotonic increases or decreases in these beliefs over a specified number of previous samples.

If the trust beliefs have been consistently increasing or decreasing, the confidence level is incremented, reflecting increasing confidence in the reliability of the observed trend. Conversely, if the trust beliefs exhibit fluctuating or inconsistent patterns, the confidence level may be adjusted downwards to reflect uncertainty in the underlying trust dynamics. Additionally, it applies a clipping mechanism to constrain the confidence values within the range [0, 1], preventing them from exceeding these boundaries.

4.3.2 Decision Thresholds

To determine whether the human is trusted or not, first the confidence function output is compared to the random number between 0 and 1. After that, the current trust values (W_t, C_t) are compared to the (0, (1-p) / 2) threshold. The threshold values were determined by the center of the interval within which trust is defined $[-1, 1]$. To integrate preferences into the decision-making process, we added the preference factor to the willingness threshold. If the current (W_t, C_t) values exceed these thresholds, it indicates trust in the human. Thus, all three following conditions must be true for the robot to trust the human, where u $\sim \mathcal{U}(0, 1)$:

$$\text{u} < \text{confidence} \tag{1}$$

$$\text{W}_\text{t} \geq \frac{1 - \text{p}}{2} \quad \text{and} \quad \text{C}_\text{t} \quad \geq 0 \tag{2}$$

4.4 Artificial Trust Communication

To study the contribution of real-time communication, a communication method for the artificial trust mental model of the robot was designed. Figure 2 depicts the chosen design. The y-axis is used to depict trust, while the x-axis depicts the task type. A bar chart with three bars was used to depict the trust values corresponding to each type of task. Initially, the design used six bars—two bars for each task, representing willingness and competence. However, human-computer interaction heuristics show that complex visualization methods can cognitively overload the message receiver [19]. Therefore, willingness and competence were aggregated into a single value per task. This was achieved by summing the values and mapping them from the range $[-2, 2]$ to $[-1, 1]$ by dividing by 2.

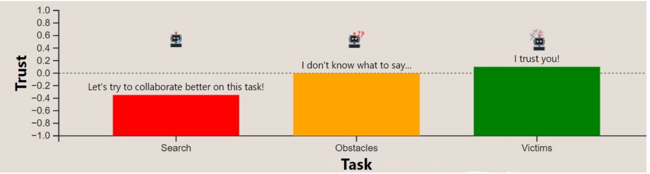

Figure 2: Design of the real-time, visual communication of the artificial trust mental model

The bars reflected the values to the user in real-time, allowing them to see the consequences of their actions and learn which actions led to increasing or decreasing artificial trust. For better visualization, a color-coding scheme was used: green bars indicated trust values greater than 0.05, red indicated values lower than -0.05, and orange indicated values between -0.05 and 0.05. Explanatory labels, following the same rules, were displayed on the top of each bar.

The domain interval was chosen based on the sum of the actual thresholds used for behavior adaptation. Thus, an indicative threshold was also added at value 0 to mark this. Consequently, orange indicated the transition between the other two states, signaling to the participant that the robot's beliefs would quickly change; green indicated a favorable trust value, while red indicated a non-favorable one.

5 Methods

This study aims to help answer the question: "How does a real-time visual (RTV) communication of the mental model of the agent's trust affect the human teammate's trust in the agent and overall satisfaction?". The user study embodies an independent measures experimental design to avoid order effects. The independent variable is the presence of the communication of the mental model formed by artificial trust beliefs to the human. Two dependent variables are studied: natural trust and overall satisfaction after playing the game. The objective was to test the two research hypotheses:
H1: Visual, real-time communication of artificial trust beliefs increases human trust in the AI agent.
H2: Visual, real-time communication of artificial trust beliefs increases overall human satisfaction.
The experiment took place in person, with participants engaging in a collaborative game with an AI agent and completing two questionnaires under the guidance and supervision of the researcher.

5.1 Participants

This research recruited 44 participants using the researchers' social networks. 22 participants were assigned to the baseline condition and 22 to the communication of trust condition. Participants were selected based on their availability and willingness to participate in the study. The most frequent age group was 18-24 (43 participants), but the ages ranged from 18 to 44 years old. All participants reside in Europe. 27 identified as male and 17 as female. 10 participants were Master's students, 33 were Bachelor's students, and one was an HBO student. 36 reported that they majored in the Computer Science field. 17 also indicated familiarity with Matrx. Most participants (20 participants) reported having moderate gaming experience, with 14 participants having a lot of experience and six participants having little experience. Only four participants reported having no gaming experience at all.

5.2 Materials

The game used in the experiment employs MATRX 2.2.0, a Python package designed for creating human-AI teaming simulations. The research team was provided with a code base containing a 2D grid-world game, which was subsequently extended. This initial boilerplate code contained no implementation of a mental model and served as a basic structure. The game was hosted locally on laptops running Windows and MacOS operating systems.

5.3 Measurements

To assess the impact of the communication strategy on natural trust and overall human satisfaction, the two factors were measured to enable later comparison between experimental conditions. Subjective metrics were obtained using a questionnaire and additional open-ended questions to evaluate both variables. Objective metrics captured participants' behavior related to trust.

5.3.1 Subjective Measurements

Both dependent variables were measured through a **post-experiment questionnaire**. The questionnaire combined two pre-validated surveys, one measuring trust and the other measuring satisfaction. Moreover, the questionnaire concludes with 4 open-ended questions.

To measure trust in the context of explainable AI, the method must be capable of detecting the emergence of distrust and mistrust [9]. Therefore, Hoffman et al. (2023) deconstructed and filtered existing scales to select appropriate items for measuring trust. An adapted version of the "Trust Scale for the XAI Context" was used to measure trust. This scale consists of 8 items, that can be answered in a Likert scale format [9].

To measure satisfaction in collaborative human-AI teaming, it is crucial to consider the degree to which users feel they sufficiently understand the system [9]. To gather participants' opinions regarding their satisfaction with communicating with RescueBot, the "Explanation Satisfaction Scale" proposed by Hoffman et al. (2023) was adapted and used. It consists of seven items, that can be answered in a Likert scale format [9].

The questionnaire concluded with four optional open-ended questions. These questions were added to gain deeper insights into the participants' opinions, as answering standard questionnaires might limit their input. The first question inquired about what was missing in the collaboration setup presented to the user. Questions two and three asked participants to pinpoint the most and least preferred aspects of collaborating with the AI agent. The final question explored the participants' perceptions of the artificial trust beliefs towards them and the feelings these perceptions generated.

5.3.2 Objective Measurements

Self-reported measurements might reflect personal bias and the participant's immediate considerations when taking the survey, which could slightly differ from their feelings at the moment of action [13]. Therefore, the use of objective measures to complement the subjective measures is suitable to measure trust [13]. Data collection included behavioral logs of the participants' observable actions to measure trust throughout or at the end of each game.

Compliance, defined as the number of times a user follows the AI's suggestions [13], is proposed by A. Krausman et al. as an objective measure of trust. In the context of this experiment, compliance was defined by the number of times the human respected the RescueBot's suggestions of coming to help at a certain location. Other behavioral indicators of trust included the **ratio of jointly and independently completed actions**, based on the assumption that a higher frequency of collaboration indicates a greater level of trust [26]. Additionally, the **number of messages sent by participants** was measured, with the assumption that more communication might reflect higher trust. The final values of **artificial trust** per participant were also logged. Although not a specific metric, performance is measured as the **game duration**, or the time taken to finish the task, as research indicates a link between trust and performance [28].

5.4 Procedure

The entire procedure lasted approximately 25 minutes per participant. The first step involved participants reading and signing the informed consent and ethics review checklist forms, proposed by the TU Delft Human Research Ethics Committee. Following this, participants completed an anonymized survey to collect potential confounding factors, including age group, region of residence, gender, highest level of education completed, prior knowledge of MATRX, gaming expertise, and whether they majored in the Computer Science field. Then, participants were given a scripted explanation of the concept of artificial trust. This was followed by a game tutorial to familiarize them with the game rules. After the tutorial, participants played the game. Finally, they completed a post-experiment questionnaire, which measured the two dependent variables: trust and overall satisfaction.

6 Results

This section presents the results of the user study. Two groups were compared: baseline and real-time visual (RTV) communication. The predictor variable is categorical, and the outcome variable is quantitative. Therefore, the Independent Sample t-test and the Mann-Whitney U test were considered to compare the average scores per metric. Each dataset was checked for normality using the Shapiro-Wilk test and for homogeneity of variances using Levene's test. Since no metric satisfied both the normality and homogeneity of variances conditions, the Mann-Whitney U test was exclusively used for the comparisons. The significance level was set at 0.05.

6.1 Subjective and Objective Measurements

Table 2 provides the median and interquartile range (IQR) for each metric within each group. Additionally, the table reports the p-values, z-scores, and U statistics for comparisons between conditions for each metric. Statistically significant results are denoted with an asterisk ($p < .05$)).

A Mann-Whitney U test was conducted to compare satisfaction levels between the Baseline and real-time visual (RTV) communication conditions. The median satisfaction score for the Baseline condition was 3.71 (IQR = 1.29), whereas the median satisfaction score for the RTV condition was 4.14 (IQR = 1). The results indicated a significant difference between the two conditions, U = 146, z = -2.24, p = .0251. This suggests that the RTV condition resulted in significantly higher satisfaction levels compared to the Baseline condition.

A Mann-Whitney U test was conducted to compare game duration between the Baseline and RTV conditions. The median game duration for the Baseline condition was 4877.5 seconds (IQR = 899), whereas the median game duration for the RTV condition was 5258.5 seconds (IQR = 629). The results indicated a significant difference between the two conditions, U = 154, z = -2.05, p = .0403. This suggests that the

Table 2: Summary of Median and Standard Deviation for each measurement and condition (Baseline and Real-time Visual Communication), along with Mann-Whitney U Test results comparing the two conditions.

Type	Measurement	Condition	Median	IQR	p-value	z-score	U
Subjective	Trust	Baseline	3.68	1	0.10524	-1.61961	172.5
		RTV	3.87	0.75			
	Satisfaction	Baseline	3.71	1.29	**0.0251***	-2.24164	146
		RTV	4.14	1			
Objective	**Game duration**	Baseline	4877.5	899	**0.0403***	-2.05385	154
		RTV	5258.5	629			
	Compliance	Baseline	2.5	2	0.88866	-0.14084	235.5
		RTV	3	2			
	Collaboration frequency	Baseline	0.67009	0.23263	0.5157	0.6455	214
		RTV	0.6207	0.2812			
	Number of human messages	Baseline	18.5	6	0.11642	-1.57267	174.5
		RTV	20.5	7			
	Artificial trust	Baseline	4.65	1.5	0.34722	-0.93891	201.5
		RTV	5.066868	1.2			

RTV condition resulted in significantly longer game durations compared to the Baseline condition.

6.2 Correlations

Testing the correlation between gender and expertise in computer science with trust and satisfaction was performed using a Chi-square non-parametric test. Trust values were split into two categories: values smaller than or equal to 3, and values greater than 3. This was necessary because one assumption for the Chi-square test was not met: the expected value of cells should be 5 or greater in at least 80% of cells. However, this analysis did not yield any significant results.

7 Discussion

This section interprets the results concerning the main research question, evaluating how a real-time visual communication strategy of a mental model of artificial trust impacts human trust and overall satisfaction. Additionally, the limitations of the study and potential future work are discussed.

7.1 RTV Communication Impact on Natural Trust

The reported results do not validate hypothesis **H1**, which proposed that RTV communication of the artificial trust mental model increases natural trust. The only significant result was the game duration measurement, which indicated that RTV communication led to a longer task completion time. Table 2 shows that, although statistically significant, the p-value is relatively high, indicating weak evidence for this conclusion. Moreover, while research suggests a link between performance and trust, this sole metric cannot validate the hypothesis.

Considering this, we discuss possible factors that might have impacted performance. Firstly, the RTV condition was fully deployed on laptops running Windows as an operating system, whereas the Baseline condition included both MacOS and Windows. Participants using MacOS were possibly faster due to the higher frame rate, while Windows users experienced lag. Secondly, the need for participants to process additional information, even when simplified through a visual representation, might have increased cognitive load, leading to longer completion times.

Furthermore, the high ratio of participants majoring in Computer Science could explain why no significant changes in trust between conditions were reported, though overall satisfaction improved. For individuals who understand the underlying mechanisms of AI, trust might be influenced more by observable actions than by communication. This is reflected in the participants' responses; when asked what they think the agent thinks of them, three participants indicated that they do not believe the agent literally ”thinks” about them. This suggests that they recognize the agent operates based on predetermined logic and cannot form its own decisions.

Previous work suggests that people outside the computer science field often view AI as a black box, leading to distrust [25]. RTV communication of the mental model adds an extra layer of transparency, which is needed to mitigate the black box effect that leads to distrust. Greater knowledge fosters greater understanding, and thus the additional communication may not be necessary to increase trust in this case. However, satisfaction may be influenced by communication, as it might be perceived as a new feature of the agent.

7.2 RTV Communication Impact on Overall Satisfaction

We discuss the findings regarding the impact of RTV communication on overall human satisfaction in a human-AI collaborative context. The results of the subjective measurements of overall satisfaction confirm the second research hypothesis (**H2**), that real-time visual communication of the artificial trust mental model increases overall human satisfaction. This conclusion is supported by some participants' responses to the post-experiment questionnaire open questions.

When asked what they think the RescueBot thinks of them and how that makes them feel, one participant indicated that “from the trust metrics the bot seems to have a very good opinion of me” and that this makes them feel “happy.” This statement explicitly shows that the additional communication enhances their overall understanding and provides insight into the artificial agent's rationale, positively impacting their satisfaction with the collaboration. This aligns with previous research discussed in subsection 2.2 that associated enhanced communication with human satisfaction [7].

7.3 Limitations

This section outlines the limitations encountered during the development of the study. Firstly, the use of different computers led to varying game experiences, which might have introduced bias. For example, participants who played the game on a MacOS laptop might have finished the game sooner due to the enhanced visibility provided by a higher frame rate. This discrepancy not only offered a performance advantage but also caused frustration for some participants, potentially impacting their overall satisfaction or trust.

Secondly, the study involved 44 participants, divided into two conditions, resulting in 22 participants per condition. This limitation was due to time constraints; however, an increased sample size of 40 participants per condition would be advisable [2]. Additionally, the diversity of the sample also has an impact. As mentioned in section **??**, only one participant was outside the 18-24 age group, all participants resided in Europe, and 36 participants studied Computer Science. Greater diversity in the sample would help reduce bias.

Finally, the environmental setup posed its restrictions. Studying trust and overall satisfaction in a fixed environment is a limitation. For example, varying the number of agents or assigning them different goals might have yielded different results concerning trust and overall satisfaction.

7.4 Future Work

For future work, it would be beneficial to fine-tune the hyperparameters used in the mental model, specifically the willingness and competence thresholds, as well as the increase factor I. This would allow for a better assessment in scenarios where distrust is appropriate. The current values were selected through manual testing, but they could be optimized further.

Additionally, it would be valuable to compare the effect of RTV communication on the trust and overall satisfaction of computer scientists versus non-computer scientists. This comparison could provide deeper insights into making AI more explainable and trustworthy for individuals who do not understand the underlying mechanisms.

8 Responsible Research

This section reflects on the ethical aspects of this research and discusses the reproducibility of the experiments. Current research raises awareness of potential ethical issues regarding human-AI collaboration. To address these issues, I examined the possible risks that the study poses to the participants, guided by the "Ethics review checklist" proposed by the Human Research Ethics Committee at TU Delft.

The primary issues of the user study concern privacy and data security. To address these concerns, a thorough risk assessment was conducted. The identified issues involved the collection, processing, and/or storage of directly identifiable PII (Personally Identifiable Information) and PIRD (Personally Identifiable Research Data). However, we believed there was no significant risk associated with these practices. Any PII was collected separately through an informed consent form, which was only accessible to the research team. Additionally, we believed there was no risk associated with collecting PIRD data, as it was solely used to describe samples and was anonymized. Considering these assessments and the fact that participants were provided with an informed consent form outlining the study's requirements and any potential risks, the study was approved by the Human Research Ethics Committee of TU Delft.

Reflecting on the method's reproducibility, Odd Erik Gundersen argues that reproducibility does not only refer to following the same experimental methods but also to achieving the same results [8]. It is crucial to highlight the subjective aspect of the research: the variability of human opinions. Although one might use identical tools (same code, same questionnaires, etc.), responses can differ due to psychological, social, and contextual factors between individuals.

Even the same person might experience the same game differently at different times, influenced by their mood, external circumstances, or prior experiences. Consequently, this variability poses challenges not only to the reproducibility of such experiments but also to their applicability. Regardless of the research team, achieving consistent results across different sets of participants may be unattainable.

9 Conclusion

This study examined how real-time visual (RTV) communication of an AI agent's trust mental model impacts human trust and overall satisfaction. We implemented a mental model of artificial trust beliefs based on two aspects: competence and willingness, while also considering environmental factors. Specifically, the research aimed to determine whether an RTV communication of such a model enhances the natural trust that human teammates place in the AI agent and their overall satisfaction.

Using an independent measures design, 44 participants were divided into two groups: one experiencing RTV communication of the robot's mental model, and a control group without this communication. Participants collaborated with an AI agent in a 2D grid-based Urban Search and Rescue game. The impact was assessed through both subjective measures (surveys and open-ended questions) and objective measures(game duration, compliance, collaboration frequency, communication volume, and final artificial trust value).

The results did not support hypothesis **H1**, which proposed that real-time visual (RTV) communication of the artificial trust mental model would increase natural trust. The only significant finding was an increase in task completion time, indicating that RTV communication led to longer game durations. Although this result was statistically significant, the relatively high p-value suggests weak evidence, and that metric is not a standalone argument for denying the hypothesis.

Several factors may have influenced these results. The RTV condition was fully run on laptops using the Windows operating system, whereas the Baseline condition included both MacOS and Windows users. MacOS users experienced higher frame rates and smoother game performance, thus Windows users' performance was impacted. Additionally, the need for participants to process extra information, even when simplified through visual representation, might have increased cognitive load, leading to longer completion times.

Furthermore, the high proportion of participants pursuing Computer Science could explain the lack of significant changes in trust between conditions. For individuals with a strong understanding of AI mechanisms, trust may be influenced more by observable actions than by communication. Prior research suggests that people outside the computer science field often view AI as a black box, leading to distrust. The RTV communication of the mental model adds transparency that may minimize this effect on less knowledgeable users but may not significantly impact those already familiar with AI.

However, results found satisfaction to be influenced by the RTV communication. This is because, as opposed to the previous case, communication might be perceived as an extra feature by people familiar with the AI field and thus influence satisfaction in both cases of participants. This was reflected in the subjective measurements, as statistically more participants were satisfied with the extra communication condition.

In conclusion, while RTV communication did not significantly enhance trust, it did improve overall satisfaction. Future work could explore larger and more diverse samples, adjust the experimental setup to ensure uniform performance conditions, and further investigate the impact of the real-time communication of an artificial trust mental model on different user groups.

References

[1] Joseph A Barbera and Anthony Macintyre. Urban search and rescue. *Emergency Medicine Clinics*, 14(2):399–412, 1996.

[2] Raluca Budiu and Kate Moran. How many participants for quantitative usability studies: A summary of sample-size recommendations. *Nielsen Normal Group*, 2021.

[3] Ralph Chami and Connel Fullenkamp. Trust and efficiency. *Journal of banking & finance*, 26(9):1785–1809, 2002.

[4] Rino Falcone and Cristiano Castelfranchi. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pages 740–747. IEEE, 2004.

[5] Michael Floyd, Michael Drinkwater, and David Aha. Trust-guided behavior adaptation using case-based reasoning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[6] Susan R Fussell, Robert E Kraut, F Javier Lerch, William L Scherlis, Matthew M McNally, and Jonathan J Cadiz. Coordination, overload and team performance: effects of team communication strategies. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 275–284, 1998.

[7] Vijai N Giri and B Pavan Kumar. Assessing the impact of organizational communication on job satisfaction and job performance. *Psychological Studies*, 55:137–143, 2010.

[8] Odd Erik Gundersen. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, 379(2197):20200210, 2021.

[9] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[10] Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. Artificial trust for decision-making in human-ai teamwork: Steps and challenges. In *Proceedings of the HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI)*, 2023.

[11] Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. How should an ai trust its human teammates? exploring possible cues of artificial trust. *ACM Transactions on Interactive Intelligent Systems*, 14(1):1–26, 2024.

[12] Carolina Centeio Jorge, Emma M van Zoelen, Ruben Verhagen, Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. Appropriate context-dependent artificial trust in human-machine teamwork. In *Putting AI in the Critical Loop*, pages 41–60. Elsevier, 2024.

[13] Andrea Krausman, Catherine Neubauer, Daniel Forster, Shan Lakhmani, Anthony L Baker, Sean M Fitzhugh, Gregory Gremillion, Julia L Wright, Jason S Metcalfe, and Kristin E Schaefer. Trust measurement in human-autonomy teams: Development of a conceptual toolkit. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(3):1–58, 2022.

[14] Lindsay Larson and Leslie A DeChurch. Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The leadership quarterly*, 31(1):101377, 2020.

[15] Michael Lewis, Katia Sycara, and Phillip Walker. The role of trust in human-robot interaction. *Foundations of trusted autonomy*, pages 135–159, 2018.

[16] Matthew B Luebbers, Aaquib Tabrez, Kyler Ruvane, and Bradley Hayes. Autonomous justification for enabling explainable decision support in human-robot teaming. *Proceedings of Robotics: Science and Systems. Daegu, Republic of Korea. https://doi.org/10.15607/RSS*, 2023.

[17] Bertram F Malle and Daniel Ullman. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*, pages 3–25. Elsevier, 2021.

[18] D Harrison McKnight and Norman L Chervany. What is trust? a conceptual analysis and an interdisciplinary model. 2000.

[19] Jakob Nielsen. Ten usability heuristics. 2005.

[20] Heather L O'Brien, Jaime Arguello, and Rob Capra. An empirical study of interest, task complexity, and search behaviour on user engagement. *Information Processing & Management*, 57(3):102226, 2020.

[21] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *Plos one*, 15(2):e0229132, 2020.

[22] Caroline Player and Nathan Griffiths. Improving trust and reputation assessment with dynamic behaviour. *The Knowledge Engineering Review*, 35:e29, 2020.

[23] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. Let's think together! assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–29, 2022.

[24] Helge Svare, Anne Haugen Gausdal, and Guido Möllering. The function of ability, benevolence, and integrity-based trust in innovation networks. *Industry and Innovation*, 27(6):585–604, 2020.

[25] Warren J Von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34(4):1607–1622, 2021.

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

[27] X Jessie Yang, Christopher Schemanske, and Christine Searle. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, 65(5):862–878, 2023.

[28] Akbar Zaheer, Bill McEvily, and Vincenzo Perrone. Does trust matter? exploring the effects of interorganizational and interpersonal trust on performance. *Organization science*, 9(2):141–159, 1998.

[29] Rachid Zeffane, Syed A Tipu, and James C Ryan. Communication, commitment & trust: Exploring the triad. *International Journal of Business and Management*, 6(6):77–87, 2011.

[30] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. Investigating ai teammate communication strategies and their impact in human-ai teams for effective teamwork. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–31, 2023.