Communicating trust-based beliefs and decisions in human-AI teams using real-time visual explanations

Elena Dumitrescu[1]
Supervisors: Myrthe 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 Dumitrescu
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Carolina Centeio Jorge, Ujwal Gadijaru

Abstract

To collaborate effectively, humans and AI agents need to trust each other. Communication between teammates is an essential component to achieve this, as it makes the AI system more understandable to humans. However, previous research lacks a focus on ways to build an AI agent's trust in its human teammate and, consequently, on how the AI's beliefs can be communicated to the human. As such, this study explores how real-time visual explanations of the AI agent's trust in its human teammate influence human trust and overall satisfaction. Through a user experiment (n=46) conducted on an Urban Search and Rescue simulation, integrating trust explanations was compared against a baseline containing no such information. Results show a statistically significant increase in both human trust and satisfaction when the explanations are provided, highlighting the need for further exploration into methods of communicating trust.

1 Introduction

Artificial agents are increasingly designed to collaborate with humans, by assisting them in tasks and improving efficiency. In such a setting, the AI and human agents work together towards a common goal, forming human-AI teams (HATs). Mutual trust is essential in this interaction [1]: humans need to trust artificial agents to collaborate effectively (natural trust), while artificial agents should incorporate trust into their decision process to determine how and with whom to engage [2] (artificial trust). Thus, it is imperative to research ways of consolidating the human-AI trust from both perspectives, i.e. both the natural and the artificial trust.

To build artificial trust, the AI agent should be able to perceive human characteristics and assess whether they are a cue for trustworthiness [3]. This requires the AI to construct an accurate mental model of the human teammate and possibly adjust its behaviour based on it. Mental models are, however, a complex concept, encompassing all information related to environment dynamics, responses to these dynamics, team goals, and team role interdependencies [1]. This poses a problem, as there is limited research on building artificial trust in human-AI team scenarios. While existing studies offer conceptual models of constructing trust, empirical research on the advantages of building artificial trust and practical implementations of such models remain underexplored.

Communication has proven to be essential in building and maintaining natural trust, as an explainable system improves the user's trust in the algorithm [4]. When designing human-AI communication mechanisms, decisions about which aspects of the AI model to communicate and how to communicate them are pivotal, as they directly influence team trust and performance [5]. However, as empirical research on building artificial trust is scarce, there is consequently a gap in studying the effects of communicating this aspect of the AI model. As such, this research analyses the impact of communicating the AI agent's trust in its human teammate, which aims to increase the human's understanding of the AI algorithm. Moreover, previous literature lacks a focus on the different types of communication the human-AI teams can incorporate [5]. Thus, this study will focus on a particular communication type, specifically real-time visual explanations of trust. There is evidence supporting this choice: explanations can affect humans' perception of an AI's ability to help [6], visual information is preferred by users over textual alternatives [7], and collaborative tasks are considered an opportunity for real-time feedback and updates [1].

This research aims to consolidate human-AI collaboration in team-based activities by building and communicating artificial trust. To achieve this, the study focuses on 1) constructing a trust model of the AI agent and 2) implementing real-time visual explanations of

the AI's trust in the human teammate. The following research question has been formulated to support this development:

How do real-time visual explanations of the mental model of the AI agent's trust in its human teammate affect the human's trust in the AI agent and overall satisfaction?

To effectively address the research question, several sub-questions have been composed:

1. What are the main components of the mental model of the AI's trust in its human teammate? How are they formally defined?
2. How can the AI agent's trust be visually explained in a real-time environment?
3. How can human trust and satisfaction be measured?
4. How does the inclusion of real-time visual explanations impact the trust and satisfaction of the human teammate compared to a baseline model?

These sub-questions serve as a guide through the research process, which includes three main phases: building the trust model (1), implementing the communication method (2), and assessing the performance of this method by conducting a user study (3, 4). The first two sub-questions are answered by performing a literature study supported by implementation, while the last two are discussed based on the setup and results of the user study.

This research has the potential to improve support for building trust models in practical scenarios and to promote natural trust via real-time visual explanations. The main focus is developing trust models that reflect real-life contexts and promoting communication styles that feel more trustworthy. Rather than viewing AI agents as tools, the aim is to encourage a perspective where they are seen as legitimate teammates. This shift in mindset has the potential to greatly enhance team performance and effectiveness [5].

The rest of the paper is structured as follows. Section 2 presents the background literature used to develop the trust and communication models, followed by a formalization of the trust model in section 3. The methodology and experiment design are described in section 4, and an overview of the results is presented in section 5. A discussion on responsible research is presented in section 6 and the experiment results are discussed subsequently in section 7. The conclusions of this research are highlighted in section 8.

2 Background

2.1 The Trusting Process

According to Sabater-Mir and Vercouter (2013), trusting another involves a process that can be divided naturally into two stages: trust evaluation and trust decision [2]. The first stage, **trust evaluation**, assesses the trustworthiness of the trustee (party to be trusted) based on all mental states, values, and beliefs accumulated by the trustor (trusting party) [2]. For artificial agents, especially in short-lived teams, this information might only be collectable from directly observable cues and behaviours. Falcone et al. (2011) define these observable signs as manifesta, and the internal properties that can be derived from them as krypta [8]. Previous literature has outlined multiple frameworks for modelling the krypta, such as the ABI (Ability-Benevolence-Integrity) Model developed for human-human teams [9] or the Socio-Cognitive Model of Trust [10].

The Socio-Cognitive Model states that the trustor can form two basic beliefs regarding the trustee: competence and willingness. The competence belief is related to the ability of the trustee to perform a given task, or how useful the trustee is to achieve the trustor's goal. The willingness belief represents how much the trustor thinks the trustee will perform the given task, or how willing the trustee is to achieve the trustor's goal. Based on these concepts, the agent can construct different sets of beliefs for each task domain, following the model proposed by Paglieri et al. (2013) about building trust based on the message quality of the trustee in an argumentative scenario [11]. They explain that trustworthiness in one domain does not necessarily imply trustworthiness in another: a doctor is considered competent in the health domain, but not necessarily when suggesting a restaurant [11].

The second stage in the trusting process is the **trust decision**, which determines whether the trustee will be trusted with a given task [2]. This decision involves not only the trust evaluation but also the context of the interaction [2]. Based on the trust decision, the trustor might adjust its **behaviour** towards the trustee, and the process repeats. In the context of artificial agents, not trusting the human can refer to performing more tasks alone instead of asking for assistance, or adjusting the task allocation.

Previous studies proposed that both the trust evaluation and decision stages are influenced by context, as it impacts the krypta and manifesta of a teammate [12]. Preferences are a significant and "often ignored" part of this context [13, p. 386], as they influence humans' engagement with tasks. For instance, humans may choose not to assist the AI agent due to a dislike for a specific task rather than a reflection of their krypta, and the AI should account for this when assessing trust. Centeio Jorge et al. (2024) propose a conceptual framework to integrate context into trust models, composed of task and team configurations [12]. A notable component of task configuration is the set of stimuli, which can influence humans' motivation to perform a task and provide a better understanding of its perceived complexity. Preferences can arguably be considered a part of this set, as they convey similar aspects. Other essential elements of this framework include task workload, criticality, and team lifespan and composition [12].

2.2 Communicating the Trust Model

As a broader concept, communication is considered to be a central point in human-AI team processes and a facilitator of shared knowledge [5, 12]. Communication can thus be viewed as an effort towards Explainable AI (XAI), an artificial intelligence branch focused on making systems more understandable to humans [14]. There are several past experiments which focused on the impact AI communication has on users. Zhang et al. (2023) studied how AI communication strategies impact human-AI teaming processes, concluding that proactive communication enhances human trust and situation awareness [5]. The experiment by Verhagen et al. (2022) analyses the effect of various AI communication styles (silent, transparent, explainable, adaptive) on teamwork performance, across different levels of interdependence [15]. Their findings generally indicated increased levels of human trust and understanding when the robot communicated. Le Guillou et al. (2023) analysed the impact of the AI agent providing intention-based explanations on user trust and acceptability, concluding that such explanations enhance trust [16]. Overall, it is noticeable that there is a strong connection between appropriate AI communication and natural trust.

Although certainly valuable, the communication types investigated in most studies are environment- or task-related, with no focus on implementing or communicating artificial trust. Therefore, following the advice of Zhang et al. (2023), "Research should gear towards

understanding the nuances and different types of communication that the current state-of-the-art HATs afford, and that the future HATs should afford" [5, p .281:5], this research focuses on the potential utility of communicating the AI agent's trust beliefs and related behaviour.

As the communication model outlined in the research question is explanation-based, it is worth clarifying the distinction between explainable and transparent systems. Following the framework proposed by Verhagen et al. (2021), transparency refers to disclosing knowledge about the system functionality, or answers to "what"-questions [14]. On the other hand, explainability provides answers to "why" or "how"-questions, clarifying relations between system elements and thus supporting human understanding [14]. The explanations of trust should, therefore, include not only plain information but also the reasoning behind it. The choice of integrating explanations is supported by past studies, which demonstrate that providing reasoning information increases trust [6, 17]. Moreover, a measure of interest for the research question is human satisfaction. Therefore, following the conclusions of Szymanski et al. (2021) that users prefer visualization-based explanations over textual ones, the trust model of the AI agent should be presented graphically [7]. However, since the notion of explainability implies providing a reason when communicating knowledge, the visualizations should be enhanced with textual reasonings. This decision is also supported by Szymanski et al. (2021), who concluded that enhancing visual explanations with textual additions improves user performance [7]. Finally, collaborative tasks are considered to be an opportunity for more frequent feedback and updates [1]. This advantage can be used by presenting the trust beliefs of the AI agent in a real-time manner, on every trust update or behaviour change.

3 Trust Model Formalization

To communicate the AI agent's trust beliefs, this study compiled a context-dependent trust model for building artificial trust, following the concepts presented in subsection 2.1.

In the evaluation stage, the AI agent's beliefs are updated according to the perceived competence and willingness of the human. Formally, when observing the nth behaviour cue of the human H for a certain task $t \in D$ from task domain D, the AI agent's beliefs in the human krypta value $\phi \in \{competent, willing\}$ are updated as:

$$\begin{cases} B_n(\phi(H,D)) = B_{n-1}(\phi(H,D)) + \Delta(t) \\ B_0(\phi(H,D)) = 0 \end{cases} \tag{1}$$

where $B_n(x) \in [-1,1]$ is the belief of the AI agent on an arbitrary value x after n observations and $\Delta(t)$ is the amount of adjustment (positive/negative) for a task t. The amount of adjustment depends on the importance and interdependence level of the task. Having a different set of beliefs for each individual task instead of aggregated domains was also considered. However, due to the short lifespan of the human-AI collaboration chosen for the user study, it was decided to opt for constructing task domains instead. This ensures that the beliefs have enough time to consolidate and converge towards the end of the collaboration.

For tasks t that can be dependent on human preferences, an additional preference factor $P(t)$ (positive/negative) is considered when computing the willingness belief. This study considers that preferences can only impact the willingness of a human to perform a certain task, independent of their competence. The human preferences were, thus, considered in both the evaluation and decision stages, serving as a context factor. Integrating preferences

into the evaluation stage is, to the best of the author's knowledge, not an experimented area in the current literature. This inclusion aims to form more accurate trust beliefs, which should be better understood and accepted by humans when communicated. Formally, the preference factor is included as:

$$B_n(willing(H, D)) = B_{n-1}(willing(H, D)) + \Delta(t) + P(t) \tag{2}$$

To compute the preference factor, this research considered a heuristic-based approach, centred on the idea that humans prefer to do less difficult tasks. There is evidence supporting this heuristic. O'Brien et al. (2020) discuss in their experimental study regarding web search engagement that complex tasks are associated with negative emotions and are deemed less engaging [18]. Moreover, Jorge et al. (2024), in their experiment about which manifesta should an AI agent take into account as cues of the human's krypta, discussed that most participants chose tasks with the least effort associated to them (the easiest products to collect in a supermarket environment) [3].

For trust decisions, aggregating trust beliefs is a necessary step. Following the study by Paglieri et al. (2013), this research follows the hypothesis that a source is trusted if it is believed to be both competent and willing [11]. Formally, the trust decision $\tau_n(t) \in \{0, 1\}$ on task t after n observations is:

$$\tau_n(t) = B_n(competent(H, D)) \geq T_c \wedge B_n(willing(H, D)) \geq T_w \tag{3}$$

where T_c, T_w are predefined thresholds and $\tau_n(t) = 1$ represents trust, $\tau_n(t) = 0$ represents distrust. For the willingness belief, T_w is dependent on the preference factor $P(t)$ of the task t for which the decision is taken.

Since the study is dealing with short-lived collaboration, blindly making decisions based on trust in early stages can negatively impact the overall goal. Thus, to model the consolidation of beliefs over time, confidence was introduced. This addition aims to mimic how users' trust in intelligent systems changes over time as they gain more experience, a point highlighted in earlier studies [19]. Formally, when making a decision, the AI agent trusts its own beliefs and, consequently, uses its trust decision $\tau_n(t)$ with confidence C, or else considers the human as trustworthy by default. Confidence can therefore be viewed as a probability measure, $C \in [0, 1]$.

Updating the confidence value is based on the monotonicity of the previous trust adjustments. For example, if the human manifesta produced only trust increases/decreases in the last few assessments, the confidence increases as well since the AI's opinion of the human is solidified. A similar concept is described by Paglieri et al. (2013), which explain that information should only change a source's assessment when something new is learned, either consolidating the judgement or revealing the previous opinion to be wrong [11]. The confidence is updated per domain, taking into account separately both beliefs. Formally, adjusting the confidence on domain D for belief ϕ with adjustment $\delta(\phi)$, considering the last k trust assessments out of the total of n assessments, is computed as:

$$C_n(D) = \begin{cases} C_{n-1}(D) & \text{if } \forall i \; n-k+1 \leq i < n, \; B_i(\phi) = B_{i+1}(\phi), \text{ otherwise} \\ C_{n-1}(D) + \delta(\phi) & \text{if } \forall i \; n-k+1 \leq i < n, \; B_i(\phi) \leq B_{i+1}(\phi) \\ C_{n-1}(D) + \delta(\phi) & \text{if } \forall i \; n-k+1 \leq i < n, \; B_i(\phi) \geq B_{i+1}(\phi) \\ C_{n-1}(D) - \delta(\phi) & \text{otherwise} \end{cases} \tag{4}$$

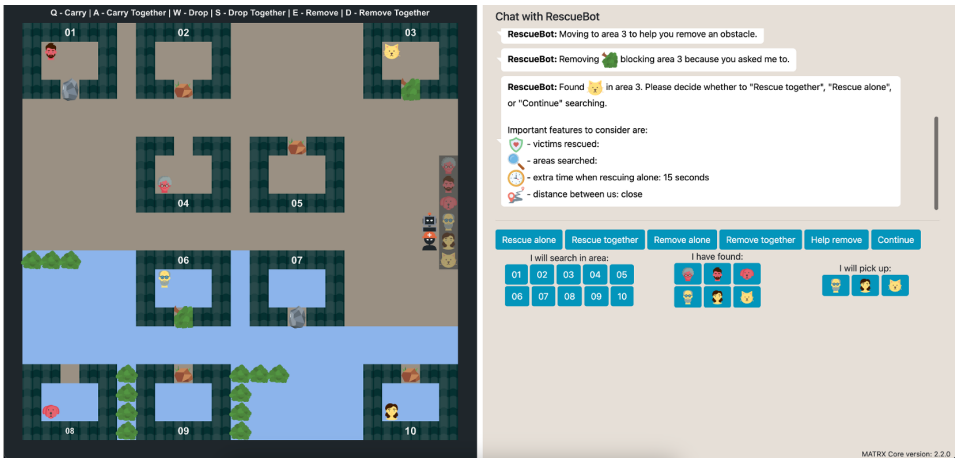


Figure 1: Environment used during the experiment. Left side shows the initial map configuration in "God" view. Right side shows the chat area.

4 Method

The goal of this research is to assess whether communicating the AI agent's trust model increases the human's trust in the AI and overall satisfaction. To achieve this, a between-subject experiment was conducted in person, collecting both observable human behaviours through game logs and self-reported measures of trust and satisfaction using validated questionnaires.

4.1 Participants

To conduct this experiment, 46 participants were recruited using the author's personal networks. All participants resided in Europe and most belonged to the 18-24 age group (41). Ages were between 18 and 44. Most participants had an academic Computer Science-related background (38) and were Bachelor students or graduates (34). Regarding gender, 33 participants identified themselves as men and the rest as women. Only some of them had experience with the MATRX software (12). The gaming experience ranged from no experience (4), very little experience (6), some experience (17), to a lot of experience (19).

4.2 Environment

Simulation. The experiment used a simulated Urban Search and Rescue environment, adapted from an existing implementation [20] and developed using the MATRX Software [21]. The environment features a two-dimensional map with 10 areas where the robot (RescueBot) and the human participant navigate and interact, along with a chat area where the teammates exchange information (Figure 1). There are 6 victims to be rescued, with three being critically injured (red) and three mildly injured (yellow). The critically injured victims could only be rescued with the AI's help, while the mildly injured ones could be rescued by only one teammate (but working together improved efficiency). Some areas were obstructed by 8 obstacles, such as rocks, trees, and stones, which need to be cleared to access those rooms. Clearing the obstacles required different interdependence levels with

the robot. Rocks (grey) demanded cooperation from both teammates to clear them. Trees (green) could only be removed by the robot itself. Stones (brown) were the most flexible obstacle - either teammate could handle them, although working together improved efficiency. The RescueBot, victims, and obstacles were only visible to the human if close to their avatar (as opposed to the "God" view showcased in Figure 1).

Task. The goal of this simulation was to successfully find and transport all 6 victims to the rescue zone. Completing this task required a high level of interaction with the robot, both communication-wise and when performing joint tasks. Each simulation lasted 10 minutes, after which the environment shut down. The task domains based on which the AI formed its trust beliefs were Search (announcing searches and properly searching rooms), Obstacles (clearing obstacles efficiently), and Victims (finding and collecting victims).

Communication. In the chat area, the participant could only communicate with the robot using predefined phrases, presented as buttons. This interface allowed for:

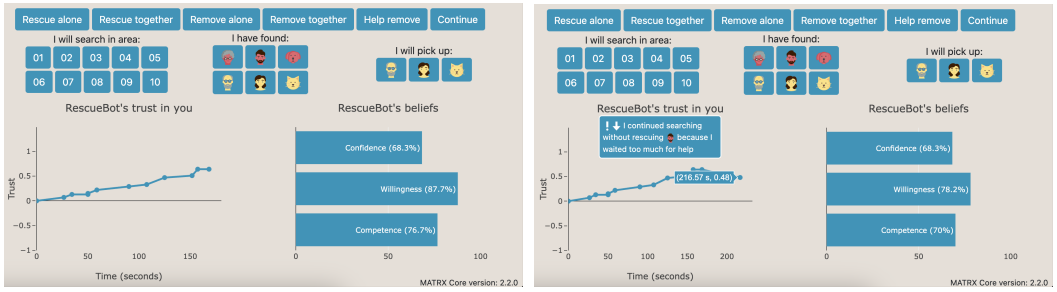
- sharing decisions about searching ("I will search in area X")
- requesting help with removing obstacles ("Help remove at X")
- answering questions ("Remove alone/together", "Rescue alone/together", "Continue")
- announcing the discovery and rescue of victims ("I have found X", "I will pick up X").

Preferences. To elicit human preferences, the environment was designed to give the tasks varying engagement levels. Notably, a flooded area was included that covers half of the map, which takes longer for the user to navigate compared to the non-flooded area. Moreover, the map contains special victims (elderly victims) whose rescue requires more time. The distance between the human and the robot was also considered. All three aspects (flooded areas, special victims, distance) were used to compute the preference factor, based on the heuristic that humans tend to avoid complex tasks. Additionally, the distance was also taken into account when determining the robot's waiting time for human response or assistance, as part of the behaviour adaptation process.

4.3 Conditions

This experiment compares the changes in human trust and satisfaction when incorporating real-time visual explanations about the AI agent's trust model (Trust Explanations group), compared to a baseline without this information (Baseline group). To fulfil this, two trust-related plots were integrated into the chat area of the TE group, offering explanations and insights into the trust model (Figure 2). The choice to use plots for visualization is supported by previous studies that adopted the same method [7]. Moreover, for the trust model to be easily understandable, only the components deemed the most important were displayed: the trust assessment and decisions, the human's krypta model, and the AI agent's confidence.

The left plot shows the change in the AI agent's overall trust in the participant over time, thus leveraging the real-time aspect of the collaboration. The trust value was averaged across all domains (Search/Obstacles/Victims) and beliefs (competence/willingness), to make the plot more comprehensible. Upon hover, each data point displays a brief explanation of why the trust value changed (Figure 2b), such as a specific human action that caused the



(a) Default view

(b) View upon hovering on a data point

Figure 2: Trust plots displayed for the TE condition.

Table 1: Trust-based explanations for the TE condition.

Trust Increase	Trust Decrease	Behaviour Adaptation
	Found V/O in room searched by you	I trust you, so I continued searching without rescuing/removing V/O
You searched a room	You searched 2 rooms in <5 seconds	I trust you with rescuing V
You want to remove together O	You searched a room already searched by me/you before	I trust you to rescue V together
You asked for help with O and are here	You forgot to announce search/found	I continued searching without rescuing/removing V/O because I waited too much for an answer/for help
You want/asked me to remove O	You asked for help with O but were not there	I rescued/removed V/O alone because I waited too much for an answer/for help
We removed O together	Found V but you said you collected it	I will help you with rescuing V
Found V because you told me	Victim V not found in room you said	I rescue/remove V/O alone because I think it is more efficient
You want to rescue together V	I rescued/removed V/O alone because I waited too much for an answer/for help	I rescue/remove V/O alone because I do not think you prefer this task
We rescued V together	I continued searching without rescuing/removing V/O because I waited too much for an answer/for help	
You found/collected V		

O = obstacle, V = victim.

increase or decrease in trust (marked with up/down arrows), or notable events related to the agent's behaviour (marked with an exclamation mark). This inclusion aims to fulfil the explainability aspect of the research question. A comprehensive list of all trust-based explanations can be found in Table 1. The right plot is a bar chart displaying the values of the two components the AI agent's trust is based on, competence and willingness, as well as the AI agent's confidence in its own beliefs. The latter was added for better insights into the trust decision process. These values are averaged on all domains and are presented as percentages, aiming to make the plot clearer.

4.4 Procedure

Each participant was first asked to read a research overview and complete a consent form before proceeding with the experiment. They were randomly assigned to one of the two conditions after consenting. Then, each participant was asked to complete a personal information survey in which they stated their age group, gender, region, level of education,

game experience, knowledge of the MATRX Software, and whether they major(ed) in a Computer Science-related field. The participant was then instructed on how to navigate the environment and interact with the robot, by playing a tutorial in a toy environment. In the tutorial, the robot prompted the user to perform different tasks and explained the most relevant environmental aspects. Regardless of condition, participants were also informed that the robot contains a trust model, based on their competence and willingness, and that the robot adapts its behaviour based on it. For the TE condition, the participants also received a verbal explanation of how to use the trust-related plots. Afterwards, participants proceeded with the official task. During the task, they were allowed to ask for technical support, such as game commands, but no other help was provided. After the task was finished, each participant completed a questionnaire regarding their perception of the RescueBot.

4.5 Metrics

4.5.1 Subjective Measures

To measure the participant's self-reported (SR) trust and satisfaction, two validated questionnaires proposed by Hoffman et al. (2023) were used: the Trust Scale for the XAI Context and the Explanation Satisfaction Scale [22]. Both questionnaires were based on a 5-point Likert scale and were adapted slightly to fit the topic of this research. The survey also contained an exploratory, optional section in which participants could respond to four open-ended questions regarding their perception of RescueBot. The integral questionnaires can be found in Appendix A.

4.5.2 Objective Measures

During the simulation, objective metrics were recorded and collected in the background for analysis. Specifically, the following events were logged for each simulation: the number of messages received from the participant, the number of human actions (both individual and joint), the mouse movements of each user, the number of successful joint tasks proposed by the AI, and the completion time of the simulation in ticks. The number of human actions and the completion time were recorded using MATRX's environment loggers. Ten ticks represent approximately one second and are the time unit used by the MATRX software. These logs were then used to compute multiple metrics potentially capable of indicating trust and satisfaction:

- The number of human messages was divided by the total completion time in seconds to assess the participants' **communication rate**. This measure was recorded due to evidence suggesting that a higher communication rate typically indicates increased trust [23].
- The ratio of joint actions relative to total human actions was computed using the number of joint and individual human actions, to assess the human's **level of interaction** with the robot.
- The **mouse movements** were recorded to monitor the interaction of the participants with the trust plots in the TE condition. A **focus** metric was also calculated for the plots, by dividing the number of mouse movements over the plots by the total number of movements. These measures are inspired by previous studies, which utilized eye and mouse tracking to assess user satisfaction. Generally, it is reported that users concentrate their mouse cursor or gaze on elements that attract them [23, 24].

Table 2: Pearson's correlations between self-reported measures and objective metrics.

	Communication Rate	**Level of Interaction**	**Focus**	**Compliance**
SR Satisfaction	0.34*	-0.287	0.095	-0.423*
SR Trust	0.247	-0.113	0.182	-0.088

* Statistically significant at $p < 0.05$ level (green).

Table 3: Comparison test results for assessing differences across the two conditions for the dependent variables trust, satisfaction, communication rate, compliance.

Metric	**Statistical Test**	**P-value**	**Condition**	**Mean (μ)**	**SD (σ)**
SR Trust	Independent Samples Welch's T-test	< 0.001*	TE	4.261	0.31$^\diamond$
			Baseline	3.511	0.624$^\diamond$
SR Satisfaction	Independent Samples Welch's T-test	0.002*	TE	4.344	0.41$^\diamond$
			Baseline	3.688	0.873$^\diamond$
Communication rate	Mann-Whitney U test	0.011*	TE	0.049	0.011
			Baseline†	0.042	0.014
Compliance	Mann-Whitney U test	0.216	TE†	2.864	1.66
			Baseline	3.09	1.311

* Statistically significant at $p < 0.05$ level (green).
† Non-normality (orange).
◇ Heteroscedasticity (yellow).

- The number of successful joint tasks proposed by the robot was used to assess the **compliance** of the human with the robot's requests. Compliance reflects the human's inclination to accept system guidance or decisions, providing insight into their trust in the system [23].

5 Results

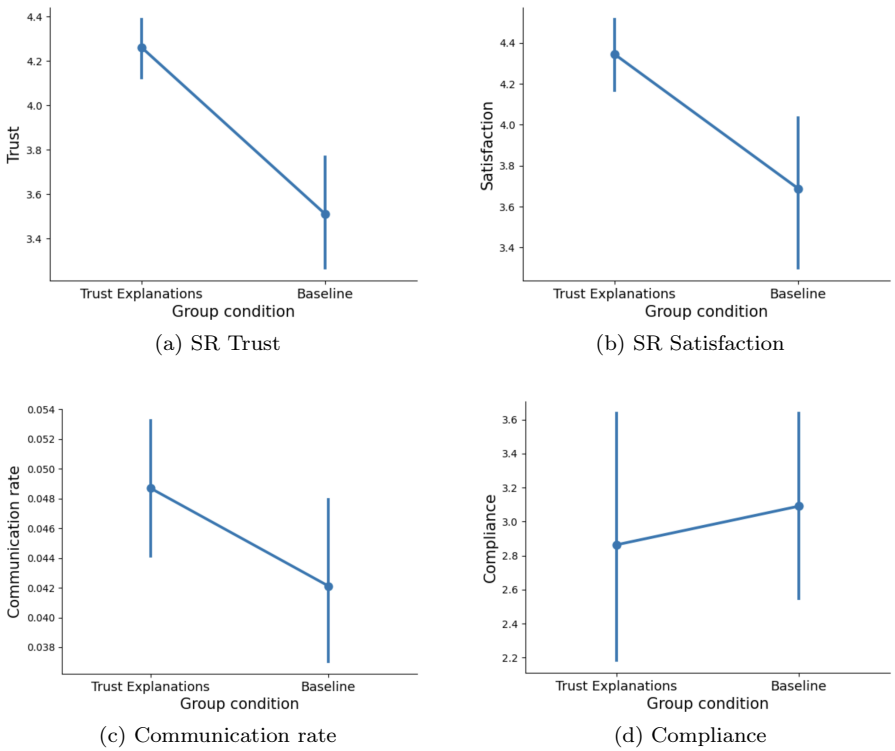
Based on the collected data, a statistical analysis was performed on both subjective and objective metrics using the SciPy Python library[1]. One outlier from the TE group was removed due to its significant deviation from the rest of the sample, as it negatively impacted parametric assumptions. This decision is supported by previous studies [25]. Consequently, a randomly chosen Baseline result was removed to maintain a balanced design, bringing the final experiment to 44 participants. Two hypotheses were formulated to guide the analysis:

H1 Incorporating real-time visual explanations of the AI agent's trust in its human teammate increases natural trust.

H2 Incorporating real-time visual explanations of the AI agent's trust in its human teammate increases overall satisfaction.

Correlations. The relationships between self-reported and objective measures were analysed using Pearson's Correlation coefficient. Table 2 showcases a summary of the results, along with their assessed statistical significance at $p < 0.05$ level.

[1]SciPy: https://scipy.org/



(a) SR Trust

(b) SR Satisfaction

(c) Communication rate

(d) Compliance

Figure 3: Interaction plots for assessing differences across the two conditions for the dependent variables trust, satisfaction, communication rate, compliance.

Comparison tests. Comparison tests were performed to assess differences between the TE and Baseline groups, using both subjective and objective metrics as dependent variables (Figure 3). Only the objective measures that correlated with the self-reported ones were included in the analysis. Before selecting the appropriate comparison test, each group's data was tested for parametric assumptions: normality, homogeneity of variances, no outliers, independence. The independence assumption was ensured by the experiment's between-subject design. Normality was assessed using the Shapiro-Wilk test, with significance level $\alpha = 0.05$. Bartlett's test was used to further check for the equal variances assumption if normality was met. After assessing the parametric assumptions, a suitable comparison test was chosen for each measure. Because datasets exhibited heteroscedasticity, an independent samples Welch's T-test was performed if the other assumptions were met. When parametric assumptions were not met, a Mann-Whitney U test was performed instead. Table 3 presents the results for each measure, along with their statistical significance at $p < 0.05$ level.

Mouse Movements. To visually evaluate whether the participants interacted with the trust plots, a heatmap was created, showing the average mouse movements of all participants in the TE group. Figure 4 shows both the plain generated plot (left) and an enhanced version with a sample image from the environment overlaid on the heatmap (right).

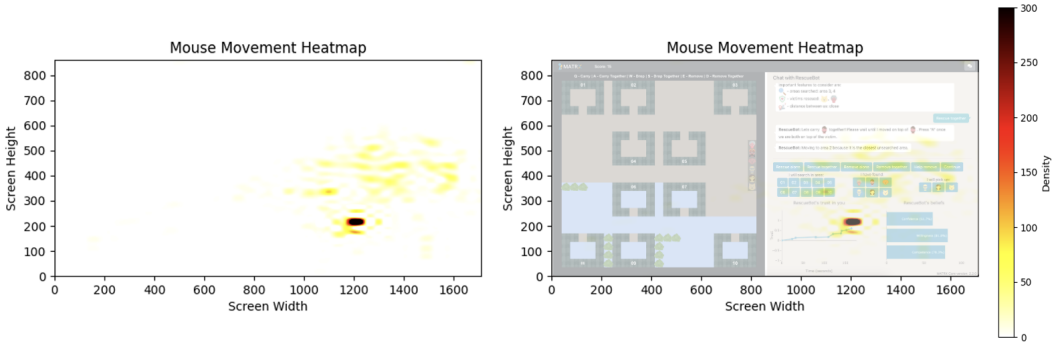


Figure 4: Heatmap of the aggregated mouse movements in the TE group.

6 Responsible Research

This research received ethical approval from the TU Delft HREC, id. 4043. Before conducting the experiment, all participants were informed about the research overview and potential risks. Following this, they signed an informed consent form and an ethics review checklist for human research. The primary risk identified was related to data processing, thus several measures were taken to mitigate any potential breaches. Personally identifiable information (PII), including full names and signatures, was collected solely as part of the consent form and ethics checklist and is accessible only to the research team. Moreover, the Personally Identifiable research data (gender, age group, region, education, Computer Science experience, MATRX Software experience, and game experience) was collected to describe the data samples and was anonymized for each participant. The data collection was carried out using Microsoft Forms[2], which is GDPR-compliant. These methods ensured the privacy of the participants, thus adhering to ethical research standards.

Regarding data transparency and experiment reproducibility, the full codebase of the simulation, the collected data (both raw and processed), and the scripts used for the statistical analysis are all available in a Gitlab repository provided by TU Delft[3]. The baseline version of the simulation environment is also provided in a separate repository[4]. All processing techniques, including the removal of one outlier and the statistical analysis performed, were described in section 5. The exact questionnaires used to collect self-reported metrics are available in Appendix A. Participants were encouraged by the research team to respond honestly, as there were no right or wrong answers, to preserve the integrity of this study. Lastly, the potential limitations of this research are objectively presented in subsection 7.3, to improve the transparency of this study further. These measures comply with the FAIR (Findable, Accessible, Interoperable, Reusable) principles for responsible research.

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

[3]Full Gitlab repository: https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Tielman_Jorge/edumitrescu-Communicating-trust-based-beliefs-and-decisions-i

[4]Baseline repository: https://gitlab.ewi.tudelft.nl/cjorge/rp2024artificialtrust

7 Discussion

This section presents a discussion of possible interpretations of the experiment results, highlighting the most relevant aspects. The structure follows the two measures of interest for the research question, natural trust and overall satisfaction. The experiment limitations and future work are presented at the end of this section.

7.1 Natural Trust

In terms of self-reported trust, the comparison test revealed a significantly higher mean for the TE group compared to the Baseline group. These findings suggest that integrating real-time visual explanations of artificial trust significantly increases humans' trust in the AI agent. Another plausible inference is that the system and its underlying AI algorithm become more understandable through the integration of trust-based explanations. The results thus **support hypothesis H1**.

7.2 Overall Satisfaction

The overall satisfaction was first assessed in terms of self-reported satisfaction, which was found to be significantly higher in the TE group compared to the Baseline group. This shows that incorporating visual explanations of the AI agent's trust increases the human's satisfaction, thus **supporting hypothesis H2**.

Moreover, a statistically significant positive correlation was found between SR satisfaction and communication rate, suggesting that this objective metric can be an indication of overall satisfaction. This is surprising, as the communication rate was mostly assessed as a metric indicating trust by past studies [23]. However, this difference may be attributed to the questionnaire's focus on explanation satisfaction, which is inherently a component of communication. Comparison tests were thus performed for the communication rate as well, as it was considered to be an indirect communication satisfaction measure. It was found that the communication rate is higher when including the visual explanations of trust, further supporting H2.

A statistically significant correlation was also found between SR satisfaction and compliance. However, this correlation is negative, as opposed to previous studies that analysed compliance in trust-related contexts [23, 24, 25]. One possible explanation for this result is that the computed metric might reflect the perceived complexity of the overall simulation. Compliance increased when participants agreed to perform tasks together with the robot; however, excessive dependence on the robot's assistance may indicate a higher perceived complexity. As humans' confidence in their abilities decreases when task complexity increases [23], this could have negatively impacted their satisfaction. Further comparison tests between TE and Baseline groups in terms of this metric revealed no statistical significance, with the Baseline group having a slightly higher mean than the TE group.

Finally, the mouse movements were tracked as a measure of engagement and, thus, satisfaction with the trust plots in the TE condition. Although aggregated measures based on the movements revealed no statistical significance, the aggregated heatmap presented in Figure 4 clearly indicates a higher focus on the hovering graph compared to the rest of the screen. An individual analysis of each participant's mouse movements (Appendix B) revealed that, while some participants were engaged with the hovering explanations, others were not paying much attention to them. However, both subgroups had similar mouse movement density on the rest of the screen, which might have produced the visible difference

in aggregated heatmap density. What can be concluded is that, although having visual explanations increased overall satisfaction as assessed by previous tests, the supplementary textual explanations were only considered engaging by some of the participants.

This duality can also be inferred from the open-ended section of the questionnaire. Two participants from the TE group suggested integrating the trust-based explanations in the chat area instead of as a hovering feature because the explanations were "*hard to look*" at (P10), which might indicate information overload. This is not surprising, as past experiments also resulted in information overload if the volume of explanations was perceived as too dense [5, 6]. On the other hand, two other participants in the TE group mentioned the real-time feedback/plots as a favourite part of their collaboration with RescueBot, which "*helped in having more confidence in the robot*" (P22).

7.3 Experiment Limitations

A potential limitation of this experiment is the hardware used for the TE group compared to the Baseline group when conducting the simulation. For the TE group, a Macbook laptop was used, which ran the simulation environment smoothly, with very little latency. In contrast, the Baseline group used both Macbook and Windows laptops, with the latter experiencing slower performance within the environment. Thus, the trust and satisfaction levels of Windows participants could have been negatively affected by this difference. This limitation can also be an explanation for the heteroscedasticity of the data, as both self-reported trust and self-reported satisfaction exhibited significant variance in the Baseline group compared to the TE group.

Another limitation that could have impacted the results is the homogeneity of participants. Analyzing the demographics reveals that over 75% of participants have at least some gaming experience, and 82.6% have an academic background related to Computer Science. This uniformity in participant backgrounds likely influenced the study's outcomes, contributing to consistently high performance and positive attitudes towards the simulation and AI. Moreover, the high performance of participants limited the potential effectiveness of some objective metrics, such as simulation scores and game completeness, which were not suitable for meaningful analysis.

7.4 Future Work

In the future of this research, it would be interesting to further explore the impact of communicating the AI agent's trust from multiple perspectives. Different ways of communicating trust, such as textual or visual, or explainable versus transparent-based, should be investigated to allow for comparisons and find the most appropriate methods. Looking back at the potential information overload highlighted by the participants, a solution also worth exploring is adding trust explanations as part of the chat area, shifting towards hybrid communication.

Moreover, future research should further explore potential metrics that assess human trust and satisfaction. Due to experiment limitations, the current study only focused on an accessible range of metrics. Measures related to team performance were not suitable for analysis due to participant homogeneity and hardware inconsistencies. Additionally, more advanced tracking techniques, which would have provided a more complete picture of the user engagement, were less accessible to the research team. For instance, eye tracking can provide a more accurate picture of users' interaction with the trust plots. Analysing the

plots visually without reading the trust explanations is also a reflection of engagement, which was not captured by mouse movement analysis alone.

Lastly, future studies should generally gear more towards empirical studies on artificial trust. The literature presents various conceptual frameworks and techniques to build an AI agent's trust model, however, there is not much effort towards automatizing this process. In an ideal scenario, trust model implementations should be readily available for research and should be flexible enough to sustain multiple contexts. The model presented in this study has the potential to be reused in other similar experiments, and a future direction would be to further improve its adaptability and effectiveness.

8 Conclusions

This study analysed the impact of communicating the AI agent's trust beliefs and related behaviour to a human teammate, in terms of human trust and overall satisfaction. A context-dependent model was proposed to build the AI's trust in its teammates, which combines multiple frameworks from past literature. The communication method was designed as real-time, visual, and explanation-based, and was presented as trust-related plots enhanced with small textual reasonings. The influence of this communication on human trust and satisfaction was studied by conducting an empirical experiment with 46 participants, using both subjective and objective measures for analysis. Results showed a significant increase in both human trust and overall satisfaction when adding the communication method, compared to a baseline with no such information. Moreover, the communication rate was found to be a relevant objective measure of human satisfaction and was higher when trust communication was integrated. The results also hinted towards potential information overload when integrating explanations directly into the trust plots, which can be a suggestion of hybrid communication as a more performant method. Overall, this research takes a step towards a more empirical exploration of artificial trust, by providing a practical implementation and communicating it to human participants.

References

[1] Eduardo Salas, Dana Sims, and Shawn Burke. "Is there a "Big Five" in Teamwork?" In: *Small Group Research* 36 (Oct. 2005), pp. 555–599. DOI: 10.1177/1046496405277134.

[2] Jordi Sabater-Mir and Laurent Vercouter. "Multiagent systems". In: ed. by Gerhard Weiss. MIT Press, 2013. Chap. 9.

[3] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. "How Should an AI Trust its Human Teammates? Exploring Possible Cues of Artificial Trust". In: *ACM Trans. Interact. Intell. Syst.* 14.1 (Jan. 2024). ISSN: 2160-6455. DOI: 10.1145/3635475.

[4] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems". In: *ACM Trans. Interact. Intell. Syst.* 11.3–4 (Sept. 2021). ISSN: 2160-6455. DOI: 10.1145/3387166.

[5] Rui Zhang et al. "Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork". In: *Proceedings of the ACM on Human-Computer Interaction* 7 (2023), pp. 1–31.

[6] Erin K. Chiou et al. "Towards Human–Robot Teaming: Tradeoffs of Explanation-Based Communication Strategies in a Virtual Search and Rescue Task". In: *International Journal of Social Robotics* 14.5 (2022), pp. 1117–1136. DOI: 10.1007/s12369-021-00834-1.

[7] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. "Visual, textual or hybrid: the effect of user expertise on different explanations". In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*. IUI '21. Association for Computing Machinery, 2021, pp. 109–119. ISBN: 9781450380171. DOI: 10.1145/3397481.3450662.

[8] Rino Falcone et al. "From Manifesta to Krypta: The Relevance of Categories for Trusting Others". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 4 (Mar. 2011). DOI: 10.1145/2438653.2438662.

[9] Roger C. Mayer, James H. Davis, and F. David Schoorman. "An Integrative Model of Organizational Trust". In: *The Academy of Management Review* 20.3 (1995), pp. 709–734. ISSN: 03637425.

[10] Rino Falcone and Cristiano Castelfranchi. "Trust Dynamics: How Trust Is Influenced by Direct Experiences and by Trust Itself". In: vol. 2. Feb. 2004, pp. 740–747. ISBN: 1-58113-864-4. DOI: 10.1109/AAMAS.2004.286.

[11] Fabio Paglieri et al. "Trusting the messenger because of the message: Feedback dynamics from information quality to source evaluation". In: *Computational and Mathematical Organization Theory* 20 (Aug. 2013). DOI: 10.1007/s10588-013-9166-x.

[12] Carolina Centeio Jorge et al. "4 - Appropriate context-dependent artificial trust in human-machine teamwork". In: *Putting AI in the Critical Loop*. Academic Press, 2024, pp. 41–60. ISBN: 978-0-443-15988-6. DOI: https://doi.org/10.1016/B978-0-443-15988-6.00007-8.

[13] Matthew Johnson and Jeffrey M. Bradshaw. "Chapter 16 - The role of interdependence in trust". In: *Trust in Human-Robot Interaction*. Ed. by Chang S. Nam and Joseph B. Lyons. Academic Press, 2021, pp. 379–403. ISBN: 978-0-12-819472-0. DOI: https://doi.org/10.1016/B978-0-12-819472-0.00016-2.

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

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

[16] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. "Trusting Artificial Agents: Communication Trumps Performance". In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '23. International Foundation for Autonomous Agents and Multiagent Systems, 2023, pp. 299–306. ISBN: 9781450394321.

[17] Thomas O'Neill et al. "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature". In: *Human Factors* 64.5 (2022), pp. 904–938. DOI: 10.1177/0018720820960865.

[18] Heather L. O'Brien, Jaime Arguello, and Rob Capra. "An empirical study of interest, task complexity, and search behaviour on user engagement". In: *Information Processing Management* 57.3 (2020), p. 102226. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2020.102226.

[19] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. "User Trust in Intelligent Systems: A Journey Over Time". In: Mar. 2016, pp. 164–168. DOI: 10.1145/2856767.2856811.

[20] Ruben Verhagen. *Human-Agent Teamwork for Search and Rescue*. 2020. URL: https://github.com/rsverhagen94/TUD-Collaborative-AI-2024.

[21] Tjalling Haije Jasper van der Waa. *MATRX: Human Agent Teaming Rapid Experimentation software*. Version 2.3.2. July 2023. DOI: 10.5281/zenodo.8154912.

[22] Robert R. Hoffman et al. "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance". In: *Frontiers in Computer Science* 5 (2023). ISSN: 2624-9898. DOI: 10.3389/fcomp.2023.1096257.

[23] Andrea Krausman et al. "Trust Measurement in Human-Autonomy Teams: Development of a Conceptual Toolkit". In: *J. Hum.-Robot Interact.* 11.3 (Sept. 2022). DOI: 10.1145/3530874.

[24] Spencer Kohn et al. "Measurement of Trust in Automation: A Narrative Review and Reference Guide". In: *Frontiers in Psychology* 12 (Oct. 2021), p. 604977. DOI: 10.3389/fpsyg.2021.604977.

[25] Matthew Luebbers et al. "Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming". In: July 2023. DOI: 10.15607/RSS.2023.XIX.002.

A Questionnaires

The following questionnaire (Table 4) was used to assess the participants' self-reported trust and satisfaction, based on a 1-5 Likert scale. The questionnaire is adapted from two scales proposed by Hoffman et al. (2023), the Trust Scale for the XAI Context and the Explanation Satisfaction Scale [22].

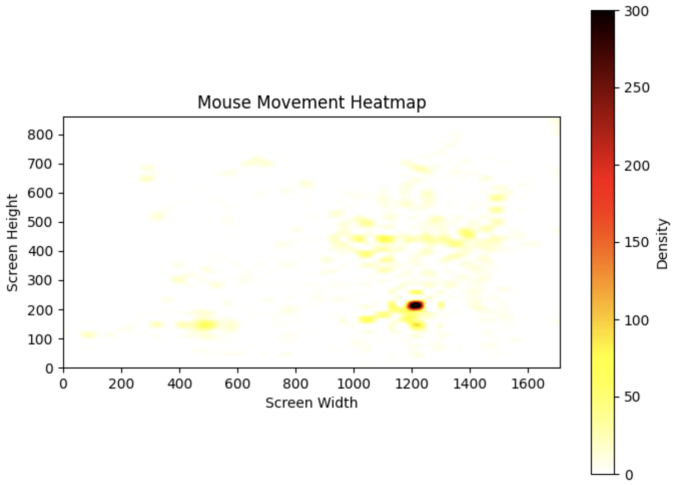
Table 4: Questionnaire used to assess self-reported measures, split by sections.

Trust	I am confident in RescueBot. I feel that it works well.
	The outputs (communication, decisions) of RescueBot are very predictable.
	The RescueBot is very reliable. I can count on it to be correct all the time.
	I feel safe that when I rely on RescueBot I will get the right result.
	RescueBot is efficient and works very quickly.
	I am wary of the RescueBot.*
	The RescueBot can perform a task better than a novice human user.
	I like using the RescueBot's guidance for decision making.
Satisfaction	From RescueBot's explanations, I know how it works.
	The RescueBot's explanations of how it works are satisfying.
	The RescueBot's explanations of how it works have sufficient detail.
	The RescueBot's explanations of how it works seem complete.
	The RescueBot's explanations of how it works tell me how to use it.
	The RescueBot's explanations of how it works are useful to my goals.
	The RescueBot's explanations show me how accurate the system is.
Open Questions	What information would you have liked the RescueBot to provide but was missing?
	What did you like most about your collaboration with RescueBot?
	What did you like least about your collaboration with RescueBot?
	What do you think RescueBot thinks of you? How does that make you feel?

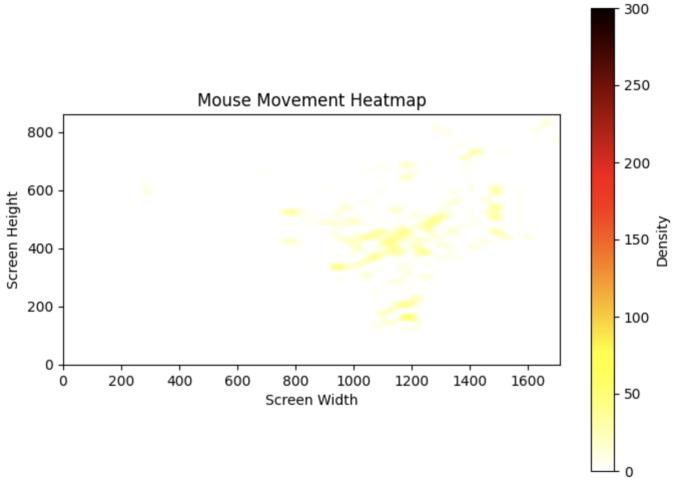
* Represents distrust. The score needs to be inverted when aggregating results.

B Individual Heatmaps

Figure 5 presents two examples of the heatmaps generated for each participant. Image (a) belongs to the group which exhibited a lot of interest in the trust explanations, while image (b) corresponds to the less interested group. The mouse movements on the rest of the screen are fairly similar in terms of density.



(a) Lots of engagement with trust plot



(b) Little engagement with trust plot

Figure 5: Examples of two individual heatmaps from the mouse movement analysis.