How Should Your Artificial Teammate Tell You How Much It Trusts You?

Centeio Jorge, Carolina; Dumitrescu, Elena; M. Jonker, Catholijn; Loghin, Razvan; Marossi, Sahar; Uleia, Elena; L. Tielman, Myrthe

DOI
10.1145/3717511.3747086

Publication date
2025

Document Version
Final published version

Published in
IVA 2025 - Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents

How Should Your Artificial Teammate Tell You How Much It Trusts You?

Carolina Centeio Jorge
Interactive Intelligence, INSY, EEMCS
Delft University of Technology
Delft, Netherlands
C.Jorge@tudelft.nl

Elena Dumitrescu
Interactive Intelligence, INSY, EEMCS
Delft University of Technology
Delft, Netherlands
E.I.Dumitrescu-1@student.tudelft.nl

Catholijn M. Jonker
Interactive Intelligence, INSY, EEMCS
Delft University of Technology
Delft, Netherlands
LIACS
Leiden University
Leiden, Netherlands
c.m.jonker@tudelft.nl

Razvan Loghin
Interactive Intelligence, INSY, EEMCS
Delft University of Technology
Delft, Netherlands
R.Loghin@student.tudelft.nl

Sahar Marossi
Interactive Intelligence, INSY, EEMCS
Delft University of Technology
Delft, Netherlands
S.Marossi@student.tudelft.nl

Elena Uleia
Interactive Intelligence, INSY, EEMCS
Delft University of Technology
Delft, Netherlands
E.Uleia@student.tudelft.nl

Myrthe L. Tielman
Interactive Intelligence, INSY, EEMCS
Delft University of Technology
Delft, Netherlands
m.l.tielman@tudelft.nl

Abstract

Mutual trust between humans and interactive artificial agents is crucial for effective human-agent teamwork. This involves not only the human appropriately trusting the artificial teammate, but also the artificial teammate assessing the human's trustworthiness for different tasks (i.e., artificial trust in human partners). Literature indicated that transparency and explainability is generally beneficial for human-agent collaboration. However, communicating artificial trust potentially affects human trust and satisfaction, which impact team dynamics. Towards studying these effects, we developed an artificial trust model and implemented five distinct communication approaches which varied in modality (visual/graphical and/or text), level (communication and/or explanation), and timing (real-time or occasional). We evaluated the effects of the different communication styles through a user study (N=120) in a 2D grid-world Search and Rescue scenario. Our results show that all our artificial trust explanations improved human trust and satisfaction, but the mere graphical communication of it did not. These results are bound to the specific scenario and context in which this study was run and require further exploration. As such, this work presents a first step towards understanding the consequences of communicating and explaining to a human teammate their assessed trustworthiness.

IVA '25, Berlin, Germany
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1508-2/25/09
https://doi.org/10.1145/3717511.3747086

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

human-AI teamwork, explainable AI, communication, artificial trust

ACM Reference Format:
Carolina Centeio Jorge, Elena Dumitrescu, Catholijn M. Jonker, Razvan Loghin, Sahar Marossi, Elena Uleia, and Myrthe L. Tielman. 2025. How Should Your Artificial Teammate Tell You How Much It Trusts You?. In *ACM International Conference on Intelligent Virtual Agents (IVA '25), September 16–19, 2025, Berlin, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3717511.3747086

1 Introduction

With the current expansion of artificial intelligence, interactive artificial agents are evolving from mere tools to being perceived as true teammates [31], in a variety of embodiments and scenarios, from robots in search and rescue [16, 49] and domestic cases [23], to virtual agents in cooperative video games [47]. Team trust is essential in collaborative interactions [55]. This is affected by two teammates' mutual trust, i.e., trust relationships in which the trustor is a human (*natural trust*), and where the trustor is an artificial agent (*artificial trust*) [55].

On the one hand, humans need to trust artificial agents appropriately to collaborate effectively [34]. On the other hand, recent literature suggests that artificial agents should incorporate trust into their decision process to determine how and with whom to engage (artificial trust) [3, 6]. Both can impact safety, efficiency, and the overall success of joint operations. Mutual appropriate trust is intertwined with communication and, consequently, the team's shared mental models [36, 51]. In fact, human trust in artificial agents (including AI agents) is often built on the artificial agent's communication such as explanations [4, 60, 61]. Ideally, artificial agents should also communicate and explain their beliefs, in order to explain their actions and decisions. Similarly, as artificial agents use artificial trust for decision-making, they should explain their beliefs about human trustworthiness. However, communicating trust may have different effects depending on how it is done. In this paper we explore how the communication of artificial trust by an artificial agent impacts the natural trust of the human in that agent, as well as their satisfaction, and collaborative behaviour in human-agent teams[1].

Artificial trust (AT) has recently been used as a tool for decision-making in human-agent teams [6, 9], for decision-making or task selection. When building artificial trust, artificial agents form beliefs in human characteristics that are cues for trustworthiness [11, 19]. Beliefs about trustworthiness can be divided into belief in a teammate's competence (i.e., *can they do it?*), and belief in willingness (i.e., *will they do it?*) [18]. With an accurate mental model of these human teammate's characteristics, the agent adjusts its behaviour towards effective collaboration.

Artificial agents should communicate the beliefs underlying their decisions to human teammates. When these beliefs and decisions are based on artificial trust, the agents should communicate their assessment of human trustworthiness which led to such decisions. Communicating AT, how it is affected by human actions, and how it consequently affects the agent's actions may make the interaction more transparent and understandable. Although transparency can lead to higher trust [17, 29], some users can be happier when not knowing the whole truth [48]. We imagine that the communication of perceived human trustworthiness should be transparent, but, just like in trust repair strategies, it may be tricky. Research shows that while communication of trust positively impacts teamwork, the impact of communicating distrust is not as straightforward [28, 35]. Imagine, for example, that you have an artificial teammate and you need to collaboratively pilot a plane. Imagine that your artificial teammate (accurately) believes that you are not competent to pilot the plane in a certain context (e.g., in fog) and decides to call for help, but you firmly believe that you are. While transparency is important, communicating (dis)trust may or may not affect how the human feels, and consequently, quality of the teamwork. Furthermore, communication of beliefs can be done in several ways, from verbal to visual, from real-time to occasional communication, and from mere communication to full explanation. This work is a first step towards exploring how the communication of AT beliefs should be done and what the impact is of such communication on the human teammate's emotions and behaviour.

In this paper, we compare five AT communication methods, as well as a baseline without AT communication. This paper makes three contributions: 1) we developed an artificial agent with an artificial trust model which is updated in real time (presented in Section 3.2), 2) we developed five different methods to communicate artificial trust (presented in Section 3.4), and 3) designed and ran a user study (N=120) on a 2D grid-world Search and Rescue scenario to evaluate the effect of the communication styles on the participants' trust and satisfaction (presented in Section 3.3). The results are presented in Section 4 and discussed in 5. This work was executed as a part of a final project of a Computer Science BSc, where five students each designed and tested a different type of AT communication and evaluated their effect on natural trust and satisfaction.

2 Background

2.1 Artificial Trust

Although artificial trust is relatively unexplored for human-machine teams, it has been vastly used in multi-agent systems, i.e., in systems with only artificial agents. However, these works do not necessarily call this artificial trust, but rather trust [18], as in human societies, or computational trust, see e.g. [56]. The term artificial trust is recent [6]. It appears with the need to distinguish between the trust relationships in which the trustor is an artificial agent (artificial trust) and a human (natural trust).

According to Sabater-Mir and Vercouter (2013), the first step for an artificial agent to trust another agent is trust evaluation [50]. The **trust evaluation** phase consists on modelling the trustworthiness (or krypta [19]) of the trustee based on accumulated available information [50], such as directly observable cues and behaviours (also known as manifesta [19]). Previous literature has outlined multiple frameworks for modelling the trustworthiness, such as the ABI (Ability-Benevolence-Integrity) Model developed for human-human teams [39] or the Socio-Cognitive Model of Trust [18], used mainly for multi-agent systems.

The Socio-Cognitive Model states that the trustor can form an evaluation of trust through two beliefs regarding the trustee: *competence* belief and *willingness* belief. The *competence* belief is related to the ability, e.g., set of skills, of the trustee to perform a given task. The *willingness* belief represents how much the trustor thinks the trustee is willing, e.g., intention, to perform the given task. Trust and trustworthiness, as well as the beliefs of competence and willingness, are task and context dependent. Trustworthiness in one domain or task does not necessarily imply trustworthiness in another [3, 45]. For example, a doctor is considered competent in the health domain, but not necessarily when suggesting a restaurant. Similarly, willingness can be affected by the teammate's preferences (e.g., preferring to do one task over another), environmental factors (e.g., going for a task that is physically closer), or strategy (e.g. going for tasks they have seen before) [11, 41]. Based on these concepts, the agent can form beliefs of trustworthiness regarding each teammate's through the evaluation of competence and willingness for each task.

Forming competence and willingness beliefs from behavioural cues, in particular human teammates, is quite challenging and not directly present in literature. However, there is research on cues for detection of intentions [63], natural trust [1, 24] in interaction with embodied AI. With studies in 2D grid-worlds, literature presents metrics to assess teamwork fluency, such as metrics of performance or task completeness [8, 59]. Similarly, we can find metrics for ability, benevolence and integrity, such as speed, favouritism, and commitment, respectively, also in a 2D grid-world [11]. Finally, [7] presents a model that learns the human teammate's sequential

[1] In this context, we use the terms human-agent, human-AI, and human-machine interchangeably.

behaviour, using reinforcement learning. Overall, the modelling of human trustworthiness, as well as the consequent evaluation of trust, are underdeveloped and we attempt to advance them in this paper.

Once trust is evaluated, Sabater-Mir and Vercouter (2013) identify the second stage of the trusting process as the **trust decision**, which determines whether the trustee will be trusted with a given task [50]. Based on the trust decision, the trustor may adjust its *behaviour.* For example, if one agent trusts that another agent can and will perform a certain task successfully, it may proceed to the next task. On the other hand, if they do not trust the agent, then they may decide to help or suggest a different allocation of tasks. The works [3, 6] show through a simulation that trust based on human teammates' capabilities can be used to efficiently allocate tasks in human-AI teamwork scenarios. In human-machine teams, decisions based on an appropriate level of trust in the human teammate may increase performance, safety, and overall human satisfaction [27].

2.2 Communication in human-machine teams

Closed-loop communication is important to share the mental models among teammates and to guarantee mutual trust [51]. For mutual and appropriate trust, the agent should be transparent, and able to explain its decisions [64]. This means that when using artificial trust to make decisions, the agent should then be able to communicate its trust model to the human teammate. As a broader concept, communication is seen as a central point in human-AI team processes and a facilitator of shared knowledge [67]. The current literature highlights its critical role in supporting cognitive [21] and affective processes [52], while also enhancing job satisfaction [22]. Communication can thus be viewed as an effort towards making systems more transparent and understandable to humans [62].

There are multiple ways to communicate an artificial agent's beliefs and decisions to a human teammate. In terms of **modality**, human-agent communication can include, for example, textual explanations [20], summaries [14], visual/graphical representations [33], audio [2], or mixed modalities [5, 46]. Research shows that humans can process visual representations faster compared to textual ones [53] and that it can be an effective modality for communication within human-machine teams [43]. However, other works suggest that mixed modalities, combining visual and another modality, such as text or audio, can be more effective [46, 54].

Furthermore, the **timing** of communication [12], such as real-time or summary-based, can influence team dynamics [67]. In general, collaborative tasks are considered to be an opportunity for more frequent feedback and updates [51]. 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. On the other hand, information overload, which real-time communication is prone to due to the high message frequency, should be avoided, especially in situations with high stakes such as urban search and rescue environments [37, 38].

Finally, it is worth discussing the **level** of information that should be provided in human-AI communication. While previous studies showed that providing reasoning information can significantly increase natural trust [15, 44], communication can also serve the purpose of simply informing the human teammate, without offering any justifications or explanations. This distinction aligns with the broader contrast between explainable and transparent systems. While transparency discloses knowledge about the system functionality, or answers to *what*-questions, explainability provides answers to *why* or *how*-questions, clarifying relations between system elements and thus supporting human understanding [62]. Thus, communicating trust can involve not only presenting factual information but also providing the reasoning behind any changes. There is a vast variety of approaches, and current research has not reached a general consensus on which communication strategy is most effective. These choices are also domain specific and they have not been tested to communicate artificial trust in human-agent teams. This motivates us to investigate how different strategies impact on the human teammate's emotions and behaviour.

3 Methodology

3.1 Environment

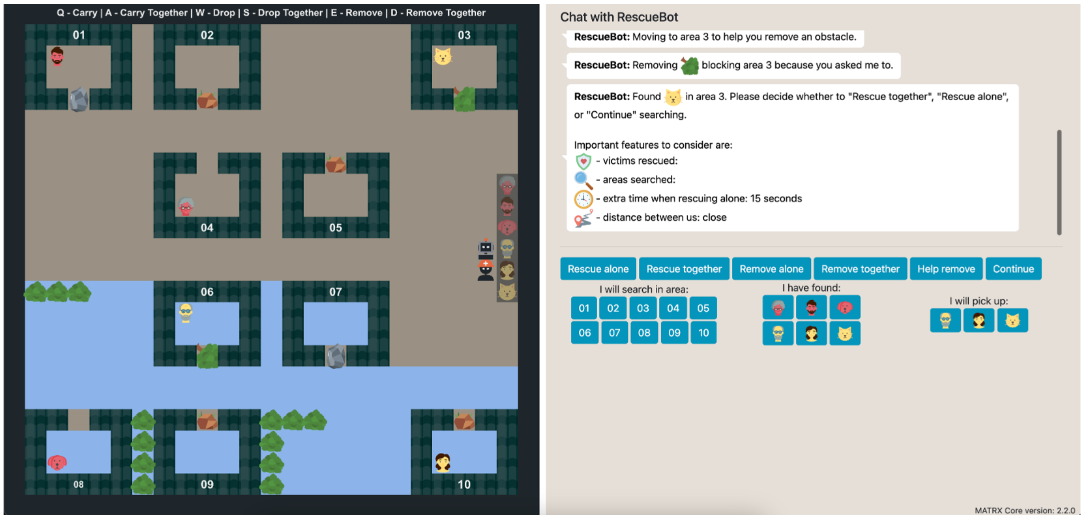
Scenario: The experiment used a 2D grid-world simulated Urban Search and Rescue environment, adapted from an existing implementation [58] and developed using the MATRX Software [57]. The environment features a map with 10 areas where a virtual robot (RescueBot) and the human participant navigate and interact, along with a chat area where the teammates exchange information (see Figure 1). There are 6 victims to be rescued, with variable level of injury severity. The critically injured victims (red) could only be rescued with the RescueBot's help, while the mildly injured ones (yellow) could be rescued by only one teammate, but working together improved efficiency. There are obstacles, such as rocks, trees, and stones, which needed to be cleared to access some rooms. Clearing rocks (grey) demanded cooperation from both teammates to clear them, while 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 visibility was restricted to close obstacles only and the mission lasted ten minutes. The goal of this simulation was to successfully find and transport all victims to the rescue zone.

Communication: In the chat area, the participant could 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").

3.2 Artificial Trust Model

Although this study does not focus on developing a trust model, we need to explain the model we used to guide the communication and explanations. Based on the previously mentioned Socio-Cognitive Model [18], we developed a task-dependent artificial trust (AT) model based on the perceived competence and willingness of the human teammate for this search and rescue scenario. An



RescueBot: Moving to area 3 to help you remove an obstacle.

RescueBot: Removing 🌳 blocking area 3 because you asked me to.

RescueBot: Found 🐱 in area 3. Please decide whether to "Rescue together", "Rescue alone", or "Continue" searching.

Important features to consider are:
- victims rescued:
- areas searched:
- extra time when rescuing alone: 15 seconds
- distance between us: close

Figure 1: Environment used during the experiment. Left side shows the initial map configuration (without visibility restrictions). Right side shows the chat area.

agent may trust a human teammate for a certain task at time t as $Belief(AT_{task,t}) = Belief(Can_{task,t}) + Belief(Will_{task,t})$, where AT is artificial trust, and Can and $Will$ are the competence and willingness beliefs, respectively. For simplicity of trust calculation, we consider three tasks: searching rooms, removing obstacles, and rescuing victims. At the start of the search and rescue mission ($t = 0$), the values of Can and $Will$ are 0 for all three tasks. The AT belief is bound between -1 and 1.

The beliefs in competence and willingness are updated by adding a factor f, through behaviour and context observation. This factor depends on the importance and interdependence level of the task. The value of f can be small (0.1), medium (0.2), or large (0.4), and in a positive (+) or negative (-) direction. For example, lying about a critically injured victim's location has a large and negative (-0.4) f adjustments on both competence and willingness, while lying about a mildly injured victim's location has a medium and negative (-0.2) f.

Furthermore, an additional preference factor $P(task)$ is added when computing the willingness belief at a given time: $Belief(Will_{task,t}) = Belief(Will_{task,t-1}) + f(will, task) + P(task)$. This study considers that preferences can impact the willingness of a human to perform a certain task, independent of their competence. 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. For example, complex tasks are associated with negative emotions and are deemed less engaging [42], which may lead participants to choose tasks with the least effort associated to them [11].

Finally, we consider a confidence value that reflects the level of consolidation of each belief. Since the study is dealing with short-lived collaboration, blindly making decisions based on trust in early stages can negatively impact the overall goal. 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 [26]. The value of confidence, C is $\in [0, 1]$ and increases when consistent observations related to one belief occur, i.e., more observations with the same tendency, more confidence in the belief.

To decide on whether to engage in a trust relationship, the agent computes the trust decision. This decision is made by 1) comparing the artificial trust belief with the set thresholds for a required task and 2) verifying whether the confidence is higher than the confidence threshold. More information on the code can be found in Section 6.

3.3 User study

To test the effects of different types of artificial trust communication, we designed a between-subject user study. This study was approved by Delft University of Technology's ethical committee (HREC), with ID 5063, and the dataset can be found in Section 6. We compare the effect of six artificial trust communication types (baseline, real-time textual explanations, real-time visual communication, real-time visual communication and explanations, occasional textual communication and explanations, and occasional visual communication and explanations) on the human's trust and satisfaction.

3.3.1 *Participants.*

To conduct this experiment, 120 participants were recruited using the author's personal networks. All participants resided in Europe and most belonged to the 18-24 age group (96). Ages were between 18 and 54. Most participants had an academic Computer Science-related background (89) and were Bachelor students or graduates (65). Regarding gender, 86 participants identified themselves as men, 31 as women, and 1 as non-binary. Some of them had experience with the MATRX software (31). The gaming experience ranged from no experience (8), very little experience (21), some experience (37), to a lot of experience (54). In terms of the simulation environment, 76 of the participants used macOS as their operating system, and 44 used Windows.

3.4 Communication conditions

Basic communication was already available on the environment chosen, which was used as a baseline (B). As shown in Fig. 1, this included messages from the RescueBot (in white) regarding its actions, or requests of help. The RescueBot informs which in area they will search, which victims they have found where, and whether they will rescue any victim (and take them to the safezone). This communication does not include information regarding the RescueBot's AT, nor any justification/explanation for it. As such, five types of AT communication were developed, besides the baseline communication (six communication conditions in total). These conditions were designed to provide a broad exploration of different possible ways of communicating, varying in timing, modality, and level.

*Real-time short textual explanations (**RTE**):*. This type of communication is the closest to the baseline. It adds to the baseline by sharing information with the human teammate whenever the artificial trust beliefs are altered based on an action. For example, if the participant mentions they found a victim but the victim is not there, (action) then the RescueBot will mention that the trust regarding searching victims decreased because of that (explanation). Some of these explanations can be found in Table 1.

*Real-time visual (graphical) communication (**RV**):*. In this condition, the AT beliefs are presented to the human through a bar chart (see Fig. 3). They are updated in real-time, as the mission happens. The bars have different colours depending on whether the human is trusted (green) for a task, not trusted (red), or in between (yellow), and are accompanied by an emoticon. This method is transparent but not explainable.

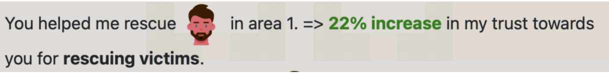
*Real-time visual (graphical) communication + explanations (**RVE**):*. Similar to RV, this method presents bar charts of the artificial trust beliefs (see Fig. 4). However, this condition presents the values of competence, willingness and confidence separately but not per task, without the presence of colours or emoticons. In addition, this condition presents a line plot with overall AT values and its changes in time. The user can hover for further textual explanations regarding the changes.

*Occasional textual summary of changes (**OTE**):*. Instead of providing information in real-time, this approach provides two text-based summaries during the mission and one at the end. Each summary provides an extensive description of the status of the game, as well as the AT levels. Furthermore, these summaries explain how human actions impacted AT beliefs, as well as how AT beliefs impacted the RescueBot behaviour (an example of the latter is showed in Fig. 2). Text regarding positive and negative impacts was displayed in green and red, respectively. During each progress point, the game was paused, and the summary was displayed as a pop-up, covering most of the screen. After checking the fours pages of summary, the player could resume the game by clicking on a closing button.

*Occasional visual (graphical) summary of changes (**OVE**):*. Occasional visual (graphical) summary of changes: Similar to OTE, this communication of AT beliefs is done through an occasional visual summary, as in Fig. 5. This visual summary allowed hovering for textual explanations.

3.4.1 Measures.

Subjective Measures: To measure the participant's self-reported *trust* and *satisfaction*, two validated questionnaires were used: the Trust Scale for the XAI Context and the Explanation Satisfaction Scale [25]. 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.



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

Figure 2: Part of the communication of AT in OTE condition.

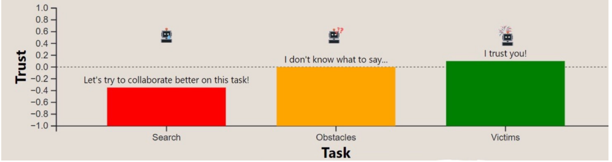
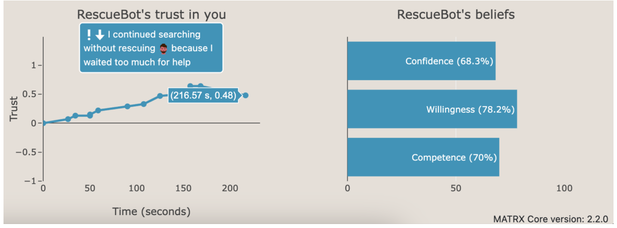
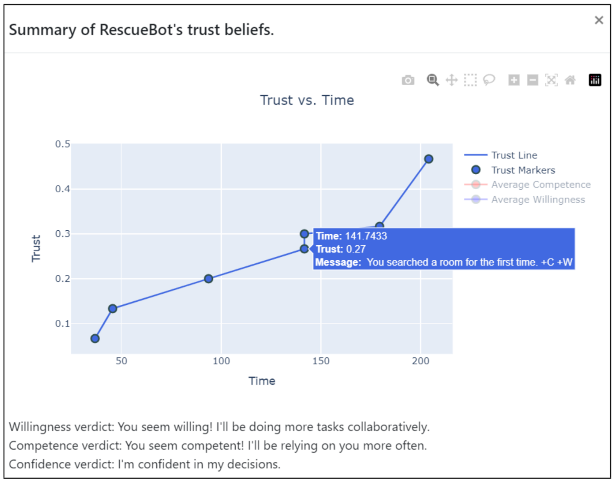


Figure 3: Communication of AT in RV condition.



! ↓ I continued searching without rescuing because I waited too much for help

Figure 4: Communication of AT in RVE condition showing explanation on hovering


Summary of RescueBot's trust beliefs.

Time: 141.7433
Trust: 0.27
Message: You searched a room for the first time. +C +W

Willingness verdict: You seem willing! I'll be doing more tasks collaboratively.
Competence verdict: You seem competent! I'll be relying on you more often.
Confidence verdict: I'm confident in my decisions.

Figure 5: Communication of AT in OVE condition showing explanation on hovering.

Table 1: Overview of some action-explanation pairs present in RTE communication.

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

Objective Measures: During the simulation, objective metrics were recorded and collected in the background for analysis. Specifically, we logged the number of messages received from the participant, the number of human and RescueBot actions (both individual and joint), the number of successful joint tasks proposed by the RescueBot, the completion time of the simulation in ticks, and the artificial trust beliefs of the RescueBot at the end of the simulation (**competence**, **willingness**, **confidence**). These logs were then used to compute multiple metrics potentially capable of indicating trust and satisfaction. Particularly, we computed the final **artificial trust**, the **communication rate**, the **action rate**, the **level of collaboration** in tasks, and **compliance**, i.e., the proportion of tasks suggested by RescueBot that the human takes. Both high communication rate and compliance may indicate trust [30].

3.5 Procedure

The participants were first asked to read a research overview and complete a consent form before proceeding with the experiment. After consenting, they were randomly assigned to one of the experimental conditions. 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 participants then followed a tutorial in a toy environment to familiarize themselves with the tasks, controls, and chat system. After the completion of the tutorial, a brief explanation of the trust model was given. They were informed about the mental model of the artificial agent, the definitions of trust (competence/willingness/confidence), and the behavioural adaptations. Depending on the condition, participants were familiarized with communication method. Following that, the official task would begin. The user was instructed to collaborate with the RescueBot during the search and rescue mission. Once the game was completed, the user would fill in the questionnaires.

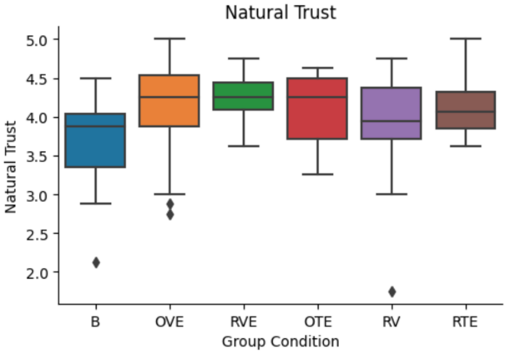


Figure 6: Distribution of the results, per condition, regarding (natural) trust.

4 Results

In order to compare trust and satisfaction measures across conditions, we ran a Kruskal-Wallis analysis, since the data did not meet the normality assumption. Kruskal-Wallis showed a significant difference among conditions for the metrics of **natural trust** ($H = 13$, $p = 0.02$), **satisfaction** ($H = 16$, $p < 0.01$), **competence** ($H = 12.3$, $p = 0.03$). The distribution per condition of these three metrics can be found in Fig. 6, 7 and 8, where trust is specified as natural trust, and competence as artificial competence, to demonstrate it is a computed belief, part of artificial trust. For these three metrics, we proceeded towards a pairwise analysis, which can be found in

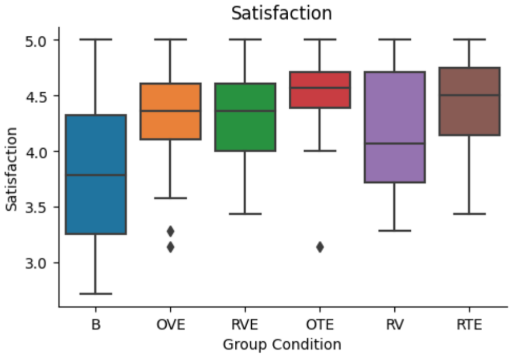


Figure 7: Distribution of the results, per condition, regarding satisfaction.

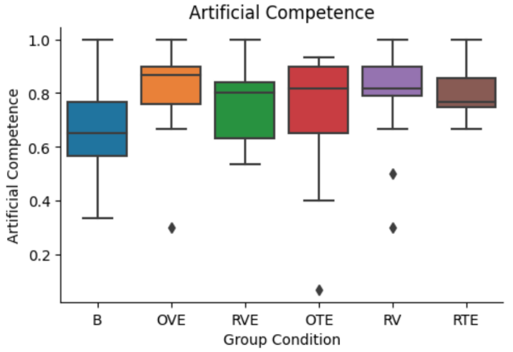


Figure 8: Distribution of the results, per condition, regarding (artificial) competence.

table 2. Depending on the normality of the conditions involved, we ran either a t-test (parametric) or a Mann-Whitney U test (non-parametric). We found statistically significant averages differences between some of the conditions and the baseline, but not between any of the other conditions. Particularly, all conditions except RV present a significant difference both in trust and satisfaction when compared to the baseline. Three conditions (OVE, RV, and RTE) show a significant difference in competence when compared to the baseline.

We also found differences across conditions in the action rate metrics. However, this was discarded since this metric seemed to be confounded by several variables: the operative system in which the environment was run, as well as the age, game experience, and educational background of the participants. All other metrics were either not significant.

5 Discussion

5.1 Results

Our results suggest that explaining the artificial trust values, and not simply being transparent about them, improves the human teammate's trust and satisfaction. Accordingly, all conditions with this *level* of information, (RVE, RT, OTE, and OVE) seem to have outperformed the baseline in terms of trust and satisfaction. Existing literature corroborates that explanations, not just transparency, increase trust, see e.g., [4, 61]. Similarly, other works also found explanations of mental models to increase satisfaction [32, 65]. The only condition without explanations, the RV condition (real-time visual communication), was the only condition that did not show a significant difference when compared to baseline. However, we can also see that this condition has a higher mean for participant competence than the baseline. This suggests the communication was not harmful to the performance, and even improved it. In any case, this condition did not show any significant difference when compared with the other communication conditions (besides baseline), which suggests it was also not significantly worse than the others.

In terms of *modality*, [54] suggests hybrid explanations are preferred by users, and RVE and OVE can both be considered hybrid. Besides the above-mentioned works, [46] also showed that the visual modality was less effective when compared to text, audio, or hybrid approaches. The RV condition, which did not show a significant difference in trust and satisfaction, was only visual. From our results, we cannot conclude which modality or timing is better, since AT communication conditions did not show significant differences when compared against each other. These results are not unexpected as the conditions were not designed to isolate the effects of modality and timing.

Although most objective metrics (potentially indicative of trust) did not show any statistically significant difference among conditions, competence did. Competence, or artificial competence, is one of the beliefs that constitutes artificial trust, besides willingness. Three conditions (OVE, RV, and RTE) showed a significant higher competence than baseline. This suggests that the human behaviour was affected by the communication of artificial trust. Research shows that effective communication, including transparency and explanations, can affect team efficiency and operator performance [13, 40, 66]. In particular, this may mean that the human behaviour was affected by the artificial trust itself, as suggested by [9]. This may be because certain feelings were triggered, or simply because the participants took the AT communication as a guide for how to behave.

5.2 Limitations and Future Work

Our main limitation was that the conditions were not very comparable between each other. As mentioned before, each condition was developed by a different CS student, which increased the differences in implementation. This means that the effects of modality, timing and level were not isolated, making it hard to control for these variables. For example, we could not conclude whether real-time or occasional explanations worked better, since there were other manipulations among the conditions. In contrast, our contributions are more diverse and can serve as baseline for further exploration of modalities, timing, and level in the communication of AT in human-AI teams.

In our study, the artificial trust values were consistently high across conditions. This was possibly because our experimental setup naturally encouraged trustworthy behaviour from participants. Having mainly positive communication of AT poses a limitation, introducing questions about the communication of distrust assessments. Future research should explore how humans react to distrust signals from AI (e.g., “I don't trust you due to past failures”)

Table 2: Statistically significance pairwise differences between baseline and other communication conditions. The test column specifies which test was used, either t-test or Mann-Whitney U test (MW U). The statistic column presents the Z value for Mann-Whitney and t-statistic for t-test.

Variable	Condition 1	Condition 2	Test	Statistic	p-value
Trust	B (M=3.66, SD=0.59)	OVE (M=4.1, SD=0.64)	MW U	111.50	<0.05
		RVE (M=4.24, SD=0.32)	t-test	-3.82	<0.01
		OTE (M=4.09, SD=0.44)	MW U	109.5	<0.05
		RTE (M=4.12, SD=0.4)	t-test	-2.86	<0.01
Satisfaction	B (M=3.79, STD=0.68)	OVE (M=4.28, STD=0.53)	t-test	-2.51	<0.05
		RVE (M=4.31, STD=0.42)	t-test	-2.87	<0.01
		OTE (M=4.51, STD=0.41)	MW U	72	<0.01
		RTE (M=4.41, STD=0.47)	t-test	-3.31	<0.01
Competence	B (M=0.66, STD=0.18)	OVE (M=0.82, STD=0.15)	MW U	96	<0.01
		RV (M=0.8, STD=0.16)	MW U	104	<0.05
		RTE (M=0.8, STD=0.09)	t-test	-2.94	<0.01

in different contexts. While withholding information or even deception might improve the human-AI relationship, as suggested by Rogers's work on deception [48], we believe this approach is ethically problematic. However, explicitly expressing distrust could also damage human-AI collaboration. This creates a significant challenge that requires further investigation to balance transparency with effective teamwork.

Another limitation is the demographics of our sample. Given that this was a Computer Science BSc final project, many participants were from similar circles of background, age, area of studies, etc. This may have given us biased results and should be further investigated with, for example, participants with lower knowledge of computers and AI. Moreover, the study focused on short-term trust assessments, having the users play a single mission and report their subjective experiences. Longitudinal studies could explore how the artificial trust model, as well as its communication, may impact trust over extended periods of interaction with the system.

6 Conclusion

In this paper we present the effects of artificial trust (AT) communication in human-AI teamwork. We ran a user study (N=120) to evaluate five different AT communication methods that vary in level of information, modality and timing. For this, we implemented an AT model which is based on the beliefs of competence and willingness, and that takes into account human preferences and overall confidence in the beliefs. Our results show that explanations of AT during human-agent teamwork can improve both trust and satisfaction. Furthermore, results also suggest that some conditions impact human behaviour, increasing their competence. This work marks the first step towards understanding the impacts of AT and its communication in human-agent teamwork, and raises questions regarding the communication of distrust in teamwork. Both communication and mutual appropriate trust are major pillars for effective collaboration between humans and advanced interactive artificial intelligence and need to be further research.

Acknowledgments

This paper is the result of a BSc final project from the CSE BSc of Delft University of Technology. We thank all the participants who made this possible, and acknowledge the support of the Delft AI Initiative (AI*MAN lab), and The Dutch Research Council (NWO) (Grant 1136993).

Author contributions

CCJ and MLT conceptualized the RQs and study design, building upon a project developed together with CMJ. CCJ developed the artificial trust (AT) model conceptually. ED, EU, RL, and SM were students doing their BSc final project, closely supervised by CCJ and assessed by MLT. They jointly implemented the AT model's core functionality, with individual contributions to specific experimental conditions: ED implemented RVE, EU implemented RV, RL implemented OTE, and SM implemented OVE. The experiment execution and data collection were conducted by ED, EU, RL, and SM. Data analysis was performed by CCJ, ED, EU, RL, and SM. CCJ wrote the published version of the manuscript, incorporating initial separate reports prepared by ED, EU, SM, and RL. The manuscript was improved by MLT, ED, and SM, and approved by EU, RL, and CMJ. Although not an author of the manuscript, we would like to acknowledge the contributions of Tamer Sahin in implementing and running the experimental condition RTE.

Supplementary Materials & Information

The code and data used in this study can be found on this *Github repository* and on 4TU.ResearchData [10], respectively.

References

[1] Ighoyota Ben Ajenaghughrure, Sonia Claudia DaCosta Sousa, and David Lamas. 2020. Measuring Trust with Psychophysiological Signals: A Systematic Mapping Study of Approaches Used. *Multimodal Technol. Interact.* 4, 3 (2020), 63. doi:10.3390/MTI4030063

[2] Alican Akman and Björn W. Schuller. 2024. Audio Explainable Artificial Intelligence: A Review. *Intelligent Computing* 3 (2024), 0074. doi:10.34133/icomputing.0074 arXiv:https://spj.science.org/doi/pdf/10.34133/icomputing.0074

[3] Arsha Ali, Hebert Azevedo-Sa, Dawn M Tilbury, and Lionel P Robert Jr. 2022. Heterogeneous human–robot task allocation based on artificial trust. *Scientific Reports* 12, 1 (2022), 15304.

[4] Alessa Angerschmid, Kevin Theuermann, Andreas Holzinger, Fang Chen, and Jianlong Zhou. 2022. Effects of Fairness and Explanation on Trust in Ethical AI. In *Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August 23–26, 2022, Proceedings* (Vienna, Austria). Springer-Verlag, Berlin, Heidelberg, 51–67. doi:10.1007/978-3-031-14463-9_4

[5] Lilit Avetisyan, Jackie Ayoub, and Feng Zhou. 2022. Investigating Explanations in Conditional and Highly Automated Driving: The Effects of Situation Awareness and Modality. *CoRR* abs/2207.07496 (2022). doi:10.48550/ARXIV.2207.07496 arXiv:2207.07496

[6] Hebert Azevedo-Sa, X Jessie Yang, Lionel P Robert, and Dawn M Tilbury. 2021. A unified bi-directional model for natural and artificial trust in human–robot collaboration. *IEEE robotics and automation letters* 6, 3 (2021), 5913–5920.

[7] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2019. *On the utility of learning about humans for human-AI coordination.* Curran Associates Inc., Red Hook, NY, USA.

[8] Carolina Centeio Jorge, Nikki H. Bouman, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Frontiers Robotics AI* 10 (2023). doi:10.3389/FROBT.2023.1143723

[9] Carolina Centeio Jorge, Ewart J De Visser, Myrthe L Tielman, Catholijn M Jonker, and Lionel P Robert. 2024. Artificial Trust in Mutually Adaptive Human-Machine Teams. In *Proceedings of the AAAI Symposium Series*, Vol. 4. 18–23.

[10] Carolina Centeio Jorge, Elena Dumitrescu, C.M. Jonker, Razvan Loghin, Sahar Marossi, et al. 2025. Dataset of User Study "Artificial Trust Communication in a 2D grid-world Collaborative Search and Rescue Scenario". https://doi.org/10.4121/ace287c9-7a02-4d1f-aef7-8b306448edd5.v1. 4TU.ResearchData, dataset.

[11] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. How Should an AI Trust its Human Teammates? Exploring Possible Cues of Artificial Trust. *ACM Trans. Interact. Intell. Syst.* 14, 1 (2024), 5:1–5:26. doi:10.1145/3635475

[12] Cheng Chen, Mengqi Liao, and S. Shyam Sundar. 2024. When to Explain? Exploring the Effects of Explanation Timing on User Perceptions and Trust in AI systems. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems* (Austin, TX, USA) *(TAS '24)*. Association for Computing Machinery, New York, NY, USA, Article 10, 17 pages. doi:10.1145/3686038.3686066

[13] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[14] Yuanzhe Chen, Panpan Xu, and Liu Ren. 2018. Sequence Synopsis: Optimize Visual Summary of Temporal Event Data. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 45–55. doi:10.1109/TVCG.2017.2745083

[15] Erin K. Chiou, Mustafa Demir, Verica Buchanan, Christopher C. Corral, Mica R. Endsley, Glenn J. Lematta, Nancy J. Cooke, and Nathan J. McNeese. 2022. Towards Human–Robot Teaming: Tradeoffs of Explanation-Based Communication Strategies in a Virtual Search and Rescue Task. *International Journal of Social Robotics* 14, 5 (2022), 1117–1136. doi:10.1007/s12369-021-00834-1

[16] Joachim De Greeff, Tina Mioch, Willeke Van Vught, Koen Hindriks, Mark A Neerincx, and Ivana Kruijff-Korbayová. 2018. Persistent robot-assisted disaster response. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 99–100.

[17] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J Hepenstal, Christopher A Miller, and Dylan D Schmorrow. 2019. Trust engineering for human-AI teams. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 322–326.

[18] Rino Falcone and Cristiano Castelfranchi. 2004. Trust Dynamics: How Trust Is Influenced by Direct Experiences and by Trust Itself. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2004* 2, 740–747. doi:10.1109/AAMAS.2004.286

[19] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. 2011. From Manifesta to Krypta: The Relevance of Categories for Trusting Others. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4 (03 2011). doi:10.1145/2438653.2438662

[20] Laurent Frering, Gerald Steinbauer-Wagner, and Andreas Holzinger. 2025. Integrating Belief-Desire-Intention agents with large language models for reliable human–robot interaction and explainable Artificial Intelligence. *Engineering Applications of Artificial Intelligence* 141 (2025), 109771. doi:10.1016/j.engappai.2024.109771

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

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

[23] Cedric Goubard and Yiannis Demiris. 2023. Cooking Up Trust: Eye Gaze and Posture for Trust-Aware Action Selection in Human-Robot Collaboration. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (<conf-loc>, <city>Edinburgh</city>, <country>United Kingdom</country>, </conf-loc>) *(TAS '23)*. Association for Computing Machinery, New York, NY, USA, Article 34, 5 pages. doi:10.1145/3597512.3597518

[24] Cedric Goubard and Yiannis Demiris. 2023. Cooking Up Trust: Eye Gaze and Posture for Trust-Aware Action Selection in Human-Robot Collaboration. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems, TAS 2023, Edinburgh, United Kingdom, July 11-12, 2023*. ACM, 34:1–34:5. doi:10.1145/3597512.3597518

[25] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023). doi:10.3389/fcomp.2023.1096257

[26] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User Trust in Intelligent Systems: A Journey Over Time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 164–168. doi:10.1145/2856767.2856811

[27] Carolina Centeio Jorge, Ewart J De Visser, Myrthe L Tielman, Catholijn M Jonker, and Lionel P Robert. 2024. Artificial Trust in Mutually Adaptive Human-Machine Teams. In *Proceedings of the AAAI Symposium Series*, Vol. 4. 18–23.

[28] Angelos Kostis, Maria Bengtsson, and Malin H Näsholm. 2022. Mechanisms and Dynamics in the Interplay of Trust and Distrust: Insights from project-based collaboration. *Organization Studies* 43, 8 (2022), 1173–1196.

[29] Esther S. Kox, Juul van den Boogaard, Vesa Turjaka, and José H. Kerstholt. 2024. The Journey or the Destination: The Impact of Transparency and Goal Attainment on Trust in Human-Robot Teams. *J. Hum.-Robot Interact.* 14, 2, Article 23 (Dec. 2024), 23 pages. doi:10.1145/3702245

[30] 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. 2022. Trust Measurement in Human-Autonomy Teams: Development of a Conceptual Toolkit. *J. Hum.-Robot Interact.* 11, 3, Article 33 (sep 2022), 58 pages. doi:10.1145/3530874

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

[32] Bryan Lavender, Sami Abuhaimed, and Sandip Sen. 2023. Relative Effects of Positive and Negative Explanations on Satisfaction and Performance in Human-Agent Teams. In *The International FLAIRS Conference Proceedings*, Vol. 36.

[33] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. 2023. Trusting Artificial Agents: Communication Trumps Performance. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 299–306.

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

[35] Paul Benjamin Lowry, Ryan M Schuetzler, Justin Scott Giboney, and Thomas A Gregory. 2015. Is trust always better than distrust? The potential value of distrust in newer virtual teams engaged in short-term decision-making. *Group Decision and Negotiation* 24 (2015), 723–752.

[36] Matthew B Luebbers, Aaquib Tabrez, Kyler Ruvane, and Bradley Hayes. 2023. Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming.. In *Robotics: Science and Systems.*

[37] James Mardell. 2015. *Assisting search and rescue through visual attention.* Imperial College London.

[38] Richard Mayer, William Bove, Alexandra Bryman, Rebecca Mars, and Lene Tapangco. 1996. When Less Is More: Meaningful Learning From Visual and Verbal Summaries of Science Textbook Lessons. *Journal of Educational Psychology* 88 (03 1996), 64–73. doi:10.1037/0022-0663.88.1.64

[39] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734.

[40] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.

[41] Ali Noormohammadi-Asl, Ali Ayub, Stephen L. Smith, and Kerstin Dautenhahn. 2023. Adapting to Human Preferences to Lead or Follow in Human-Robot Collaboration: A System Evaluation. In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28-31, 2023.* IEEE, 1851–1858. doi:10.1109/RO-MAN57019.2023.10309328

[42] Heather L. O'Brien, Jaime Arguello, and Rob Capra. 2020. An empirical study of interest, task complexity, and search behaviour on user engagement. *Inf. Process. Manage.* 57, 3 (May 2020), 19 pages. doi:10.1016/j.ipm.2020.102226

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

[44] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* 64, 5 (2022), 904–938. doi:10.1177/0018720820960865

[45] Fabio Paglieri, Cristiano Castelfranchi, Celia da Costa Pereira, Rino Falcone, Andrea Tettamanzi, and Serena Villata. 2013. Trusting the messenger because of the message: Feedback dynamics from information quality to source evaluation. *Computational and Mathematical Organization Theory* 20 (08 2013). doi:10.1007/s10588-013-9166-x

[46] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *UMAP '22: 30th*

ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022, Alejandro Bellogín, Ludovico Boratto, Olga C. Santos, Liliana Ardissono, and Bart P. Knijnenburg (Eds.). ACM, 223–233. doi:10.1145/3503252.3531311

[47] José Bernardo Rocha and Rui Prada. 2025. Procedural Content Generation for Cooperative Games - A Systematic Review. *IEEE Transactions on Games* (2025), 1–14. doi:10.1109/TG.2025.3530419

[48] Kantwon Rogers, Reiden John Allen Webber, Jinhee Chang, Geronimo Gorostiaga Zubizarreta, and Ayanna Howard. 2024. Lie, Repent, Repeat: Exploring Apologies after Repeated Robot Deception. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 602–610. doi:10.1145/3610977.3634980

[49] Elie Saad, Koen V. Hindriks, and Mark A. Neerincx. 2018. Ontology Design for Task Allocation and Management in Urban Search and Rescue Missions. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence, ICAART 2018, Volume 2, Funchal, Madeira, Portugal, January 16-18, 2018*, Ana Paula Rocha and H. Jaap van den Herik (Eds.). SciTePress, 622–629. doi:10.5220/0006661106220629

[50] Jordi Sabater-Mir and Laurent Vercouter. 2013. *Multiagent systems*. MIT Press, Chapter 9.

[51] Eduardo Salas, Dana Sims, and Shawn Burke. 2005. Is there a "Big Five" in Teamwork? *Small Group Research* 36 (10 2005), 555–599. doi:10.1177/1046496405277134

[52] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. 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 (2022), 1–29.

[53] Ashish Sharma. 2012. Consumer Perception and Attitude towards the Visual Elements in Social Campaign Advertisement. *IOSR Journal of Business and Management* 3 (01 2012), 6–17. doi:10.9790/487X-0310617

[54] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. 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, 109–119. doi:10.1145/3397481.3450662

[55] Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra, and Myrthe Tielman. 2024. Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. *European Journal of Work and Organizational Psychology* 33, 2 (2024), 158–171.

[56] Joana Urbano, Ana Paula Rocha, and Eugénio C. Oliveira. 2011. A Dynamic Agents' Behavior Model for Computational Trust. In *Progress in Artificial Intelligence, 15th Portuguese Conference on Artificial Intelligence, EPIA 2011, Lisbon, Portugal, October 10-13, 2011. Proceedings (Lecture Notes in Computer Science, Vol. 7026)*, Luis Antunes and Helena Sofia Pinto (Eds.). Springer, 536–550. doi:10.1007/978-3-642-24769-9_39

[57] Jasper van der Waa and Tjalling Haije. 2023. MATRX: Human Agent Teaming Rapid Experimentation software (version 2.3.2). doi:10.5281/zenodo.8154912

[58] Ruben Verhagen. 2020. Human-Agent Teamwork for Search and Rescue Github Repository, https://github.com/rsverhagen94/TUD-Collaborative-AI-2024. https://github.com/rsverhagen94/TUD-Collaborative-AI-2024

[59] Ruben S Verhagen, Alexandra Marcu, Mark A Neerincx, and Myrthe L Tielman. 2024. The Influence of Interdependence on Trust Calibration in Human-Machine Teams. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, 300–314.

[60] Ruben S Verhagen, Siddharth Mehrotra, Mark A Neerincx, Catholijn M Jonker, and Myrthe L Tielman. 2022. Exploring Effectiveness of Explanations for Appropriate Trust: Lessons from Cognitive Psychology. *arXiv preprint arXiv:2210.03737* (2022).

[61] Ruben S. Verhagen, Mark A. Neerincx, Can Parlar, Marin Vogel, and Myrthe L. Tielman. 2023. Personalized Agent Explanations for Human-Agent Teamwork: Adapting Explanations to User Trust, Workload, and Performance. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh (Eds.). ACM, 2316–2318. doi:10.5555/3545946.3598919

[62] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. 2021. 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, 119–138.

[63] Samuele Vinanzi and Angelo Cangelosi. 2022. CASPER: Cognitive Architecture for Social Perception and Engagement in Robots. *CoRR* abs/2209.01012 (2022). doi:10.48550/ARXIV.2209.01012 arXiv:2209.01012

[64] Michael Winikoff. 2017. Towards Trusting Autonomous Systems. In *Engineering Multi-Agent Systems - 5th International Workshop, EMAS 2017, Sao Paulo, Brazil, May 8-9, 2017, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 10738)*, Amal El Fallah Seghrouchni, Alessandro Ricci, and Tran Cao Son (Eds.). Springer, 3–20. doi:10.1007/978-3-319-91899-0_1

[65] Julia L Wright, Jessie YC Chen, and Shan G Lakhmani. 2019. Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems* 50, 3 (2019), 254–263.

[66] Désirée Zercher, Ekaterina Jussupow, and Armin Heinzl. 2023. When AI joins the team: a literature review on intragroup processes and their effect on team performance in team-AI collaboration. In *European Conference on Information Systems (ECIS) 2023*.

[67] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan J. Mcneese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork. *Proceedings of the ACM on Human-Computer Interaction* 7 (2023), 1–31.

A Questionnaires

The following questionnaire (Table 3) 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 [25]. The items with ∗ represents distrust, i.e., the score needs to be inverted when aggregating results.

Table 3: 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?