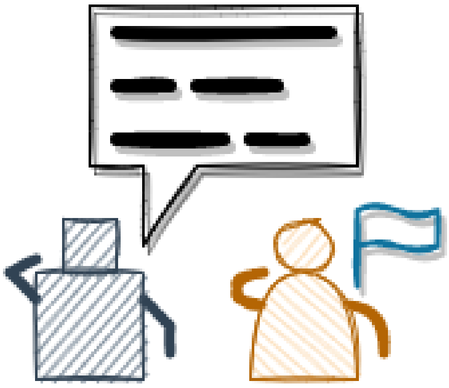


The Influence of Adapting an Agent's Explanation Style to a Human Team Leader Role on Human-Agent Teamwork during a Simulated Search and Rescue Task

by
Ryan Kap

Delft University of Technology
Faculty EEMCS
MSc Computer Science
Software Technology

Student Number:	4426436
Thesis Defence Date:	18 October 2022

Preface

Topic

The topic of this thesis is Human-Agent Teaming. The thesis will study the effect of adapting an explanation style to the team leader role in Human-Agent Teaming (HAT).

Context

The thesis is carried out at the Delft University of Technology at the faculty of Electrical Engineering, Mathematics, and Computer Science (EEMCS) under the research department of Intelligent Systems (INSY) with the research group Interactive Intelligence (II).

Main Findings

The main findings demonstrate that an agent adapting its explanation style to the team leader role using this thesis' HAT design - rather than using randomised explanation styles - significantly increases the team leader's trust in the agent, understandability of the agent, perception of the agent's user-awareness, and satisfaction of the agent's explanations.

Thesis Committee

The thesis committee consists of Prof.dr. Mark Neerincx (thesis advisor), Dr. Myrthe Tielman (daily supervisor), Ruben Verhagen, PhD candidate (daily co-supervisor), and Dr. Pablo Cesar (external committee member).

Acknowledgements

I want to thank my thesis committee for providing me with the opportunity to graduate within the Interactive Intelligence research group. In particular, I want to thank Ruben Verhagen for his constant advice and feedback which kept me motivated throughout this process to improve my knowledge and academic writing. Furthermore, I want to thank Dr. Myrthe Tielman for her advice and insights on how to do proper research and on how the academic world works. Moreover, I want to thank prof.dr. Mark Neerincx for his critical questions and general advice which made me step up my game to deliver a better thesis. Next, I want to thank dr. Pablo Cesar for taking the time and making an effort to be involved in my thesis. Additionally, I want to thank my family and friends, for I could not have completed this thesis without their support. Above all, I want to thank Elisa Fuhrmann, as she was my main pillar of support throughout this entire process.

Abstract

Communication is one of the main challenges in Human-Agent Teams (HATs). An important aspect of communication in HATs is the use of explanation styles. This thesis examines the influence of an explainable agent adapting its explanation style to a supervising human team leader on team performance, trust, situation awareness, collaborative fluency, explanation satisfaction, understandability, and user-awareness. To perform a simulated Search And Rescue (SAR) task, a HAT is designed. With this design, a pre-study is then conducted using a questionnaire to discover the best-ranked explanation styles in the most important situations of the SAR task. Next, the user-study is carried out with 46 participants, using the HAT design and analysed data from the pre-study. There are two conditions: the agent adapting the explanation style to the human team leader and the baseline condition where the explanation style is randomised. The results show that the subjective measurements of trust, understandability, explanation satisfaction, and perceived user-awareness are significantly higher in the adaptive agent condition group. The same cannot be concluded for objective measurements such as team performance and situation awareness.

Keywords human-robot teamwork, human-agent teamwork, explainable AI, explainability, explanation styles, user-study, search and rescue, user-awareness, personalised explanations.

List of Abbreviations

AI	Artificial Intelligence
BDI	Belief-Desire Intent
CDM	Coactive Design Method
CSM	Coactive System Model
FP	Folk Psychology
HAT	Human-Agent Team
HREC	Human Research Ethics Committee
IA	Interdependence Analysis
INSARAG	International Search and Rescue Advisory Group
K9	Canine
LOA	Level of Automation
MATRX	huMan-Agent Teaming Rapid eXperimentation
MT	Medic Team
OPD	Observability, Predictability, and Directability
RT	Rescue Team
SA	Situation Awareness
SAGAT	Situation Awareness Global Assessment Technique
SAR	Search and Rescue
SAT	Situation Awareness-based Agent Transparency
SMM	Shared Mental Model
ST	Search Team
TDP	Team Design Pattern
ToM	Theory of Mind
USAR	Urban Search and Rescue
XAI	eXplainable Artificial Intelligence

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Human-Agent Teams (HATs) are teams that contain at least one human as well as any form of agent, such as a robot, an Artificial Intelligence (AI) system, or any intelligent machine. To have an effective HAT, it is essential that the complex interactions between the human and the agent and the interaction between the agent and their environment is successful [3]. It is crucial for agents in HATs to be understandable and predictable to their human team members for the team to work effectively and efficiently together [4]. Understandability and predictability can be achieved by explainability, which is what the field of eXplainable AI (XAI) focuses on.

HATs are emerging in different domains, but especially in Defense (military) [5], Healthcare (e-Health) [6], and disaster response (e.g., Search and Rescue [7]). In (urban) Search and Rescue (SAR), humans and agents (in this case robots) work together to achieve the same goal: rescue as many people (but also pets if need be) as possible. The benefits of using robots in SAR tasks are that robots never get tired, can get into tiny spaces, and go into dangerous areas where humans cannot go. Conversely, humans are better at recognising danger and have an ethical responsibility, whereas robots do not. Combining these advantages of both humans and robots can give HATs a great advantage over solely human teams, if applied correctly.

One of the main challenges in HATs is communication. Specifically, communication of agents explaining their actions and their reasons for making certain decisions to their team members. These agents are called explainable agents. Previous studies on explainable agents have suggested further investigating communication, evaluation, and personalisation within HATs, since these are still understudied aspects within the field of explainable agents [8]. For instance, van der Waa et al. [9] suggest that within the design of a HAT, explanations that fit with the assigned team roles should be included in the design of the team to make better use of the explanations and their potential. Furthermore, Harbers et al. [10] and Neerincx et al. [11] argue that aspects, such as the role of an agent, should be taken into account when developing a HAT to make the explanations better understandable. Lastly, Verhagen et al. [12] suggest exploring the adaptation of information based on different team roles or tasks. These suggestions show that there is a research gap regarding explanations and team roles within HATs. Adaptation of behaviour is crucial for effective collaboration. Therefore, it would be interesting to research agents adapting their explanation style by using different types of explanations tailored towards team roles.

1.1. Research Questions

Based on the discussed research gaps, the following main research question has been formulated:

How does adapting an agent's explanation style to a human team leader role influence human-agent teamwork during a simulated search and rescue task?

This main research question can be divided into the following subquestions:

RQ1. What is the difference in average team performance, situation awareness, trust in the agent, and both objective and subjective collaborative fluency between an adaptive explainable agent based on the team leader role and a baseline explainable agent providing random explanations?

RQ2. What is the difference in average subjective user-awareness and understandability of the agent, and subjective satisfaction with the agent's explanations between an adaptive explainable agent based on the team leader role and a baseline explainable agent providing random explanations?

RQ3. What is the effect of humans' self-reported gaming experience, user-awareness of the agent, understandability, and satisfaction with the agent's explanations on team performance, situation awareness, perceived trust, and collaborative fluency?

To answer the main research question and the subquestions, one should first be able to answer the following questions:

1. What does the collaboration design of the human-agent team look like for a simulated search and rescue task?

2. How does one adapt the agent's explanations to the human team leader role of a HAT performing a search and rescue task?

1.2. Thesis Outline

After the current introductory chapter, this thesis will continue with background information and related work in Chapter 2. Then, the design of the HAT used in the pre-study and user-study will be described in Chapter 3. Next, the methodology and results of the conducted pre-study will be described in Chapter 4. Then, the research methodology for the user-study will be described in Chapter 5. Afterwards, the results of the user-study will be displayed in Chapter 6. Lastly, Chapter 7 presents the discussion and corresponding conclusion of this thesis, followed by suggestions for future research.

Chapter 2

Background & Related Work

2.1. Human-Agent Teaming

In this section, concepts such as joint activities, shared mental models, interdependence, team roles, and team design patterns are described and discussed. These topics are deemed relevant information for this thesis.

2.1.1. Human-Agent Teams

Teaming intelligence is argued to be the most prominent gap for intelligent systems [13]. Knowledge, skills, and strategies for managing interdependence are examples of teaming intelligence. The use of HATs could be the solution to fill this gap of teaming intelligence. One reason is that HATs have certain advantages over human teams and teams of agents. HATs can combine the best of both worlds because human reasoning and algorithmic reasoning are characteristically different [14]. For example, agents are typically better at handling huge amounts of data in comparison to humans, whereas humans are usually better at recognising danger than agents [10].

As mentioned before, communication is one of the main challenges in HATs. The field of XAI is trying to tackle this problem by making AI systems, which can be agents in HATs, more understandable and predictable to humans. Verhagen et al. [4] propose a framework that defines how an incomprehensible system becomes an interpretable system through transparency and how an interpretable system becomes understandable through explainability. This leads to the idea that system understandability should be the main goal when developing agents for HATs.

In addition to system understandability, it is paramount to have good coordination between humans and agents during joint activities in order to combine the best of both worlds [15]. Joint activities are a set of actions performed by a group of coordinating entities [16]. Bradshaw et al. [15] state that there are different kinds of joint activities, namely co-allocation, cooperation, and collaboration. Collaboration happens when all entities involved are trying to achieve the same goal and there is interdependence of actions. There are requirements for joint activities, such as members of a team being mutually predictable and maintaining common ground [16]. Predictability concerns the ability to predict behaviour. However, Dragan et al. [17] argue that predictability is also about understanding others' actions. Maintaining common ground concerns the exchange of information about each other's goals, intentions, and observations.

Johnson et al. [1] name different requirements for joint activities than Klein et al. [16], namely Observability, Predictability, and Directability (OPD). Observability concerns the status and particular knowledge of an agent, which is observable to others and helps with coordination in teamwork. Directability concerns the ability to direct the behaviour of team members and vice versa, such as team role assignment and task allocation. In addition, one can direct behaviour in a subtle way by providing advice and giving warnings [1].

2.1.2. Shared Knowledge

One way to structure shared knowledge between team members is through a Shared Mental Model (SMM). A mental model is the internal representation of the knowledge someone has of the environment [18]. In turn, a SMM refers to the collection of the mental models of all members of a team. SMMs appertain to how similar or compatible the knowledge structures of team members are. Mental models enable humans to describe, understand, explain, and predict the system environment and the actions of their fellow team members, communicate more effectively, and better coordinate the team's behaviour. [18, 10, 19]. Salas et al. [20] propose using SMMs as a coordinating mechanism for effective teamwork as compatible knowledge structures will increase team performance [21, 22]. Moreover, the use of mental models is expected to enhance trust, satisfaction with the team, and system performance [23]. In case there is a lack of common ground, it could result in confusion and failed expectations. Consequently, this could hurt the performance of a team [24].

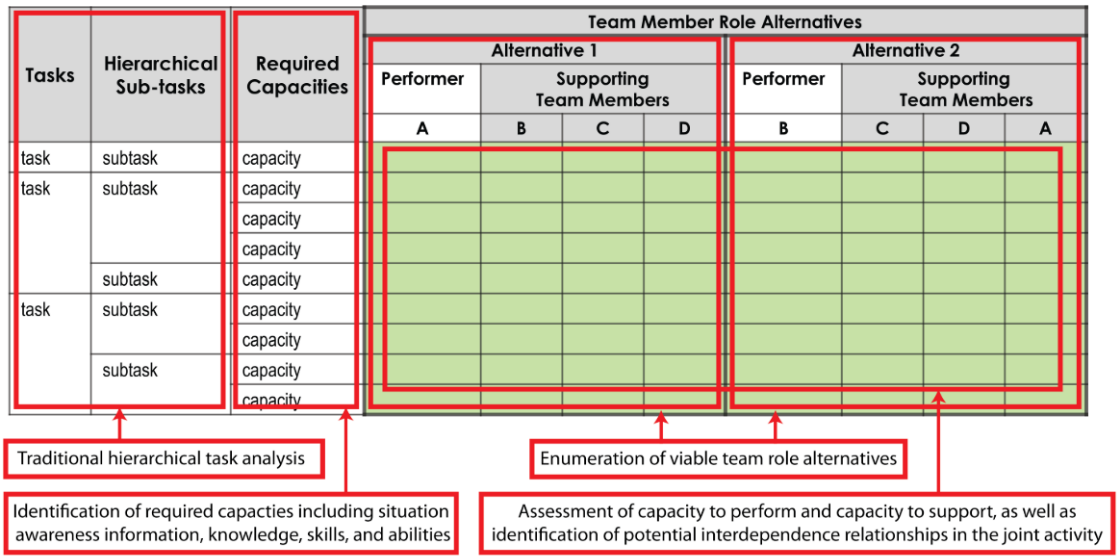
2.1.3. Interdependence

Different models of teamwork have various categories, characteristics, and properties, but the importance of interdependence is prevalent in all of them [13]. For example, in the study of Bradshaw et al. [15], they defined intention and interdependence as criteria for joint activity. In addition, it is argued that teamwork must have some level of interdependence [15, 21, 22]. Interdependence is often oversimplified as team members having mutual dependence, but interdependence rather deals with the set of relationships that manage the dependencies between team members. Johnson et al. [25] define interdependence as the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activities. Hard interdependence comes from a lack of capability and soft interdependence comes from an opportunity to be more effective or efficient together. For instance, a hard dependency is when a stuck survivor can only be rescued by two team members rather than one. A soft dependency is when carrying a survivor alone is possible but carrying the survivor with two team members is faster and more reliable. In other words, interdependence is about the combination of the capacity of the members (e.g., knowledge and abilities), the complementary relationships between them, and the joint activities being performed by the members [13]. Thus, with a deep understanding of the interdependence relationships between team members, one will have more insight into what kind of coordination is required for joint activities [25].

2.1.4. Coactive Design

Bearing the importance of interdependence relationships within a team in mind, Johnson et al. [25] propose the Coactive Design Method (CDM), which is useful for understanding the interdependence relationships within a HAT. This approach is motivated by the desire for agents to be more like human interdependent team members [1]. CDM is based on the Coactive System Model (CSM) which deals with supporting necessary interdependence relationships in teamwork through the aforementioned OPD requirements. In CSM, the interdependence of team members in joint activities is a critical factor for the design of human-agent systems. The goal of the design method is to create a system that has just enough OPD to support the interdependence in HATs.

Table 2.1 depicts the Coactive Design Method's three main processes, namely an identification process, a selection and implementation process, and an evaluation of change process. This method fills in a table called the Interdependence Analysis (IA) table. Table 2.1 displays the the colours to fill in the IA table. Having identified the interdependence relations, these relationships are then used in the design, making the achievement of the objectives of coordination, collaboration, and teamwork possible in an easier way.



Tasks	Hierarchical Sub-tasks	Required Capacities	Team Member Role Alternatives							
			Alternative 1				Alternative 2			
			Performer	Supporting Team Members			Performer	Supporting Team Members		
			A	B	C	D	B	C	D	A
task	subtask	capacity								
task	subtask	capacity								
		capacity								
		capacity								
	subtask	capacity								
task	subtask	capacity								
		capacity								
	subtask	capacity								
		capacity								

Traditional hierarchical task analysis

Identification of required capacties including situation awareness information, knowledge, skills, and abilities

Enumeration of viable team role alternatives

Assessment of capacity to perform and capacity to support, as well as identification of potential interdependence relationships in the joint activity

Observability, Predictability, and Directability (OPD) requirements derive from the role alternatives the designer chooses to support, their associated interdependence relationships, and the required capacities.

Figure 2.1: Explanation of the different areas of the Interdependence Analysis (IA) table, from Johnson et al. [1].

Team Member Role Alternatives	
Performer	Supporting Team Member
I can do it all	My assistance could improve efficiency
I can do it all but my reliability is $<$ 100%	My assistance could improve reliability
I can contribute but need assistance	My assistance is required
I cannot do it	I cannot provide assistance

Table 2.1: Legend for colours in interdependence analysis table.

2.1.5. Team Roles

Now that joint activities and interdependence within HATs have been described, the next relevant concept to describe is team roles. Teams need to have properly defined team roles to be able to efficiently work together [26, 27]. There are three main reasons to have clearly defined roles in HATs: no tasks are repeated, no tasks are overlooked, and no time is wasted on repeated assignments of agents [28]. It is important to acknowledge your team members' roles, as an acknowledgement of team members' assigned team roles forms expectations concerning how to interact with one another [15]. Moreover, as previously mentioned in section 2.1.2, knowledge about the interaction of a team, such as their roles and interdependence relationships, enables team members to predict interaction reciprocally, which in turn improves teamwork [29].

Generally, team roles are defined through the different responsibilities and goals of all team members [30]. These team role definitions need to be accepted by each team member to ensure a smooth cooperation [31]. For human teams, there are different ways to categorise team roles. For example, Burke et al. [30] showcase different taxonomies of team roles, ranging from Bales [32]'s task roles (e.g., asking for or giving introductions, opinions, and suggestions) to Mathieu et al. [33]'s task roles (e.g., organiser, do-er, challenger, and innovator). The idea of connecting roles to a task in some way seems to be included in nearly all proposed models for team roles.

The team roles that are being used in HATS usually look like roles from solely human teams, even if the interactions in HAT differ from solely human teams [28]. Fortunately for humans, the roles filled by humans will likely not disappear in the foreseeable future mainly due to the human's ability to understand behaviour patterns, (human) intentions, large implications, and ethical responsibilities, whereas agents do not [14].

There are three roles defined for agents in HATs according to Sycara and Sukthankar [22]. The first role is individual assistant (e.g., expert systems and recommender systems) and focuses on supporting a single human. The second role is team assistant (i.e., an agent that cooperates with the entire team). A team assistant agent mostly supports coordination activities within the team. Harbers et al. [10] mention that they are not aware of explanation approaches for team assistant agents. The third role is equal team member. In this role, an agent aims to apply the same reasoning and perform the same tasks as that of an equal, human team member. Different roles of agents in HATs yield different types of explanations [10]. The role of the agent with respect to the human is also worth considering. For example, during a search and rescue task, when a robot comes across a survivor and if that robot is assisting medical personnel, it will have to take into account its own role in relation to the survivor and the medical personnel when communicating [34].

Currently, the possibility of assigning roles to agents in HATs is somewhat limited because of a lack of available technology, but as teaming intelligence is constantly being developed, the complexity and possibility for agents to take on more roles will continue to grow. This diversification of agent roles in HATs will require clearly defined roles [28].

2.1.6. Team Design Patterns

Now that it is clear what is important for collaboration in HATs and how the definitions of team members' roles can aid the team, one should look at designing solutions for problems that HATs can solve. A Team Design Pattern (TDP) is a method for designing solutions for HATs that are simple, general, descriptive, and well-structured [2]. TDPs provide an intuitive method to aid the discussion on HAT solutions, even for non-experts. TDPs can be used to visualise the collaboration design between humans and agents, while also describing the requirements, advantages, and disadvantages of such designs.

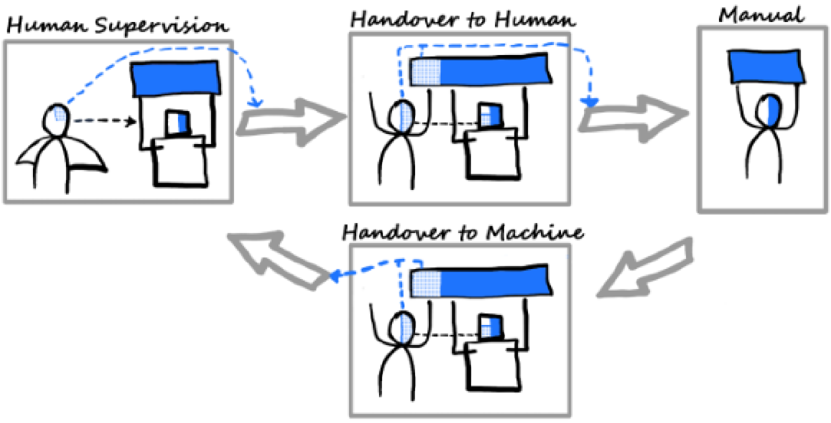


Figure 2.2: Example of a TDP, from Van Diggelen and Johnson [2] (Figure 6)

In TDPs, a distinction is made between different types of work: direct work, indirect work, and off-task work. Additionally, distinctions are made between types of joint work, namely physical, cognitive, and asymmetric joint work. Van Diggelen and Johnson [2, p. 3] define joint work as “a set of direct and/or indirect work carried out by two or more actors related to the same common goal”. The TDP method categorises three types of communication: physically proximal, not physically in range but there is remote communication possible, and inhibited communication. Lastly, there are ways to describe transitions between different TDPs. For example, transitioning between the type of work, the degree of engagement, or the work division. A related example of TDP can be found in the work of van der Waa et al. [9]. They have created three different TDPs for the level of control the human has in HATs: data-driven decision support, dynamic task allocation, and supervised autonomy.

2.2. Important Factors in HATs

Now that relevant concepts for this thesis relating to HATs have been described, important factors for HATs will be discussed. There are additional factors to the ones previously mentioned that are crucial for the success of agents in HATs: situation awareness, mutual trust, and collaborative fluency. In addition, one should look at transparency, explainability, and user satisfaction to make the agents understandable, as proposed by Verhagen et al. [4]’s framework. These are factors that might be influenced by adapting the explanation style.

2.2.1. Team Performance

One of the first factors regarding HATs one might think of is performance. To measure team performance, performance criteria are needed. Jonker et al. [35] mention that typical criteria for HATs are the time to complete a task, how successful the task was completed, and the manner in which the task was completed. Measuring the way a task was completed includes different elements, such as intentional or unexpected side effects of completing the task, the circumstances while the task was being performed, and the interaction between the team members.

2.2.2. Collaborative Fluency

A factor related to team performance is collaborative fluency: the quality of joint activities within human teams or HATs [36]. The term collaborative fluency is closely related to task efficiency, but it is not the same. A study in which this subtle difference became apparent is Hoffman and Breazeal [37]. They measured fluency and efficiency in task completion for two groups, where one group's fluency was rated higher than the other group. However, no difference was found in efficiency levels between the two groups.

Hoffman [36] mention two ways to measure collaborative fluency, namely subjectively and objectively. Subjective fluency is based on the perception of a person and objective fluency is more quantifiable in a given interaction. For subjective fluency, Hoffman [36] use a questionnaire consisting of six parts: Human-Robot Fluency, Robot Contribution, Trust in Robot, Robot Teammate Traits, Working Alliance for Human-Robot Teams, and Improvement. Moreover, examples of Hoffman's objective measurement of collaborative fluency are robot idle time and human idle time. These are measured as the percentage that a team member was idle during a task.

In Hoffman [36]'s study, the researchers found that the human idle time percentage positively correlated with subjective fluency. They argue that this correlation might be explained by the participants' focus on the agents' contribution and that the participants considered their idle time as pleasant.

2.2.3. Situation Awareness and Transparency

Situation Awareness (SA) is an important factor for HATs. According toEndsley [38], SA refers to the perception of elements in the environment, the comprehension of their meaning, and the projection of their status. Proper SA is especially important in remote robot operations, meaning the robot operator cannot see the robot from their control terminal (e.g., as in some cases with SAR). Inspired by Endsley [38], Chen and Barnes [14] have developed the Situation Awareness-based Agent Transparency (SAT) model. The goal of this model is to organise the issues related to supporting the awareness agent and human team members have of each other. System transparency is needed to regulate the human team members' trust in the agents and for maintaining SA [14].

SAT consists of three independent hierarchical levels: goals and actions, reasoning, and projections. Each of these levels describes the information an agent needs to communicate in order to maintain a transparent interaction. The cumulative result of these three levels results in SA. At the first level (SAT 1), the agent communicates their current status, action, or plan. At the second level (SAT 2), the agent communicates their reasoning process. At the third and final level (SAT 3), the agent communicates their projections or predictions, where the level of certainty is key.

A commonly used method to measure SA is called the Situation Awareness Global Assessment Technique (SAGAT) [39], which measures SA globally and in a direct manner, rather than concluding SA via other variables. This removes any problems with measuring SA after the completion of an experiment. SAGAT requires the random freezing of a simulation. During this pause, participants have to answer questions on the different SAT levels. SAGAT is best for measuring individual SA in real-time with human-in-the-loop situations such as a simulated search and rescue task on the computer.

2.2.4. Trust

Trust is also an important factor for HATs. In human-technology interaction, trust has a considerable influence on emotions and thereby affects the interaction. Furthermore, trust plays an important part in accepting technology as part of a human-technology interaction. Moreover, an emotional response to technology can contribute to positive performance and safety levels [40]. Lee and See [40, p. 51] define operator trust as “the attitude that an agent will help achieve an individual’s goals in a situation characteri[s]ed by uncertainty and vulnerability”. In addition, Mayer et al. [41] state that the factors concerning perceived trustworthiness are ability, benevolence, and integrity.

There are different ways to evaluate a user’s trust in a system both objectively and subjectively. A challenge regarding the objective measurement of trust arises in formulating trust in HATs mathematically. Examples of measuring objective trust are reliance or compliance. Within HATs, one could measure how reliant an agent is or how often agents comply with team members. Contrastively, subjective trust is easier to measure.Hoffman et al. [42] recommend a scale for measuring trust in XAI systems by asking users whether they are confident in the system and whether they found the system to be predictable, efficient, reliable, and believable. Their recommendation is largely adapted from Cahour and Forzy [43].

2.2.5. System Understanding

System understanding can be measured through system performance. If a user understands the system well, it can perform better or more efficiently. Moreover, Madsen and Gregor [44] argue that understandability is one of five factors that can be used for analysing trust. They measured understandability through a questionnaire. Regrettably, their work did not indicate whether a Likert scale was used to measure understandability. Hellström and Bensch [45] define the term ‘understandable robots’ as a robot’s ability to make itself understood to its human interlocutor, resulting in the human’s capability to explain and predict the behaviour of the robot.

The use of explanations will improve a user’s understanding of the system, which in turn increases user satisfaction, trust, and performance [23]. Swartout [46] argue that, among other things, understandability is a property of effective explanations. Therefore, explanation satisfaction could be a valuable indicator of effective explanations.

2.2.6. User-Awareness

Anjomshoae et al. [8] performed a systematic literature study on explainable agents. They found that research into context- and user-aware agent explanations have been meagre so far. Moreover, the studies that did investigate such explanations focused mostly on context-aware explanations. This shows that personalised or user-aware agent explanations have been investigated insufficiently. Therefore, the next factor discussed is user-awareness.

User-aware explanations and context-aware explanations sometimes get confused with each other or are used interchangeably, since some interpret user-aware explanations as explanations that look at the context of a user. For example, Brezillon [47], Doyle et al. [48] include the environment of the user in their definition of ‘user context’, which raises the question of whether explanations using the environment of the user are context-aware explanations, user-aware explanations, or both. Although the two explanation levels are closely related, the difference is that user-aware explanations consider anything that has to do with the user (e.g., their age, skill, and role in the team) [23, 8], whereas context-aware explanations use everything outside of the user-context (i.e., the situation and environment).

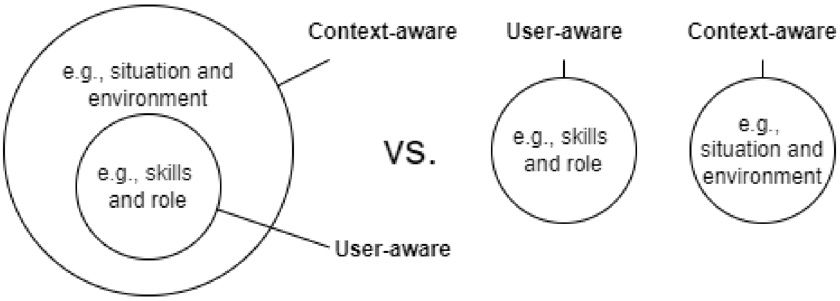


Figure 2.3: The distinction between user-aware and context-aware explanations.

Nonetheless, good explanations should take the receiving user (and their context) into account because context awareness and personalisation are key factors for explainable agents [6].

2.3. Explaining Behaviour

As has been made clear, one of the most important factors for effective collaboration within a team is how well team members explain their behaviour. Therefore, this chapter will cover this aspect in more depth. Firstly, the social sciences and psychological background are discussed. Secondly, the relationship between beliefs, desires, and intentions is discussed. Thirdly, adaptiveness and explanations are discussed. Fourthly, different explanation types or styles are discussed.

2.3.1. Social Sciences and Psychological Background

When collaborating, one needs to explain their behaviour and there are multiple ways in which one can go about this. According to Anjomshoae et al. [8], Folk Psychology (FP) [49, 50, 51] and Theory of Mind (ToM) [8, 52, 50, 53] are the most commonly used social sciences and psychological background for explanation methods in modelling explainable agents to explain their behaviour. FP concerns the ability to attribute mental states to agents [51, 50, 53]. Mental states include beliefs, desires, and intentions. ToM regards the idea of assigning mental states to agents from FP and uses these mental states to predict and explain behaviour [54, 49]. ToM also relates to explaining and predicting human behaviour via attitudes such as believing, wanting, and hoping [49]. Thus, you can explain behaviour by something called the intentional stance [55], where you assume that the action of an agent is the result of its intentions. When the intentional stance is applied to non-human entities, it is also called Anthropomorphism.

Intention is formed by having three things, namely, that the behaviour is based on desire, a belief that the behaviour can be performed, and a belief that the desire can be achieved [53]. An action is called intentional if an action is performed by an agent that has the ability and the awareness of performing that action. Different explanation theories use different terms for people's explanations of intentional human behaviour. Nonetheless, whether called intentional stance, FP or ToM, they all refer to explanations in terms of mental concepts like beliefs, intentions, and goals [56]. Therefore, a common framework to programme agents for HATs using these models is to use a Belief-Desire Intent (BDI) explanation model [8].

2.3.2. Beliefs, Desires, Intentions

Agents that use the BDI model accommodate the possibility of generating understandable and useful explanations for a human team member [11]. Examples of these can be found in tactical combat teams [57, 58] and fire-fighting teams [59]. Harbers et al. [60] argue that the use of the BDI model is fitting for developing self-explaining agents. According to Harbers et al. [10], there are four typical steps to take for a BDI agent. Firstly, perceiving the world and updating the agent's internal beliefs and goals. Secondly, using its current goals and beliefs to select a plan. Thirdly, selecting an intention. Lastly, selecting a new plan if the intention is a sub-goal or performing the intention if the action is atomic.

2.3.3. Adaptive Explanations

According to Stowers et al. [61], adaption in HATs can be investigated in two ways. The first way is adaptability (i.e., human-controlled adaptation). For example, the adaptation of the level of automation [62]. The second way is adaptiveness (i.e., a machine-controlled adaptation of the level of automation). Using an adaptive system rather than an adaptable system can result in better competence and trust [63]. An example of adaptability is when humans in HATs decide the order of task execution for their agent team member, such as choosing the rescue plan it has to execute. However, research has shown that humans do not always allocate tasking in the most effective way [64]. To confine this limitation of humans, agents could use adaptiveness to take control or to only give control to humans in specific situations. Adaptiveness can also be applied to explanations of agents, e.g., providing explanations based on the receiver's team role.

2.3.4. Explanation Types

Ideally, agents in HATs use different explanation types. Explanations types and explanations styles are used interchangeably, but they all reference the same idea. Harbers et al. [10] argue that agents should be equipped with explanation capabilities in order to perform well in HATs. Effective explanations can improve the performance of a human-AI system [23]. Additionally, the use of explanations improves a user's mental model and their understanding of the mental model [65]. The use of explanations advocates for a justified trust and mistrust in systems and simultaneously reduces unjustified trust and mistrust [66]. On the other hand, an agent should not explain everything and explain constantly, since explanations should not be too long [6]. Otherwise, the information might not be fully understood or received. The following explanation types are considered for this thesis: goal-based explanations, belief-based explanations, confidence explanations, feature attributions, contrastive explanations, and counterfactual explanations.

Goal-based Explanations Goal-based explanations show the (sub)goal related to the action performed [67, 6]. Usually, the selected goal is the first subgoal related to the action instead of the main goal. For example, "I am going to carry this survivor because *I want* to rescue as many survivors as possible".

Belief-based Explanations Belief-based explanations contain a belief an agent has related to the action [6]. For example, "I am going to carry this survivor now because *I believe* this survivor is critically injured".

Confidence Explanations A confidence explanation uses a confidence percentage (between 0% and 100%) for certain external sensors. For example, the agent could say "I am *80% certain* I heard something in room B". Confidence explanations are connected to the certainties of agents. Knowing agents' certainties will improve HATs performance [68, 69, 70, 71].

Feature Attributions Feature attributions display additional features, which are deemed important for leading to a certain decision. For example, "I am going to rescue this survivor now because *my sensors indicate that this area is dangerous*". Furthermore, Ribeiro et al. [72] argue that feature attributions are not necessarily interpretable data as they do not provide a deep understanding of why the chosen features were important. However, these features are sometimes quick to interpret and can be easily visualised [9].

Contrastive Explanations Contrastive explanations display why a certain decision is taken over another decision. An agent compares the current situation with another situation of concern. Contrastive explanations state a part of the agent's reasoning. These explanations can be used to answer the question: Why not X? [53] An example of a contrastive explanation is: "I am going to rescue this survivor now because *continuing the search might result in not finding possible survivors*". van der Waa et al. [9] studied the effect of contrastive explanations and found that the use of contrastive explanations improves humans' understanding of agents' reasoning and makes the agent's reasoning more predictable.

Counterfactual Explanations Counterfactual explanations highlight certain features that, if they had been different, would have resulted in the agent taking another action than the one it has now taken [73]. For example, "I am going to rescue this survivor now. *If this survivor was not critically injured, I would have rescued them later*".

2.4. Search and Rescue

Search and Rescue is an operation that involves the search and rescue of people or other living beings whom the SAR team believe to be in distress, lost, sick, or injured either in remote or difficult-to-reach areas, such as mountains, deserts, forests, or seas [74]. Most SAR teams are categorised by their location of operation, such as the sea (e.g., coastguards), mountain (e.g., the Österreichischer Bergrettungsdienst which rescues people in skiing areas), or combat (e.g., the United States Air Force).

The International Search and Rescue Advisory Group (INSARAG) is a global network which belongs to the United Nations [75]. INSARAG consists of more than 90 countries and organisations. These deal with urban SAR-related issues. Therefore, they aim to establish a minimum international standard for SAR teams. INSARAG have written guidelines for (international) Urban Search and Rescue (USAR) operations. These guidelines mention three teams that can be relevant to the current thesis: The Search Team (ST), the Rescue Team (RT), and the Medic Team (MT).

2.4.1. Materials

Presently, dogs (K9s) are used in SAR missions. The main reason is that they can do certain things humans cannot. For example, they can crawl into small spaces and can sense things better via their sense of smell. This is an excellent example of team members with different capabilities working together, just like in HATs. Human SAR teams have the ability to use heavy tools such as drilling equipment to free people from collapsed buildings. The use of an agent (e.g., a robot) could benefit human teams by doing the heavy lifting and performing tasks that are dangerous to humans.

2.4.2. Roles

According to the 2020 INSARAG guidelines, there are different roles at play during a SAR: management, medical, logistics, search, and rescue. The following roles are the ones most teams have in common: team leader, medic, communications, technical searcher, and technical rescuer. The human role in SAR teams, where the human is remote, is usually supervisor or operator. When the human is in the field together with the agent(s), the human and agent(s) are usually peers [34].

2.4.3. HAT

(Urban) SAR is considered a high-profile HAT challenge. The highly unstructured nature of SAR environments imposes severe challenges on agents' mobility, communication, (internal) map-building, and situation awareness [34]. However, there are some present examples of HATs in SAR missions. For example, The Rising Sprawl-Tuned Autonomous Robot is a search and rescue robot developed by researchers at the Ben-Gurion University of the Negev in Israël [76]. Another example is SmokeBot, which is a robot designed to help firefighters by entering smoke-filled buildings and creating maps that allow them to move quickly through areas with extremely limited visibility [77]. A third example is the water rescue robot EMILY, which strives to save lives in beach, ocean, river, and flood situations and also aids in search and recovery missions using sonar technology [78].

Schneider and Wildermuth [79] analysed requirements for robots in SAR. They argue that searching robots need navigation and mapping capabilities and casualty identification. They also argue for the requirements of a rescue robot in SAR, namely, communication, support, and remote mobile manipulation. Teams of semi-autonomous robots can provide valuable assistance in USAR missions by efficiently exploring cluttered environments and searching for potential victims. Their advantage over solely teleoperated robots is that they can address the task handling and situation awareness limitations of human operators by providing some level of autonomy to the multi-robot team [80].

2.4.4. MATRX

To conduct experiments with SAR teams without bringing any danger to participants, one can simulate SAR tasks. An example of a programme that can be used for simulations is MATRX, which stands for huMan-Agent Teaming Rapid eXperimentation. MATRX is a software package designed for team tasks with HATs. The programme was developed to fill the gap of a team task library for HAT research. MATRX supports TDP, building tasks, basic autonomous agents, a basic user interface, and data logging.

Chapter 3

HAT Design

This chapter presents the HAT design that is used in this thesis. As mentioned in Chapter 1, there are two questions which need answering before one is able to answer the main research question. The goal of this chapter is to answer the first of these two questions: What does the collaboration design of the human-agent team look like for a simulated search and rescue task? This HAT design will then be used to answer the second question in Chapter 4: Pre-study.

In a HAT design, certain aspects need to be considered. Goodrich and Schultz [34] name five attributes that might influence the interaction between humans and agents in a HAT: the structure of the team, the level of autonomy each team member possesses, the shaping of the task, the adaption, learning, and training of humans and agents, and the way of information exchange. Therefore, the first thing that has to be described in this HAT design is the structure of the team (e.g., which roles the human and agents have). This will include the level of autonomy. Second, the shaping of the tasks of the team and any information needed to perform these tasks will be described. This information can be used to reveal the most important scenarios needed to collect information on adapting the explanation style. The information exchange will take place via the chat box system of MATRX and the learning aspect will be offered through a tutorial map.

3.1. Team Roles

For assigning team roles, inspiration is drawn from Section 2.4: Search and Rescue. The INSARAG mentions two teams for SAR teams: the search team and the rescue team, which both have a team leader. The focus of the current HAT design lies on the combination of these two teams. The main focus of the search team is the risk assessment of an area and the subsequent search. The main focus of the rescue team is making and executing a rescue plan. The INSARAG also mentions a medic team; since that is not the focus of this thesis, such a team is not used in the current HAT design, but a safety zone is used in its place.

Given the complexity of implementing and analysing more than one human role in a HAT and the limited time frame for this thesis, only one human role will be applied. Since the effect of adapting explanation styles will be clearer if communication takes place between a pair of team members, the human will be part of a team with one agent. Additionally, an overload of explanations by agents on the human will be reduced considerably in this manner. However, there may be some scenarios where multiple agents are needed. For example, carrying a critically injured person or lifting heavy objects requires the assistance of at least one more agent. In these cases, the extra agents will not explain their behaviour to ensure that a potential overload of information will not be an issue.

3.1.1. The Human Role

The human will take on the role of supervising team leader in both phases of the search and rescue (SAR). van der Waa et al. [9] mention that if the workload for the human team member is too high, the explanations are not really used even though they were perceived as useful in hindsight. Moreover, if the human has a supervising role, the number of explanations given by the agent can be more controlled in opposition to the human also active in the field. For example, when the human is also in-field, participants could be using different strategies due to moving to complete the same task. This can lead to a broad range of situations, which in turn could lead to a big difference in the type and amount of explanations given by the agent. Therefore, the role of the human will be of supervising team leader that will not go into the field themselves.

3.1.2. The Agent Role

The agent will take on the team role of supporting individual team members or supporting the team as a whole [22]. As mentioned in Section 3.1, there are no known explanation types for this role. This is one of the reasons for conducting the pre-study. The agent will aid the team in assisting with coordination among the human team members, aiding focus or attention, and by aiding communication between team members. The agent will not take on their role as an equal team member, because their levels of reasoning and performing tasks are not on the same level as that of their human team members. For example, the reliability and capability of assessing whether a human is injured or judging whether an area is safe can be improved by a human supervisor. Agents can take three actions within HATs. The first action is that the agent should give the human advice on certain decisions. For example, whether to call for medical assistance. Secondly, the agent can ask the team leader for advice. For example, when the agent is not certain about a situation. Thirdly, the agent will take autonomous actions when they are certain of the result that will follow their action or certain of their sensors that lead to making this action.

The three actions the agent can take can be described as different Levels of Automation (LOA) [81]. Taking an autonomous action can be described as level 7, where the computer executes automatically and then necessarily informs the human. The agent giving or needing advice can be described as level 3, where the computer narrows the selection down to a few choices. Another LOA is where the agent is on equal footing with the other human team members, but that level is not part of this thesis.

3.2. Tasks of the Team

The tasks of the team can be split up into two phases: the search phase and the rescue phase. The following table provides an overview of the different tasks and capabilities of the search robot and rescue robot in their respective phases.

Task	Explore Disaster Site	Clear Obstacles	Clear Entries	Access Areas	Explore Safe Areas	Explore Dangerous Areas	Carry Healthy Survivors	Carry Injured Survivors	Carry Stuck Survivors
Search Phase	✓	✓	✓	✓	✓	?	✓	?	?
Rescue Phase	×	×	×	×	✓	✓	✓	✓	?

Table 3.1: Difference of capabilities of the agents in their corresponding phase.

The check mark means that the agent in that phase can perform the task. The cross mark means that the agent in that phase cannot perform the task. The question mark means that the agent in that phase can perform the task, but at the cost of a penalty. For example, the agents can free a stuck survivor alone, but during the rescue phase, another rescue robot can help to reduce the rescue time in half. Agents in the search phase are called search robots and agents in the rescue phase are called rescue robots in this thesis.

3.2.1. Search Phase

During the search phase, the search robot searches the disaster area for subareas where survivors could be. The entries to these subareas might be blocked and there might be obstacles on the roads between subareas. Only the search phase agent can explore the disaster site, clear these obstacles, and clear entries.

If a search robot enters a subarea, it assesses the safety of that area. If the area is dangerous and the search robot explores it, there is a chance of getting damaged and having to recover. A solution would be to let the rescue robot search this dangerous area during the rescue phase, since they will not get damaged because of their special equipment for instance. If the search robot comes across a survivor during the search phase and the survivor has been assessed to be healthy (i.e., not critically injured or stuck, meaning that it does not need additional medical attention to get to safety), the search robot can take the survivor to safety autonomously. In all other instances, the search robot can let the rescue robot rescue this survivor during the rescue phase. The search robot has a 50% chance to fail to rescue critically injured survivors. In contrast, the rescue robot will not fail to rescue a critically injured survivor, because of its special equipment for instance. Lastly, if a survivor is stuck, the search robot can free the stuck survivor alone, but this will take twice the amount of rescue time in comparison to when two rescue robots free the survivor during the rescue phase. The search robot can choose to let the rescue robot rescue a stuck survivor. Only during the rescue phase can the rescue robot call another rescue robot to free the stuck survivor together without the time penalty of freeing a stuck survivor alone.

If all areas have been explored, the search robot will go back to things that the team may have skipped over, such as blocked entries or rescuing healthy survivors. This going back to skipped steps continues until time has run out or until all areas have been explored.

3.2.2. Rescue Phase

Once the search robot has returned safely to the safety zone, the search robot will be replaced by a rescue robot. During the rescue phase, the team leader devises a plan to rescue the survivors using the information gained from the search phase. Afterwards, the rescue robot will execute the items on the rescue plan in order. Whenever the rescue robot comes across a healthy survivor, it will automatically rescue them. If the rescue robot comes across a critically injured survivor, it will automatically rescue them as well, but sometimes it will advise to move the rescuing of that survivor to the back of the rescue plan if the rescue time would become too long. Lastly, if the rescue robot finds a stuck survivor, it will advise to call another rescue robot to rescue this stuck survivor together to reduce the rescue time in half. The robot will give this advice based on the time remaining on the clock and the estimated rescue time. If the rescue robot is not certain of something it will ask for advice from the team leader. When the rescue phase time is up or when every item on the rescue plan has been executed, the search and rescue task is finished. The team leader will then be able to determine the team performance score (i.e., the number of rescued survivors).

3.2.3. Coactive Design

In order to determine the interdependence relationships within the team performing the tasks of SAR, the coactive design method will be used as described in Section 2.1.4. The following table is the result of the Coactive Design method:

Tasks	**Hierarchical Subtasks**	**Required Capacities**	**Team Member Role Alternatives**	
			Alternative 1	
			Performer	Support
			Agent	**Team Leader**
Search for areas	Remove obstacle	Recognise obstacle	🟩	🟨
		Strength to remove obstacle	🟩	🟥
		Interpret if obstacle is obstructing	🟩	🟨
		Decide to remove obstacle	🟨	🟨
	Assess entry	Recognise entry	🟩	🟨
		Interpret if entry is accessible	🟩	🟨
		Calculate time to complete	🟩	🟥
		Strength to clear entry	🟩	🟥
Assess area	Search the area	Travel around area	🟩	🟥
		Recognise survivor	🟨	🟨
		Remember places been in area	🟩	🟥
Rescue survivor	Assess survivor	Recognise survivor	🟨	🟨
		Assess injuries	🟨	🟨
	Pick-up survivor	Have right equipment	🟩	🟥
Go to area	Go to area	Locate area entry	🟩	🟥
		Travel	🟩	🟥
Make rescue plan	Decide on order of items	Prioritise items	🟧	🟧

Table 3.2: Overview of interdependence analysis.

Next, we can obtain the ODP requirements from Table 3.2 using Table 2.1: the agents have the strength to execute all tasks and the human team leader is not strictly needed. However, the human team leader can improve the level of reliability in the recognising, interpreting, and assessing done by the robot.

3.2.4. Information Needed for the Tasks

The next aspect to consider is what information is needed for the agent to make a decision. The following table provides an overview of all the information that is needed for the agent to take an action, give advice, or ask for advice:

Category	Information
Disaster site	*Areas*: Number of areas in total, number of areas left to find, and number of areas already found *Obstacles*: Obstacle locations, whether they are blocking the path, and estimated removal time *Survivors*: Survivors found, survivors already rescued, and survivors left to rescue
Area	Distance to area, assessed or not, dangerous or safe, number of survivors inside, area explored or not
Entry	Entry assessed or not, entry clear or blocked
Survivor	Critically injured or not, stuck or not
Agent	*Capabilities*: Clear obstacles, assess entries, clear entries, assess areas, enter unsafe areas, carry injured survivors

Table 3.3: Overview of all the information that is needed for the agent to take an action, give advice, or ask for advice.

The next step is to use the information in the table in the relevant situations and determine when the agent can perform an action, give advice, and ask for advice. The agent will take an autonomous action if it is confident in its beliefs and if this action would contribute the most to its goal. If the agent thinks that the best thing to do, given the available information, is to divert from an expected action (i.e., removing an obstacle when finding an obstacle), then the agent gives or needs advice. For example, the robot has found a non-injured survivor, but there are still rooms left to explore and there is little time left to do so; the robot can either continue with the search phase and leave rescuing this survivor for the rescue phase or it can rescue the survivor immediately, taking up more time during the search phase. Whether the robot gives advice or needs advice is based on the confidence the robot has in its beliefs that lead to the decision. When the agent's confidence level is below a certain threshold, it will ask for advice; otherwise, it will give advice.

3.2.5. Team Design Pattern

When the agent takes an autonomous action and then afterwards gives advice on something, it switches its level of automation. When the agent switches its level of automation, this can be displayed using a TDP. The following figure displays the dynamic level automation of the robot and the relationship between the team leader and the search and rescue robots:

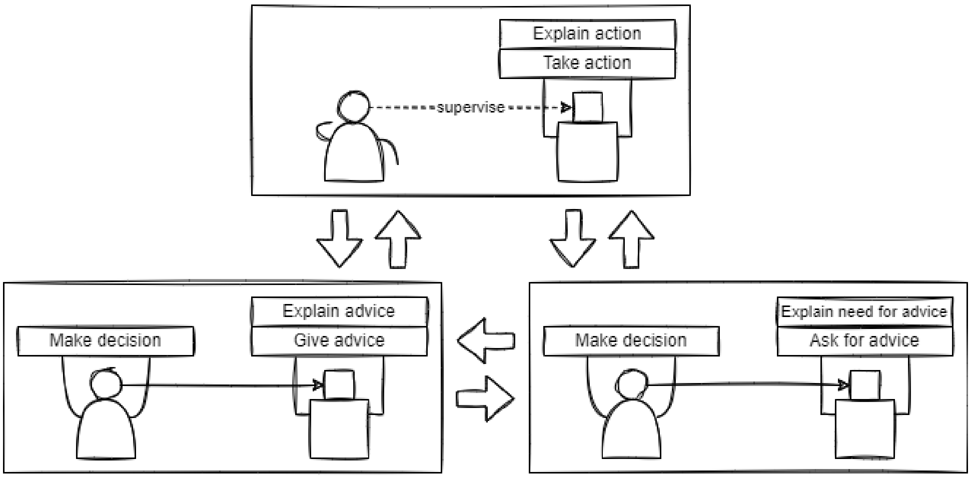


Figure 3.1: TDP of the dynamic level automation within the HAT.

In Figure 3.1, one can see how the team leader and the robots change dynamically between the robot taking an autonomous action and the team leader supervising, the robot giving advice and the team leader making a decision, and the robot asking for advice and the team leader making a decision.

3.2.6. Scenarios

It is important to consider the most important and occurring scenarios from the task in order to predict how the explanations are formed throughout the task. The scenarios where the agent wants to give advice or needs advice are the most relevant for the current thesis, with a focus on scenarios where giving explanations is relevant. There are many scenarios possible during the user-study. Four scenarios were chosen to be considered relevant enough for exploring the explanations for adapting the agents' explanation style to the team leader role: the agent has found an obstacle, the agent has found an entry to an area, the agent has found an unassessed area, and the agent has found a survivor. After re-evaluation, the scenarios where the agent finds an obstacle and where the agent finds an entry to an area were merged because the explanations were too similar.

With regard to the explanations, the focus was put on the time it takes to complete an action. It is important that the human team leader makes informed decisions, no matter the explanation type. All the scenarios with the different explanation types for each phase in the task can be found in Appendix B. This chapter has answered the following question: What does the collaboration design of the human-agent team look like for a simulated search and rescue task? Using this design, the next question can be answered via the pre-study in Chapter 4: How does one adapt the agent's explanations to the human team leader role of a HAT performing a search and rescue task?

Chapter 4

Pre-study

This chapter presents an online study investigating what type of explanations fit the role of the team leader mentioned in Chapter 3: HAT Design. The goal of the current chapter is to answer this question by exploring if and in which situations confidence-, contrastive-, counterfactual-, goal-based explanations, and feature attributions fit the team leader role of the HAT performing a search and rescue task. The results of this pre-study were used in the 'adaptive' condition of the user-study, which will be discussed in Chapter 5.

4.1. Design

To determine if and in which situations the aforementioned explanation styles fit the team leader role, the baseline in which there is no adapting of explanation styles needed also to be considered. This has raised questions such as: What does the baseline look like? How can one try to minimise external factors influencing the results of the user-study? Answering these questions was paramount in designing the pre-study. It can be difficult to verify that any possible effect is solely caused by adapting the explanation style and not by external factors. An external factor could be that one condition provides more information than the other, which could lead to either a better understanding of the explanation or not reading all the provided information and therefore less understanding of the explanation [82]. Hence, the amount of information provided in the two conditions that are being compared should be the same. Additionally, the explanation styles in the two conditions must be distinct enough to be able to measure any effect. For example, belief-based explanations and confidence explanations were found to be too similar and therefore, belief-based explanations were removed from the list of explanation types for this thesis. Furthermore, it is worth considering how to determine if any measured effect is caused by the adaptation of the explanation style and not caused by the given explanations being perceived as 'good'. With these considerations, the pre-study was designed as follows.

In an online study, there are multiple ways to measure a preference via a questionnaire. For example, choosing the best option from a list, rating each option, or ranking all options. The question type 'ranking' was chosen for this pre-study because of its ability to see the relationship between explanation types. To strengthen confidence in the results of this pre-study, one can ask people with actual experience in the team role, people who have performed tasks belonging to the team role, or closely related experts.

A selection of the most important scenarios from the HAT design performing a search and rescue was made, see section 3.2.6 for more information. Exploring every preference for each possible scenario and behaviour from the agent would require too much time. To reduce the time it takes to complete the questionnaire, questions about the situations 'found obstacle' and 'found entry' were combined with 'found entry' as in these situations the behaviour of the robot and the team leader's choices were too similar. Table 4.1 displays all the different scenarios presented in the expert study:

Situation	Fact	Foil	Feature	Agent's action
Survivor	Carry the survivor	Request medical assistance	Not critically injured	Autonomous Action
Survivor	Let two robots rescue the survivor	Carry the survivor alone	Survivor is stuck	Give Advice
Survivor	Uncertainty	Carry the survivor	Not critically Injured	Need Advice
Entry	Continue search	Clear the entry	Entry blocked by big obstacle (too much time)	Give Advice
Entry	Clear entry	Enter area immediately	Entry blocked by small obstacle	Autonomous Action
Entry	Uncertainty	Clear entry	Entry blocked by small obstacle	Need Advice
Area	Search area	Request different robot	Safe to enter	Autonomous Action
Area	Request different robot	Search area alone	Dangerous	Give Advice
Area	Uncertainty	Search the area	Area is safe	Need Advice

Table 4.1: Selection of most relevant scenarios for this HAT performing SAR.

A scenario consists of a situation with a fact, foil, feature, and agent's action. The fact is the event that occurred and the foil is the event that did not. After presenting a scenario to the participants and letting them rank each explanation, a question was added asking the participants to elaborate on their chosen ranking to gain even more insight. Lastly, to strengthen the idea that the results are explanation styles that fit the team leader role and not just the situation, one can ask the same question within the same scenario, but then ask the participants to envision themselves in a different role, namely the same role as the robot, when ranking the explanations. This means that the robot would still provide its explanation to the human, the goal of the team would remain the same, and the composition of the team would also be the same. To see the full questionnaire, see Appendix I.

4.2. Participants

The pool of participants selected for this pre-study consisted of volunteers from the Lowland Rescue [83]. Lowland Rescue is a member of the UKSAR, which provides the strategic overview and organisation of search and rescue in the UK and Northern Ireland. There are about 1800 volunteers active divided over 36 teams within Lowland Rescue. All teams were emailed to participate in the pre-study, see Appendix F for the email. To keep the data as anonymous as possible, all invitations were sent at the same time. Dutch-speaking experts were excluded from the pool of participants since their level of English might have been insufficient and therefore influence the results. Three weeks after sending the invitation, there were still no responses. Consequently, non-experts were also invited to participate in this pre-study. In the end, there were 21 anonymous participants. The Human Research Ethics Committee approved this pre-study with HREC ID 2181. See Appendix G for the informed consent and Appendix H for the data management plan.

4.3. Procedure

The participants filled out the required informed consent part of the questionnaire to be able to continue with the pre-study. Then, expert participants filled out their roles within their teams. This step was not required for non-expert participants. Lastly, for each of the scenarios from Table 4.1 participants were given information about the scenario so they could envision themselves to be in that scenario and the team leader role. Participants were given five different explanations corresponding to five different explanation styles. They were then asked to rank the five explanations in response to that scenario. Afterwards, the participants were asked to elaborate on their chosen ranking. Lastly, the participants had to re-rank the explanations of that same scenario, but now from a different perspective within their HAT, namely from the same role as the robot. To see an example of the procedure, see questions 12-14 in Appendix I.

4.4. Analysis

Before analysing the data, the data was cleaned using Python. This resulted in having 9 data entries per participant with each entry having 10 ranking scores (5 explanation types times two roles) and textual answers to the elaboration questions. The data can be divided into six data sets for the six different situations, see Table 4.2. Then, for each data set, the average ranking per explanation type was calculated for the two role questions. This resulted in the six data sets containing each 10 data points per participant. Next, a student t-test was used to analyse the difference between the ranking of each explanation type for each situation. If the normality assumption was violated, the Wilcoxon test was used. If normality was found but the homogeneity assumption was violated, a Welch t-test was used. To adjust the p values, the Bonferroni method was applied. Lastly, a within-subjects ANOVA test was used to analyse the difference in ranking between explanation types within the same situation. The Friedman test was used if the data violated the normality assumption. You can find the full code in Appendix J.

4.5. Results

Table 4.2 displays the best ranked average of each explanation type in the six different situations. For each situation and team role, feature attribution was ranked best on average, except for the robot team role and situations with survivors. There, the contrastive explanation was ranked best on average. For the full table of the rankings per situation, team role, and explanation type, see Appendix K.

Situation	Team Role	Best Ranked Explanation Type	Mean	SD
Entry	Team Leader	Feature attribution	4.032	0.767
	Robot	Feature attribution	3.857	0.922
Area	Team Leader	Feature attribution	4.222	0.725
	Robot	Feature attribution	4.270	0.867
Survivor	Team Leader	Feature attribution	3.937	0.807
	Robot	Contrastive expl.	3.698	0.781
Autonomous action	Team Leader	Feature attribution	3.968	0.881
	Robot	Feature attribution	3.714	0.973
Get advice	Team Leader	Feature attribution	4.349	0.764
	Robot	Feature attribution	4.206	0.986
Give advice	Team Leader	Feature attribution	3.873	0.922
	Robot	Feature attribution	3.841	0.886

Table 4.2: Best ranked explanation type based on the six situations in the data.

Figure 4.1 shows the relation between the average ranking between the two roles for all situations per explanation style. The results of the statistical analyses between the rankings of each team role, situation, and explanation type combination are displayed in Table 4.3. It was found that no explanation type was significantly different between the two roles for all situations. For the overview of the paired analyses with the different situation data sets, see Appendix L.

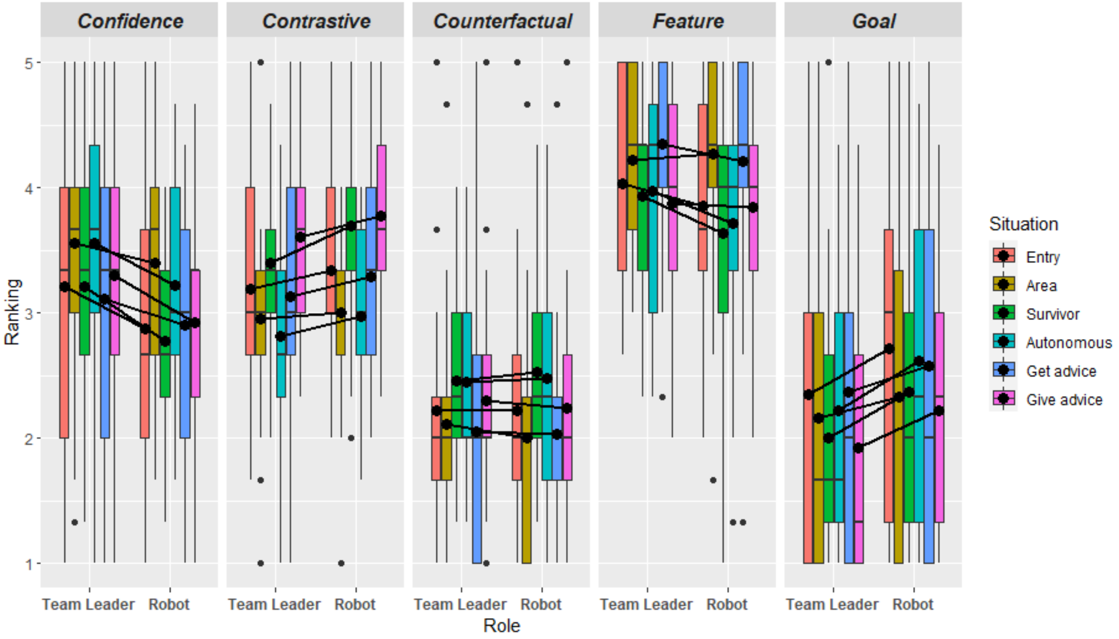


Figure 4.1: Boxplot of comparing the ranking between two roles for all situations per explanation type.

Situation	Expl. Type	Method	Statistics	Adj. p	Sign.	Effect size
Auto. action	Confidence	student t-test	1.154	0.255	ns	0.356
	Contrastive	Wilcoxon	190	0.445	ns	0.120
	Counterfactual	Wilcoxon	218	0.960	ns	0.010
	Feature	Wilcoxon	254	0.392	ns	0.134
	Goal	Wilcoxon	181.5	0.329	ns	0.153
Get advice	Confidence	student t-test	0.597	0.554	ns	0.184
	Contrastive	Wilcoxon	204	0.685	ns	0.065
	Counterfactual	Wilcoxon	213	0.857	ns	0.030
	Feature	Wilcoxon	232.5	0.763	ns	0.048
	Goal	Wilcoxon	204.5	0.690	ns	0.063
Give advice	Confidence	student t-test	1.212	0.233	ns	0.374
	Contrastive	Wilcoxon	194.5	0.516	ns	0.102
	Counterfactual	Wilcoxon	234.5	0.730	ns	0.055
	Feature	Wilcoxon	224.5	0.929	ns	0.016
	Goal	Wilcoxon	180	0.305	ns	0.160

Table 4.3: Analyses of difference in ranking score of the explanation types between the two roles (team leader and robot).

4.6. Discussion

Ideally, each situation would have resulted in a different explanation type to have a clear way to adapt the robots' explanation style throughout the task (see Table 4.2). However, the best-ranked explanation type for each situation is feature attribution. One cannot consider constantly using the same explanation style as adaptive behaviour. Unfortunately, this result could mean that these six situations or the two team roles do not affect the preferred explanation style for the search and rescue. On the bright side, the fact that this explanation type is consistently ranked as best indicates that this explanation type is preferred regardless of the role a team member has within this HAT and regardless of the specific situation of the search and rescue. As a reminder, feature attribution incorporates the most important feature in the explanation when the robot makes a decision, gives advice, or needs advice. As mentioned before, feature attributions do not provide a deep understanding of why a certain feature was chosen. Therefore, the results of the pre-study could indicate that the reason a feature is chosen is quickly interpreted in this context. In search and rescue, the time pressure is high and thus decisions need to be made quickly. Consequently, trust can be an important factor in making quick decisions without explicitly knowing why a certain feature was so important. If the participants' average trust in the robot is high, it could explain why feature attribution was ranked best on average. This assumption can be verified in the user-study.

Even when the participants were asked to assume the role of the robot instead of the team leader role in the exact same situation, feature attribution was ranked best on average for each subset of data, except in situations involving a survivor. In these situations, contrastive explanations were ranked best (M = 3.698, SD = 0.781). After using statistical analyses (see Table 4.3 or Appendix L), no significant difference between the two roles for each combination of situation and explanation type was found. This means that participants envisioning themselves in a different role (that of the robot) does not influence their preference for explanation type. This could be explained by multiple things. First, it could be hard to envision the difference between the roles as a participant even with the text stating the robot's goal and re-explaining the situation. Second, it is possible that the preferred explanation type for these particular situations could be the same even though the roles have different responsibilities and different tasks. Third, these scenarios do not represent the whole search and rescue enough, even though they were picked to be the most important. When selecting these scenarios, the main focus was to find a preference in explanation type for the team leader role, not the robot's team role. An additional study could be done where the main focus is to explore the difference in explanation type between these two roles or a study where the sole focus is the preference for the robot's role within this HAT.

4.6.1. Solution for Adapting the Explanation Style

The results from Table 4.2 cannot be used for the user-study as each situation will use the same explanation type. Two questions need answering to change this. The first question is: If the explanation type for each situation was different, in which situation should the robot change its explanation type? Right now, some situations overlap. For example, the situation could be about an area in a disaster site but also the situation where the robot needs advice. Therefore, there are three ways to solve this:

1. Adapt the explanation style in the situations where the agent finds an *entry* (or obstacle), an *area*, and a *survivor*;
2. Adapt the explanation style in the situations where the agent takes an *autonomous action*, *needs advice* from the team leader, and *gives advice* to the team leader;
3. Adapt the explanation style to all the scenarios from Table 4.1 (this is more difficult to implement and less generalised).

Given the nature of the task and the roles of the robot and the human, the second solution makes the most sense as this is more generalisable for HATs and not just specific for this particular task of search and rescue (entry, area, survivor).

The second question is: How does one solve the 'issue' where the preferences are all the same for the situations? Since the explanation types are the same for each subset, one can look at the second-best-ranked explanation type and see if they are ranked significantly different. If they are not, one could proportionately alternate between the best and the second-best ranked explanation type. This would give it the adaptive nature this thesis is looking for, while still using the information from the pre-study.

Results showed that the ranking within the 'autonomous action' situation was statistically significantly different between the explanation types, $\chi^2(4) = 20$, $p < 0.001$, with a small effect size, $W = 0.238$. The ranking within the 'give advice' situation was statistically significantly different between the explanation types, $\chi^2(4) = 30.4$, $p < 0.001$, with a moderate effect size, $W = 0.362$. The ranking within the 'give advice' situation was statistically significantly different between the explanation types, $\chi^2(4) = 33.3$, $p < 0.001$, with a moderate effect size, $W = 0.397$. Figure 4.2 shows the significant differences between the explanation types for the autonomous action questions, the 'get advice' questions, and the 'give advice' questions. Not all the rankings for the explanation types were distributed normally, therefore Friedman's test was used.

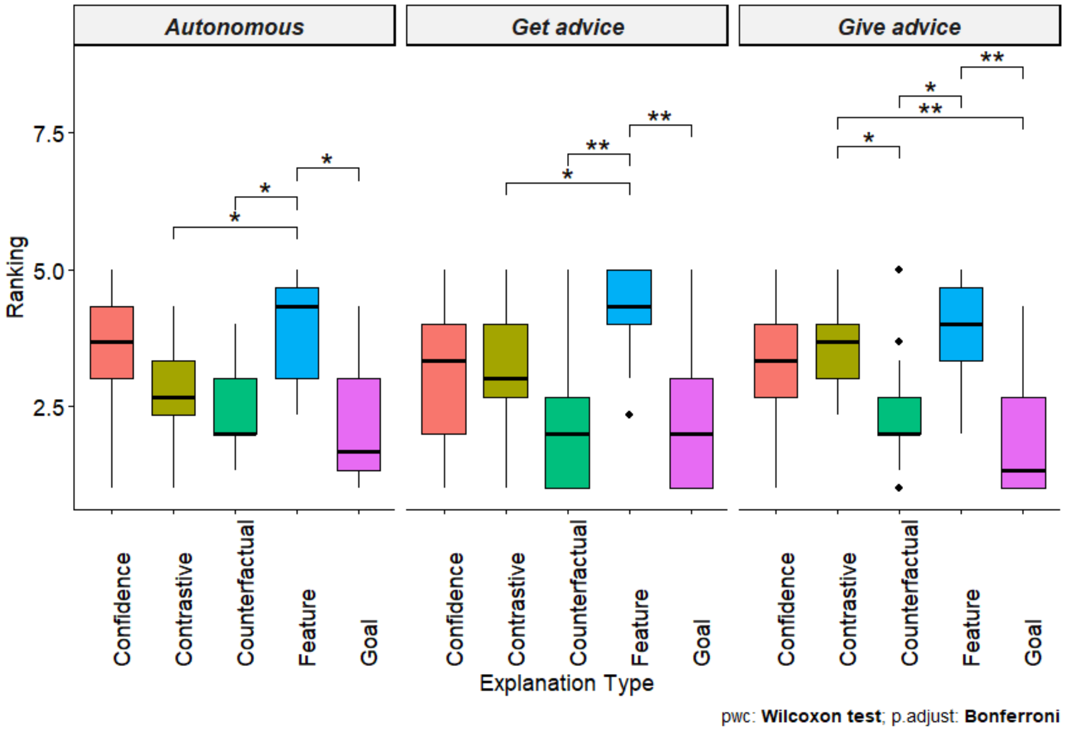


Figure 4.2: Pairwise comparisons using paired Wilcoxon signed-rank test on explanation types per situation.

None of the rankings of the explanation type 'feature attribution' were significantly different from the second-best ranked explanation type (confidence- or contrastive explanation). This means that these top two explanation types could be used interchangeably and the best approach is to use the top two ranked explanation types for the scenarios (autonomous action, get advice, and give advice) and proportionately alternate between the two. The top two ranked explanation types per situation can be viewed in Table 4.4.

Robot's action	Explanation Types	
Autonomous Action	Feature Attribution	Contrastive Explanation
Give Advice	Feature Attribution	Confidence Explanation
Get Advice	Feature Attribution	Confidence Explanation

Table 4.4: Explanation Types given to the adaptive group for the user-study.

Now that the two questions mentioned at the beginning of this chapter had been answered, the information from Table 4.4 could be used for the adaptive condition in the user- study (see Chapter 5 for more information).

Chapter 5

Methodology

This chapter presents the methodology underlying the user-study of this thesis. The user-study explores the influence of adapting the explanation style to the team leader role of a HAT during a search and rescue task. The participants performed the search and rescue task in a simulated 2D disaster site.

5.1. Participants

All demographic data were collected beforehand. The participants were students or recent graduates from different universities in the Netherlands; most participants were students at the Delft University of Technology. All participants were between 18 and 29 years old. By focusing on this specific age group and educational level of the participants, any effects of older age and non-academic backgrounds were prevented. In Table 5.1, an overview of the participants' demographic data is displayed for both condition groups. The Human Research Ethics Committee approved this pre-study with HREC ID 2245 See Appendix M for the data management plan and the informed consent.

Demographic		**Baseline**	**Adaptive**	**Total**
Age	18-21	1	0	0
	22-25	11	15	26
	26-29	11	8	19
Gender	Male	12	15	27
	Female	11	18	29
Education level	Secondary school	1	0	1
	Bachelor degree	11	15	26
	Master degree	11	8	19
Gaming experience	None	3	2	5
	A little	10	12	22
	Average	5	5	10
	A lot	5	4	9

Table 5.1: Overview of participants' (N=46), age, gender, highest completed education level, and self-reported gaming experience.

5.2. Measurements

The following topics were evaluated in the user-study: team performance, situation awareness, perceived trust, collaborative fluency, perceived user-awareness, system understandability, and explanation satisfaction. All topics except for team performance and objective collaborative fluency were subjectively measured via a questionnaire, which was filled out during and after the completion of the simulated search and rescue task. See Appendix N for the questionnaire.

In the questionnaire, the participants were asked what percentage of the explanations they had actually read on a scale of 0-100 with steps of ten to make it easy for the participants to answer. Additionally, the participants were given 2 open questions. The first question asked which behaviour of the robots had influenced the participant's answers both in the simulations (e.g., which advice they gave) and in the questionnaire so far. The second question asked whether the participant had anything left to say, both good or bad, about the experiment, the robots, or the given explanations by the robots.

5.2.1. Situation Awareness

To measure Situation Awareness, the SAGAT method [39] was used. 2.5 minutes after the start of the simulated search and rescue task, the simulation was paused automatically and the participants had to answer questions about the robot, such as its location, what it was doing, and what it was going to do before the simulation was paused. The simulation constantly saved relevant data that could be considered as 'ground truth' for the scoring of situation awareness. The questions in the questionnaire were influenced by different levels of Chen and Barnes [14]'s SAT model: the current status of the robot (SAT 1) and the prediction of the robot's level (SAT 3). The participants got a score from 1-10 based on how far their answers were off from the mentioned ground truth. For example, if the number of encountered survivors so far was 4, but the participant answered 3, the score would have been 9. See Appendix P for the code.

5.2.2. Perceived Trust

Two Likert scale questionnaires were used for measuring perceived trust. The first questionnaire came from Hoffman et al. [42, 66], which they called the Recommended Scale for XAI. This questionnaire measured predictability, reliability, efficiency, believability, and confidence in the system. The second questionnaire came from Hoffman [36] and is called Trust in Robot. This questionnaire was used in particular for measuring collaborative fluency (see Section 2.2.2). The results of the two questionnaires were analysed both separately and combined to see if a difference would be found. All Likert scales ranged from 1-7, where 1 would indicate 'Strongly Disagree' and 7 would indicate 'Strongly Agree'.

5.2.3. Collaborative Fluency

Collaborative fluency was measured in both an objective and a subjective way. To measure the objective collaborative fluency, the robots' and human's idle time were used, by looking at how long it took for the participant to react to the robots' given advice or need for advice. The time it took for the participant to come up with the rescue plan was also considered idle time for the robots. The percentage that both the robots and the human were idle during the total task time was considered as the objective collaborative fluency in the team. See Appendix P for the code that measures the objective collaborative fluency. To measure subjective collaborative fluency, the 7-point Likert scale questionnaire by Hoffman [36] was used.

5.2.4. Team Performance

To measure the team performance, the percentage of survivors saved during the task was calculated. In addition, the type of saved survivors (healthy, injured, or stuck) was measured.

5.2.5. User-Awareness

To measure perceived user-awareness, the participants were asked two 7-point Likert scale questions in the questionnaire.

5.2.6. Understandability

To measure the participants' level of understandability of the robot, Madsen and Gregor [44]'s five 7-point Likert scale questions were used. These Likert scale questions were also mentioned in the analysis of [42]. The questionnaire measured predictability, understanding of assistance, understanding of usage, and ease of use.

5.2.7. Explanation Satisfaction

To measure the participants' explanation satisfaction with the robots, two 7-point Likert scale questions were asked in the questionnaire. The questionnaire measured this by asking the participants about their liking of the explanations and their satisfaction with the explanations.

5.3. Conceptual Model

In Figure 5.1, the conceptual model, showing the relations between the different variables of this study, is depicted. The arrow between the two blue dependent variable blocks stands for the regression analysis in this study. There is one independent variable, resulting in two groups: Adaptive and Baseline.

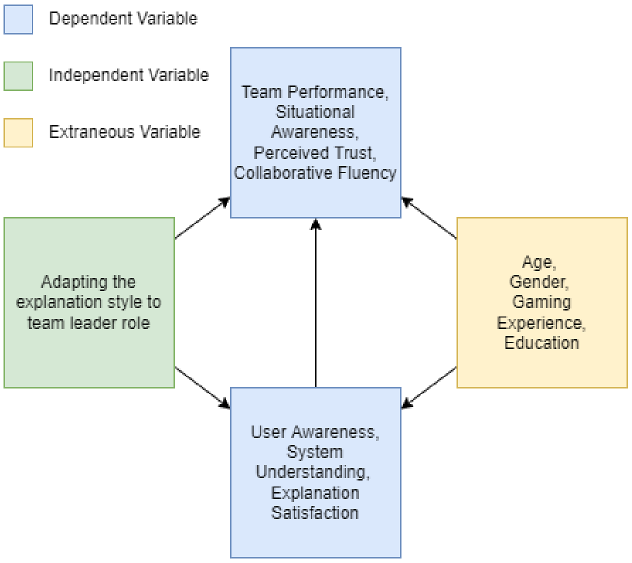

Dependent Variable
Independent Variable
Extraneous Variable
Team Performance, Situational Awareness, Perceived Trust, Collaborative Fluency
Adapting the explanation style to team leader role
Age, Gender, Gaming Experience, Education
User Awareness, System Understanding, Explanation Satisfaction

Figure 5.1: Conceptual model of the user-study.

5.4. Materials

Firstly, the consent form was printed on paper for participants to read and sign. Secondly, a Dell XPS-13 computer with an i7-1065G7 CPU @ 1.30 Ghz, 1498 Mhz, 4 cores, 8 logical cores, and 32 GB of RAM was used. The laptop ran both the server and the client side of MATRX. The questionnaire was conducted on the laptop as well. Lastly, any COVID-19-related items, such as cleaning wipes and mouth masks, were available if needed.

5.5. Procedure

Firstly, the participants started by reading and signing the informed consent form, see Appendix M for the form. After this, the participants filled out the first part of the questionnaire on their demographic data. Then, the participants were asked to play the tutorial level of the experiment to introduce the mechanics and the system to the participants, see Appendix section C.2 for more information. This tutorial was a very small version of the actual experiment and had no time limit. For the text used during the tutorial, see Appendix O. Next, the participants performed the actual task of supervising an agent as a team leader during the user-study experiment. Halfway through the task, after 2.5 minutes, the simulation paused automatically and participants had to answer questions about situation awareness in the questionnaire. When the time had run out or when all survivors had been rescued, the tasks stopped and the participants finished the questionnaire by answering questions about perceived trust, collaborative fluency, user-awareness, system understanding, and explanation satisfaction.

5.6. Task

The task consisted of a simulated search and rescue operation. The design from Chapter 3 was used for this experiment. In short, the participants would start the task with the search phase, where the team leader and the search robot would explore the disaster site randomly by looking for obstacles, entries, areas, and survivors. If they would find any of these, the robot could take an autonomous action, ask for advice, or give advice. When the search phase was done, the search robot would return to the safety zone and the rescue phase would start. The rescue phase would start with the rescue plan, which the team leader would have to create and the rescue robot would have to perform. The items on the plan consist of the choices the participants made during the search phase. For example, if the participant chose to let the rescue robot save a stuck survivor, that would become a rescue plan item. The timer for the rescue phase would already start while making the rescue plan. After all the items on the rescue plan would be executed or when the timer would run out, the simulation would stop. Here, the simulation would store all the collected information, such as collaborative fluency and team performance.

5.6.1. Implementation

For the implementation of this simulation task, MATRX was used. Figure 5.2 displays the icons for each survivor and team member that were created. Dangerous areas were coloured red and safe areas were coloured blue. Figure 5.3 shows the map of the tutorial that each participant played before starting the experiment. Figure 5.4 exhibits the map that was used for the experiment.

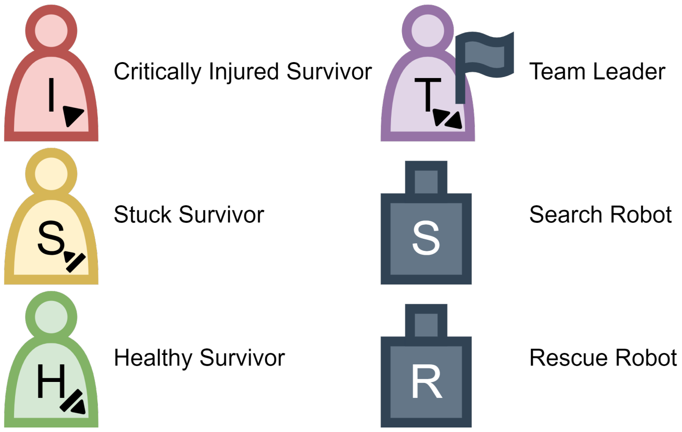


Figure 5.2: Icons of the different types of survivors, robots, and the team leader.

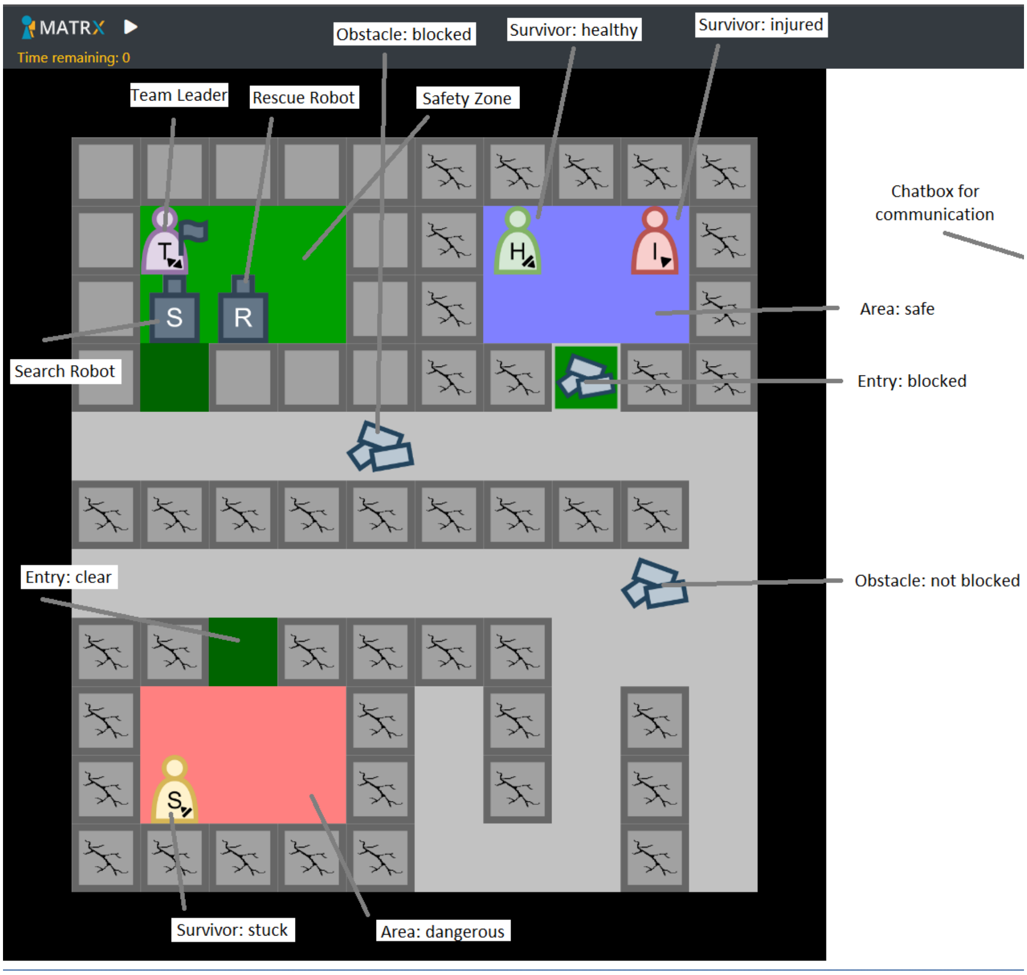


Figure 5.3: Overview of the tutorial map.

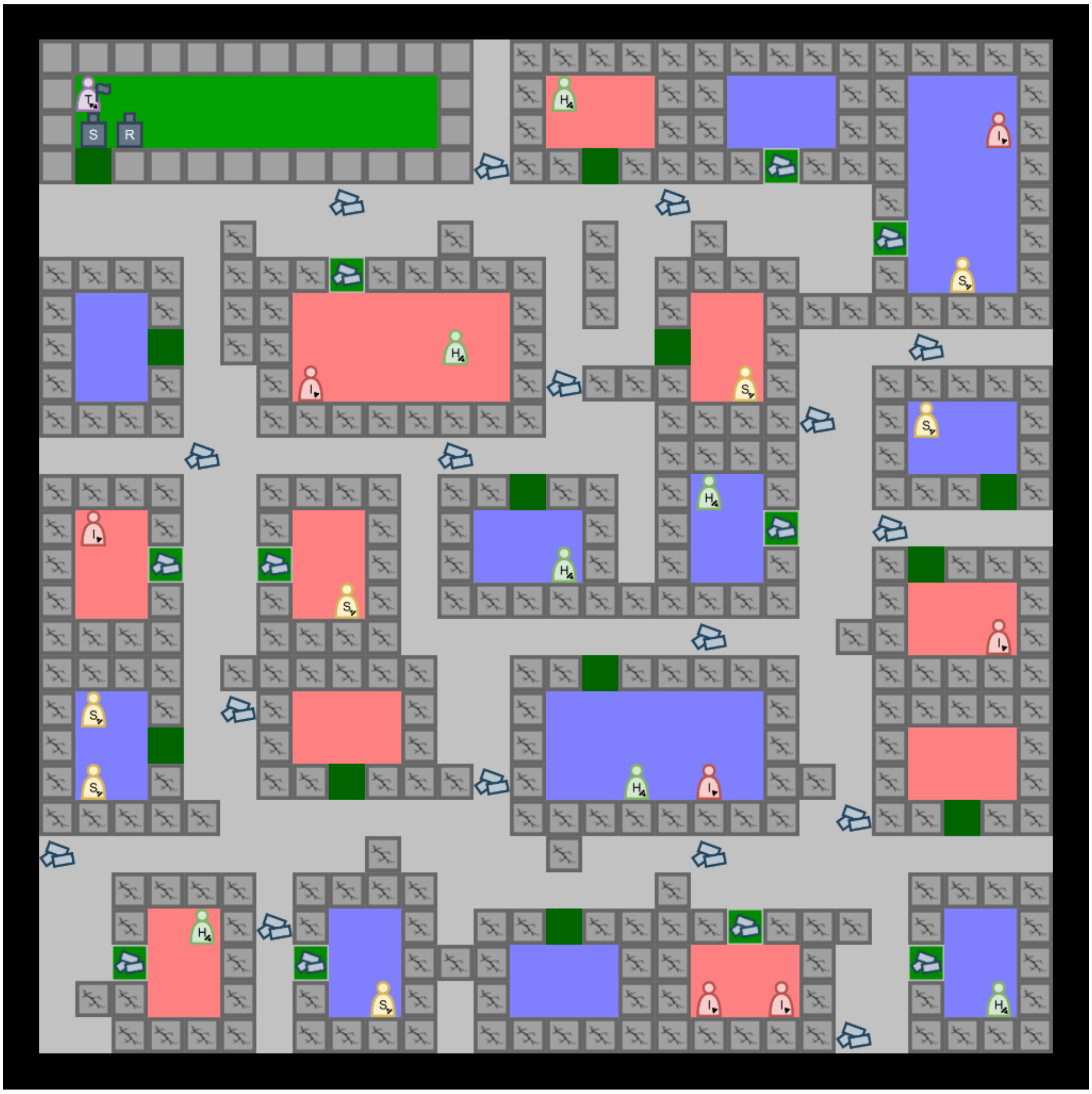


Figure 5.4: Overview of the map used in the experiment.

This map in Figure 5.4 consisted of 21 areas of which eleven were clear and ten were blocked. Moreover, on the map were 21 survivors of which seven were healthy, seven were critically injured, and seven were stuck. Additionally, the map contained 17 obstacles that were inside the Disaster Site blocking pathways (11) or doorway entrances (6). Having an imbalance between all these aspects of the map could influence the result of the study. Therefore, balancing these elements during the design of the map was key. The following aspects were balanced:

- The certainty the robot had of each item. For example, the confidence the robot had in how long it would take to rescue a survivor or to remove an obstacle;

- The time it would take to perform an action. For example, how long it would take to clear a blocked area or to free a stuck survivor. Additionally, some of these times need to be considered as ‘too long’ for the robot to react differently than when an action would not take too long (e.g., take a different action or give different advice);
- The combination of all these elements. For example, the number of different types of survivors (healthy, injured, or stuck) that were in a blocked area versus those that were in an area with a clear entry.

See Appendix D for an overview of the values used for each element. When the task would start, only the safety zone tiles and tiles where the search robot had physically been would be visible. This would add to the realistic feeling of the search phase. For the full implementation of the HAT design in MATRX see Appendix C. To view the full code, see Appendix E.

5.7. Analysis

The data was cleaned using Python and Pandas. Then, the cleaned data was analysed using R with RStudio, see Appendix P for the code. Firstly, a two-sample t-test between the means of the two groups for team performance, situational awareness, perceived trust, both objective and subjective collaborative fluency, user-awareness, understandability, and explanation satisfaction was conducted. In the case of any extreme outliers, it had to be decided whether to remove them or not. If the assumption of normally distributed data was violated, a Wilcoxon test was to be used. If the assumption of the normality was met, but the assumption of equality of variances was violated, a Welch t-test was to be used. If no assumptions were violated, a student t-test was to be used. Secondly, various multiple linear regressions between the dependent variables were carried out. Thirdly, the two aforementioned open questions in Section 5.2 were analysed using techniques from qualitative coding, where excerpts from qualitative data were systematically categorised to find themes and patterns. Qualitative coding can increase validity, decrease bias, accurately represent participants, and enable transparency [84]. For the first open question, descriptive coding was used, because the question tries to find influential behaviour. For the second open question, an adaptive version of open coding and axial coding was used to discover what themes emerged in the answers. A fellow Computer Science master student was asked to analyse the two open questions as well to see how reproducible the results were and to examine how well the coding was done.

Chapter 6

Results

In this chapter, the results of the user-study will be described and displayed. First, analyses of the demographic data will be described. Second, analyses of the differences between the two condition groups will be described. Third, analyses of the open questions from the questionnaire will be described. Fourth, the regression analyses will be described.

6.1. Demographic Data

For the differences in age, gender, education, and gaming experience between the two conditional groups, a Kurskall-Wallis test was used, since no normality was found. No significant differences were found for age ($p = 0.49$), gender ($p = 0.37$), education ($p = 0.2$), and gaming experience ($p = 0.89$) between the two conditional groups.

6.2. Differences Between Groups

6.2.1. Team Performance

In Figure 6.1, a boxplot displays the percentage of survivors saved per group. The mean percentage of survivors saved in the baseline group was 34.37 ($SD = 5.91$, median = 33.3, IQR = 7.14), whereas the mean of the adaptive group was 36.2 ($SD = 8.46$, median = 38.1, IQR = 14.3). The Shapiro-Wilk test indicated that the null hypothesis of normally distributed data could be rejected due to the baseline group's $p = 0.026$. Therefore, instead of the student t-test, the Wilcoxon test was used, which showed that the median difference between the groups was not statistically significant ($p = 0.434$, $W = 229$, effect size $r = 0.117$). The effect size indicates that adapting the explanation style to the team leader role had a small effect on team performance.

The mean number of times participants followed the advice of the robot of the baseline group was 6.17 ($SD = 2.06$, median = 5, IQR = 2), whereas the mean of the adaptive group was 4.70 ($SD = 2.57$, median = 6, IQR = 3). The Wilcoxon test revealed that the median difference was statistically significant ($p = 0.038$, $W = 358$, effect size $r = 0.307$). The effect size indicates that adapting the explanation style to the team leader role had a moderate effect on the number of accepted advice.

The mean number of times that participants decided for the robot to continue the search or rescue plan rather than performing an action (e.g., removing an obstacle, clearing a blocked entry, or immediately rescuing a survivor) of the baseline group was 9.30 ($SD = 2.51$, median = 9, IQR

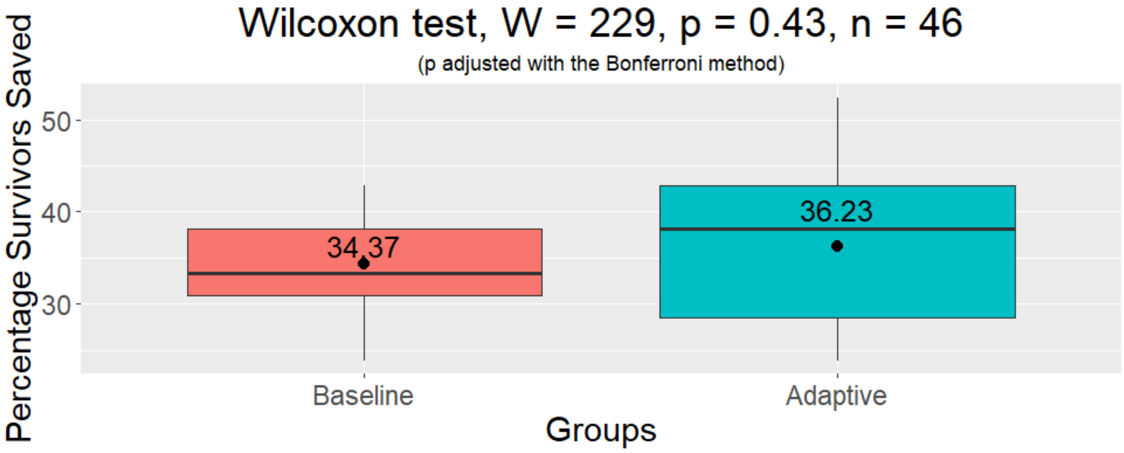


Figure 6.1: Boxplot of the percentages of survivors saved per group based on a Wilcoxon test.

$= 2.5$), whereas the mean of the adaptive group was 7.65 ($SD = 1.87$, median $= 7$, IQR $= 1$). The Wilcoxon test showed that the median difference was statistically significant ($p = 0.018$, $W = 371$, effect size $r = 0.349$). The effect size indicates that adapting the explanation style to the team leader role had a moderate effect on the number of times the robot skipped an action for the adaptive group.

6.2.2. Situation Awareness

The mean situation awareness score of the baseline group was 82.08 ($SD = 9.52$), whereas the mean of the adaptive group was 83.92 ($SD = 7.14$). No assumptions of the student t-test were violated. A two-sample student t-test revealed that the mean difference was not statistically significant, $t(44) = -0.74$, $p = 0.46$, $d = 0.218$. The Cohen's d of 0.218 indicates that adapting the explanation style to the team leader role had a small effect on the situation awareness score.

6.2.3. Trust

Both Hoffman [36]'s Trust in Robot questionnaire ($W = 162, p = 0.022, r = 0.340$) and [42]'s Recommended Scale questionnaire ($t(44) = -2.52, p = 0.016, d = 0.742$) revealed a statistically significant difference in trust between the two groups. However, only the Recommended Scale questionnaire was used for the trust analysis, because it is more extensive and used more often in research. In Figure 6.2, the boxplot depicts the perceived trust per group.

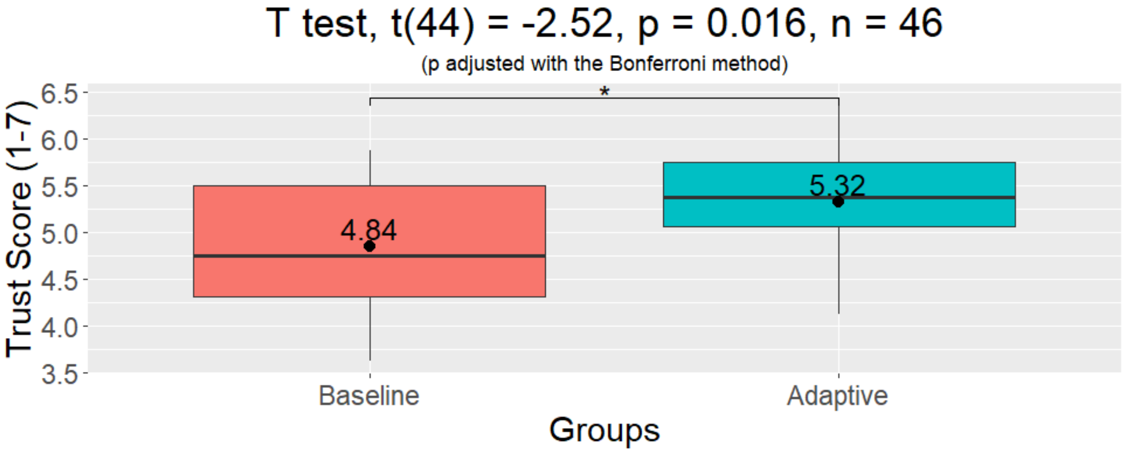


Figure 6.2: Boxplot of the perceived trust per group with student t-test.

The mean trust score for the baseline group was 4.84 ($SD = 0.706$), whereas the mean trust score for the adaptive group was 5.32 ($SD = 0.576$). No assumptions of the student t-test were violated. A two-sample student t-test revealed that the mean difference between groups was statistically significant, $t(44) = -2.52$, $p = 0.016$, $d = 0.742$. The Cohen's d of 0.742 indicates that adapting the explanation types to the team leader role had a moderate effect on trust.

6.2.4. Collaborative Fluency

The boxplot in Figure 6.3 depicts the objective collaborative fluency per group.

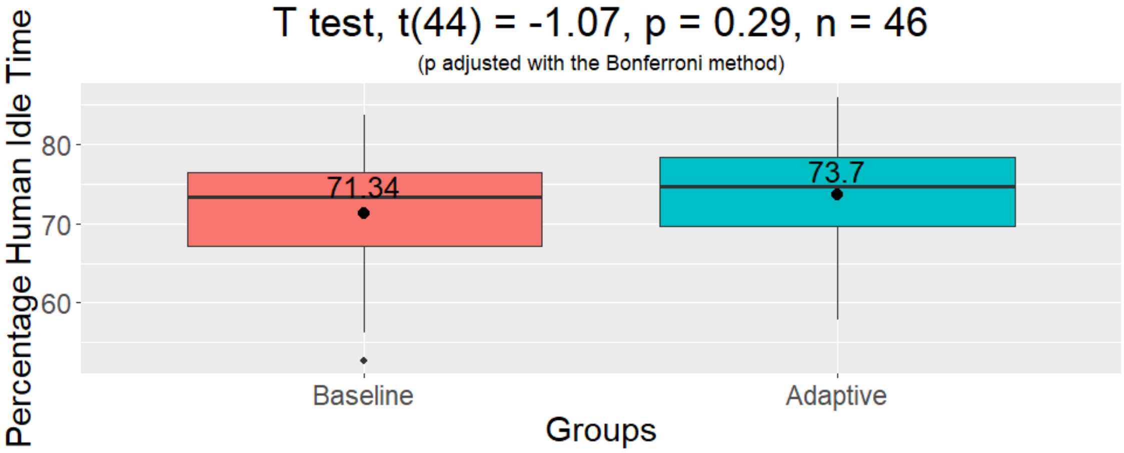


Figure 6.3: Boxplot of the objective collaborative fluency per group based on student t-test.

The mean objective fluency score for the baseline group was 71.34 ($SD = 7.71$), whereas the mean score for the adaptive group was 73.7 ($SD = 7.26$). No assumptions of the student t-test were violated. A two-sample student t-test revealed that the mean difference was not statistically significant, $t(44) = -1.07$, $p = 0.29$, $d = 0.315$. The Cohen's d indicates that adapting the explanation style to the team leader role had a small effect on the objective fluency score. The result for the robot idle time percentage is the same as for the human idle time percentage, but mirrored because these idle times are a turn-taking pattern between the team members, where either one or the other is idle at all times. The boxplot in Figure 6.4 depicts the subjective collaborative fluency per group.

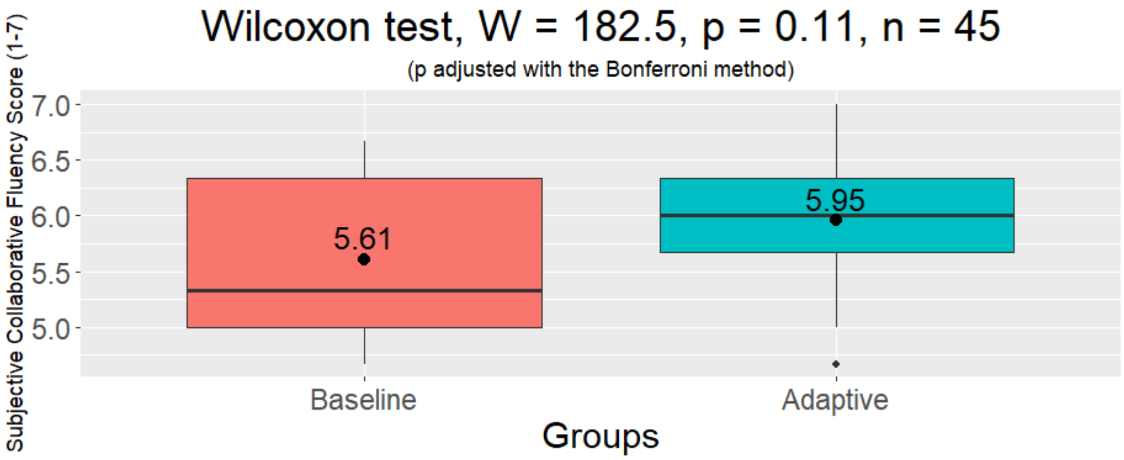


Figure 6.4: Boxplot of the subjective collaborative fluency per group with a Wilcoxon test.

The subjective collaborative fluency had an extreme outlier in the adaptive group (score = 3.667). Consequently, that data point was removed from the analysis. The removal of the data point did not influence the outcome of p-value crossing the significance threshold. The mean subjective fluency score for the baseline group was 5.61 ($SD = 0.679$, median = 5.33, IQR = 1.33), whereas the mean of the adaptive group was 5.96 ($SD = 0.724$, median = 6, IQR = 0.667). The Shaperio-Wilk test indicated that the null hypothesis of normally distributed data for subjective fluency could be rejected due to the baseline group's $p < 0.05$. Therefore, the Wilcoxon test was used for subjective collaborative fluency rather than the student t-test. The Wilcoxon test revealed that the median difference was not statistically significant, $p = 0.108$, $W = 182.5$, effect size $r = 0.241$. The effect size indicates that adapting the explanation style to the team leader role had a small effect on subjective collaborative fluency.

6.2.5. Explanation Satisfaction, Understandability, and User-Awareness

The boxplot in Figure 6.5 shows the explanation satisfaction, understandability, and perceived user-awareness scores of the participants per group.

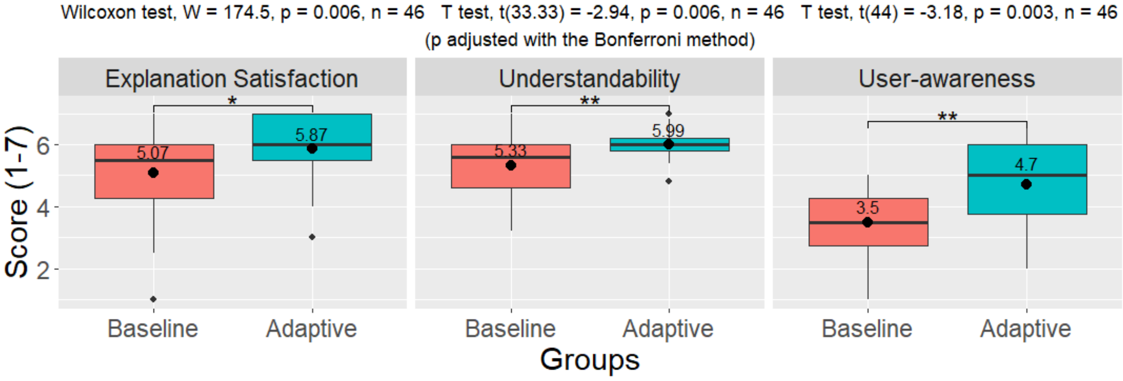


Figure 6.5: Boxplot of explanation satisfaction per group with a Wilcoxon test, understandability per group with a Welch t-test, and perceived user-awareness per group with a Wilcoxon test.

The mean subjective satisfaction of the robot's explanations for the baseline group was 5.07 ($SD = 1.49$, median = 5.5, IQR = 1.75), whereas the mean subjective satisfaction for the adaptive group was 5.87 ($SD = 1.07$, median = 6, IQR = 1.5). The Shapiro-Wilk test indicated that the null hypothesis of normally distributed data for explanation satisfaction could be rejected due to the adaptive group's $p = 0.012$. Therefore, the Wilcoxon test was used for explanation satisfaction rather than the student t-test. The Wilcoxon test revealed that the median difference was statistically significant, $p = 0.046$, $W = 174.5$, effect size $r = 0.295$. The effect size indicates that adapting the explanation style to the team leader role had a small effect on the team leader's satisfaction with the robots' explanations.

The mean subjective understandability for the baseline group was 5.33 ($SD = 0.955$), whereas the mean of the adaptive group was 5.991 ($SD = 0.503$). The Levene's test indicated that the null hypothesis of homogeneity of variance could be rejected for understandability due to the baseline group's $p = 0.003$. Therefore, the Welch t-test was used rather than the student t-test. A Welch t-test showed that the mean difference was statistically significant, $t(33.32) = -2.94$, $p = 0.006$, $d = 0.866$. This Cohen's d indicates that adapting the explanation style to the team leader role had a large effect on subjective understandability.

The mean subjective perceived user-awareness of the robot of the baseline group was 3.5 ($SD = 1.13$), whereas the mean of the adaptive group was 4.7 ($SD = 1.4$). No assumptions for the student t-test were violated. A two-sample student t-test revealed that the mean difference was statistically significant, $t(44) = -3.18, p = 0.003, d = 0.939$. This Cohen's d indicates that adapting the explanation style to the team leader role had a large effect on the subjective perceived user-awareness of the robot.

6.3. Open Questions

6.3.1. Influential Behaviour of the Robot

Figure 6.6 displays the frequency of the categories of two or higher obtained by using open coding on the answers to the question: What behaviour of the robot has influenced your answers?

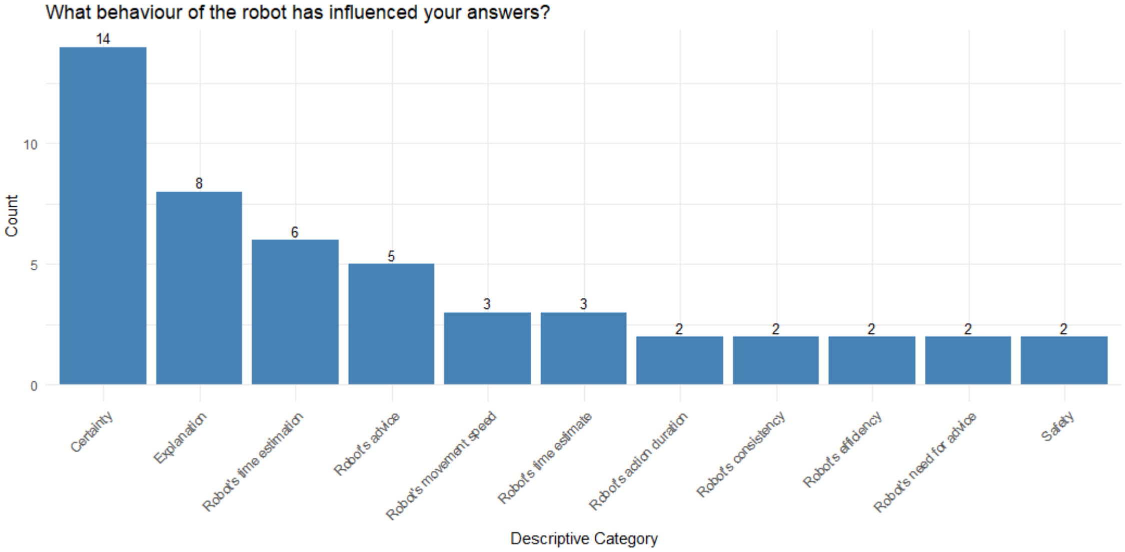


Figure 6.6: Descriptive categories from using open coding on open question number one with a frequency of two or higher.

The category that was counted most for having influenced the participants' answers and decisions in the user-study was the certainty of the robot expressed in the explanations. The category that occurred most frequently after that was the explanations given by the robot in general. After that, the robot's time estimation in the explanations influenced the answers and decisions of participants the most.

A fellow Computer Science master student performed this descriptive coding for the answers to open question 1 independently as well. Their open coding resulted in the following top four categories: confidence level, estimated time, advice, and explanations.

6.3.2. General Remarks

Figure 6.7 displays the answers to the question: Anything you want to share about the experiment or robot or explanations (good or bad)? In the middle of the figure, the circle with the bold line is the main theory or in this case question of the answers. Connected to the middle circle are categories that resulted from open coding the answers.

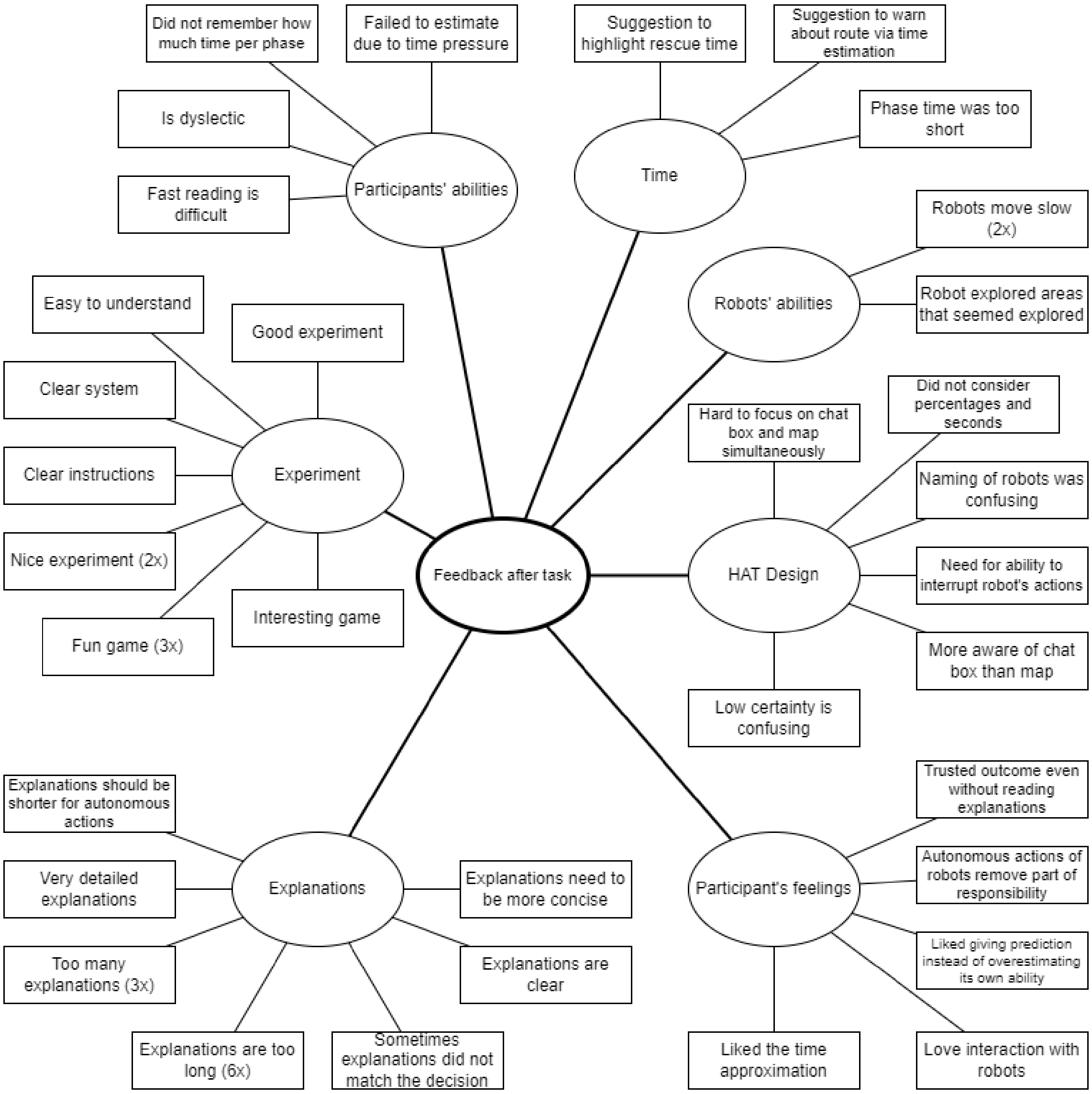

Feedback after task
Participants' abilities
Did not remember how much time per phase
Failed to estimate due to time pressure
Is dyslectic
Fast reading is difficult
Time
Suggestion to highlight rescue time
Suggestion to warn about route via time estimation
Phase time was too short
Robots' abilities
Robots move slow (2x)
Robot explored areas that seemed explored
Experiment
Easy to understand
Good experiment
Clear system
Clear instructions
Nice experiment (2x)
Fun game (3x)
Interesting game
HAT Design
Hard to focus on chat box and map simultaneously
Did not consider percentages and seconds
Naming of robots was confusing
Need for ability to interrupt robot's actions
More aware of chat box than map
Low certainty is confusing
Explanations
Explanations should be shorter for autonomous actions
Very detailed explanations
Too many explanations (3x)
Explanations are too long (6x)
Sometimes explanations did not match the decision
Explanations need to be more concise
Explanations are clear
Participant's feelings
Trusted outcome even without reading explanations
Autonomous actions of robots remove part of responsibility
Liked giving prediction instead of overestimating its own ability
Love interaction with robots
Liked the time approximation

Figure 6.7: Open and axial coding of the answers to the question: Anything you want to share about the experiment or robot or explanations (good or bad)?

After open coding the answers, seven categories were found. These categories were the explanations, how participants felt about certain elements of the task, aspects of the HAT design, negative aspects of the robots' abilities, suggestions on the time during the task, the participants' (dis)abilities, and general statements on the experiment.

6.4. Regression

6.4.1. Team Performance

To test if team performance could be significantly predicted, multiple linear regression was used. The predictor variables were gaming experience, both subjective and objective collaborative fluency, situation awareness, and trust. In addition, a potential interaction effect between the condition group and any of the predictor variables was analysed.

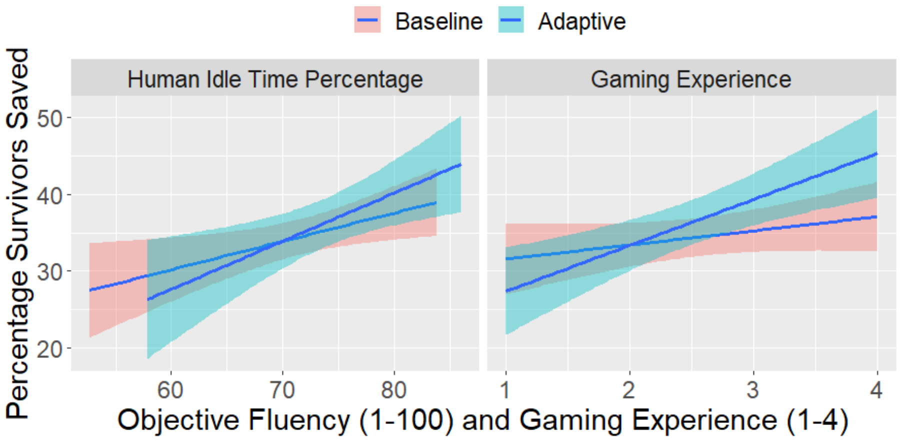


Figure 6.8: Linear regression between gaming experience, human idle time percentage, and team performance.

The overall regression was found to be statistically significant ($R^2 = 0.271, F(11, 34) = 2.52, p = 0.019$). In addition, gaming experience ($\beta = 1.032, p < 0.001$) and objective collaborative fluency ($\beta = 0.126, p < 0.001$) were found to be significant predictors of team performance. Furthermore, an interaction effect between gaming experience and the conditional group was found ($F(3, 42) = 6.416, p = 0.044$).

6.4.2. Trust

To test if trust could be significantly predicted, multiple linear regression was used. The predictor variables were explanation satisfaction, understandability, and user-awareness. In addition, a potential interaction effect between the condition group and any of the predictor variables was analysed.

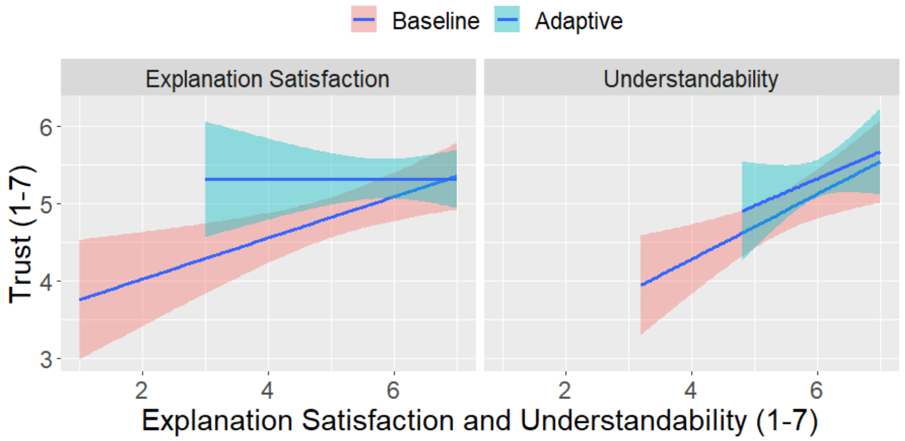


Figure 6.9: Linear regression between explanation satisfaction, understandability, and trust.

The overall regression was found to be statistically significant ($R^2 = 0.352, F(7, 38) = 4.486, p = 0.001$). In addition, understandability ($\beta = 0.126, p < 0.010$) and explanation satisfaction ($\beta = 0.086, p = 0.035$) were found to be significant trust predictors. Furthermore, no interaction effect between the predictor variables and the condition group was found.

6.4.3. Subjective Fluency

To test if subjective collaborative fluency could be significantly predicted, multiple linear regression was used. The predictor variables were explanation satisfaction, understandability, and user-awareness. In addition, a potential interaction effect between the condition group and any of the predictor variables was analysed.

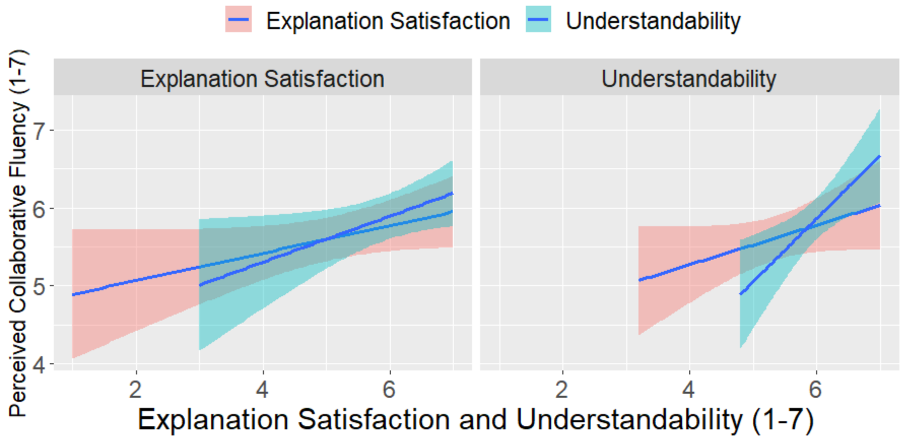


Figure 6.10: Linear regression between explanation satisfaction, understandability, and subjective fluency.

The overall regression was found to be statistically significant ($R^2 = 0.225, F(7, 38) = 2.858, p = 0.017$). In addition, understandability ($\beta = 0.116, p = 0.002$) and explanation satisfaction ($\beta =$

$0.071, p = 0.003$) were found to be significant subjective collaborative fluency predictors. Furthermore, no interaction effect between the predictor variables and the condition group was found.

6.4.4. Explanation Satisfaction and Understandability

To test if explanation satisfaction could be significantly predicted, multiple linear regression was used. The predictor variables were team performance, both subjective and objective collaborative fluency, situation awareness, and trust. In addition, a potential interaction effect between the condition group and any of the predictor variables was analysed.

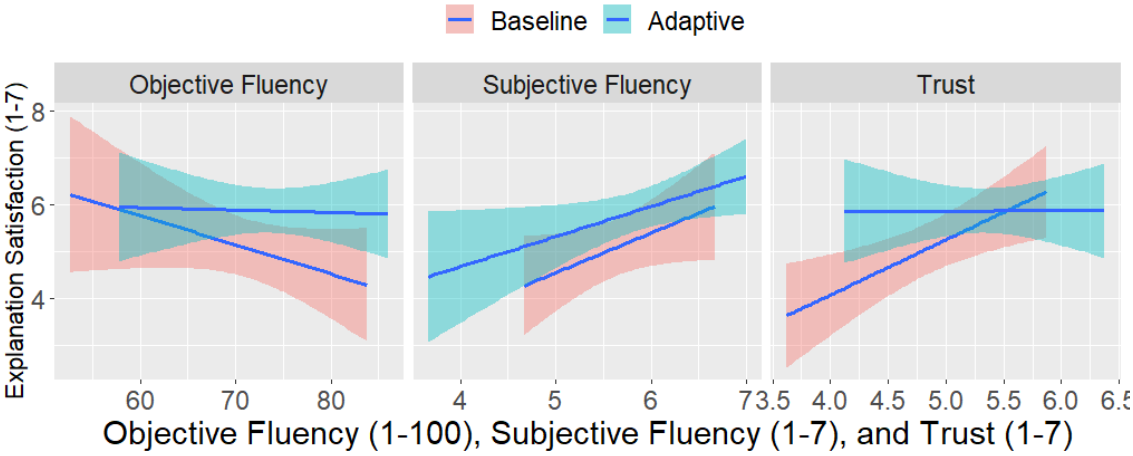


Figure 6.11: Linear regression between objective fluency, subjective fluency, trust, and explanation satisfaction.

The overall regression was found to be statistically significant ($R^2 = 0.388, F(11, 34) = 2.52, p = 0.002$). In addition, trust ($\beta = 0.328, p < 0.001$), objective collaborative fluency ($\beta = 0.034, p = 0.031$), and subjective collaborative fluency ($\beta = 0.334, p = 0.014$) were found to be significant explanation satisfaction predictors. Furthermore, an interaction effect between trust and the conditional group was found ($F(3, 42) = 5.419, p = 0.003$).

Another multiple linear regression was used to test if understandability could be significantly predicted. The predictor variables were team performance, both subjective and objective collaborative fluency, situation awareness, and trust. In addition, a potential interaction effect between the condition group and any of the predictor variables was analysed.

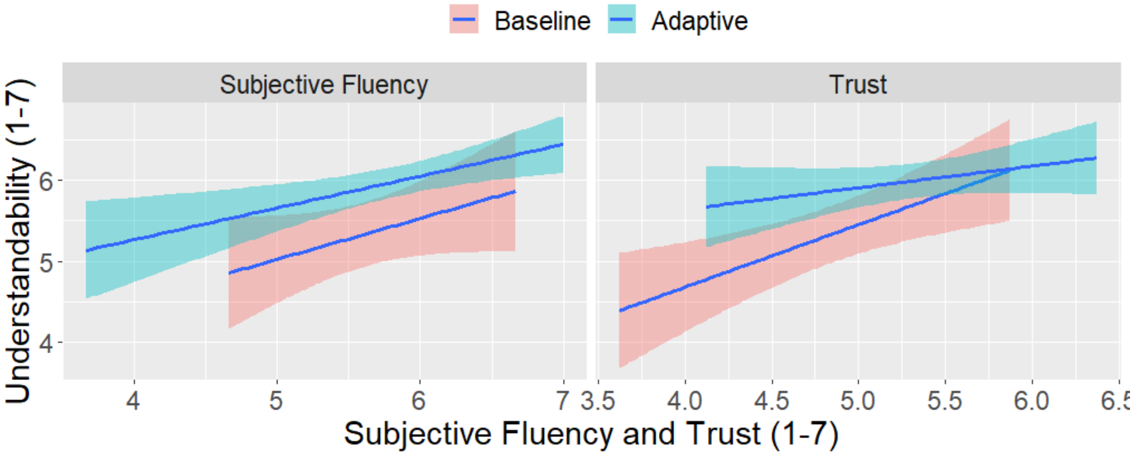


Figure 6.12: Linear regression between trust, objective fluency, and understandability.

The overall regression was found to be statistically significant ($R^2 = 0.461, F(11, 34) = 4.494, p < 0.001$). In addition, trust ($\beta = 0.189, p < 0.001$) and subjective collaborative fluency ($\beta = 0.192, p = 0.006$) were found to be significant understandability predictors. Furthermore, an interaction effect between trust and the conditional group was found ($F(3, 42) = 9.019, p < 0.001$).

Chapter 7

Discussion

In this chapter, the results of the user-study will be discussed. In addition, the limitations of this thesis will be discussed. Moreover, any suggestions regarding future work based on this thesis will be given. Lastly, this thesis and its contribution will be concluded.

7.1. Differences Between Groups

7.1.1. Team Performance

As one can see from the results in Section 6.2.1, the team performance of the adaptive group, indicating the number of survivors saved, was higher than the team performance of the baseline group, albeit not statistically significant. Therefore, we cannot confidently say that adapting the explanation style to the human team leader role of this HAT during *this* search and rescue task increases the team performance. As the percentage of survivors saved was between 23.81% and 52.38% for both groups, it does not reflect the expected outcome that both groups result in high team performance ($> 60\%$) by using explanations tailored to the task, regardless of the explanation type. In the experiment, the participants had no direct control over the robot, but only indirect via decision-making based on advice given by the robot and based on the rescue plan. One of the design choices was to make it impossible for the participants to save all survivors by adding a time constraint. This decision might explain the range of team performance between 23.81% and 52.38% for both groups. However, the lowest part of the team performance range, 23.81%, might also be explained by the possibility that this simulation reflects real-life search and rescue tasks, where tough decisions need to be made and sometimes this means that not every survivor can be brought to safety. Additionally, the effect size of adapting the explanation style to the team leader role is moderate, which suggests the idea that adapting the explanation style positively impacts team performance, but this is not supported by the statistical analyses. Furthermore, adapting the explanation style does not influence the amount of information given to the participants. This might explain why the mean percentage of survivors saved does not differ significantly between the two condition groups, as adapting the explanation style does not give you a direct advantage in saving survivors.

Notably, the mean number of times participants followed the robot's advice was significantly lower for the adaptive group than the baseline group. This might be explained by the adaptation of the explanation style by the robot in their provided advice, leading to a better understanding of the robot by the team leader.

In this thesis, a skipped scenario refers to the situation in which the team leader decides for the robot to continue its search rather than remove an encountered obstacle or clear a blocked entry for instance. The results show that the adaptive group skipped scenarios significantly less than the baseline group. An explanation for this could be that the adaptive group was more focused on exploring current situations (e.g., clearing a blocked entry) rather than trying to discover other situations (e.g., continuing the search for other areas or survivors). Another explanation might be that the robot adapting its explanation style could make the participant understand the situation and the robot's advice better, which enables the participant to make a more informed decision. In this case, that would mean that the team leader is more willing to continue to investigate the current situation. However, the number of skipped situations is derived from all situations; removing obstacles or not, clearing entries or not, and rescuing survivors or not. This makes it difficult to say something meaningful about the significant difference in relation to a specific situation or action.

7.1.2. Situation Awareness

As mentioned in section 6.2.2, there is no significant difference in situation awareness between the two groups and the effect size is also small. Every participant scored above 60 out of 100 points on situation awareness regardless of their condition group, which indicates a ceiling effect. This effect could be explained by the questions that are asked during the SAGAT being too easy and the way they were graded being too loose. During the user-study, it was fairly easy to receive a good score. For example, if the robot was at position (30,30) and the participant answered (15,15), the participant would still score a 6 out of 10. In addition, it was expected that situation awareness would decrease when the robot has a high level of automation. However, there was no significant difference in the number of autonomous actions the robots took during the search and rescue task between both groups, meaning that the level of automation was approximately the same. As a result of both the found ceiling effect among the participants and the non-significant difference in the number of autonomous actions of robots between groups, the aforementioned expectation could not be tested.

7.1.3. Trust

Section 6.2.3 shows that the mean of the self-reported trust in the adaptive group is significantly higher (5.82) than the mean of the baseline group (4.84). Therefore, one can confidently say that adapting the explanation style of the robots to the human team leader in a HAT performing this search and rescue task positively influences the team leader's trust in the robots. The results are in line with the expectation that the score will be high ($>$ 60 out of 100) for both groups owing to the robot's communication. In Verhagen et al. [12], trust significantly increases when explanations and adaptive explanations were used. Additionally, the moderate-to-large effect size of trust helps to strengthen the result of this thesis. Trust being significantly different between the two conditional groups could also be explained by being consistent in using the same explanation style for the same situation could infer more sense of reliability and therefore trust as opposed to randomly using explanation styles for every situation.

7.1.4. Collaborative Fluency

Objective collaborative fluency As mentioned in section 6.2.4, the mean of the objective fluency (human and robot idle time percentage) does not significantly differ between both groups. This does not reflect the expectation that the participants in the adaptive group would read more explanations and more thoroughly and would thus take more time during the task. This expectation could also not be measured properly as the self-reported percentage of read explanations does not differ significantly between both groups. In addition, the significant difference between objective collaborative fluency could be due to the time it takes to read the different explanation styles do not differ much from each other. Therefore, the actual time to read two different explanation styles and then make a decision should not make a huge difference. However, the effect size being small limits the confidence to say anything about the influence of adapting the explanation style to the team leader in a HAT performing this search and rescue task on objective collaborative fluency.

Subjective collaborative fluency As can be seen in section 6.2.4, the median of the subjective collaborative fluency does not significantly differ between the two groups. Therefore, it cannot be concluded that adapting the explanation style to the team leader in a human-agent team performing this search and rescue influences subjective collaborative fluency. The fact that the difference in subjective fluency between the groups is not significant suggests that, contrary to expectations, there is no transitive law between adaptive explanation styles, perceived understandability, and subjective collaborative fluency. Notably, there was a ceiling effect for subjective collaborative fluency (97.83% $>$ 4.7, which is 60% of the 1-7 range) for both groups. This could be explained by the use of explainable communication in both groups, which could give a general feeling of effective collaborative fluency.

7.1.5. User-Awareness

As one can see in section 6.2.5, user-awareness differs significantly between the two groups. Therefore, it can be concluded that adapting the explanation style to the team leader role of a HAT during this search and rescue task positively influences the team leader's perceived user-awareness of the robots. The fact that adapting the explanation style has a large effect size on the perceived user-awareness strengthens this claim. The significant difference in user-awareness is in line with the expectation that adapting the explanation style gives the team leader the feeling that the robots are more user-aware and give more personalised explanations than the robots in the baseline. In addition, it was expected that the concept of user-awareness would be deemed as confusing by the participants since the connection between user-awareness and their team role was not explained properly. The significant difference in user-awareness is also in line with the expectation that using the results of the pre-study increases the perceived user-awareness for the adaptive group.

7.1.6. Understandability

As can be seen in section 6.2.5, understandability is significantly higher for the adaptive group (5.99) compared to the baseline group (5.33), $p = 0.006$. Therefore, one can say that adapting the explanation style of the robots to the team leader role in a HAT performing this search and rescue task positively influences the human team leader's perceived understandability of the robots. The fact that adapting the explanation style has a large effect size on understandability strengthens this claim. The significant difference in understandability aligns with the expectation that using explanations that are tailored towards the team leader role increases the understandability of the receiver.

7.1.7. Explanation Satisfaction

As is depicted in section 6.2.5, there was a significant difference in explanation satisfaction between the two conditional groups. This result aligns with the expectation that adapting the explanation style to a user (in this case the role of that user) increases the satisfaction of the explanations. Mueller et al. [23] and [6] argue that good explanations take the user receiving the explanations into account and the significant difference in explanation satisfaction reflects that. In addition, the explanation satisfaction score is high (> 4.7, which is 60% of the range 1-7) for both groups. This result might be explained by assuming that the explanations are satisfactory in general and that the difference between giving a random explanation style and adapting them according to the team leader role does increase the explanation satisfaction.

7.2. Open questions

7.2.1. Influential Behaviour of the Robot

As Figure 6.6 shows, the level of confidence the robot has in its sensors, the explanations given by the robot, the robot's time estimation of an action in its explanations, and the robot's advice are the four behaviour categories of the robot most often mentioned by the participants that influenced their answers. This tells us that the confidence the robot has in something, among other things, is important for the supervising team leader in a human-agent team for this search and rescue task. This reflects the results of the pre-study, since the explanation type 'confidence' is ranked second in almost all pre-study scenarios (see section 4.5).

Moreover, explanations are also considered to have a considerable influence on the answers of and decisions made by the participants. This reaffirms that it does matter what the robot communicates and in what manner, as stated by multiple studies such as Neerincx et al. [11]. Furthermore, the robot's time estimation is likewise considered influential behaviour on the participants' answers, which is not surprising considering that this search and rescue task is time sensitive.

Lastly, the robot's advice was a behaviour that many participants considered influential to their answers and decisions. This is unsurprising, given the fact that participants are explained during the tutorial of the game that the robot's role, among other things, is to support the team leader by giving advice (see Appendix O for the tutorial's text) and that it is important to listen to your team member's ideas and advice as a team leader.

7.2.2. General Remarks

As one can see in Figure 6.7, the general remarks on the experiment, robot, or explanations can be categorised into several topics. The first category is 'experiment', where participants mainly said that they enjoyed the experiment. It could be the case that people will be less likely to say anything negative about an experiment in these manners (e.g., bad experiment or stupid game) if you are physically next to them. Nonetheless, these remarks do indicate that participants felt like the experiment went well and that it was enjoyable.

The second category is 'explanations', where participants mentioned that the explanations were too long in general and in particular instances there were too many explanations. This reflects van der Waa et al. [9] and Kaptein et al. [6], who argue that explanations should not be too long or else an information overload could happen and the explanations will be less or not at all effective. The length of the explanations has been considered during the design of the HAT in Chapter 3. Even though some explanations do feel long (see Appendix B for the explanations in different scenarios), this way they were designed gave them the same amount of information regardless of explanation style. The different explanation styles providing the same amount of information was considered to be of more importance than shortening the explanations to minimise the risk of participants finding the explanations too long. Moreover, in the current way the explanations are designed, the differences in explanation style were more obvious and therefore better suited for the experiment as this would make any result from analyses more noticeable.

The third category is ‘participant’s feelings’, where people expressed their emotions regarding an aspect of the experiment. Notably, a participant mentioned they trusted the outcome of the choice they were given without even reading the explanation given by the robot. This aligns with the results that the perceived trust was generally high. Another interesting answer was that one participant felt that the autonomous actions performed by the robots removed a part of their subjective responsibility towards the outcome of that action.

The fourth category is ‘HAT Design’, where participants made remarks about the design of the robot and the HAT in general. Two participants mentioned that it was difficult to keep an overview between the map and the chat box during the task, which was one of the important responsibilities of the team leader. Another participant mentioned that they wanted to have direct control, rather than the exclusively indirect control they had during this task. These two observations reflect aspects of a real supervising team leader and therefore strengthen the idea that the participants were actually in a supervising team leader role on the team.

The fifth category is ‘Robot’s ability’, where two participants found that the robot moved too slowly. During the design, a trait-off had to be made between the speed of the robot and the consecutive speed of the number of explanations the robot provided. If the robot moves too quickly, too many explanations might appear in the chat box and unfortunately, some participants already found this to be the case.

The sixth category is ‘time’, where some participants provided suggestions on aspects to change regarding time, such as highlighting certain aspects of the explanations or the robot providing additional suggestions about routing ideas.

The seventh and last category is ‘participants abilities’, where participants gave feedback based on their abilities (e.g., dyslectic or slow readers). Other participants felt they could not perform optimally due to time pressure or they forgot important things they learned during the tutorial. These remarks also reflect aspects of what it takes to be a team leader in a search and rescue team. One should be able to handle time pressure and act quickly. Summarising these remarks, participants enjoyed the experiment, but they expressed that explanations were too long and that there were too many explanations, they would have wanted to be able to directly influence the robot and they would have liked to see different information highlighted in the robot’s explanations.

7.3. Regression

As one can see in section 6.4, gaming experience and human idle time percentage are significant predictors of team performance. Furthermore, explanation satisfaction and understandability are significant predictors of both trust and subjective collaborative fluency. Additionally, both objective and subjective collaborative fluency and trust are significant predictors of explanation satisfaction. In this regression, there is an interaction effect between trust and the adaptive group. Furthermore, subjective collaborative fluency and trust are significant predictors of understandability. In this regression, an interaction effect between trust and the adaptive group can also be found.

Gaming experience being a significant predictor of team performance could be explained by the fact that there is skill involved in being the team leader for this search and rescue task. People with more gaming experience are more likely to perform well in game-like simulations such as this user-study. The fact that an interaction effect between gaming experience and the adaptive condition group was found, means that gaming experience has a stronger effect on team performance when the agent adapts its explanation style to the team leader in comparison to when the agent uses random explanation styles. An explanation for this could be that getting explanations tailored to the team leader role in combination with the participant having more gaming experience increases the chances that they know better what to do throughout the whole task to make better decisions that lead to saving more survivors. When the human idle time percentage is higher, it means that the robot is less idle and can thus take more actions at the same time as the other participants. Therefore, it makes sense that has a positive effect on team performance.

Moreover, understandability and explanation satisfaction being significant predictors of trust could be explained by the fact that the feeling of understanding a system and being satisfied by the explanations of that system naturally leads to higher trust in said system. Furthermore, the same could be said about understandability and explanation satisfaction both significantly predicting subjective collaborative fluency. If one feels they understand the system and is satisfied by its explanations it makes sense that one would feel that the collaboration is also successful.

In addition, both objective and subjective fluency and trust being significant predictors of explanation satisfaction could be explained by the fact that when you feel the collaboration runs smoothly, the team task runs objectively fluent, and you trust the robot, you are more likely to be satisfied by the robot's explanations. These predictors could all add to the feeling of liking the robot and therefore being more easily satisfied by its explanations.

Moreover, the interaction effect that was found between trust and the adaptive condition group means that trust has a stronger effect on explanation satisfaction when the agent adapts its explanation style to the team leader in comparison to when the agent uses random explanation styles. This could be explained by trust and explanation satisfaction being significantly higher in the adaptive condition group. Therefore, it makes sense that trust in the adaptive condition group is consistently high as one can see by the horizontal line in Figure 6.11 and also that the explanation satisfaction is consistently high as one can see by the horizontal line in Figure 6.10.

Furthermore, trust and subjective fluency significantly predicting understandability could be explained by when feeling that the collaboration runs fluently and that you trust the robot is enough by itself to make one feel that they understand the robot better, even if that is objectively wrong. Notably, the found interaction effect between trust and the adaptive condition group indicates that trust has a weaker effect on understandability when the agent adapts its explanation style to the team leader in comparison to when the agent uses random explanation styles. The understandability is high in the adaptive condition group regardless of trust scores as one can see by the line being nearly horizontal in Figure 6.12. Contrastively, the line representing the association between trust and understandability is much steeper in the baseline group. This difference between the condition groups could explain the interaction effect. If one trusts a robot and it uses the right explanation styles in the appropriate contexts, one can assume that it feels like they understand the robot or they are satisfied with the explanations even though it could be the case that they objectively do not understand the robot or that the explanations are not effective at all.

7.4. Limitations

During the pre-study, some participants could have had different characteristics that are not representative of the entire population (e.g., certain backgrounds). This could have influenced their ranking of the explanation styles. Kaptein et al. [6] discuss a similar limitation. Another limitation of this thesis is that the number of participants per group is only 24. With a higher number of participants, one could have more confidence in the results. Furthermore, the fact that no experts participated in the pre-study impacts the reliability of its results as using related experts gives better insights. However, the use of non-experts does promote exploration as their statements or conclusions are more likely to be general compared to experts. Moreover, the design of the HAT was very specific to this research question. Even though the guidelines from the INSARAG were followed during the design, it is hard to make the results of this thesis generalisable. Additionally, the generalisability of the results is limited by constraining the participant's age and educational location and level. Another limitation is that the number of times a team leader accepted the advice given by a robot is not recorded. This means that the values are less comparable between the groups as the total number is not known. These values are probably also different with each participant due to the random exploration of the robot in the search phase. Furthermore, mainly due to limited time and resources for this thesis, the adaptation of the explanation style was only explored with regard to one team role and using only one robot at a time. Lastly, the layout of the map and its features (e.g., the number of critically injured survivors and dangerous areas) were as balanced as possible. Unfortunately, this style limits the generalisability of the results to only having confidence in this particular map.

7.5. Future Work

For future work, it is recommended to improve upon the limitations mentioned in the previous section. For the pre-study, it is recommended to redo a similar study but with actual experts in search and rescue as this will validate the results for this application more. In addition, it could be beneficial to explore more scenarios than only the most important one as was done in this pre-study as a means to keep the pre-study short. Moreover, to reduce the information overload that most participants of the user-study felt, one could look at the effect of only giving detailed explanations when requested by the user, as has been researched by Kreske et al. [85] and van der Waa et al. [9]. Additionally, to have more generalisable HATS, future work could look at truly being adaptive and not relying on a pre-recorded data set as Stowers et al. [61] concludes for general HMT. This could mean adapting the explanations on the fly by constantly updating the explanation style given a situation. Furthermore, as some participants mention the difficulty of dividing attention between the chat box and the map, future work could look at using a physical embodiment of the agent. The physical embodiment of the agent could aid the focus of the team leader. This suggestion was also brought forward by Verhagen et al. [12].

In addition to improving on the before-mentioned limitations, it is recommended to also look at the effect of adapting explanation styles on the cognitive load of humans, because of the effect cognitive load has on team performance. Furthermore, to explore the effect of adapting the explanation style on team performance even further, it could be interesting to analyse a mix of communication forms, such as lights and different symbols in the communication. Next, it would be useful to look investigate whether the results would still hold up when using additional robots that work simultaneously, as that would be more complex, but it would also be more realistic for a search and rescue task. Lastly, it could be interesting to develop a similar small real-life experiment to see if the results would still hold up. This will make it more realistic for the participants and it might be easier for participants to put themselves into the shoes of an actual team leader.

7.6. Conclusion

In conclusion, the goal of the research question was to close the research gap with regard to user-awareness and personalisation in explainable agents. In addition, the research question tried to follow up on various suggestions made in previous research in the field of XAI. In doing so, the research question will provide more insights and a deeper understanding of the relationship between explanations that consider team roles and HATs within the field of XAI.

The research question mentioned at the start of this thesis was: How does adapting an agent's explanation style to a human team leader role influence human-agent teamwork during a simulated search and rescue task? To answer this research question, a pre-study and a user-study were conducted. The aim of the pre-study was to discover which explanation style is preferred in the most important situations of a search and rescue task by the human team leader of a HAT. These turned out to be feature attributions and contrastive explanations when the robot takes an autonomous action and feature attributions and confidence explanations when the robot gives or asks for advice. The results of the pre-study were then implemented into the design of the user-study, where participants had to assume the role of team leader in a HAT performing a simulated search and rescue task and rescue as many survivors as possible within a set time limit. In this simulation, the robot would adapt its explanation style in the most important situations to the team leader in accordance with the pre-study results. The results of this simulation experiment showed that no significant difference between the team's performance, the team leader's situation awareness, and the collaborative fluency (both objective and subjective) when adapting the explanation style to the team leader of a HAT was found. However, a significant difference was found between the team leader's perceived trust in the robot, the team leader's perceived understandability of the robot, the team leader's satisfaction with the robot's explanations, and the perceived user awareness of the robot by the team leader. This indicates that adapting the explanation style to the team leader positively influences human-agent teamwork. This claim is strengthened by the large effect size of adapting the explanation style to the team leader on the team leader's understandability of the robot and the team leader's perceived user-awareness of the robot. Additionally, participants' gaming experience level and the human idle time percentage were significant predictors of team performance. Also, explanation satisfaction and understandability were significant predictors of both trust and subjective collaborative fluency. Moreover, both objective and subjective fluency and trust were significant predictors of explanation satisfaction. Furthermore, subjective collaborative fluency and trust were significant predictors of understandability. Lastly, an interaction effect was found between trust and adapting the explanation style to the team leader when predicting both explanation satisfaction and understandability.

On the whole, it cannot be concluded that adapting the explanation style to the team leader role in this simulated search and rescue task influenced the objective measured factors of the HAT. However, what can be concluded is that adapting the explanation style to the team leader role in this simulated search and rescue task positively influenced the subjective measured feelings of trust in the robot, understandability of the robot, satisfaction with the robot's explanations, and perceived user-awareness of the robot by the team leader. In addition, this thesis provides insights and data regarding the explanation style preferences of the human team leader in this HAT design. As there were no known data available yet, this thesis is a good stepping stone for further researching the adaptation of explanation styles in search and rescue tasks with HATs.

Bibliography

[1] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis, “Coactive Design: Designing Support for Interdependence in Joint Activity,” *Journal of Human-Robot Interaction*, vol. 3, no. 1, p. 43, 2014.

[2] J. Van Diggelen and M. Johnson, “Team design patterns,” in *HAI 2019 - Proceedings of the 7th International Conference on Human-Agent Interaction.* Association for Computing Machinery, Inc, 9 2019, pp. 118–126.

[3] K. Stowers, J. Oglesby, S. Sonesh, K. Leyva, C. Iwig, and E. Salas, “A framework to guide the assessment of human-machine systems,” *Human Factors*, vol. 59, no. 2, pp. 172–188, 3 2017.

[4] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, “A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12688 LNAI, pp. 119–138, 2021.

[5] J. Y. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, “Situation awareness-based agent transparency and human-autonomy teaming effectiveness,” *Theoretical Issues in Ergonomics Science*, vol. 19, no. 3, pp. 259–282, 5 2018. [Online]. Available: https://doi.org/10.1080/1463922X.2017.1315750https://www-tandfonline-com.tudelft.idm.oclc.org/doi/abs/10.1080/1463922X.2017.1315750

[6] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, “Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults,” in *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, vol. 2017-Janua, 2017, pp. 676–682.

[7] R. Verhagen and M. A. Neerincx, “Communication Style and Interdependence in Human-Robot Teamwork - the Influence of Transparency and Explainability.”

[8] S. Anjomshoae, D. Calvaresi, A. Najjar, and K. Främling, “Explainable agents and robots: Results from a systematic literature review,” in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, vol. 2, no. Aamas, 2019, pp. 1078–1088.

[9] J. van der Waa, S. Verdult, K. van den Bosch, J. van Diggelen, T. Haije, B. van der Stigchel, and I. Cocu, “Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations,” *Frontiers in Robotics and AI*, vol. 8, no. May, pp. 1–20, 2021.

[10] M. Harbers, J. M. Bradshaw, M. Johnson, P. Feltovich, K. Van Den Bosch, and J. J. Meyer, “Explanation in human-agent teamwork,” in *Lecture Notes in Computer Science (including*

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7254 LNAI, 2012, pp. 21–37.

[11] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, *Using perceptual and cognitive explanations for enhanced human-agent team performance.* Springer International Publishing, 2018, vol. 10906 LNAI. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-91122-9_18

[12] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, "The influence of interdependence and a transparent or explainable communication style on human-robot teamwork," *Frontiers in Robotics and AI*, no. September, pp. 1–20, 2022.

[13] M. Johnson and A. H. Vera, "No Ai is an island: The case for teaming intelligence," *AI Magazine*, vol. 40, no. 1, pp. 16–28, 3 2019. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/2842

[14] J. Y. Chen and M. J. Barnes, "Human - Agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, 2 2014.

[15] J. M. Bradshaw, P. J. Feltovich, and M. Johnson, "Human-agent interaction," *The Handbook of Human-Machine Interaction: A Human-Centered Design Approach*, pp. 283–300, 1 2011. [Online]. Available: https://www-taylorfrancis-com.tudelft.idm.oclc.org/chapters/edit/10.1201/9781315557380-14/human\OT1\textendashagent-interaction-jeffrey-bradshaw-paul-feltovich-matthew-johnsonhttps://www.researchgate.net/publication/267819585_Human-Agent_Interaction

[16] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, "Ten challenges for making automation a "team player" in joint human-agent activity," pp. 91–95, 11 2004.

[17] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2013, pp. 301–308.

[18] C. M. Jonker, M. B. van Riemsdijk, and B. Vermeulen, "Shared mental models: a conceptual analysis." *COIN 2010 International Workshops*, no. Section 2, pp. 132–151, 2010.

[19] W. B. Rouse and N. M. Morris, "On Looking Into the Black Box. Prospects and Limits in the Search for Mental Models," pp. 349–363, 11 1986. [Online]. Available: /record/1987-09358-001

[20] E. Salas, D. E. Sims, and C. Shawn Burke, "Is there A "big five" in teamwork?" pp. 555–599, 2005.

[21] J. Cannon-Bowers, E. Salas, and S. Converse, "Shared mental models in expert team decision making Individual and group decision making," *Lawrence Erlbaurm*, pp. 221–246, 1993.

[22] K. Sycara and G. Sukthankar, "Literature Review of Teamwork Models," *Robotics*, vol. 31, no. CMU-RI-TR-06-50, p. 06–50, 2006. [Online]. Available: https://www.ri.cmu.edu/pub_files/pub4/sycara_katia_2006_1/sycara_katia_2006_1.pdf

[23] S. T. Mueller, E. S. Veinott, R. R. Hoffman, G. Klein, L. Alam, T. Mamun, and W. J. Clancey, "Principles of Explanation in Human-AI Systems," 2021. [Online]. Available: http://arxiv.org/abs/2102.04972

[24] J. A. Cannon-Bowers and E. Salas, “Reflections on shared cognition,” *Journal of Organizational Behavior*, vol. 22, no. 2, pp. 195–202, 3 2001. [Online]. Available: https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/full/10.1002/job.82https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/abs/10.1002/job.82https://onlinelibrary-wiley-com.tudelft.idm.oclc.org/doi/10.1002/job.82

[25] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, B. Van Riemsdijk, and M. Sierhuis, “The fundamental principle of coactive design: Interdependence must shape autonomy,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6541 LNAI, 2011, pp. 172–191.

[26] B. Senior, “Team roles and team performance: Is there ’really’ a link?” *Journal of Occupational and Organizational Psychology*, vol. 70, no. 3, pp. 241–258, 1997. [Online]. Available: /record/1997-05983-003

[27] M. Sierhuis, J. M. Bradshaw, A. Acquisti, R. V. Hoof, and R. Jeffers, “Human-Agent Teamwork and Adjustable Autonomy in Practice,” *Robotics*, p. 8 pp., 2003. [Online]. Available: http://www.dagstuhl.de/Materials/Files/07/07122/07122.SierhuisMaarten.Paper.pdf

[28] C. Flathmann, N. McNeese, and L. B. Canonico, “Using Human-Agent Teams to Purposefully Design Multi-Agent Systems,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 1425–1429, 11 2019. [Online]. Available: https://www.mendeley.com/catalogue/50bd34df-0774-3e11-ad48-fd53a61d2285/?utm_source=desktop&utm_medium=1.19.8&utm_campaign=open_catalog&userDocumentId=%7B46c6a9d9-aee7-3ded-8f0d-e397437c288b%7D

[29] J. E. Mathieu, G. F. Goodwin, T. S. Heffner, E. Salas, and J. A. Cannon-Bowers, “The influence of shared mental models on team process and performance,” pp. 273–283, 2000.

[30] C. S. Burke, E. Georganta, and S. Marlow, “A bottom up perspective to understanding the dynamics of team roles in mission critical teams,” *Frontiers in Psychology*, vol. 10, no. JUN, p. 1322, 2019.

[31] G. L. Stewart, I. S. Fulmer, and M. R. Barrick, “An exploration of member roles as a multilevel linking mechanism for individual traits and team outcomes,” *Personnel Psychology*, vol. 58, no. 2, pp. 343–365, 6 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1744-6570.2005.00480.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.2005.00480.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2005.00480.x

[32] R. F. Bales, “A Set of Categories for the Analysis of Small Group Interaction,” *American Sociological Review*, vol. 15, no. 2, p. 257, 4 1950.

[33] J. E. Mathieu, S. I. Tannenbaum, M. R. Kukenberger, J. S. Donsbach, and G. M. Alliger, “Team Role Experience and Orientation: A Measure and Tests of Construct Validity,” *Group and Organization Management*, vol. 40, no. 1, pp. 6–34, 2 2015.

[34] M. A. Goodrich and A. C. Schultz, “Human-robot interaction: A survey,” pp. 203–275, 2007.

[35] C. M. Jonker, B. Van Riemsdijk, I. C. Van De Kieft, and M. Gini, “Towards measuring sharedness of team mental models by compositional means,” in *AAAI Fall Symposium - Technical Report*, vol. FS-11-05, 2011, pp. 20–25. [Online]. Available: www.aaai.org

[36] G. Hoffman, "Evaluating Fluency in Human-Robot Collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 6 2019.

[37] G. Hoffman and C. Breazeal, "Cost-based anticipatory action selection for human-robot fluency," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 952–961, 2007.

[38] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," pp. 32–64, 1995.

[39] Mica R. EndsleyDaniel J. Garland, "DIRECT MEASUREMENT OF SITUATION AWARENESS: VALIDITY AND USE OF SAGAT," in *Situation Awareness Analysis and Measurement*, 2000, p. 131. [Online]. Available: https://www.researchgate.net/publication/245934995_Direct_Measurement_of_Situation_Awareness_Validity_and_Use_of_SAGAThttp://books.google.com/books?hl=zh-CN&lr=&id=WrJGDsjJakcC&pgis=1

[40] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," pp. 50–80, 2004. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15151155/

[41] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *The Academy of Management Review*, vol. 20, no. 3, p. 709, 7 1995.

[42] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measuring Trust in the XAI Context," pp. 1–26, 2018.

[43] B. Cahour and J. F. Forzy, "Does projection into use improve trust and exploration? An example with a cruise control system," *Safety Science*, vol. 47, no. 9, pp. 1260–1270, 11 2009.

[44] M. Madsen and S. Gregor, "Measuring Human-Computer Trust," *Proceedings of Eleventh Australasian Conference on Information Systems*, pp. 6–8, 2000. [Online]. Available: https://www.researchgate.net/publication/228557418_Measuring_human-computer_trusthttp://books.google.com/books?hl=en&lr=&id=b0yalwi1HDMC&oi=fnd&pg=PA102&dq=The+Big+Five+Trait+Taxonomy:+History,+measurement,+and+Theoretical+Perspectives&ots=758BNaTvOi&sig

[45] T. Hellström and S. Bensch, "Understandable robots," *Paladyn*, vol. 9, no. 1, pp. 110–123, 2 2018. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/pjbr-2018-0009/html

[46] W. Swartout, *Second Generation Expert Systems*, 1993, no. July 2014.

[47] P. J. Brezillon, "Contextualized explanations," *Proceedings of the IEEE International Conference on Expert Systems for Development*, no. April 1994, pp. 119–124, 1994.

[48] D. Doyle, A. Tsymbal, and P. Cunningham, "A Review of Explanation and Explanation in Case-Based Reasoning," Tech. Rep., 2003.

[49] B. F. Malle, "How people explain behavior: A new theoretical framework," *Personality and Social Psychology Review*, vol. 3, no. 1, pp. 23–48, 1999.

[50] ——, "How the mind explains behavior: folk explanations, meaning, and social interaction," *Choice Reviews Online*, vol. 42, no. 10, pp. 42–6170, 2005.

[51] D. Hutto and I. Ravenscroft, "Folk Psychology as a Theory," in *The Stanford Encyclopedia of Philosophy*, fall 2021 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. [Online]. Available: https://plato.stanford.edu/entries/folkpsych-theory/

[52] A. I. Goldman, "Theory of Mind," *The Oxford Handbook of Philosophy of Cognitive Science*, 1 2012. [Online]. Available: https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195309799.001.0001/oxfordhb-9780195309799-e-17

[53] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," pp. 1–38, 6 2019. [Online]. Available: https://arxiv.org/abs/1706.07269v3

[54] M. Marraffa, "Theory of Mind — Internet Encyclopedia of Philosophy." [Online]. Available: https://iep.utm.edu/theomind/http://www.iep.utm.edu/theomind/

[55] P. Kitcher and D. C. Dennett, "The Intentional Stance." *The Philosophical Review*, vol. 99, no. 1, p. 126, 1 1990.

[56] M. Harbers, J. Broekens, K. Van Den Bosch, and J. J. Meyer, "Guidelines for developing explainable cognitive models," *Proceedings of the 10th International Conference on Cognitive Modeling, ICCM 2010*, pp. 85–90, 2010.

[57] G. Taylor, K. Knudsen, and L. S. Holt, "Explaining agent behavior," in *Simulation Interoperability Standards Organization - 15th Conference on Behavior Representation in Modeling and Simulation 2006*, 2006, pp. 117–125. [Online]. Available: https://www.researchgate.net/publication/228602858_Explaining_agent_behavior

[58] S. R. Haynes, M. A. Cohen, and F. E. Ritter, "Designs for explaining intelligent agents," *International Journal of Human Computer Studies*, vol. 67, no. 1, pp. 90–110, 2009. [Online]. Available: www.elsevier.com/locate/ijhcs

[59] M. Harbers, K. Van Den Bosch, and J. J. Meyer, "Design and evaluation of explainable BDI agents," *Proceedings - 2010 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2010*, vol. 2, pp. 125–132, 2010.

[60] M. Harbers, K. Van Den Bosch, and J. J. C. Meyer, "A study into preferred explanations of virtual agent behavior," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5773 LNAI, 2009, pp. 132–145.

[61] K. Stowers, L. L. Brady, C. MacLellan, R. Wohleber, and E. Salas, "Improving Teamwork Competencies in Human-Machine Teams: Perspectives From Team Science," *Frontiers in Psychology*, vol. 12, p. 1669, 5 2021.

[62] C. a. Miller, H. B. Funk, R. P. Goldman, J. Meisner, and P. Wu, "Implications of Adaptive vs. Adaptable UIs on Decision Making: Why 'Automated Adaptiveness' Is Not Always the Right Answer," *Proceedings of the 1st International Conference on Augmented Cognition*, no. May 2014, pp. 1180–1189, 2005. [Online]. Available: http://ezproxy.net.ucf.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ega&AN=ega213477&site=ehost-live

[63] S. Schneider and F. Kummert, "Comparing Robot and Human guided Personalization: Adaptive Exercise Robots are Perceived as more Competent and Trustworthy," *International Journal of Social Robotics*, vol. 13, no. 2, pp. 169–185, 2021. [Online]. Available: https://doi.org/10.1007/s12369-020-00629-w

[64] J. Lin, G. Matthews, R. W. Wohleber, G. J. Funke, G. L. Calhoun, H. A. Ruff, J. Szalma, and P. Chiu, "Overload and Automation-Dependence in a Multi-UAS Simulation: Task Demand and Individual Difference Factors," *Journal of Experimental Psychology: Applied*, 2019.

[65] T. Kulesza, M. Burnett, W. K. Wong, and S. Stumpf, “Principles of Explanatory Debugging to personalize interactive machine learning,” in *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. 2015-Janua. Association for Computing Machinery, 3 2015, pp. 126–137.

[66] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Metrics for Explainable AI: Challenges and Prospects,” 12 2018. [Online]. Available: https://arxiv.org/abs/1812.04608v2http://arxiv.org/abs/1812.04608

[67] D. B. Leake, “Goal-based explanation evaluation,” *Cognitive Science*, vol. 15, no. 4, pp. 509–545, 1991.

[68] E. J. Bass, L. A. Baumgart, and K. K. Shepley, “The effect of information analysis automation display content on human judgment performance in noisy environments,” *Journal of Cognitive Engineering and Decision Making*, vol. 7, no. 1, pp. 49–65, 3 2013. [Online]. Available: /record/2013-06180-003

[69] J. Beller, M. Heesen, and M. Vollrath, “Improving the driver-automation interaction: An approach using automation uncertainty,” *Human Factors*, vol. 55, no. 6, pp. 1130–1141, 12 2013.

[70] J. Y. Chen, A. R. Selkowitz, K. Stowers, S. G. Lakhmani, and M. J. Barnes, “Human-autonomy teaming and agent transparency,” in *ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 91–92. [Online]. Available: http://dx.doi.org/10.1145/3029798.3038339

[71] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci, “Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management,” *Human Factors*, vol. 58, no. 3, pp. 401–415, 5 2016. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0018720815621206

[72] M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?” explaining the predictions of any classifier,” in *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*. Association for Computational Linguistics (ACL), 2 2016, pp. 97–101. [Online]. Available: https://arxiv.org/abs/1602.04938v3

[73] K. Sokol and P. Flach, “Desiderata for interpretability: Explaining decision tree predictions with counterfactuals,” in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. AAAI Press, 2019, pp. 10 035–10 036.

[74] W. Pierce, “Search and rescue,” *Prehospital and Disaster Medicine*, vol. 2, no. 1-4, pp. 21–24, 1986. [Online]. Available: https://www.sciencedaily.com/terms/search_and_rescue.htm

[75] “INSARAG GUIDELINES 2020 – INSARAG.” [Online]. Available: https://www.insarag.org/methodology/insarag-guidelines/

[76] D. Zarrouk, A. Pullin, N. Kohut, and R. S. Fearing, “STAR, a sprawl tuned autonomous robot,” in *Proceedings - IEEE International Conference on Robotics and Automation*, 2013, pp. 20–25. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.309.6908&rep=rep1&type=pdf

[77] "SmokeBot - Home." [Online]. Available: https://www.smokebot.eu/

[78] Hydronalix, "MOBE — Advanced Small Surface Robotic Systems — Hydronalix — United States." [Online]. Available: https://www.hydronalix.com/disaster-responsehttps://www.hydronalix.com/mobe

[79] F. E. Schneider and D. Wildermuth, "Assessing the search and rescue domain as an applied and realistic benchmark for robotic systems," *Proceedings of the 2016 17th International Carpathian Control Conference, ICCC 2016*, no. pre 2005, pp. 657–662, 2016.

[80] A. Hong, O. Igharoro, Y. Liu, F. Niroui, G. Nejat, and B. Benhabib, "Investigating Human-Robot Teams for Learning-Based Semi-autonomous Control in Urban Search and Rescue Environments," *Journal of Intelligent and Robotic Systems: Theory and Applications*, vol. 94, no. 3-4, pp. 669–686, 8 2019. [Online]. Available: https://link.springer.com/article/10.1007/s10846-018-0899-0

[81] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, vol. 30, no. 3, pp. 286–297, 2000.

[82] T. Nagy and C. Jorge, "Influence of shared information on predictability in human-agent teams," pp. 1–25, 2021.

[83] "Lowland Rescue." [Online]. Available: https://www.lowlandrescue.org/

[84] J. Saldana, "Essential Guide to Coding Qualitative Data," *Delve*, pp. 1–35, 2009. [Online]. Available: https://delvetool.com/guide

[85] J. Kreske, K. O'Holleran, and H. Rajaniemi, "Trusted reasoning engine for autonomous systems with an interactive demonstrator," *4th SEAS DTC Technical Conference*, no. 5, 2009.

[86] "ColorADD - Color is for ALL!" [Online]. Available: https://www.coloradd.net/en/

Appendix A

HAT Design: Scenarios

It is important to look at relevant scenarios from the task in order to see how the explanations are formed throughout the task. There are many scenarios possible during the main experiment, but there are four scenarios chosen which can be considered relevant enough for exploring the explanations for adapting their style to the team leader role: the agent has found an obstacle, the agent has found an entry to an area, the agent has found an unassessed area, and the agent has found an unassessed survivor. After reevaluation, the scenario where the agent finds an obstacle and where the agent finds an entry to an area are merged because the explanations were too similar.

The explanation type ‘confidence explanation’ is used to indicate certainty, but the rest of the explanation types for the scenarios and each phase in the task can be found in Appendix B. The focus of the explanations was put on the time it takes to complete the action. It is important that the human Team Leader made informed decisions, no matter the explanation type.

Scenario 1 - Obstacle/Entry Found Here, it needs to be determined if the obstacle is obstructing a path to survivors or safety. If (and only if) that is the case, it needs to be determined how long it will take to remove the obstacle. So, this event can have two variables which determine the belief of the agent. The agent can continue the search (maybe come back later) or remove the obstacle/clear the entry.

Too long to remove	Certain	Action	Explanation
		I need advice to continue the search or remove obstacle	I am only 35% certain that it will take 4-15 seconds to remove this obstacle.
	✓	Autonomous action: remove obstacle	I am 85% certain that it will take 6-8 seconds to remove this obstacle.
✓		I need advice to continue the search or remove obstacle	I am only 35% certain that it will take 4-15 seconds to remove this obstacle.
✓	✓	I advise to continue the search. Alternative: remove obstacle	I am 85% certain that it will take 12-15 seconds to remove this obstacle.

Table A.1: Difference in communication due to information when an agent from the search phase has assessed an obstacle/entry.

Scenario 2 - Area Assessed Here, it needs to be determined whether the area is dangerous or safe to explore. The agent can explore the area or let the agent with the right capabilities explore this area during the rescue phase.

Dangerous	Certain	Action	Explanation
		I need advice to explore this area or let the Rescue Robot search this area during the Rescue Phase	I am only 35% certain that this area is safe to explore. If I explore a dangerous area, I might get damaged and would have to recover for 1-10 seconds. The Rescue Robot will not get damaged.
	✓	Autonomous action: explore area	I am 85% certain that this area is safe to explore.
✓		I need advice to explore this area or let the Rescue Robot search this area during the Rescue Phase	I am only 35% certain that this area is too dangerous to explore. If I explore a dangerous area, I might get damaged and would have to recover for 1-10 seconds. The Rescue Robot will not get damaged.
✓	✓	I advise to let the Rescue Robot search this area during the Rescue Phase. Alternative: explore area	I am 85% certain that this area is too dangerous for me to explore. If I explore a dangerous area, I might get damaged and would have to recover for 1-10 seconds. The Rescue Robot will not get damaged.

Table A.2: Difference in communication due to information when an agent from the search phase has assessed an area.

Scenario 3 - Survivor Assessed Here, it needs to be determined whether the found survivor is critically injured, stuck, or neither (healthy). The agent can carry the survivor to safety, continue the search, call a different robot for during the rescue phase, or call an additional robot for during the rescue phase.

Critically injured	Stuck	Certain	**Action**	**Explanation**
			I need advice to carry this survivor or let the Rescue Robot rescue this survivor during the Rescue Phase	This survivor is healthy. I am only 35% certain that it will take 2-8 seconds to rescue this survivor.
		✓	Autonomous action: carry survivor	This survivor is healthy. I am 85% certain that it will take 3-5 seconds to rescue this survivor.
	✓		I need advice to carry this survivor or let the Rescue Robot rescue this survivor during the Rescue Phase	This survivor is stuck. I am only 35% certain that it would take 5-35 seconds to rescue this survivor by myself. During the Rescue Phase, the Rescue Bot can call the other Rescue Robot which will half this rescue time.
	✓	✓	I advise to let the Rescue Robot rescue this survivor during the Rescue Phase. Alternative: carry survivor	This survivor is stuck. I am 85% certain that it would take 18-20 seconds to rescue this survivor by myself. During the Rescue Phase, the Rescue Bot can call the other Rescue Robot which will half this rescue time.
✓			I need advice to carry this survivor or let the Rescue Robot rescue this survivor during the Rescue Phase	This survivor is critically injured. I am only 35% certain it will take 3-15 seconds to rescue this survivor. There is only a 50% chance I will successfully rescue this survivor. The Rescue Robot cannot fail at rescuing a survivor.
✓		✓	I advise to let the Rescue Robot rescue this survivor during the Rescue Phase Alternative: carry survivor	This survivor is critically injured. I am 85% certain it will take 6-8 seconds to rescue this survivor. There is only a 50% chance I will successfully rescue this survivor. The Rescue Robot cannot fail at rescuing a survivor.

Table A.3: Difference in communication due to information when an agent from the search phase has assessed a survivor.

Appendix B

HAT Design: Robot decision and explanations per style and situation

Link to view the table: https://tud365-my.sharepoint.com/:x:/g/personal/rkap_tudelft_nl/EaCOYJmJxQVFqOjifhqWh9ABgufPnp6w55bNyOMs6W3pMw?e=lr2o5c

Appendix C

HAT Design: Implementation

The chapter describes the implementation in MATRX of the HAT design of this thesis. To see the full code see Appendix E.

C.1. Task Map

For the task map, the following two images display the most important elements discussed in the design.

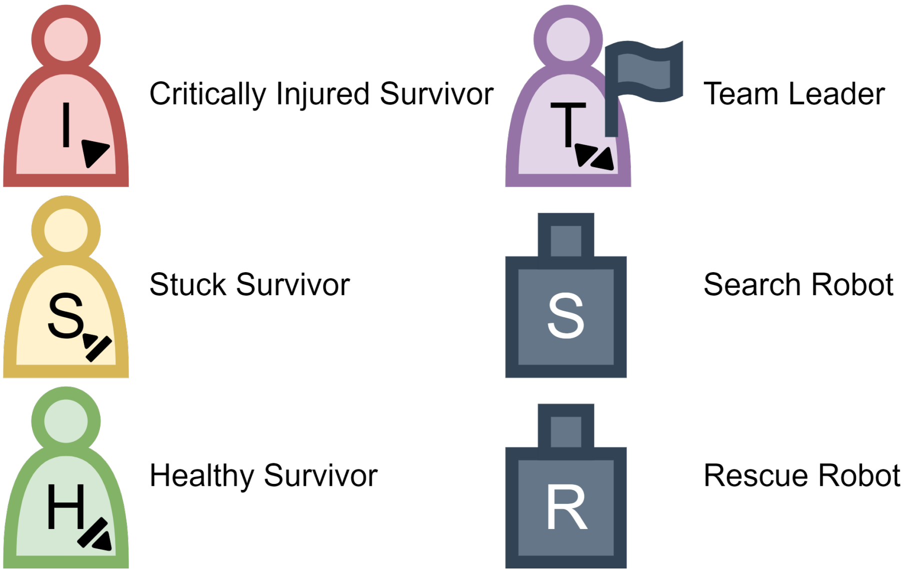


Figure C.1: Icons of the different types of survivors, robots, and the team leader.

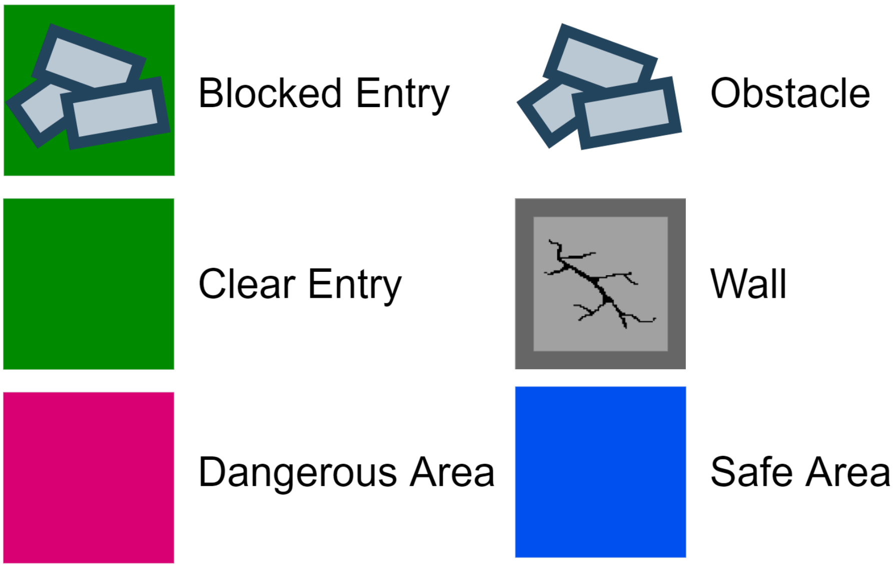


Figure C.2: Icons of a blocked and clear entry, a dangerous and safe area, an area wall, and an obstacle.

The survivors also have symbols in the bottom right of their icon which represent their colour for people who have trouble with colours [86]. First, a small test map was designed with three rooms each containing the different types of survivors (healthy, injured, and stuck) and different kinds of combinations of the safety of an area (safe or dangerous) and the status of the entries (blocked or clear). The following figure shows this map.

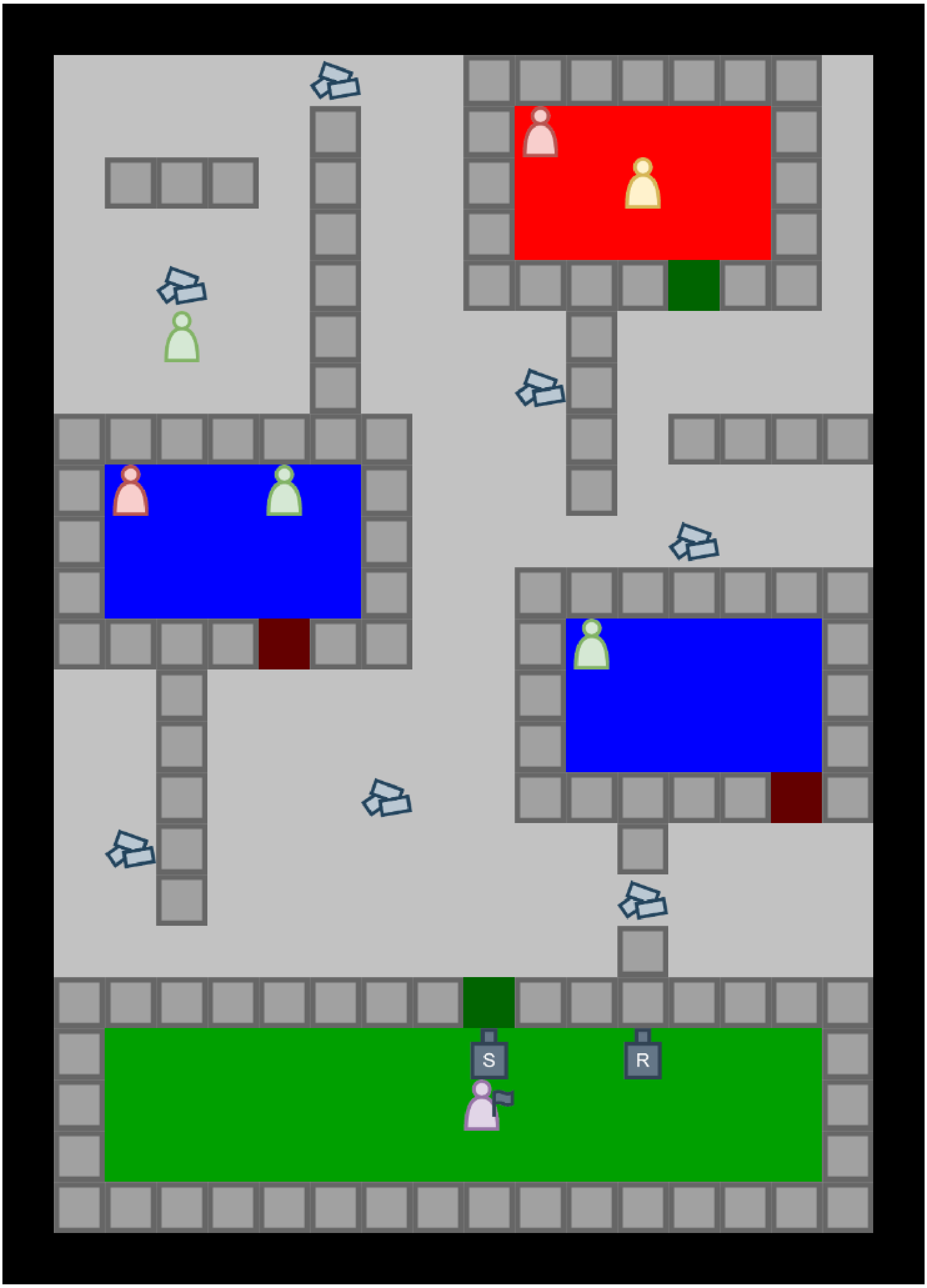


Figure C.3: Initial map design.

After a pilot study, the map needed to be bigger with more decisions to make as a team leader. The following image shows the map used during the experiment.

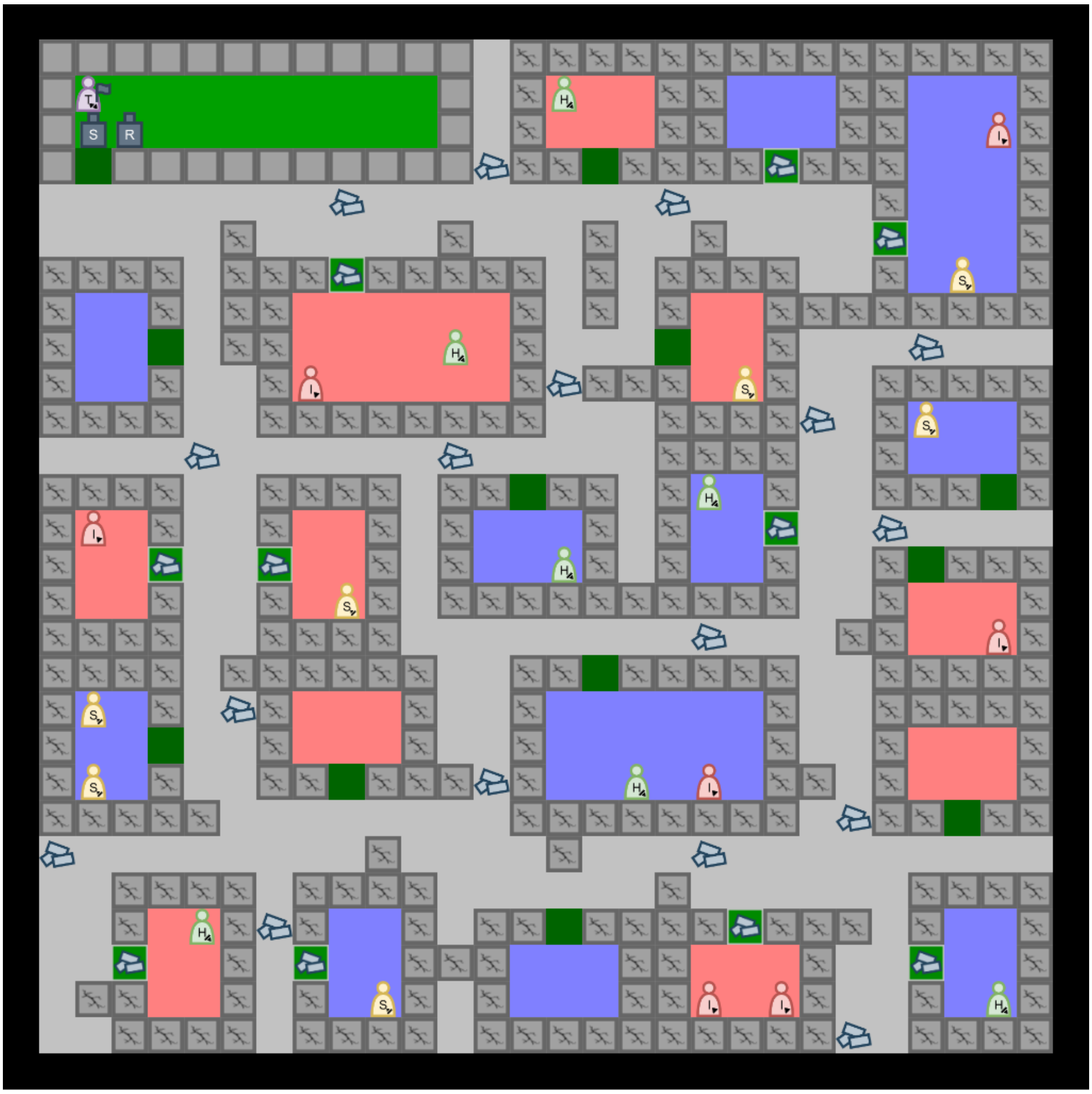


Figure C.4: Clear view of map used in the experiment.

This map has 21 areas of which 11 are clear and 10 are blocked. Moreover, the map has 21 survivors of which 7 are healthy, 7 are critically injured, and 7 are stuck. Additionally, the map has 17 obstacles inside the Disaster Site but outside the areas. In this map, 11 of the 17 obstacles are blocking the path. Having an imbalance between these elements could lead to (negatively) influencing the results. Therefore, during the design of the map, balancing these elements was key. The following aspects are balanced:

- The certainty the robot has of each item. For example, the confidence the robot has in how long it will take to rescue a survivor or to remove an obstacle.
- The time it will take to perform an action. For example, how long it will take to clear a blocked area or to free a stuck survivor. Additionally, some of these times need to be considered 'too long' for the robot to react differently (e.g., take a different action or give different advice).
- Lastly, the combination of all these elements as well. For example, the number of different types of survivors (healthy, injured, and stuck) that are in a blocked-off room versus those that are in a room with a clear entry.

See Appendix D for the overview of the values used for each element.

C.2. Tutorial

The task with all its elements and mechanics can be perceived as quite diverse and big. Therefore the initial map (see Figure C.3) was converted to a tutorial map where the robot will explain all the mechanics and give important information about the task. This way the participant will also have seen all elements and the different phases at least once all and will know what to do. It is a small version of the actual task with just two areas, two obstacles, and three survivors. The different types of areas, obstacles, entries and survivors are all present so the participant has seen them at least once. For example, a dangerous area and a safe-to-enter area. Figure 5.3 gives an overview of what the tutorial map looks like.

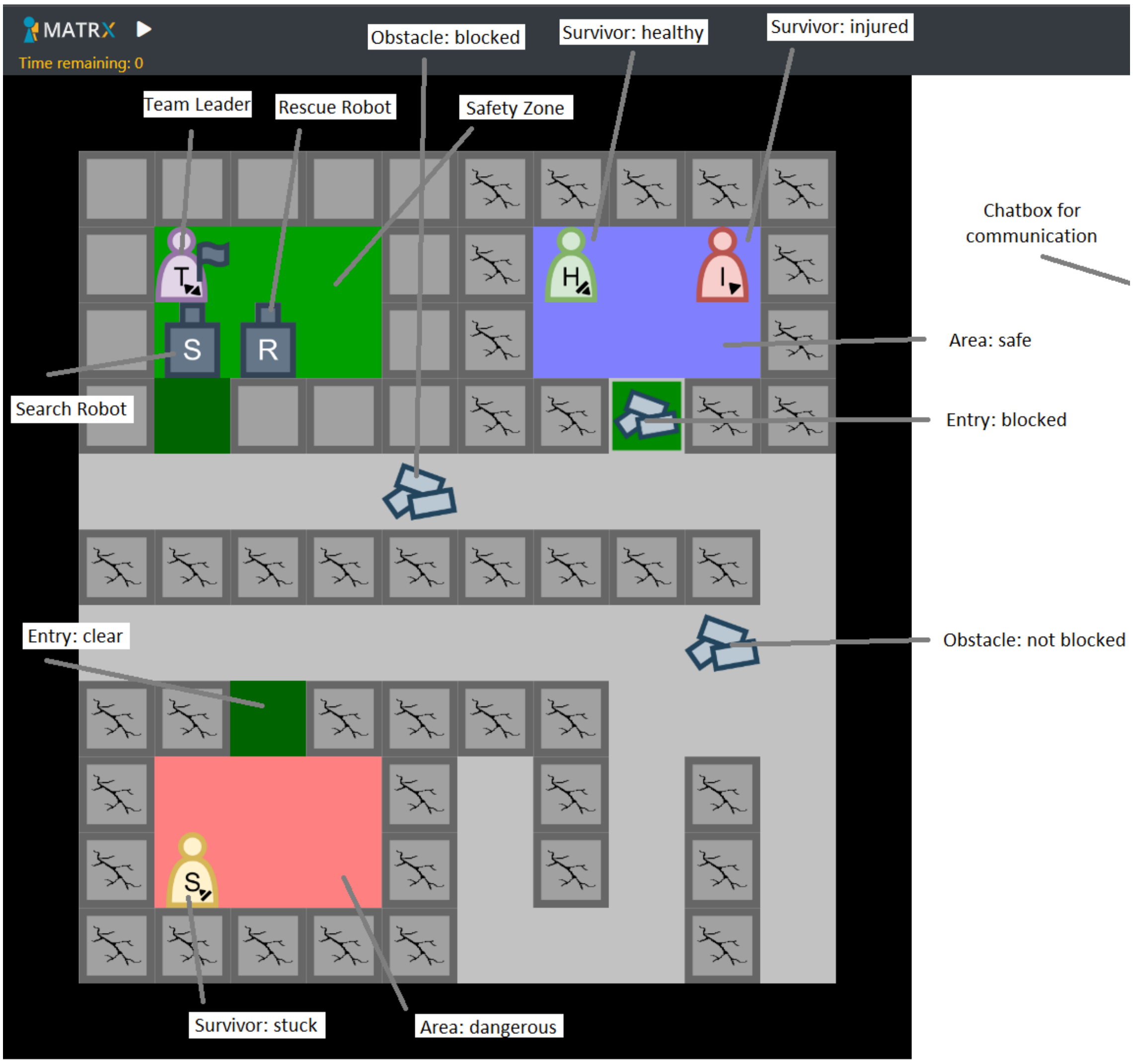


Figure C.5: Overview of the tutorial map.

C.3. Search Phase

To recap, the most important things for the search phase are the capabilities of the agent in the search phase, namely, exploring the disaster site, assessing and removing obstacles, assessing and clearing entries, assessing areas, and carrying 'healthy' survivors. Additionally, this is the phase where the human team leader can supervise the agent the most by giving advice when the agent needs it or choosing to accept advice or not when the agent gives it. The following diagram is the state diagram for the search agent in the search phase:

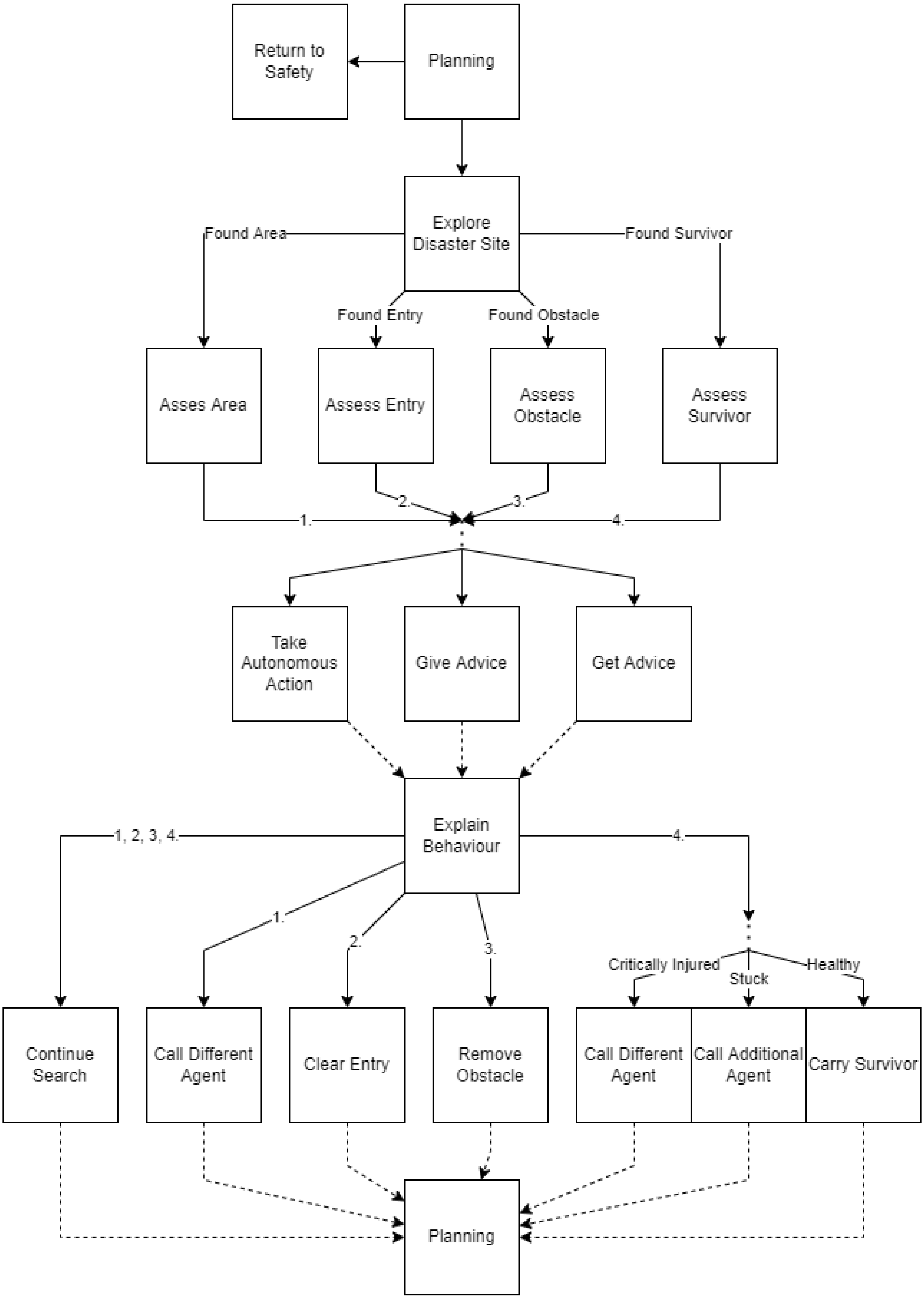


Figure C.6: State machine diagram for the agent in the search phase.

To search the disaster site, the search robot uses an adapted version of depth-first search, namely, randomly exploring a child node instead of always the most left child node. In the agent's '*decide_on_action*()' function, the agent returns the action from the state it is currently in. If the search robot finds an object (obstacle, entry, area, or survivor) it will choose to take an autonomous action, give advice, or get advice. These choices are based on a pre-determined table which has information on the situation and the confidence the agent has in that information. For example, if the search robot finds a dangerous area and it is certain of this information (> 60%) it will advise to let the Rescue Robot search this area during the Rescue Phase. This table is derived from the scenario tables in section 3.2.6, see Appendix B for an overview.

After the agent has chosen one of the three options it will explain its behaviour. The explanation type will be based on the result of the pre-study results (see table 4.4 for which explanation style the adaptive group will use based on situation information (the baseline group will pick an explanation style at random)). With the explanation type and the situation information, the robot can look at the pre-determined table to explain itself, see Appendix B for this table.

After the explanation of the behaviour of the agent, it will perform the action. Each time the agent has to wait for the user to react (give and get advice), the agent will store the name of the actions and the descriptions from the chat window in its *custom_properties* in the state of the world. The following figure shows examples of when the participant can give advice and when the participant can choose to follow the advice of the robot or not.

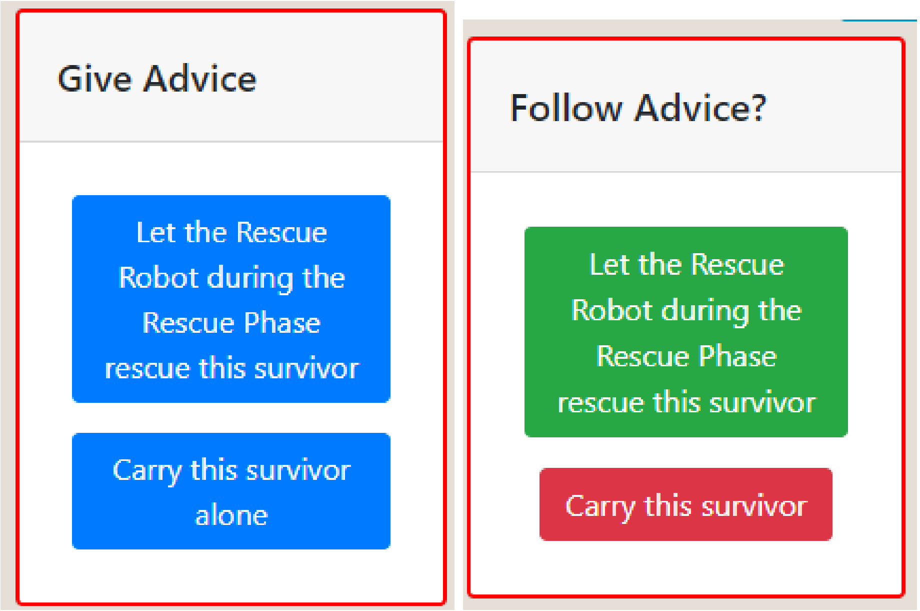

Give Advice
Let the Rescue Robot during the Rescue Phase rescue this survivor
Carry this survivor alone
Follow Advice?
Let the Rescue Robot during the Rescue Phase rescue this survivor
Carry this survivor

Figure C.7: Example of options the participants can have.

The current state can be read by the front end and therefore displayed as options. While the agent waits for an answer, it constantly checks the chat with the team leader to see if any answer corresponds to the options in the *custom_properties* to determine which option the team leader has chosen. This waiting time for the participant to react to the agent is also stored to measure objective collaborative fluency.

Any time the team leader (or the agent autonomously) chooses to continue the search (for now), the object and its information will be stored in a 'to-do' list. The agent will return to this list if there is time left in the search phase and the area has been explored where currently possible. The agent will return to the safety zone if the time has run out or the remaining ‘to-do’ list is empty. This is the moment the search phase ends.

C.4. Rescue Phase

The most important thing for the rescue phase is the rescue plan and rescuing the remaining survivors. Firstly, during the rescue phase, there are moments where an additional robot is needed to free a survivor or a different robot is needed to explore dangerous areas or carry critically injured robots. These moments are stored and displayed to the user to make a rescue plan. The following figure is the state machine of the robot in the rescue phase:

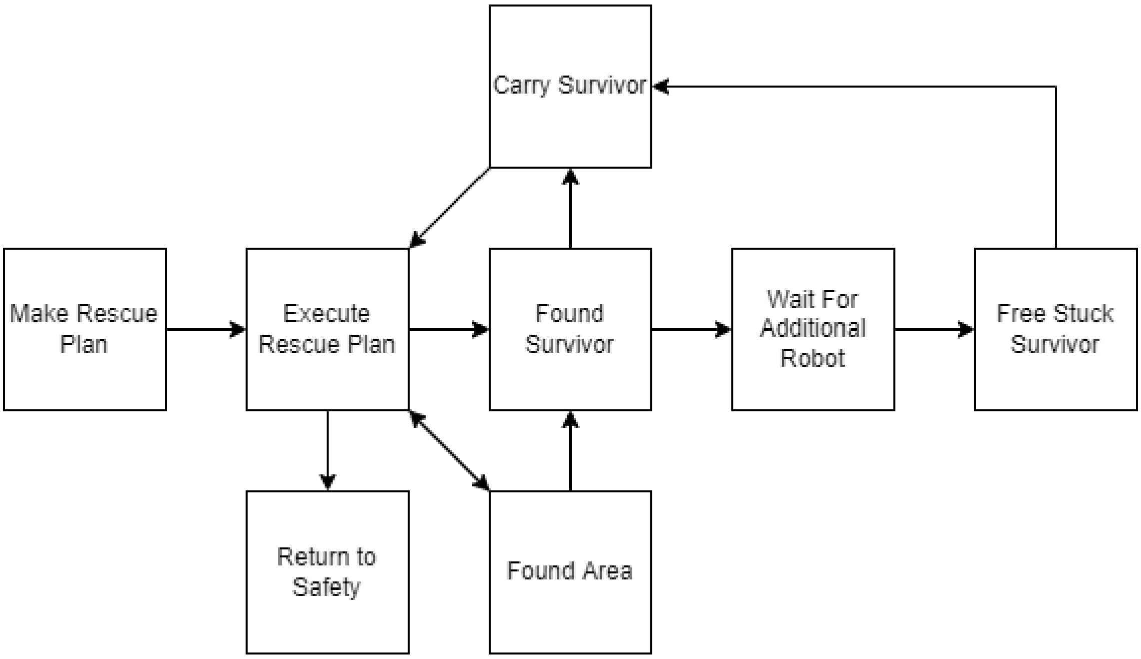
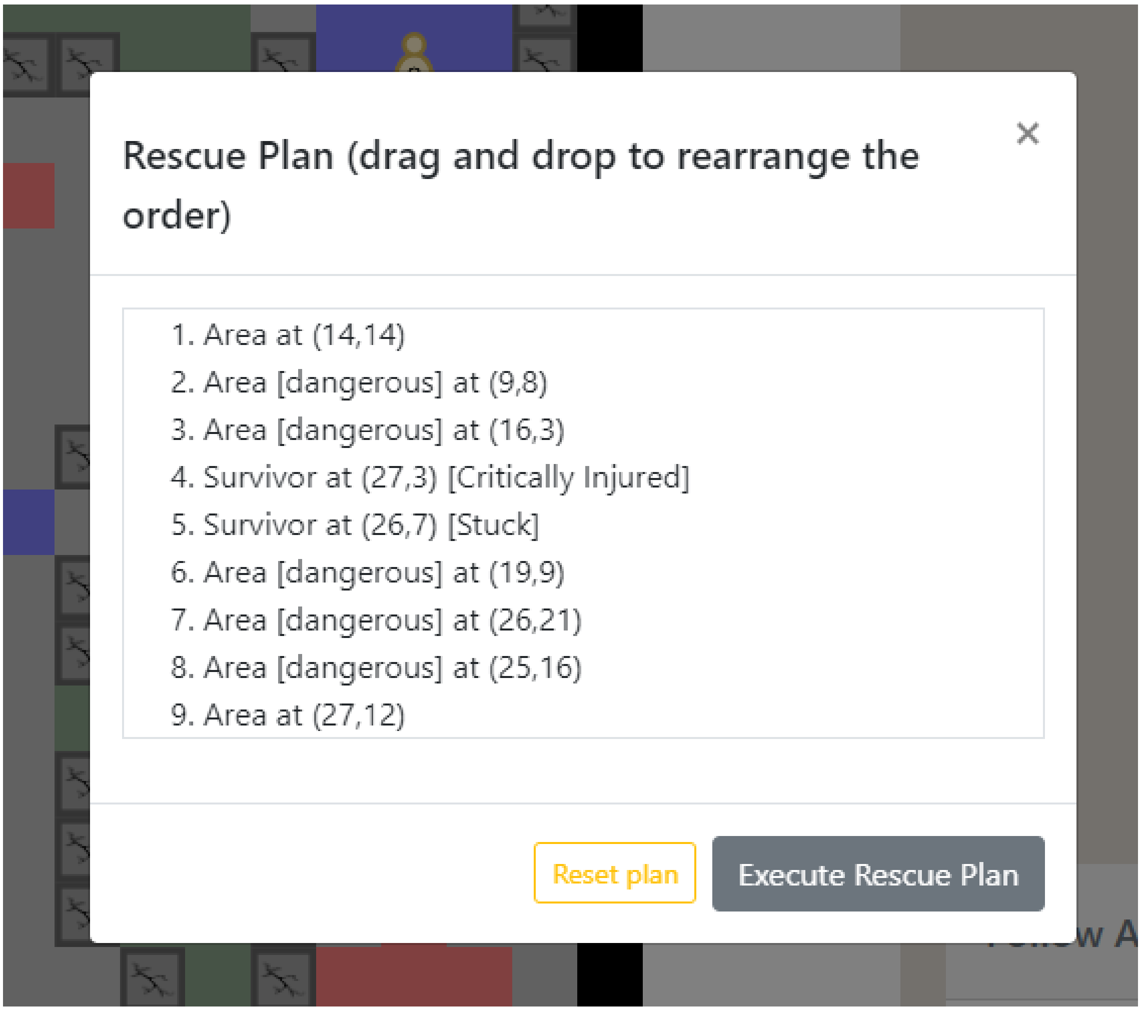


Figure C.8: State machine diagram for the agent in the rescue phase.

The rescue phase starts with making a rescue plan. The user can drag and drop the list of relevant moments from the search phase to make an ordered list. Once the participant is satisfied with the rescue plan there is a button to execute the plan, see Figure C.9.

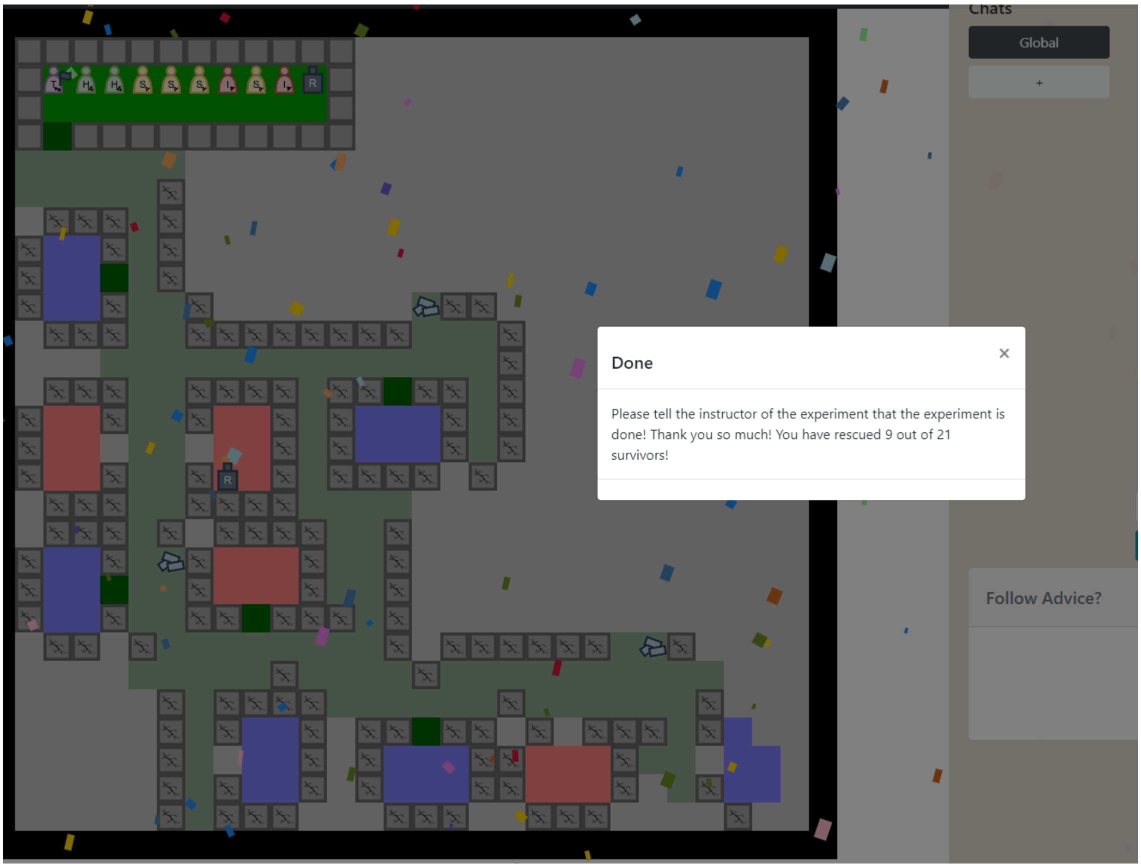


Rescue Plan (drag and drop to rearrange the order)

1. Area at (14,14)
2. Area [dangerous] at (9,8)
3. Area [dangerous] at (16,3)
4. Survivor at (27,3) [Critically Injured]
5. Survivor at (26,7) [Stuck]
6. Area [dangerous] at (19,9)
7. Area [dangerous] at (26,21)
8. Area [dangerous] at (25,16)
9. Area at (27,12)

Figure C.9: Example of a rescue plan.

This list is then sent to a singleton which stores the rescue plan data. This is done via an API call. The agent constantly tries to pull the rescue plan until the rescue plan is not *None*. As soon as the agent has a rescue plan it will go towards the location of the step. There it will either explore the area (look for survivors), carry the survivor if it is critically injured or ‘healthy’ (the same code as in the search phase), or the agent will call an additional robot which to come to its location in order to free the stuck survivor together. Once the stuck survivor is free, the additional agent will return to the safety zone and the agent will carry the survivor to the safety zone. If the agent comes across a survivor in a dangerous area, it will make that survivor a priority in the rescue plan. If the rescue plan is all executed or if the time for the rescue plan is up, the agent will return to the safety zone and the SAR task is over. The number of survivors rescued will be saved to a file along with information such as the number of times advice from the agent has been followed and how much time was remaining. The following Figure shows the end screen of the task.



Done

Please tell the instructor of the experiment that the experiment is done! Thank you so much! You have rescued 9 out of 21 survivors!

Figure C.10: Endscreen when the task is done.

Appendix D

HAT Design: Elements in the search and rescue task

Element	Range (s)
Stuck survivor	5-10
Injured Survivor	3-7
Healthy Survivor	2-5
Blocked Entry	3-10
Obstacle	3-10

Room number	**Blocked**	**Dangerous**	**Confidence**
1		✓	43
2	✓		87
3	✓		39
4			23
5	✓	✓	94
6		✓	69
7			49
8	✓	✓	66
9	✓	✓	74
10			23
11	✓		91
12		✓	55
13			70
14		✓	89
15			85
16		✓	94
17	✓	✓	82
18	✓		58
19			68
20	✓	✓	36
21	✓		77

Survivor	**Healthy**	**Injured**	**Stuck**	**Time to rescue**	**Confidence**
1	✓			2	75
2		✓		6	45
3			✓	5	81
4		✓		7	32
5	✓			3	69
6			✓	6	55
7			✓	6	91
8		✓		5	47
9			✓	8	72
10	✓			2	82
11	✓			3	50
12		✓		6	59
13			✓	6	75
14			✓	9	52
15	✓			2	88
16		✓		3	28
17	✓			5	85
18			✓	7	35
19		✓		6	75
20		✓		4	50
21	✓			2	75

Obstacle	**Blocks path**	**Time to remove**	**Confidence**
1		5	35
2	✓	3	85
3		9	55
4		3	82
5	✓	6	53
6		4	77
7		8	27
8	✓	10	69
9	✓	9	42
10	✓	7	98
11		9	78
12	✓	6	83
13	✓	3	100
14		4	23
15	✓	7	48
16	✓	3	91
17		6	61

Entry	**Blocked**	**Time to clear**	**Confidence**
1			33
2	✓	7	78
3	✓	4	45
4			32
5	✓	8	88
6			68
7			55
8	✓	3	95
9	✓	5	85
10			35
11	✓	10	82
12			40
13			49
14			74
15			91
16			92
17	✓	6	35
18	✓	4	71
19			49
20	✓	8	70
21	✓	7	81

Appendix E

MATRX code

R.B. Kap, Adapting Explanation Style to Team Roles (2022), GitLab repository, `https://gitlab.ewi.tudelft.nl/in5000/ii/matrx/adapting-explanation-style-to-team-roles`

Appendix F

Pre-study: recruitment email

Dear reader,

My name is Ryan Kap. I am a master's student in Computer Science at the Delft University of Technology (in The Netherlands). Right now, I am writing my thesis which covers collaboration between humans and agents (e.g., robots). The focus of my thesis is on the explanations of a robot directed at a team leader of a search and rescue team. As a part of this research, I have made a questionnaire. This questionnaire is the reason I am writing to you with the following question:

- Could you forward this questionnaire to people in your organisation that are part of a search and rescue team?
- Other teams that perform similar tasks or have similar goals are also allowed to fill out the questionnaire (e.g., firemen or policemen)
- The questionnaire will take around 25 minutes to complete.
- To make it as anonymous as possible, I do not wish to know who fills in the questionnaire, only if you have sent it forward or not.
- To make it as anonymous as possible, no personal information will be recorded, only the role the person has in the team.
- Taking part in this questionnaire will aid in the advancement of search and rescue teams!

The link to the questionnaire: `https://forms.office.com/r/JJdgfWq251`

I would appreciate it a lot and I want to thank you in advance.

Kind regards,
Ryan Kap
Master's student Delft University of Technology

Appendix G

Pre-study: Informed consent

Dear reader,

You are being invited to participate in a research study titled "Preferred Explanation for Team Roles in Search and Rescue". This study is being done by Master's student Ryan Kap from the Delft University of Technology (https://www.tudelft.nl/en/) under the research group Interactive Intelligence (`https://www.tudelft.nl/en/ewi/over-de-faculteit/afdelingen/intelligent-systems/interactive-intelligence`).

The purpose of this study is to see if the roles of team leader and search leader (or similar) in search and rescue (SAR) teams have preferred types of explanations in fictional scenarios of human-agent teaming and if so, which explanations are preferred by which roles. In this case, human-agent teaming means that humans and robots are working together to search and rescue people during/after a disaster. It will take approximately 26 minutes to complete. The data will be used for the master's thesis project regarding team roles and explanations in human-agent teaming. I will be asking you to read fictional scenarios from a search and rescue task with human-agent teams. After each scenario, the agent (i.e., robot) from the scenario will communicate its behaviour in different ways. After each communication, you will be presented with explanations, and it is up to you to raking the explanations for the communication to your liking.

As with any online activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by not storing any Personally Identifiable Information (PII) and making the questionnaire completely anonymous.

The participants are part of a search and rescue team (or similar such as the fire department) and have the role of team leader or search leader (or similar).

Search and Rescue operations can be stressful. Even though fictional scenarios will be described with human-agent teaming, you should be aware that there is a small risk of increased stress of reading these scenarios.

Your participation is entirely voluntary, and you have the right to refuse, stop, or withdraw from the study at any moment.

There will be no remuneration for time or compensation for travel (as this is not applicable).

You need to declare that you have read, understood, and agree with this opening statement to continue participation.

If there are any issues or comments, please feel free to contact me at
r.b.kap@student.tudelft.nl

Thank you again!

Appendix H

Pre-study: Data management plan

Plan Overview

A Data Management Plan created using DMPonline

Title: Preferred Explanations for Team Roles in Search and Rescue with Human-Agent Teaming

Creator: Ryan Kap

Affiliation: Delft University of Technology

Template: TU Delft Data Management Plan template (2021)

Project abstract:

The goal of this study is to see whether people fulfilling a leading role in a human-agent team have a preference for a specific (combination of) explanation type(s) and if so, which. Through an online questionnaire, this study will describe people from search and rescue teams (with a leading role) fictional scenarios of a search and rescue task being performed by a human-agent team. During each scenario, the agent (e.g., robot) from a scenario will communicate its behaviour to the human with a leading role. Next, the participants will answer questions about types of preferred explanations regarding this communication.

ID: 94338

Start date: 28-02-2022

End date: 27-06-2022

Last modified: 14-04-2022

Preferred Explanations for Team Roles in Search and Rescue with Human-Agent Teaming

0. Administrative questions

1. Name of data management support staff consulted during the preparation of this plan.

My faculty data steward, Santosh Ilamparuthi, has reviewed this DMP on [date].

2. Date of consultation with support staff.

2022-02-23

I. Data description and collection or re-use of existing data

3. Provide a general description of the type of data you will be working with, including any re-used data:

Type of data	File format(s)	How will data be collected (for re-used data: source and terms of use)?	Purpose of processing	Storage location	Who will have access to the data
Ranking questions about types of explanations given a situation	.csv	Questionnaire (Microsoft Forms)	Master's Thesis	Cloud	Ryan Kap + Supervisors
Open elaborative questions about their answer to the ranking question	.csv	Questionnaire (Microsoft Forms	Master's Thesis	Cloud	Ryan Kap + Supervisors

4. How much data storage will you require during the project lifetime?

- < 250 GB

II. Documentation and data quality

5. What documentation will accompany data?

- README file or other documentation explaining how data is organised

III. Storage and backup during research process

6. Where will the data (and code, if applicable) be stored and backed-up during the project lifetime?

- OneDrive

IV. Legal and ethical requirements, codes of conduct

7. Does your research involve human subjects or 3rd party datasets collected from human participants?

- Yes

8A. Will you work with personal data? (information about an identified or identifiable natural person)

If you are not sure which option to select, ask yourFaculty Data Steward for advice. You can also check with the privacy website or contact the privacy team: privacy-tud@tudelft.nl

- No

8B. Will you work with any types of confidential or classified data or code as listed below? (tick all that apply)

If you are not sure which option to select, ask yourFaculty Data Steward for advice.

- No, I will not work with any confidential or classified data/code

9. How will ownership of the data and intellectual property rights to the data be managed?

For projects involving commercially-sensitive research or research involving third parties, seek advice of yourFaculty Contract Manager when answering this question. If this is not the case, you can use the example below.

The datasets underlying the published papers will be publicly released following the TU Delft Research Data Framework Policy. During the active phase of research, the project leader from TU Delft will oversee the access rights to data (and other outputs), as well as any requests for access from external parties. They will be released publicly no later than at the time of publication of corresponding research papers.

V. Data sharing and long-term preservation

26. What data will be publicly shared?

- All data (and code) underlying published articles / reports / theses

28. How will you share your research data (and code)?

- All data will be uploaded to 4TU.ResearchData
- I will share my data and code via git(lab)/subversion and also create a snapshot in a repository

30. How much of your data will be shared in a research data repository?

- < 100 GB

31. When will the data (or code) be shared?

- As soon as corresponding results (papers, theses, reports) are published

32. Under what licence will be the data/code released?

- MIT License
- CC BY

VI. Data management responsibilities and resources

33. Is TU Delft the lead institution for this project?

- Yes, the only institution involved

34. If you leave TU Delft (or are unavailable), who is going to be responsible for the data resulting from this project?

dr. M.L. Tielman (m.l.tielman@tudelft.nl)

35. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

4TU.ResearchData is able to archive 1TB of data per researcher per year free of charge for all TU Delft researchers. We do not expect to exceed this and therefore there are no additional costs of long term preservation.

Appendix I

Pre-study: Questionnaire

Search and Rescue: Teamwork - Explanations: `https://forms.office.com/r/JJdgfWq251`

Your Team Role

11. Which title describes your role within your team best? *

If you do not work within a SAR-team or if a role does not fit, please choose the 'Other' option.

○ Team Leader

○ Search Leader

○ Rescue Leader

○ Searcher/Search Technician

○ Rescuer/Rescue Technician

○ Other

Scenario 1 - Survivor Assessed

Imagine you are a **team leader** of a Search and Rescue (SAR) team.

The goal of the team is to **rescue as many survivors** within a **time limit** (e.g., very very bad weather is coming up, making it impossible to continue).

This SAR consists of a human (you) and a robot with a **supporting team role**. The robot does the searching for survivors and rescuing of survivors and **you supervise the robot** from a safety zone.

Your goal as team leader is to **make sure the goal of the team is reached**.

During this scenario, the robot has found a survivor and has **assessed the survivor** for whether it is critically injured and stuck.

12. The robot communicates: "I am carrying the survivor to safety because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to rescue this survivor

My sensors indicate no critical injuries

I am 95% certain that this survivor is not critically injured

Requesting medical assistance is not needed

If this survivor was critically injured, I would have requested medical assistance

13. Please elaborate on your **ranking** from the **previous** question *

14. How would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I am carrying the survivor to safety because ..."
*

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to rescue this survivor

My sensors indicate no critical injuries

I am 95% certain that this survivor is not critically injured

Requesting medical assistance is not needed

If this survivor was critically injured, I would have requested medical assistance

15. The robot communicates: "I advise to let two robots rescue this survivor because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to rescue this survivor

My sensors indicate that this survivor is stuck

I am 95% certain that this survivor is stuck

I am unable to rescue this survivor alone

If this survivor was not stuck, I would have rescued this survivor alone

16. Please elaborate on your **ranking** from the **previous** question *

17. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I advise to let two robots rescue this survivor because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to rescue this survivor

My sensors indicate that this survivor is stuck

I am 95% certain that this survivor is stuck

I am unable to rescue this survivor alone

18. The robot communicates: "I need advice on whether to carry this survivor because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to make the best decision for this survivor and the team

My sensors are unable to accurately predict this survivor's injuries

I am only 40% certain that this survivor is not critically injured

I am not confident enough to make an autonomous decision about carrying this survivor or requesting medical assistance

If I was more than 80% certain that this survivor is not critically injured, I would have

19. Please elaborate on your **ranking** from the **previous** question *

20. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I need advice on whether to carry this survivor because ..."
*

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to make the best decision for this survivor and the team

My sensors are unable to accurately predict this survivor's injuries

I am only 40% certain that this survivor is not critically injured

I am not confident enough to make an autonomous decision about carrying this survivor or requesting medical assistance

Scenario 2 - Assessed entry to an area

Imagine you are a team leader of a Search And Rescue (SAR) team. This SAR consists of a human (you) and a robot with a supporting team role. The robot does the searching and rescuing and you supervise the robot.

During this scenario, **the robot has assessed an entry** to an area where survivors may be.

21. The robot communicates: "I advise to continue the search phase because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to rescue as many survivors as possible within the time limit

My sensors indicate that this entry is blocked by a big obstacle

I am 95% certain that this entry is blocked by a big obstacle

Clearing this entry will take too much time

If this entry was blocked by a small obstacle, I would have made the autonomous

22. Please elaborate on your **ranking** from the **previous** question *

23. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I advise to continue the search phase because ..."
*

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to rescue as many survivors as possible within the time limit

My sensors indicate that this entry is blocked by a big obstacle

I am 95% certain that this entry is blocked by a big obstacle

Clearing this entry will take too much time

24. The robot communicates: "I am going to clear this entry because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to assess this area

My sensors indicate that this entry is blocked by a small obstacle

I am 95% certain that this entry is blocked by a small obstacle

Entering the area is not possible

If this entry was not blocked, I would have entered the area immediately

25. Please elaborate on your **ranking** from the **previous** question *

26. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I am going to clear this entry because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to assess this area

My sensors indicate that this entry is blocked by a small obstacle

I am 95% certain that this entry is blocked by a small obstacle

Entering the area is not possible

If this entry was not blocked, I would have entered the area immediatel

27. The robot communicates: "I need advice on whether to clear this entry because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to make the best decision for the team

My sensors are unable to accurately predict the obstacle size

I am only 40% certain that this entry is blocked by a small obstacle

I am not confident enough to make an autonomous decision about clearing this entry or continuing the search phase

If I was more than 80% certain that this entry is blocked by a small obstacle, I would

28. Please elaborate on your **ranking** from the **previous** question *

29. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I need advice on whether to clear this entry because ..."
*

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to make the best decision for the team

My sensors are unable to accurately predict the obstacle size

I am only 40% certain that this entry is blocked by a small obstacle

I am not confident enough to make an autonomous decision about clearing this entry or continuing the search phase

Scenario 3 - Area Assessed

Imagine you are a team leader of a Search And Rescue (SAR) team. This SAR consists of a human (you) and a robot with a supporting team role. The robot does the searching and rescuing and you supervise the robot.

During this scenario, **the robot has assessed the safety of an area** where survivors may be.

30. The robot communicates: "I am going to search this area because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

- I want to rescue as many survivors as possible within the time limit
- My sensors indicate that this area is safe to enter
- I am 95% certain this area is safe to enter
- Requesting a different robot is not needed
- If this area was dangerous, I would have requested a different robot

31. Please elaborate on your **ranking** from the **previous** question *

32. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I am going to search this area because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to rescue as many survivors as possible within the time limit

My sensors indicate that this area is safe to enter

I am 95% certain this area is safe to enter

Requesting a different robot is not needed

If this area was dangerous I would have requested a different robot

33. The robot communicates: "I advise request a different robot to search this area because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to rescue as many survivors as possible within the time limit

My sensors indicate that this area is dangerous

I am 95% certain this area is dangerous

Searching this area alone is too risky

If this area was safe I would have searched this area alone

34. Please elaborate on your **ranking** from the **previous** question *

35. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I advise request a different robot to search this area because ..."
*

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to rescue as many survivors as possible within the time limit

My sensors indicate that this area is dangerous

I am 95% certain this area is dangerous

Searching this area alone is too risky

36. The robot communicates: "I need advice on whether to search this area because ..." *

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

I want to make the best decision for the team

My sensors are unable to accurately predict this area's safety

I am only 40% certain that this area is safe to enter

I am not confident enough to make an autonomous decision about searching this area or requesting a different robot

If I was more than 80% certain that this area is safe to enter, I would have made the

37. Please elaborate on your **ranking** from the **previous** question *

38. Would your ranking have **changed** if your **role was the same as the robot** in this scenario?
The robot communicates: "I need advice on whether to search this area because ..."
*

Please **rank** the options as explanations that you would **prefer to achieve the team goal**, where the top option (number 1) is **most** preferred and the bottom option (number 5) being the **least** preferred option.

The robot's role is the **same as before.**

Reminder: the robot's goal is to **search for survivors** and **rescue them** if possible within a time limit. It can make **autonomous** decisions, does **need advice** from the team leader sometimes, and **gives advice** to the team leader as a **supporting role**.

In this scenario, the robot **communicates with you** and not with the team leader.

I want to make the best decision for the team

My sensors are unable to accurately predict this area's safety

I am only 40% certain that this area is safe to enter

I am not confident enough to make an autonomous decision about searching this area or requesting a different robot

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

Microsoft Forms

Appendix J

Pre-study: analysis code

R.B. Kap, Adapting Explanation Style to Team Roles (2022), GitLab repository, `https://gitlab.ewi.tudelft.nl/in5000/ii/matrx/adapting-explanation-style-to-team-roles/-/tree/pre-study`

Appendix K

Pre-study: result - average ranking of explanation type per situation and team role

Situation	**Team Role**	**Best Ranked Explanation Type**	**Mean**	**SD**
Entry	Team leader	Confidence expl.	3.206	1.190
		Contrastive expl.	3.190	0.866
		Counterfactual expl.	2.222	0.921
		Feature attribution	**4.032**	**0.767**
		Goal expl.	2.349	1.190
	Robot	Confidence expl.	2.873	1.123
		Contrastive expl.	3.333	0.738
		Counterfactual expl.	2.222	0.945
		Feature attribution	**3.857**	**0.922**
		Goal expl.	2.714	1.326
Area	Team leader	Confidence expl.	3.556	1.029
		Contrastive expl.	2.952	0.791
		Counterfactual expl.	2.111	0.884
		Feature attribution	**4.222**	**0.725**
		Goal expl.	2.159	1.302
Area	Robot	Confidence expl.	3.397	0.904
		Contrastive expl.	3.000	0.667
		Counterfactual expl.	2.000	0.925
		Feature attribution	**4.270**	**0.867**
		Goal expl.	2.333	1.378
Survivor	Team leader	Confidence expl.	3.206	1.014
		Contrastive expl.	3.397	0.629
		Counterfactual expl.	2.460	0.749
		Feature attribution	**3.937**	**0.807**
		Goal expl.	2.000	1.065
	Robot	Confidence expl.	2.778	0.784

Situation	Team Role	Best Ranked Explanation Type	Mean	SD
		Contrastive expl.	**3.698**	**0.781**
		Counterfactual expl.	2.524	0.779
		Feature attribution	3.635	0.960
		Goal expl.	2.365	1.238
Autonomous action	Team leader	Confidence expl.	3.556	1.013
		Contrastive expl.	2.810	0.786
		Counterfactual expl.	2.444	0.832
		Feature attribution	**3.968**	**0.881**
		Goal expl.	2.222	1.082
	Robot	Confidence expl.	3.222	0.852
		Contrastive expl.	2.968	0.657
		Counterfactual expl.	2.476	0.986
		Feature attribution	**3.714**	**0.973**
		Goal expl.	2.619	1.240
Get advice	Team leader	Confidence expl.	3.111	1.203
		Contrastive expl.	3.127	0.980
		Counterfactual expl.	2.048	1.066
Get advice	Team leader	**Feature attribution**	**4.349**	**0.764**
		Goal expl.	2.365	1.366
	Robot	Confidence expl.	2.905	1.028
		Contrastive expl.	3.286	0.871
		Counterfactual expl.	2.032	0.823
		Feature attribution	**4.206**	**0.986**
		Goal expl.	2.571	1.517
Give advice	Team leader	Confidence expl.	3.302	1.069
		Contrastive expl.	3.603	0.757
		Counterfactual expl.	2.302	0.869
		Feature attribution	**3.873**	**0.922**
		Goal expl.	1.921	1.021
	Robot	Confidence expl.	2.921	0.966
		Contrastive expl.	3.778	0.770
		Counterfactual expl.	2.238	0.938
		Feature attribution	**3.841**	**0.886**
		Goal expl.	2.222	1.102

Table K.1: Best ranked explanation type based on different situations in the data.

Appendix L

Pre-study: result - different ranking questions

Situation	Expl. Type	Method	statistic	Adj. p	Sign.	Effect size
All combined	Confidence	student t-test	1.093	0.281	ns	0.337
	Contrastive	Wilcoxon	188	0.419	ns	0.127
	Counterfactual	Wilcoxon	224	0.940	ns	0.014
	Feature	Wilcoxon	237	0.686	ns	0.064
	Goal	Wilcoxon	189	0.434	ns	0.123
Area	Confidence	student t-test	0.531	0.598	ns	0.164
	Contrastive	Wilcoxon	207	0.737	ns	0.054
	Counterfactual	Wilcoxon	232	0.776	ns	0.046
	Feature	Wilcoxon	201	0.635	ns	0.075
	Goal	Wilcoxon	209	0.777	ns	0.045
Auto. action	Confidence	student t-test	1.154	0.255	ns	0.356
	Contrastive	Wilcoxon	190	0.445	ns	0.120
	Counterfactual	Wilcoxon	218	0.960	ns	0.010
	Feature	Wilcoxon	254	0.392	ns	0.134
	Goal	Wilcoxon	181.5	0.329	ns	0.153
Entry	Confidence	student t-test	0.934	0.356	ns	0.288
	Contrastive	Wilcoxon	199.5	0.602	ns	0.082
	Counterfactual	Wilcoxon	220	1.0	ns	0.002
	Feature	Wilcoxon	239	0.646	ns	0.073
	Goal	Wilcoxon	183.5	0.351	ns	0.146
Get advice	Confidence	student t-test	0.597	0.554	ns	0.184
	Contrastive	Wilcoxon	204	0.685	ns	0.065
	Counterfactual	Wilcoxon	213	0.857	ns	0.030
	Feature	Wilcoxon	232.5	0.763	ns	0.048
	Goal	Wilcoxon	204.5	0.690	ns	0.063
Give advice	Confidence	student t-test	1.212	0.233	ns	0.374
	Contrastive	Wilcoxon	194.5	0.516	ns	0.102
	Counterfactual	Wilcoxon	234.5	0.730	ns	0.055

	Feature	Wilcoxon	224.5	0.929	ns	0.016
	Goal	Wilcoxon	180	0.305	ns	0.160
Survivor	Confidence	student t-test	1.533	0.133	ns	0.473
	Contrastive	Wilcoxon	166	0.165	ns	0.216
	Counterfactual	Wilcoxon	212	0.839	ns	0.033
	Feature	Wilcoxon	260.5	0.315	ns	0.157
	Goal	Wilcoxon	180.5	0.316	ns	0.157

Overview analyses ranking score between the two roles.

Appendix M

User-study: data management plan and informed consent

Plan Overview

A Data Management Plan created using DMPonline

Title: Adapting Explanations to a Team Leader Role in Search and Rescue with Human-Agent Teaming

Creator: Ryan Kap

Affiliation: Delft University of Technology

Template: TU Delft Data Management Plan template (2021)

Project abstract:

The goal of this study is to explore the influence of adapting explanations to a team leader during a search and rescue within a human-agent team. Participants will have to work together with virtual agents (e.g., robots) on a laptop using the platform MATRX (https://matrx-software.com/). The participants will have to perform the task of search and rescue of survivors in a virtual 2D disaster site. The participants will take the role of the team leader, supervising the robots. The robot will explain its behaviour during actions and/or decisions. There will be two groups of participants, one where the robot adapts the explanations it gives to the role of the participant and one where the robot does not adapt its explanations to the role of the participants. After the search and rescue task, participants answer questions about the task and their subjective feelings about the team via an online questionnaire (user satisfaction, perceived trust, perceived user-awareness, perceived transparency, system understandability, situational awareness, and team performance (observability, predictability, directability)).

ID: 97829

Start date: 17-04-2022

End date: 01-08-2022

Last modified: 10-05-2022

Adapting Explanations to a Team Leader Role in Search and Rescue with Human-Agent Teaming

0. Administrative questions

1. Name of data management support staff consulted during the preparation of this plan.

My faculty data steward, Santosh Ilamparuthi, has reviewed this DMP on [date].

2. Date of consultation with support staff.

2022-04-08

I. Data description and collection or re-use of existing data

3. Provide a general description of the type of data you will be working with, including any re-used data:

Type of data	File format(s)	How will data be collected (for re-used data: source and terms of use)?	Purpose of processing	Storage location	Who will have access to the data
Number of survivors saved	.json	via MATRX	Objective evaluation	OneDrive	Ryan Kap + Supervisors
Situational Awareness of participant	.csv	Microsoft Forms	Subjective evaluation	OneDrive	Ryan Kap + Supervisors
Trust between human and agent	.csv	Microsoft Forms	Subjective evaluation	OneDrive	Ryan Kap + Supervisors
System understanding	.csv	Microsoft Forms	Subjective evaluation	OneDrive	Ryan Kap + Supervisors
Perceived user-awareness of the agent	.csv	Microsoft Forms	Subjective evaluation	OneDrive	Ryan Kap + Supervisors
User satisfaction of the agent	.csv	Microsoft Forms	Subjective evaluation	OneDrive	Ryan Kap + Supervisors
Age, Gender, Education Level, Gaming experience	.csv	Microsoft Forms	Demographics	OneDrive	Ryan Kap + Supervisors
Perceived transparency of agent	.csv	Microsoft Forms	Subjective evaluation	OneDrive	Ryan Kap + Supervisors
Team performance	.csv	Microsoft Forms	Subjective evaluation	OneDrive	Ryan Kap + Supervisors
Informed consent	.csv	Microsoft Forms	Informed consent	OneDrive	Ryan Kap + Supervisors

4. How much data storage will you require during the project lifetime?

- < 250 GB

II. Documentation and data quality

5. What documentation will accompany data?

- README file or other documentation explaining how data is organised

III. Storage and backup during research process

6. Where will the data (and code, if applicable) be stored and backed-up during the project lifetime?

- OneDrive
- Git(lab)/subversion repository at TU Delft

The code which runs the search and rescue simulation will be on a GitLab repository from the TU Delft and the questionnaire answers on OneDrive with my TU Delft account

IV. Legal and ethical requirements, codes of conduct

7. Does your research involve human subjects or 3rd party datasets collected from human participants?

- Yes

8A. Will you work with personal data? (information about an identified or identifiable natural person)

If you are not sure which option to select, ask yourFaculty Data Steward for advice. You can also check with the privacy website or contact the privacy team: privacy-tud@tudelft.nl

- Yes

Age, Gender, Education Level, and Gaming Experience

8B. Will you work with any types of confidential or classified data or code as listed below? (tick all that apply)

If you are not sure which option to select, ask yourFaculty Data Steward for advice.

- No, I will not work with any confidential or classified data/code

9. How will ownership of the data and intellectual property rights to the data be managed?

For projects involving commercially-sensitive research or research involving third parties, seek advice of yourFaculty Contract Manager when answering this question. If this is not the case, you can use the example below.

The datasets underlying the published papers will be publicly released following the TU Delft Research Data Framework Policy. During the active phase of research, the project leader from TU Delft will oversee the access rights to data (and other outputs), as well as any requests for access from external parties. They will be released publicly no later than at the time of publication of corresponding research papers.

10. Which personal data will you process? Tick all that apply

- Email addresses and/or other addresses for digital communication
- Other types of personal data - please explain below

- Data collected in Informed Consent form (names and email addresses)
- Gender, date of birth and/or age

Education Level and Gaming Experience

11. Please list the categories of data subjects

The participants are students or ex-students (maximum of 5 years graduated) (age between 18 and 29) from different universities in the Netherlands.

12. Will you be sharing personal data with individuals/organisations outside of the EEA (European Economic Area)?

- No

15. What is the legal ground for personal data processing?

- Informed consent

16. Please describe the informed consent procedure you will follow:

All study participants will be asked for their consent via an online form for taking part in the study and for data processing before the start of the experiment.

Informed Consent

Dear reader,

You are being invited to participate in a research study titled "Adapting Explanations to a Team Leader Role in Search and Rescue with Human-Agent Teaming". This study is being done by master's student Ryan Kap from the Delft University of Technology (https://www.tudelft.nl/en/), under the research group Interactive Intelligence (https://www.tudelft.nl/en/ewi/over-de-faculteit/afdelingen/intelligent-systems/interactive-intelligence).

The purpose of this study is to explore the influence of adapting explanations to a team leader during a search and rescue within a human-agent team. It will take approximately **30** minutes to complete. The data will be used for the master’s thesis project regarding team roles and explanations in human-agent teaming. I will be asking you to answer demographic questions such as your age, gender, education level, and gaming experience. Next, I will ask you to perform the task of search and rescue (SAR) where you rescue as many survivors as possible in a simulated 2D disaster site. You will take the role of the team leader, supervising the robots. Afterwards, you will be asked to complete a questionnaire about your performance, trust in the system/team, system understanding, perceived user awareness, situational awareness, perceived transparency, team performance, and user satisfaction.

As with any online activity, the risk of a breach is always possible. To the best of my ability, your answers in this study will remain confidential. I will minimize any risks by anonymising data where possible (e.g., replacing your name with a randomly generated number) and storing the data in the secured cloud of the TU Delft.

The participants are students or ex-students (maximum of 5 years graduated) (age between 18 and 29) from different universities in the Netherlands.

Search and Rescue operations can be stressful. Even though fictional scenarios will be described with human-agent teaming, you should be aware that there is a small risk of increased stress from reading these scenarios.

Any data will be stored with 4TU.ResearchData for 10 years.

You need to declare that you have read, understood, and agree with this opening statement to continue participation.

If there are any issues or comments, please feel free to ask now or contact me at r.b.kap@student.tudelft.nl

Thank you again!

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICIPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated ________________10/05/2022, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	□	□
3. I understand that taking part in the study involves: taking part in a simulated 2D search and rescue task as a team leader of a team with humans and robots, and then filling out a questionnaire with demographic questions and questions about my perceived experience of the task and team.	□	□
4. I understand that I will **not** be compensated for my participation	□	□
5. I understand that the study will end after approximately **30** minutes or whenever I choose to stop without the need to give a reason.	□	□
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
6. I understand that taking part in the study involves the following risks: light stress. I understand that these will be mitigated by being properly instructed and being reminded that it is a simulated 2D scenario.	□	□
7. I understand that taking part in the study also involves collecting specific personally identifiable information (PII): age range, gender, education level, and gaming experience with the potential risk of my identity being revealed.	□	□
8. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach: anonymise data where possible and secure cloud storage of TU Delft	□	□
9. I understand that personal information collected about me that can identify me, such as age and gender will not be shared beyond the study team.	□	□
10. I understand that the (identifiable) personal data I provide will be destroyed before the end of the year 2022	□	□
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
11. I understand that after the research study the de-identified information I provide will be used for the master's thesis of Ryan Kap	□	□
D: (LONGTERM) DATA STORAGE, ACCESS, AND REUSE		
11. I give permission for the de-identified age, gender, education level, and gaming experience that I provide to be archived in 4TU.research repository so it can be used for future research and learning.	□	□

Signatures
________________________ ________________ _______ Name of participant [printed] Signature Date
I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting. __R. B. Kap________________ ________________ _______ Researcher name [printed] Signature Date
Study contact details for further information: R.B. Kap, r.b.kap@student.tudelft.nl

17. Where will you store the signed consent forms?

- Same storage solutions as explained in question 6

18. Does the processing of the personal data result in a high risk to the data subjects?

If the processing of the personal data results in a high risk to the data subjects, it is required to perform a Data Protection Impact Assessment (DPIA). In order to determine if there is a high risk for the data subjects, please check if any of the options below that are applicable to the processing of the personal data during your research (check all that apply).
If two or more of the options listed below apply, you will have to complete the DPIA. Please get in touch with the privacy team: privacy-tud@tudelft.nl to receive support with DPIA.
If only one of the options listed below applies, your project might need a DPIA. Please get in touch with the privacy team: privacy-tud@tudelft.nl to get advice as to whether DPIA is necessary.
If you have any additional comments, please add them in the box below.

- None of the above applies

22. What will happen with personal research data after the end of the research project?

- Anonymised or aggregated data will be shared with others

The collection of gender, a range of ages (e.g. 18-21, 22-25), education level, and gaming experience can be considered anonymous since they are neither directly nor indirectly identifiable

25. Will your study participants be asked for their consent for data sharing?

- Yes, in consent form - please explain below what you will do with data from participants who did not consent to data sharing

If they do not consent to anonymised data sharing, I will stop the experiment, erase any data created during the session, and they are free to go.

V. Data sharing and long-term preservation

27. Apart from personal data mentioned in question 22, will any other data be publicly shared?

- All other non-personal data (and code) produced in the project
- All other non-personal data (and code) underlying published articles / reports / theses

29. How will you share research data (and code), including the one mentioned in question 22?

- All anonymised or aggregated data, and/or all other non-personal data will be uploaded to 4TU.ResearchData with public access
- I will share my data and code via git(lab)/subversion and also create a snapshot in a repository

30. How much of your data will be shared in a research data repository?

- < 100 GB

31. When will the data (or code) be shared?

- At the end of the research project

32. Under what licence will be the data/code released?

- MIT License
- CC BY

VI. Data management responsibilities and resources

33. Is TU Delft the lead institution for this project?

- Yes, the only institution involved

34. If you leave TU Delft (or are unavailable), who is going to be responsible for the data resulting from this project?

dr. M.L. Tielman (m.l.tielman@tudelft.nl)

35. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

4TU.ResearchData is able to archive 1TB of data per researcher per year free of charge for all TU Delft researchers. We do not expect to exceed this and therefore there are no additional costs of long term preservation.

Appendix N

User-study: questionnaire

Adapting explanations to the Team Leader in a Search and Rescue scenario - Questionnaire: `https://forms.office.com/r/SmThkH5gZc`

Adapting Explanations to the Team Leader in a Search and Rescue Scenario - Questionnaire

The survey will take approximately 5 minutes to complete.

* Required

* This form will record your name, please fill your name.

User Data

1. Did you fill out the informed consent form given by the instructor? If not, please do so and return to this question *

   ○ Yes

2. What is your anonymised ID? *

3. Which group are you in? *

○ Group 1

○ Group 2

4. What is your current age? *

○ 18-21

○ 22-25

○ 25-28

5. What is your gender? *

○ Woman

○ Non-binary

○ Man

○ Prefer not to say

6. What is your education level (most recent completed) *

○ Secondary education (High school)

○ Bachelor's or equivalent

○ Master's or equivalent

○ Doctorate or equivalent

7. What is your gaming experience? *

○ No experience

○ A little

○ Average

○ A lot

The Search and Rescue Task

Please, before continuing ask the supervisor to show you the task.

Situational Awareness

8. Where is the robot currently located? (e.g., (5, 6)) Part 1: X coordinate *

9. Where is the robot currently located? (e.g., (5, 6)) Part 2: Y coordinate *

10. Which action is the robot currently executing? *

○ Removing an obstacle

○ Searching the disaster site

○ Searching an area

○ Carrying a survivor

○ Clearing an entry

11. Which action will the robot perform next? *

○ Removing an obstacle

○ Searching the disaster site

○ Searching an area

○ Carrying a survivor

○ Clearing an entry

12. How many survivor did the robot find so far? *

○ 0

○ 1

○ 2

○ 3

○ 4

○ 5

○ 6

○ 7

○ 8

○ 9

○ 10

13. How many obstacles (not entries to areas) did the robot find so far? *

○ 0

○ 1

○ 2

○ 3

○ 4

○ 5

○ 6

○ 7

○ 8

○ 9

○ 10

14. How many entries did the robot find so far? *

○ 0

○ 1

○ 2

○ 3

○ 4

○ 5

○ 6

○ 7

○ 8

○ 9

○ 10

15. How many areas did the robot find so far? *

○ 0

○ 1

○ 2

○ 3

○ 4

○ 5

○ 6

○ 7

○ 8

○ 9

○ 10

Continue the task

Please, before continuing ask the supervisor to show you the task.

Perceived trust

16. Please fill out how you feel about the statements. *

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree
I am confident in the robots. I feel that they work well	○	○	○	○	○	○
The outputs of the robots are very predictable	○	○	○	○	○	○
The robots are very reliable. I can count on them to be correct all the time	○	○	○	○	○	○
I feel safe that when I rely on the robots I will get the right answers	○	○	○	○	○	○
The robots are efficient in that they work very quickly	○	○	○	○	○	○
I am wary of the robots	○	○	○	○	○	○
	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree
The robots can perform the task						

the task better than a novice human user	○	○	○	○	○	○
I like using the robots for decision making	○	○	○	○	○	○

17. Trust in Robot *

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree
I trusted the robot to do the right thing at the right time	○	○	○	○	○	○
The robot was trustworthy	○	○	○	○	○	○

Collaborative Fluency

18. Human-Robot Fluency *

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree
The human-robot team worked fluently together	○	○	○	○	○	○
The human-robot team's fluency improved over time	○	○	○	○	○	○
The robot contributed to the fluency of the interaction	○	○	○	○	○	○

Perceived User-awareness

19. Question *

	Strongly Disagree	Disagree	Slightly Disagreee	Neutral	Slightly Agree	Agree
I feel the robot is user aware	○	○	○	○	○	○
I feel the robot gave personalised explanations	○	○	○	○	○	○

Perceived Understandability

20. Question *

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree
I know what will happen the next time I use the system because I understand how it behaves.	○	○	○	○	○	○
I understand how the system will assist me with decisions I have to make.	○	○	○	○	○	○
Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.	○	○	○	○	○	○
It is easy to follow what the system does.	○	○	○	○	○	○
I recognize what I should do to get the	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree

do to get the advice I need from the system the next time I use it.	○	○	○	○	○	○

User Satisfaction

21. User satisfaction of the explanations given by the robot *

	Strongly disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree
I liked the explanations given by the robot	○	○	○	○	○	○
The explanations given by the robot were satisfactory	○	○	○	○	○	○

Open Questions

22. Which behaviour of the robot influenced your answers? *

23. How much did you read the explanations during the experiment? *

○ 0-10%

○ 11-20%

○ 21-30%

○ 31-40%

○ 41-50%

○ 51-60%

○ 61-70%

○ 71-80%

○ 81-90%

○ 91-100%

24. Anything you want to share about the experiment or robot or explanations? (good or bad) *

Appendix O

Tutorial text

Hello! Welcome to the tutorial! Let me introduce the team to you. I am the Search Robot () of this team, the Rescue Robot () of this team is next to me, and you are the Team Leader (). Your job is to supervise the team to achieve the team goal: rescue as many survivors as possible within the time limit. The robots have a supporting role within your team in trying to achieve its goal. We are currently in a small version of the actual experiment. Here, we will explain how the experiment will go down and how you can interact with me during the task. In this chat box, team members (you and me) will communicate with each other. Whenever I say something to you, I will also inform you of my coordinates [x, y] (x: from left to right, y: from top to bottom). If you understand, please press the 'I understand' button to continue.

On the left side of your screen, you see several tiles and/or items within a large black square. This is called the "Disaster Site" and its size is 30 by 30 tiles. Inside the Disaster Site, different areas can be found to explore. Right now, we are inside the "Safety Zone" (in the top left corner of the Disaster Site). The Safety Zone can be recognised by the green tiles, which are surrounded by walls (). This is our starting point and also the zone where we bring in any survivors that we will find and rescue. The rest of the Disaster Site will contain areas, which will have an entry and which will be surrounded by walls (). These areas can either be dangerous () or safe to explore (). In addition, the entries to these areas can be clear to enter () or they can be blocked by obstacles (). These obstacles () could also be blocking pathways between the areas in the Disaster Site. Lastly, survivors in the Disaster Site can be healthy (), injured (), or stuck (). If you understand, please press the 'I understand' button to continue.

There are 2 phases during this Search and Rescue task: the Search Phase and the Rescue Phase. Both phases will have their own timer (in seconds) in which you carry out different tasks. During the Search Phase, only I () will be able to explore the Disaster Site. I am able to explore the Disaster Site, remove obstacles (), clear blocked entries (), explore safe areas (), and rescue healthy survivors (). At the start of the Rescue Phase, you will have to make a Rescue Plan using the information you gathered during the Search Phase and then execute it. You will be given a list where you can drag and drop items on the Rescue Plan to make an order you deem most suitable for execution. You will see a Rescue Plan later in the tutorial. I () will leave the Disaster Site at the start of the Rescue Phase and a new Rescue Robot 1 () will appear. Rescue Robot 2 () ,that is currently standing next to me, will only go out into the Disaster Site if requested by that new Rescue Robot 1 (). Rescue Robot 1 () can carry (critically) injured people (), explore dangerous areas (), and free stuck survivors () with the help of the other Rescue Robot 2 () or alone but at the cost of a penalty. If you understand, please press the 'I understand' button to continue.

If I () come across something while I am exploring the disaster site, I (as a robot) can do three things: make an autonomous decision, give you some advice, or ask for your advice (when I am not certain (enough) of something). Below the chat box, you can see two boxes ‘Follow Advice?’ and ‘Give Advice’. In these boxes, options will appear given a certain situation. When you click on them you will communicate your decision to me and I will act accordingly. If you understand, please press the 'I understand' button to continue.

There is also a timer in the top left corner of the screen (below the matrx logo) which will tell you how much time you have left during the current phase. During this tutorial, the timer will do nothing, but during the actual experiment it will matter! If you see a red explanation mark () on top of my image, it will mean that I am busy performing an action. Last advice: during the actual experiment (as well as during this tutorial) it is important to constantly look at the chat box whenever something is happening and before you give or take advice, since the robots will always explain their behaviour during events. Keeping an overview between the map and the chat box is your responsibility as Team Leader ()! If you want to reread some information, you can scroll through the chat box before starting the tutorial. If you understand, please press the 'I understand' button. Once you press 'I understand', the search phase of this tutorial will start!

Appendix P

User-study: analysis code

R.B. Kap, Adapting Explanation Style to Team Roles (2022), GitLab repository,
`https://gitlab.ewi.tudelft.nl/in5000/ii/matrx/adapting-explanation-style-to-team-roles/-/tree/user-study`