Using explanations to improve subjective experience of control in situations of delayed control effects

Huishun Zhang

Responsible Professor: Mark Neerincx
Supervisors: Myrthe Tielman, Ruben Verhagen

To obtain the degree of Master of Science in Computer Science at the Delft University of Technology to be defended on Monday August 29, 2022 at 14:00

Delft University of Technology
Faculty EEMCS
MSc Computer Science
August 17, 2022

Abstract

Problem Statement. The responsibility problem in the AI field has been taken seriously. The point of the problem is how to prevent AI systems from making moral decisions for whom they cannot be held accountable. One of the directions to address the responsibility problem is Meaningful Human Control (MHC). Most of the work focused on theoretical definition and measurement explorations and few researchers investigate it by combining application scenarios. **Research Question.** One of the previous works identified that explanations could be used to improve the subjective feeling of MHC when the effect of control is delayed. Therefore, we wanted to study what type of explanations could facilitate the improvement and how explanations achieved that. **Method.** We conducted an expert study to obtain advice on selecting explanation types by a ranking question questionnaire and structured questions. Due to the findings that information sharing could improve Situation Awareness (SA), we conducted a pilot study and a user study to study the effect of explanations on different levels of SA and the human feeling of MHC. We used the Situation Awareness Global Assessment Technique (SAGAT) and a Five-point Likert scale questionnaire to measure the effect of explanations on SA and MHC. Moreover, we compared the effect of sub-explanations we used in the user study (consequential and counterfactual explanations) with a questionnaire. **Results.** It was shown that the explanations help to get more overall SA and SA in the projection level. It was also shown that higher SA scores are associated with a better feeling of MHC. However, our result showed that the explanations have no significant effect on the subjective experience of MHC. The findings also indicated that a high frequency of playing computer games can result in a good subjective experience of control. Comparisons of the sub-explanations on qualitative and quantitative analysis demonstrated that counterfactual explanations made a better impression on participants in most respects. **Discussion and Conclusion.** From the results, we concluded that explanations do increase the degree of SA of the task to a certain extent, but they do not affect the experienced control. We also found that the computer gaming experience may provide higher cooperation engagement and cohesion, resulting in increased experienced control. As for the sub-explanations, the counterfactual ones were overall better than the consequential ones by providing more information and making the participants feel the robot's intelligence.

Dedication

I want to express my sincere gratitude to my daily supervisor Myrthe Tielman. Thanks for urging me to plan my dissertation from the start. Thanks for listening to my immature designs in every meeting and encouraging me to keep going. Besides that, I would like to thank Ruben Verhagen. Thanks for introducing the research topic in the beginning. Every time I got stuck during the research, he gave excellent advice. When I didn't clearly understand what the other supervisor suggested, he re-explained it to me. For the thesis document, he also gave many advice and grammar check. I also would like to thank Prof. Mark Neerincx for his valuable advice on my research. Additionally, I would like to thank Xucong Zhang for taking the time and effort of being part of the thesis committee. Moreover, I would like to thank the expert who participated in my study and gave useful suggestions.

Next, I want to thank my family and friends. Thanks to my parents supporting me all the time. I couldn't finish my study without you. Thanks to those friends who volunteer to participate in my experiment. Thanks to Xiaohan and Dawei for being with me every time when I am down. This paper is the result of the efforts of all of the above, thanks to all.

Contents

1

Introduction

Artificial Intelligence (AI) is increasingly employed in various industries and businesses. Gartner Survey shows 37 percent of organizations have implemented AI in some form in 2019, which grew 270 percent in the past four years [38]. Also in daily life, AI systems detect their surroundings, interpret what they see, solve issues, and take action to assist in the completion of tasks, making life simpler. AI-powered voice assistants are already used by 97 percent of mobile users [20]. However, with the rapid development of AI at the same time there are many problems, such as human trust, data privacy, and security. One of the biggest problems is the problem of responsibility.

More and more AI agents are used to solve tasks that involve moral decisions, like autonomous driving [24] and autonomous weapon systems [8]. There may be problems where the attribution of responsibilities in the automation system is not clear yet at the legal and moral level, which is called the "responsibility gap" [45]. Matthias [73] first raised this problem in 2004 and the issue has received considerable critical attention these years [45, 16]. It has been reported that the manufacturer or operator cannot be responsible for the machines if they cannot predict the action of the machines. If we want to use these machines, we have to face the "responsibility gap". Therefore, there is an urgent need to find a way to address the responsibility gap in moral practice and legislation. Gunkel [45] gave three directions to close or remediate the gap. The first one is Instrumentalism 2.0, which suggests that only human beings should possess rights and responsibilities. The second one, Machine Ethics, holds the view that the machine should be capable of making ethical determinations. The third one, which is called hybrid responsibility, utilizes a mixed strategy in the middle of the two above.

More practical approaches continue to be explored along the lines of the theoretical ideas above. The theory of Artificial Moral Agents (AMA) is a continuation of the idea of Machine Ethics. Cervantes et al. [21] gave the definition "an AMA is a virtual agent (software) or physical agent (robot) capable of engaging in moral behavior or at least of avoiding immoral behavior". So it is required to consider how to introduce morality into the agent. Allen et al. [1] provided two strategies to develop AMA, which are *top-down strategy* and *bottom-up strategy*. Agents based on top-down strategies include ethical guidelines that are often developed from a certain ethical framework. This ethical framework is essentially the AMA's ethical decision-making norm. Bottom-up ethical agents, on the other hand, do not impose a set of ethical principles drawn from a certain ethical theory. Instead, they have a learning strategy by rewarding appropriate decisions. However, it was also mentioned that the bottom-up strategy is more

difficult to achieve because the goal of learning, i.e., "justice" or "ethics", is difficult to determine. In contrast, the implementation of a top-down strategy would be more realistic.

Sio and van den Hoven [27] gave a solution to the responsibility gap philosophically by raising Meaningful Human Control (MHC), and the core of this idea is the same as the top-down strategy. They argued that humans should ultimately remain in control of the whole system. They concluded the idea into two aspects: "decision-making systems should *track* human-moral reasons" and "the actions of the system should *trace* back to humans who have a proper moral understanding of the system". This way, the decision-making agents could follow moral guidelines, which are defined by humans.

However, the concept was rather theoretical and just provided a research direction. Many following studies were based on this theory for further theoretical definition and measurement explorations [25, 30, 37]. After these theoretical attempts, some research also discussed and applied this concept under certain Human-Agent Teamwork (HAT) context practically [50, 19, 97]. In order to operationalize the concept of MHC and achieve a measurable notion, van der Waa et al. proposed a measurement of MHC in HAT and performed simulation experiments [97]. They provided three measurable components: the subjective experience of human control in HAT, the behavioral compliance with ethical guidelines, and the behavioral compliance with moral values. In their experiment investigating the effects of different Team Design Patterns (TDP) on MHC, they found that the experience of human control depends on how immediate the observed effects were. In their experiment, the participants had poor feedback on the situation that the reaction from the agent is not timely after human control. Therefore, the researchers claimed that when the effects of control are delayed, the agents can explain the consequences of the exercised control. We propose that such explanations can improve the human awareness of the future states of the team environment, which is the projection level of Situation Awareness (SA).

SA was interpreted as three levels: perception, comprehension, and projection. The perception level contains the elements in the environment in a certain time and space, the comprehension level includes the meaning of those elements, and the projection part is presented as people being able to project the near future actions of the elements in the environment [32]. To clarify the relationship between SA and MHC, Calvert et al. claimed that SA is not part of control but a key aspect of attaining MHC [16]. Combining this relationship and the experiment result of van der Waa et al. [97], we hypothesize that the delayed effects of control might negatively affect the projection of future events and states, which is part of SA, and then reduce the MHC feeling. Some research showed that information sharing can improve the SA of individuals in a team or system [91, 57]. From the perspective of humans, intuitively the explanation from agents is a type of information sharing. Lewis [67] also defined explanation: "It is a quantity of information about that event's causes". Therefore, we propose that explanations could help humans get more SA. Also, in the experiment of van der Waa et al. [97], they argued that explanations of the consequences of

exercised control can improve the subjective experience of MHC.

Based on all this, we believe that when the effects of control are delayed, explaining the effects of exercised control of the agents (such as explanations of future states and events) will improve the projection level of SA, and also the subjective experience of MHC. The goal of this work is to find a type of explanation that could accomplish this improvement. Therefore, the main research question is:

How can we use explanations to improve the subjective experience of meaningful human control in delayed control situations?

And related sub-questions are:

- *Why and when do humans need explanations in this context?*
- *What do explanations that could facilitate the improvement look like?*

The rest part of the report is structured as follows: First, the related work will be shown in Chapter 2. All the related concepts and background knowledge are explained and given. In Chapter 3 Expert Study part, we investigated what kinds of explanations are beneficial for humans to get a better experience of control by asking the expert from the USAR field. We were also concerned with improving the rationality of the background knowledge of the USAR task. In Chapter 4 Experimental Study, we provided a simulated USAR scenario with delayed control situations developed with MATRX in the experimental study. We also investigated how the explanations affect the experience of MHC and the SA of humans. In Chapter 5, the experimental results are presented, including all the output from quantitative and qualitative analysis. Those results are further discussed in Chapter 6. In the discussion, we will explore what the results represent, the possible reasons for such results, and our limitations and future work, which will lead to the conclusion in Chapter 7.

2

Related Work

2.1 Human-Agent Teamwork (HAT)

2.1.1 Teamwork

Teamwork is the collaborative effort to achieve common performance goals [36]. Researchers from several fields have defined "team" in various ways. Katzenbach and Smith [59] defined the team as "a small number of people with complementary skills who are committed to a common purpose, set of performance goals, and working approach for which they hold themselves mutually accountable". They separated teams from work groups, which frequently lack important performance requirements, real interdependence, and shared responsibility. Cohen and Levesque [22] raised the joint intention theory in 1991, in which they explained teamwork as "a joint activity performed by individuals sharing certain specific mental properties". Furthermore, they gave a more operational definition of the team as "a set of agents having a shared objective and a shared mental state" [23]. The definition is based on the notion of joint intention, which requires an agent to commit to telling other team members anytime it detects that the shared objective has already been achieved, has become unattainable, or has become irrelevant.

Lenoxt et al. [66] gave the characteristics of successful teams, including self-awareness, within-team interdependence, feedback, performance monitoring, clear communication of intentions, and assisting other team members when necessary. These factors are not only applicable to human teamwork but also to human-agent teamwork. In our work, we also applied some of them, such as clear communication of intentions.

2.1.2 Human-Agent Teamwork

With the development of technology, agents such as software and robots are increasingly involved in cooperation with humans [12]. Researchers began to think about whether this form of collaboration met the requirements of teamwork and what role the agents played in teamwork. According to Sycara and Lewis [86], agents interacting with people have three key roles: supporting individual team members in the completion of their own tasks, supporting the team as a whole, and being equal team members.

Software in HAT

Van der Waa and Haije [95] developed the Man Agent Teaming Rapid eXperimentation Software package (MATRX) in 2019. As it facilitates the construction of activities that involve cooperation, it enables fast trial of novel HAT concepts. As Haije concluded, the architecture of MATRX consists of four components: the GridWorld, the Objects, the Agents, and Scenarios [46]. The GridWorld is presented as a 2D environment and the objects are items that can be placed in the environment. Agents are autonomous entities that can perceive and act on the environment and a scenario can be seen as a running environment after configuration.

After the release of MATRX, many researchers have used this package to complete HAT experiments. In the experiment of Haije [46], military unmanned aerial vehicles (Agents) need to fly to a specific area (one part in the GridWorld), perform reconnaissance, and return to their starting point. The agent must learn how to discover a way to the reconnaissance site in a variety of scenarios that best approximate the person anticipated to accomplish that mission. Since MATRX is easy to set up, Haije could change the context variables and constraints to test all conditions quickly. Moreover, van der Waa et al. [97] used MATRX to perform a simulation experiment of a triage task to test the effect on MHC with different Team Design Patterns (TDP) [98]. They built a 2D top-down environment and enabled communication between the users and the triage robot. We also adopt MATRX to develop our experiment environment.

Human-Agent Interaction

Intelligent agents are more skilled in numerical or symbolic processing (e.g., information fusion, data mining), whereas humans are experts in cognitive computing, such as anomaly or emergency handling, multiple meta-level reflections, and integrated situation assessment [36]. As humans and agents specialize in different areas, and some of the functions even complement each other, it allows the human-agent team to achieve better performance under proper cooperation and interaction [51].

As agents become more sophisticated and autonomous, the challenge of accomplishing a proper human-agent interaction increases. For example, the activities appearing in human interactions with basic teleoperated robotic platforms are no more than controlling the robots from one location to another one. The destination and, more significantly, the reasons for the journey remain entirely in the operator's head, who maintains continuous manipulation of the platform. However, when more deductions and navigation about how to accomplish the task are assigned to the robot itself, meanwhile the operator only provides occasional supervisory feedback, there is a greater difficulty to maintain a smooth interaction between the human and the robot [12]. Bradshaw et al. [12] also mentioned that coordination in HAT cannot take place without a sufficient basis for shared situation awareness, and this need for situation awareness increases as the degree of autonomy of agents increases.

2.1.3 Situation Awareness

Situation awareness (SA) plays a significant role here to facilitate the collaborative process in HAT [55]. In terms of the individuals in the human-AI system, SA is the perceptual and cognitive framework of the individual human operators working within the human-AI system. SA is a term derived from aviation psychology to define the components of tactical flight operations including the pilot's comprehension [32]. The definition of SA is interpreted by different researchers, depending on how one goes about using it. As Durso and Gronlund [29] concluded, at the most general level, scholars have used the term SA to refer to the cognitive tasks required to function in or control a dynamic environment. Sarter and Woods viewed SA as "a variety of cognitive processing activities" in a broad way [88].

However, most researchers approached SA from a more detailed view. Vidulich et al. [99] gave a more concrete definition as: "Continuous extraction of environmental information, integration of this information with previous knowledge to form a coherent mental picture, and the use of that picture in directing further perception and anticipating future events". Further, according to Endsley, SA was defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future". She also divided SA into three levels according to the definitions above, which are perception, comprehension, and projection. Among all definitions, we adopt the one defined by Endsley in further discussion.

Situation Awareness Measurement

This information processing-based three-level model is generic and provides an intuitive description of SA, but the most important part is the simplicity and division of the SA into three layers, which allows it to be measured easily and efficiently [87]. Endsley raised a corresponding measurement, Situation Awareness Global Assessment Technique (SAGAT), while she developed the three-level SA model. One of the methods to access operator SA is freeze probe techniques [33]. Typically, a task is frozen at random, all displays and screens are blanked, and a series of SA questions about the present state at the moment of the freeze is administered. Participants must respond to each question using their knowledge and comprehension of the situation at the time of the freeze. At the end of the trial, participant answers are compared to the ground truth state of the environment at the time of the freeze, and an overall SA score is computed. The main advantage is their supposedly direct and objective nature, which eliminates the problems associated with collecting SA data after the experiment. However, these methods are also known for their degree of interference with mission performance (i.e., during mission freeze). Real-time probe approaches were developed to avoid the high amount of task interference imposed by freeze probe techniques. They entail the verbal administration of SA-related inquiries online, but with no freezing of the task under analysis. Self-rating procedures

involve the use of rating scales to obtain subjective judgments of participant SA. They indicate how aware participants considered themselves to be while doing a task.

The methods used today for SA measurement can be broadly divided into two categories: objective and subjective measurements. The method that includes SAGAT and its variants is called objective measurement, which is a highly sensitive, reliable, and predictable SA measurement that can be used in a variety of fields and experimental settings [34]. Subjective opinions on SA are easy to collect in a wide variety of settings. Behaviorally-based subjective rating scales such as SABARS may provide a more suitable basis for ongoing observer ratings of SA [11]. In the latest studies, SA has also been measured by physiological indicators, but for now, it is only in the first stage [11]. In our study, we choose objective methods to measure the SA because of its quantifiability.

In the measurement of SA, researchers are also concerned about the relationship between SA and task load [65, 61]. In the study by Lee et al. [65]. they analyzed the relationship between air controllers' SA and their workload at three air traffic levels. They concluded that SA and workload are inversely related and more closely correlated when task load increases. Bolstad and Endsley [10] also found that their shared displays can help to obtain more SA at high task loads. It was shown that SA changed more significantly at high task loads for the same conditions. Therefore, we will to use a task with a higher task load, search and rescue, as our scenario, since it is always under time pressure [92].

2.2 Search and Rescue

Search and Rescue (SAR) has multiple definitions, one of which is given by the United States Defense Department. As the name of SAR suggests, it consists of two parts, Search and Rescue. Where Search means "An operation using available personnel and facilities to locate persons in distress", and Rescue means "An operation to retrieve persons in distress, provide for their initial medical or other needs and deliver them to a place of safety" [78]. In the following subsections, we will introduce how these two operations have evolved and how they are now integrated to serve people with the help of robots. The SAR tasks include many specialty sub-fields determined by the type of terrain the search is conducted over. These include mountain SAR, urban SAR (USAR), cave SAR, Maritime SAR, etc. In the next sections, we cover only the USAR as our target task.

2.2.1 Search

An effective and efficient search theory is essential to the whole SAR task. According to the report of the U.S. Coast Guard [94], in addition to those SAR missions where the location is directly known, the process of determining

the location in the remaining missions adds a great deal of uncertainty and complexity to the search and rescue mission.

Koopman [60] first raised some math models about searching theory, which are used in the U.S. Navy's searching strategy in World War II. The basic principles of search theory "classical search planning method" (CSPM) were first applied to SAR planning around 1957 when the U.S. Coast Guard published its search planning doctrine in a search and rescue manual. Because of the limitations of computers at that time, the computation of the searching was simplified to be feasible for hand calculation. Later, after continuous modifications, this principle was able to be applied to more complex scenarios. The first implementation of computer support for search programs occurred around 1970, when Richardson and Stone [84] successfully used a computer to calculate the distribution of the Scorpion submarine. They first got a prior probability distribution using Monte-Carlo procedures calculated by a computer. Then they updated the posterior probability distribution by unsuccessful search results.

USAR scenarios include, but are not limited to, natural disasters such as earthquakes and tsunamis and man-made disasters such as explosions and fires [28]. Indoor search is a large part of these scenarios and the search strategies are different, so indoor search theory is also what we need to consider. Traditional search strategies like Deep First Search (DFS) and Breadth First Search (BFS) can be applied in this scenario in the area of single-person search [72]. When a rescuer adopts the BFS strategy in an operation, he will explore the neighbors first, before moving to the next level neighbors. So it can be used to find the shortest path to the target, which is also verified in [72]. DFS strategy starts from the start point and explores as far as possible along each path before backtracking. It can cover a larger search area compared to BFS. For multiple-person strategies, Nguyen et al. [80] considered the firefighter's route through the building as an optimization problem and set modeling based on the coordination between the firefighters.

2.2.2 Rescue

Rescue in USAR is a time-demanding task. According to research, the majority of survivors of earthquake-induced building collapses are rescued during the first 24 hours of the incident [56]. In the 1980 earthquake in Southern Italy, for example, 94 percent of individuals were saved within the first 24 hours [26]. It[69] is also claimed that the first 72 hours after the earthquake are the golden hours to rescue. For fire, rescue corresponding time is more important. According to a report by [42], on the one hand, the size of fire becomes exponentially larger with time, on the other hand, the data proves that people's survival rate decreases rapidly after 20 minutes, and if the response time can be reduced by five minutes, the number of deaths due to fire in the UK can be reduced by seven percent every year. In section 3.1.2, we considered this time-demanding factor in our design.

2.2.3 SAR as HAT

SAR is never done by humans alone. Humans usually cooperate with other agents, such as dogs, drones, and robots, and work as a human-agent team. Definitions and concepts related to HAT can be found in section 2.1.

SAR with Dogs

It can be traced to 200 years ago when humans first used dogs to help with SAR [15]. The employment of dogs in SAR operations takes advantage of dogs' acute sense of smell. The scent is more intense where it originates (i.e., the victim). As the scent spreads, it gets less intense, generating a scent cone [15]. This way, the dog can gradually find more dense places according to the scent and finally find the survivors. Further research also combined dogs with other advanced technology equipment to enhance the effectiveness of searching. Zeagler et al. [103] raised wearable and mobile interfaces to help enable SAR dog and handler teams to work together more effectively.

However, the inherent disadvantages of dogs cannot be avoided. First, dogs that can engage in SAR need to pass 16-18 months of training, which is a long time and does not guarantee good training [15]. Secondly, the dog's sense of smell is greatly affected by the environmental weather and so on, so there will be limitations in the search terrain [41]. Drones and robots can avoid these two drawbacks well.

SAR with Drones

The first drones were used for SAR to replace search helicopters. Its advantages are that the operator does not need any SAR piloting experience, drones are less affected by weather conditions and the drones can cruise automatically [40]. Karaca et al. also compared searching by drones to searching on foot and by snowmobile. The results showed that when compared to the traditional approach, drones could search a larger region faster [58]. The workflow of the search is that these drones can monitor the situation from the air using a variety of sensors and equipment, and then forward the information collected to operators at ground stations for further action. The search technology of drones mostly concerns how to identify objects. There are several broad categories: thermal imaging to distinguish living and non-living things, computer vision for object recognition, and the use of cell phone signals to locate humans in cities [44].

SAR with Robots

The main purpose of using SAR robots is to protect rescuers in a dangerous and uncertain environment after a disaster. Moreover, the robot can explore more areas that are inaccessible to humans due to its design [77]. The use of the robot can be divided into four types: search robots, extraction robots, evacuation robots, and field treatment robots [101]. Search robots have the highest technology maturity among these types, they will try to locate and report the

location of any injured person. For extraction robots, they will carry the injured person out of the disaster zone. Then, injured people will be evacuated to a medical assistance location by evacuation robots, and then taken care of by field treatment robots.

According to our experiment design part in Chapter 4, we will discuss extraction robots. One of the first applied extraction robots is PackBot [102] developed in 2004. It was a modification of Unmanned Ground Vehicles (UGV), adding a flexible stretcher. When the operator remotely maneuvers the robot to the injured person, the injured person rolls onto the stretcher and is taken by the robot to a safe area. However, the feasibility of this design depends on whether the injured person or someone at hand is capable of securing the injured person within the stretcher. Later robots, such as Battlefield Extraction-Assist Robot (BEAR) [6], used robotic arms to avoid this problem. BEAR could reach its arms under the injured person and hold him to a safe area. The disadvantage of such robots with precisely operated robotic arms is that they are highly dependent on communication effects, which are usually not guaranteed in post-disaster scenarios. Furthermore, it is not safe enough to work in complex and unstructured environments [101].

The autonomy of robots necessarily needs to be improved if we want them to be more adaptable to complex and unstructured environments [9]. For example, Sun et al. [93] built an ontology model that enables SAR robots to understand how to make intelligent decisions. The robots can infer the task to be performed based on the state of the environment while obtaining semantic information about the victim.

2.2.4 USAR as HAT

Rescue robotics has been cited by the DARPA/NSF study on human-robot interaction as one of only two "Grand Challenge" applications [85]. Murphy et al. have been working in this area since 2002 and have presented many theoretical foundations and practical experiences [18, 77]. They summarized the robotics applications in USAR and the issues that arise in the study in 2004.

They divided USAR into six groups, such as search, rescue, medical, etc. Moreover, they said that the USAR physical and working environment consists of three zones: the hot zones, the warm zones, and the cold zones. The Hot Zone is the area of actual devastation and the Warm Zone is the surrounding area where the rescuers assemble and prepare their equipment. The Cold Zone is for the Incident Command headquarters, the press, and the media liaison. The entire USAR process can be summarized as follows: When an incident occurs, the first responders assume control of the site and establish the Hot, Warm, and Cold Zones. They will protect from a second terrorist attack or an aftershock and assess the site for additional safety risks. The first activity is to search the Hot Zone by the search team. If survivors are found, The Rescue team would be dispatched to extricate the victim based on optimal use of resources. After finding a survivor, the Search team is likely to continue searching.

In these subtasks, robots were only used in the searching part due to the

technological limitation by then. However, Liu and Nejat [70] have already mentioned the use of rescue robots which use simultaneous localization and mapping (SLAM) based algorithms in 2013. Moreover, one robot needs to be operated by two characters: the *robot operator* and *the problem holder*. The actions performed by the *robot operator* are navigation-related and the *problem holder* is expected to detect victims and maintain an understanding of the relevant state of the world as opposed to the state of the robot.

As for the issues that need to be solved, they mentioned that reducing the human-robot ratio is an important one. One proposed solution is to transfer sections of both roles to the robot so that it may be performed by one actor. This setting is also used in our subsequent designs. The solution also relies on a higher level of autonomy. However, as robots become more autonomous, the ethical issues that arise from them deserve our consideration.

2.2.5 Ethics in SAR

Rescue activities in disaster zones can be plagued with ethical problems. The usage of robots will certainly add new ethical issues. We will only discuss those issues raised by using robots in SAR. Battistuzzi, Recchiuto, and Sgorbissa [7] gave a review about ethical concerns in rescue robotics in 2021. They divided the ethical issues into seven categories: fairness and discrimination, false or excessive expectations, labor replacement, privacy, responsibility, safety, and trust. Among all these issues, we elaborate on *fairness and discrimination*, and *responsibility* as they are closely related to our work.

Fairness and Discrimination

Fairness and discrimination issues are unavoidable since search and rescue tasks include the order in which victims are rescued. Amigoni and Schiaffonati [3] mentioned fairness in their ethical framework for robot systems. They said that benefits and risks should be allocated equitably among the subjects involved to eliminate the possibility of certain subjects experiencing only risks while others enjoy only benefits. They also gave an example for the claim in the SAR task: The robots may make decisions about prioritizing the order in which the detected victims are reported to the human rescuers or about which detected victim it should try to transport first. Brandão [13] proposed a new fairness-aware method for coverage path planning, which mainly focuses on the speed and order of covering social groups. In their previous work [14], they provided a practical illustration of the problem. The authors describe hypothetical scenarios in which drones are used to search for victims and deliver medications after a disaster. Because the distribution of urban residents is not uniform in terms of density, age, race, and gender, the authors went on to state that planned drone routes would have a skewed distribution of these characteristics. It will particularly target young people, who are more likely to survive than elderly individuals residing in other areas. As a result, while the drone will be effective in discovering as many individuals as possible, it will not adhere to the principle

of distributive justice. We will cover solutions to fairness and discrimination in the section 3.1.2.

Responsibility

The assignment of responsibility for human-robot cooperation in SAR is tricky, especially when it comes to the ethical part. Harbers et al. [49] focused on responsibility assignment problems, which can arise when robots act with no human supervision. If the robot malfunctions, makes a mistake or causes harm, it may be unclear who is responsible for the damage caused: the operator, the programmer, the manufacturer, or the robot itself. They also mentioned that the problem will be more difficult when the robots are getting more degrees of autonomy, self-learning, and decision-making capabilities. This unclear attribution of responsibilities in the automation system at the legal and moral level is called the "responsibility gap" [45], which was first raised by Matthias in 2004 [73]. Our task assignment of humans and robots in SAR can be seen in section 3.1.3.

2.3 Meaningful Human Control

In order to solve this "responsibility gap", different researchers gave different insights. Carlsen et al. [17] introduced the concept of "man in the middle". Previous owners, as well as the designers, manufacturers, and users of such robots, may be held responsible for any issues they create. Furthermore, the concept of "Meaningful Human Control"(MHC) is raised to avoid the responsibility gap from a control perspective.

2.3.1 Definition

Although MHC has already been studied in many areas, such as autonomy systems [25] and HAT [97], its definition is still not uniformly agreed upon, and each researcher has a corresponding interpretation in his field. Since the concept of MHC came from the field of automated weapons at the beginning, the very first definitions are all in this field. Article36 [5] listed the requirements for meaningful human control over individual attacks as contextual information, positive action from humans, and accountability. Another definition from the Center for a New American Security [52] claimed that MHC has three key components, which are informed and conscious decisions, sufficient information, and weapons are designed and tested and humans are well trained. With the study by Sio and van den Hoven in 2018 [27], MHC has evolved as a concept that can be applied to autonomous systems in general. They provided a philosophical account of how autonomous systems can be designed in such a way as to make MHC possible. Specifically, they argued that humans should ultimately remain in control of the whole system. They concluded the idea into two aspects: "decision-making systems should *track* human-moral reasons" and "the

actions of the system should *trace* back to humans who have a proper moral understanding of the system".

However, according to the finding of Amoroso and Tamburrini [4], the flaws in these uniform control approaches imply that addressing the MHC problem with a single formula is implausible. Differentiated policies for MHC are introduced into many fields, such as automated weapons [4] and surgical robots [37]. According to the degree of autonomy of the surgical robot, Ficuciello et al. [37] divided the MHC into five levels, and they also claimed that in the future, the autonomy level of the robot can be automatically selected according to the surgical needs, dynamically adjusting the human vigilance but not reducing the MHC.

After these theoretical attempts, not much research discussed and applied this concept under certain HAT contexts practically [50, 74]. Mecacci et al.[74] demonstrated how MHC can be designed and embedded into automated systems and systematically applied MHC theory to dual-mode vehicles, with some case studies.

2.3.2 Measurement

As for the measurement of MHC, very few studies proposed concrete methods. To operationalize the concept of MHC and achieve a measurable notion, van der Waa et al. [97] raised three measurable components of MHC, which are subjective experience of control by humans in HAT (experienced MHC), behavioral compliance with ethical guidelines, and behavioral compliance with moral values. They said that the second component compares the entire behavior of HAT with the ethical guidelines of the task. And for the other two components, user interviews are required for the measurement. In our study, we focus mainly on measuring the subjective experience of MHC by questionnaires. In their experiments, explanations were used to influence control. They claimed that explanations allow humans to better estimate when and which control should be performed and thus achieve MHC.

2.4 Explanation

Lewis [67] defined explanation as "a quantity of information about that event's causes". This information is beneficial to HAT in different ways but the main goal is to improve coordination in human-agent teams [48]. Neerincx et al. [79] also emphasized the need for mutual communication between humans and agents about the intent and basis of their actions, and explanations as necessary for humans to understand the performance of the agents. We will discuss explanations from two aspects: content and format.

2.4.1 Explanation Content

In this section, we listed all popular explanations used by other researchers and elaborate on why and why not they fit in our study. One of the purposes of our study is to find explanations that could improve the subjective experience of MHC.

Consequential Explanations

Since we wanted to provide extra information to help humans obtain control in the delayed control situations, it was intuitive to explain the unobserved events in the near future, which are the consequential explanations. Van der Waa et al. [97] mentioned that when the effect of human control is delayed in HAT, the agent could explain the consequence of exercised control of humans, which is the unobserved event. However, the effect of this explanation has not been verified, which would also be part of our work.

Contrastive Explanations

Lipton analyzed contrastive explanations in detail from a philosophical point of view in 1990 [68]. He claimed that a contrastive phenomenon consists of a fact and a foil, and the same fact may have several different foils. When we ask contrastive why-questions, we actually not only need the fact (the answer "Why"), but also need one specific foil to answer "Why the fact rather than the foil". The advantage of contrastive explanations is that they are more inherently intuitive to humans to both produce and comprehend [54]. Although the ability to explain a decision contrastively is claimed to lead to responsible decision-making [31], the downside is that identifying a foil from multiple foils can be difficult [96]. Agents are now unable to effectively infer the contrast from the open-ended question "Why this decision?" due to the intricacy of morally important activity [97].

Counterfactual Explanations

Counterfactual explanations describe events or states of the world that have not occurred and implicitly or explicitly contradict factual world knowledge and then give the consequence of states or events change. According to Ginsberg [39], a counterfactual is a conditional statement of the form "If P, then Q" where P is "expected to be false". This kind of explanation can be seen as contrastive by nature. Counterfactual explanations specify necessary minimal changes in the input so that a contrastive output is obtained. The difference is that the antecedent P did not occur in reality. Compared to the foil in contrastive explanations, antecedent P is easier to implement and more suitable for our task, which we will describe in detail in section 3.1.4.

Feature Attribution

Feature attribution methods are one of the most popular approaches for explaining the decisions of complex AI models. The methods evaluate how strongly each feature contributed to the model's choice for each given instance. With feature attribution methods, the relevant features can be obtained and presented as explanations about why the model made certain decisions [47].

Confidence Explanations

Confidence Explanations are used to provide confidence estimations that help humans to decide whether to trust the agent. Larasati et al. [64] examined the impact of interpreter trust in AI medical support using four explanations, including confidence explanations. However, we don't plan to include this type. Miller claimed that appealing to probabilities or statistical correlations is not as effective as referring to causes [75]. Besides, our task is not focused on human trust in robots.

2.4.2 Explanation Format

As Mohseni et al. concluded [76], identifying suitable explanation formats for the desired system and user group is the first step in delivering explanations to end-users. The design process can take into account various levels of complexity, duration, and presentation state (e.g., permanent or on-demand). The type of target user also needs to be taken into account, for example, Lage et al. [63] proved that AI novices prefer more basic explanation and representation interfaces.

Schoonderwoerd et al. [89] compared different explanations and corresponding UI design patterns used on decision support systems in medical diagnosis. In these design patterns, they use techniques such as the parallel coordinates technique [53] to visualize data, allowing features and other data to be displayed visually.

Besides, the combination can also be seen as a format. Researchers use hybrid explanations to gain the benefits of different explanations [62, 90]. We also used combinations of different explanations as our potential options in the expert study. As for the interface of explanations, we use some visual images to replace text information and use both images and text for important information. Paivio [81] found that images have two codes: verbal and visual, and each is stored in a different place in the brain. The dual-coding character could increase the probability that people will remember.

3

Expert Study

Our expert study was concerned with improving the design of explanations and the rationality of the background knowledge of the USAR task. The study obtained permission from the Human Research Ethics committee of the Technical University of Delft. The participant was a firefighter who had basic SAR knowledge and experience with USAR tasks. In this study, we provided the grounded theory and the background of our task and gave different explanation examples in the scenario. We collected the expert's responses by questionnaires and structured questions.

3.1 Design

We provide design and background knowledge relevant to the task in this section.

3.1.1 SAR Organization

We introduced the organization for a better understanding of the whole task and the parties involved. Referring to Murphy's rescue organization and personnel structure [77], we simplified the structure to obtain the organization shown in Figure 3.1. There were two teams led by the task leader: Rescue Team and Search Team, which were separately led by two team leaders. Each team had its own crew, including professionals (humans) and robots.

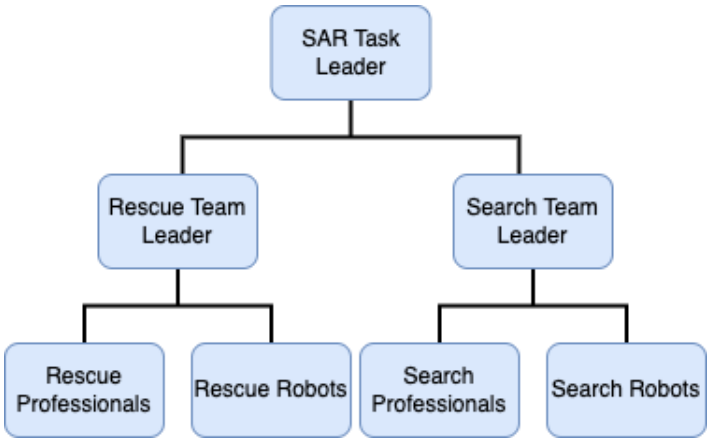


Figure 3.1: The organization of SAR

3.1.2 Moral Value Elicitation and Victim Characteristics

As we mentioned in section 2.3, humans should be able to remain in control of HAT, which means humans should make moral-related decisions. In the SAR task, the rescue part is more explicit about the ethical requirements, since prioritizing certain victims over others is strongly related to making moral decisions.

We mentioned that the rescue task is time-critical and can be slow if a human completes the decision for each rescue. So we referred to van der Waa et al. [97] in the patient triage task and used a moral value elicitation approach to reduce the workload of humans and improve moral-decision-making efficiency.

Victim Characteristics

To enable the rescue team to set moral values, we selected five ethically relevant characteristics that rescuers are concerned about for each victim: gender, age, vital sign, difficulty to rescue, and distance [2, 71, 83]. For each characteristic, there are several categories so that certain types of victims can be prioritized for rescue according to different types of moral value elicitation. The levels are shown in Figure 3.2.

characteristic	categories				
gender	male	female			
age	1-20	21-40	41-60	61-80	81-100
vital sign	low	middle	high		
distance	short	middle	long		
difficulty to rescue	low	middle	high		

Figure 3.2: Characteristic categories

Moral Value Elicitation

In our specific task, we made the following settings: Before the rescue team started to rescue, the rescue team leader would rank these five characteristics by priority according to his/her moral values. For each characteristic, there were two options. For the age, there were older preferred and younger preferred. For the gender, the two options were male preferred and female preferred. For the vital sign, difficulty to rescue, and distance, low/short and high/long options were provided. The moral value elicitation workbench is shown in Figure 3.3.

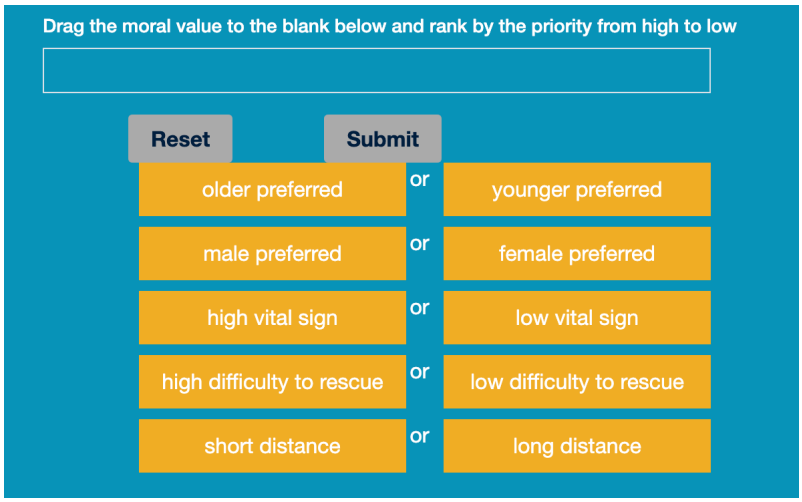


Figure 3.3: Moral value elicitation workbench

3.1.3 SAR Workflow

After introducing the parties involved in the SAR task and the most concerning part, we briefly presented the workflow to clarify the assignment of responsibilities of humans and robots in SAR. Referring to the workflow of USAR defined by Murphy [77], we developed our workflow shown in Figure 3.4. From the search team's perspective, first, the search team went inside the devastating area and tried to find the victims. Once a victim was found, the search team sent the location and information of the victim back to the rescue team, who stayed in the surrounding area. The search crew kept searching until completing all the coverage. We assumed that the search task was completed by other teams and we mainly focused on the rescue task in our study. From the rescue team's point of view, the rescue team leader deployed the team in the surrounding area, and the leader started moral value elicitation. When the leader could detect the existence of victims on the remote screen, he/she could arrange robots to rescue them. The robots followed the moral values set by the leader and selected the highest priority victim to rescue first. The whole task would be finished until all victims were rescued. In our further study, we focused mostly on the rescue phase and rescue team.

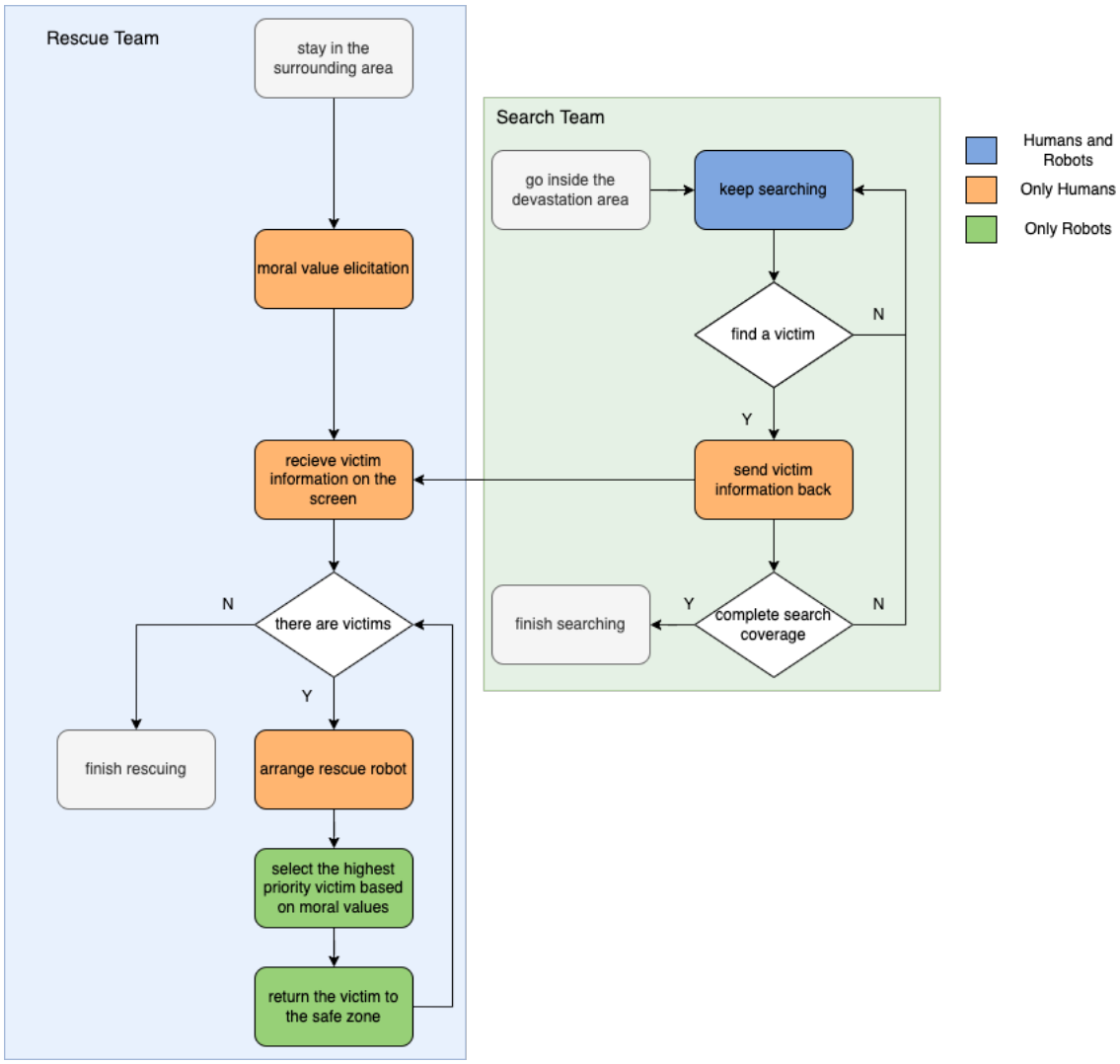


Figure 3.4: The workflow of SAR

3.1.4 Explanations

The purpose of providing explanations was to let the robot give feedback on its understanding of the moral value elicitation and what actions it would take based on it, ideally giving the team leader a better subjective control experience.

Explanation Types

We chose consequential explanations as our first basic type since we want to verify if explaining the consequence of exercised control of humans could help with the subjective experience of control. Further, we chose counterfactual explanations, instead of contrastive ones, as the other basic type. Events that did not occur but were contrary to the knowledge of the factual world in counterfactual explanations could be mapped to hypothetical changes in moral values. These changes, while not taking place in the real world, provide more information and may have implications for the future moral setting of the same person, which we will also look into later.

As for the combinations, we used consequential explanations to combine with other types we mentioned in section 2.4 except for the confidential ones, because it was said that confidence explanations were not effective in referring causes. We provided the following explanation types to our expert.

- **Consequential:** Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A.
- **Counterfactual:** If you prioritized younger age over gender, my decision would have been rescuing B rather than A.
- **Consequential + Contrastive:** Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A because you prioritize distance the most and victim A has a closer distance than victim B.
- **Consequential + Feature attribution:** Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A. For the most important value – distance, A has a short distance.
- **Consequential + Counterfactual:** Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A. If you prioritized younger age over gender, my decision would have been rescuing B rather than A.

Time of Explanation

We also considered the effect of the timing of the explanations on MHC. To explain the delayed effect of control, the robot could provide explanations after the moral value elicitation, during the whole task, and at the end of the task. We asked our expert to discuss and judge different task settings, and then determined a suitable one.

3.2 Participants

We recruited one expert from the firefighter group in the Netherlands on LinkedIn. The expert was a part-time firefighter who had basic USAR knowledge and years

of experience with USAR tasks.

3.3 Scenario

The expert played the role of rescue team leader in our task. When a disaster struck, he led his rescue team to the scene, deployed outside the danger zone on standby, and began the moral value elicitation on the robots. Meanwhile, search teams continued finding victims and marking their locations. He could see the location and status of victims remotely and then he arranged for a robot to rescue them. The robot would follow his moral value and choose victims by priority to rescue. The task would finish until all victims were safely rescued.

The robot would give its own explanation of moral value elicitation at three different times. The explanation included the information of two example victims and a statement of the robot's actions and possible reasons for choosing them, shown in Figure 3.5.

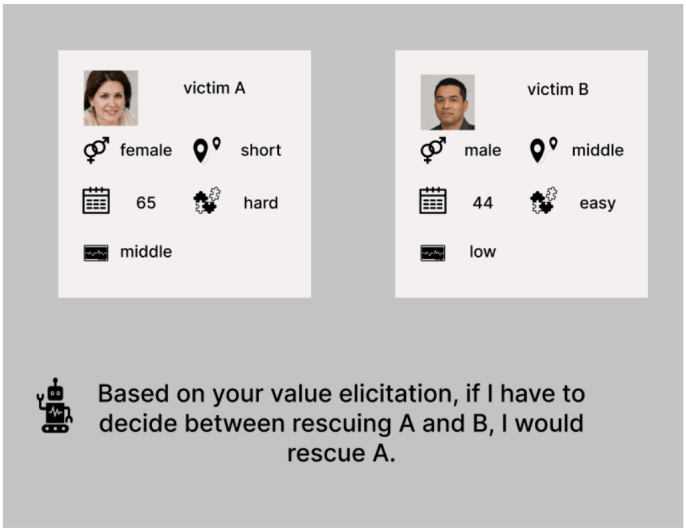


Figure 3.5: Robot explanation example

3.4 Measurement

Since one of the main purposes of the expert study was to select the appropriate explanation types for subsequent user experimentation studies, we created a questionnaire based on that used by van der Waa et al. [97] to examine the usefulness of explanations (the questionnaire is shown in appendix A). The expert was asked to rank the explanations above by how well they matched the description of the questions.

Ranking Question Questionnaire
This explanation provides the most new information
The explanation provides the most reasons about decisions made by the robot
The explanation helps me to understand the causality
I can understand the explanation best
The explanation shows the robot's understanding of my moral value elicitation best
This explanation matches my expectations best

Table 3.1: Questionnaire in expert study

For the structured questions in Table 3.2, we hoped that we could get the expert's subjective thoughts and direct feedback on the design.

Structured Questions in Expert Study
What do you think are the differences between moral value elicitation and direct control?
What information do you think is important for the team leader to know in the rescue task?
Do these explanations cover the information?
Which type of explanation do you think is better to help increase the subjective experience of human control? Why?
What do you think is the proper explanation time? Why?

Table 3.2: Structured Questions in Expert Study

3.5 Procedure

The expert study was conducted in person lasting about 30 minutes and was divided into two parts: the introduction session and the interview session.

In the introduction session, the expert first filled in the informed consent forms. Then the instructor introduced the design of settings, such as the process and organization structure of SAR and the moral value elicitation. Moreover, the instructor presented the scenario that the participants would face and some different types of explanations that would appear in the scenario.

The interview session started with a ranking question questionnaire about explanations. The expert considered what type of explanations fit the scenario and matched the description of the questions, and ranked the explanations from the most relevant one to the least relevant one. Besides, the instructor also asked some structured questions to the expert, which are shown in table 3.2.

3.6 Results

3.6.1 Explanation Choice

As for the ranking questions questionnaire, the rank of explanations on each question corresponded to a score, for example, the first place corresponded to

five points, the second place corresponded to four points, and so on. Then, we got the chart that presented the total scores of all the explanations shown in Figure 3.6.

Out of a total of seven questions, The Consequential + Counterfactual explanation was ranked first in four of them, including: "This explanation provides the most new information", "The explanation provided the most reasons about decisions made by the robot", "I can predict what the robot will do after sending it for rescue best" and "I can feel the most control about the system with the explanation". For questions "I can understand the explanation best" and "The explanation matches my expectations best", the Consequential explanation was probably ranked first due to its simplicity. In contrast, the performance of the Consequential + Counterfactual explanation on these two questions was not ideal, only ranking fifth and third respectively. Finally, for the question "The explanation shows the robot's understanding of my moral value elicitation best", the experts felt that the Consequential + Contrastive explanation performed best.

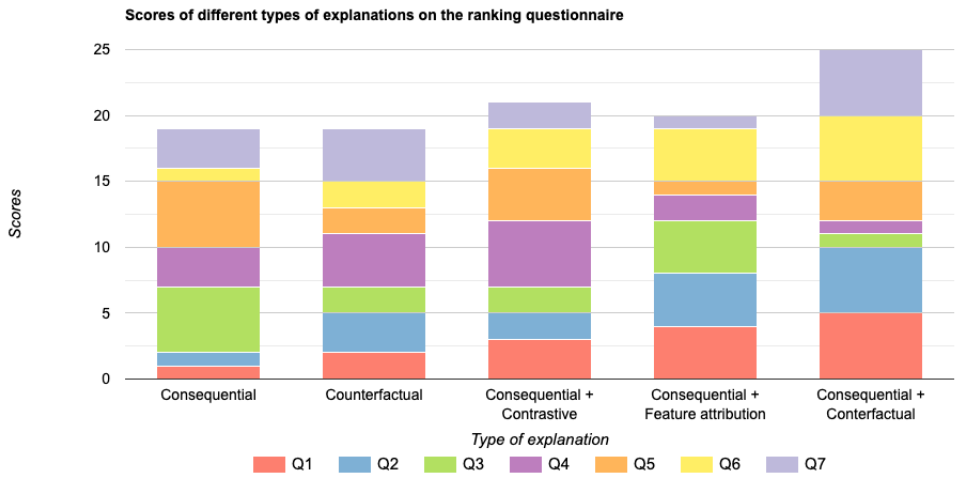


Figure 3.6: Scores of all types of explanations on the ranking questionnaires, the details about the questions are in the appendix A

3.6.2 Feedback

On a general level, the evaluation of the overall design from the expert was positive. The expert said that the design of moral value elicitation would "provide some improvement in terms of efficiency" but required accuracy in the setup. He also thought that the kinds of victim information I provided, such as difficulty to rescue and level of injury, were important to rescuers and could "help with rescue decisions". For the explanations, he thought the explanations basically

covered the information and the Consequential + Counterfactual one helped him increase the subjective experience of control most. For the proper time to provide explanations, he said it was not that useful to provide them at the end of the task, the leader needed more help with explanations during the task than after the task was over.

Besides, the expert gave some useful inputs for improving the design and user-friendliness. For example, I was using distance rather than the difficulty to reach in the design part, however, he said that the distance could not reflect the difficulty well since there might be many obstacles during the short distance to the victim. Moreover, he suggested that a unified measurement for all characteristics (low, middle, and high) was useful and friendly for the rescue team leader to understand the victim information, but until then the measurements I used were different (short and long, low and high, easy and hard). In order to distinguish the levels more clearly, we could use different colors to mark them.

characteristic	categories				
gender	male			female	
age	1-20	21-40	41-60	61-80	81-100
level of injury	low		middle		high
difficulty to reach	low		middle		high
difficulty to rescue	low		middle		high

Figure 3.7: Characteristic categories after expert study

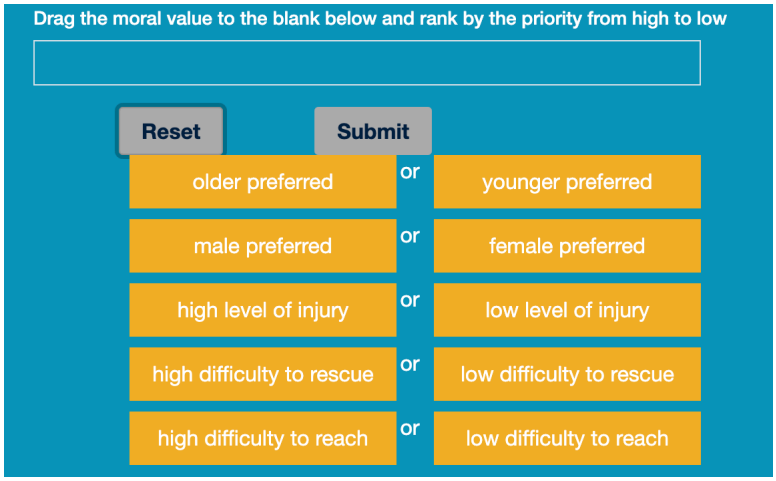


Figure 3.8: Moral value elicitation workbench after expert study

4

Experimental Study

In the experimental study, we first conducted a pilot study on a small number of participants (two for each group) to identify the flaws in our design and improved it according to the feedback. Then, we recruited more people for the user study.

4.1 Pilot Study

After the pilot study, we found a *ceiling effect*, which meant every participant could achieve top performance on the task. We have three rounds in the experiment and investigate one level of SA at the end of each round. All the participants could get a full score on the SA level1 in the first round. We figured that there were two reasons: The first one was that the task itself was too easy for the participant to keep an overview, such as only a few victims existed at the same time and the moving speed of the robot was low. The second one was that some questions in the questionnaire were too straightforward and lacked a process that allowed participants to think. Finally, the questions did not differentiate between the two groups of participants.

To solve the problem, we iterated our design:

- Hid the victim information, the participants could only see the information by hovering over the victims (see Figure 4.4).
- Increased the workload by increasing the speed of the robots and the number of victims from three to four in the first two rounds.
- Not allowed the participants to recheck the situation after pausing the task prior to SA measurement, instead directly asked the questions in the first two rounds and only leave ten seconds to check the situation in the third round.
- Changed the straightforward questions in the questionnaire, for example, instead of asking the participants to estimate the time that the next rescue task will cost, asked them how long it would take for the robot to reach, rescue, and bring the next victim back.

Null hypothesis (H0)	Alternative hypothesis (H1)
Providing combination explanations does not increase the subjective experience of MHC	Providing combination explanations can increase the subjective experience of MHC

Table 4.1: Hypothesis

4.2 Design

4.2.1 Conceptual Model and Hypothesis

To design our experiment, we abstracted our research into a conceptual model shown in Figure 4.1. There was one independent variable (explainability) with two levels (consequential + counterfactual explanations vs. no explanations). The mediating variable was the Situation Awareness of participants and the dependent variable was their subjective experience of MHC. We also listed some potential moderating variables: age, gender, education level, and computer game experience.

The experiment followed a between-subject design. According to our independent variables, we constructed two conditions: There were two groups of participants, one performed the task without explanations (baseline group) and the other conducted it with the combination of two explanations (conditional group).

Further, we give our null hypothesis and alternative hypothesis shown in Figure 4.1.

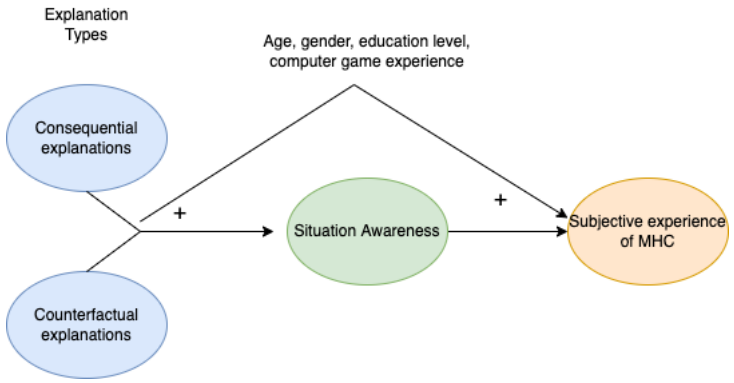


Figure 4.1: Conceptual Model

4.2.2 Environment Architecture

As we mentioned in section 2.1.2, the architecture of MATRX consists of four components: scenarios, GridWorld, objects, and agents. In our experiment, the

scenario was the rescue task performed by the rescue team leader and rescue robots. The GridWorld was the danger zone where the victims were searched and rescued. The objects were the items shown in the GridWorld, like the obstacles.

There were four agents in this experiment, which were the robot, the rescue team leader, the search team, and the god.

Rescue team leader: The rescue team leader set the moral values and supervised the task in the workbench environment and was the role taken by the participants.

Robot: The robot automatically rescued the victims following the moral values set by the rescue team leader. The logic was: Among all the victims found so far, check the highest priority characteristic first in the selected moral values of each victim, if a prior victim can be selected according to it, then the robot will go to rescue him/her; if not, then proceed to check the second priority characteristic in selected moral values, and so on.

Search team: The search team would find the victims in the danger zone and send their information to the workbench for the rescue team leader, here we only kept the part where the information was sent back since we focus on the rescue task.

God: The god could be seen as the instructor who could control the experiment and do some configurations.

4.2.3 Environment Interface

Each agent had an interface view on the experiment, but the god view (Figure 4.2) and rescue team leader view (Figure 4.3) were what we mainly used in the experiment. The instructor could monitor the danger zone, start/pause the experiment and configure the explanations type in the god view.

Compared to the god view, there were more components in the rescue team leader view such as the menu panel, victim information, moral values, robot explanation, and a view of the danger zone.

Menu panel: In the menu panel, there was a button "Moral Value" that could be tapped for moral value elicitation, as shown in Figure 3.8. Besides, the timer would count the total time spent on the task so far.

Victim information: Once a victim was found by the search team, his/her information was shown here, and the information disappeared after he/she was rescued by the robot. The colors represented different levels, low levels were green, middle ones were yellow and high ones were red.

Moral values: The moral values of the rescue team leader were shown here after the elicitation. The purpose was to give a reminder to the leader.

Robot explanation: According to the expert study, we chose Consequential + Counterfactual explanations as explanations for the rescue robots. In the consequential sub-explanation, the robot explained which victim it would rescue among all the victims and which characteristic it was based on to make the decision. In the counterfactual sub-explanation, the robot explained another situation in which it would rescue another victim if the moral values would have changed.

View of the danger zone: In this view, the rescue team leader could supervise the robot, and check the location of the victims found by the search team.

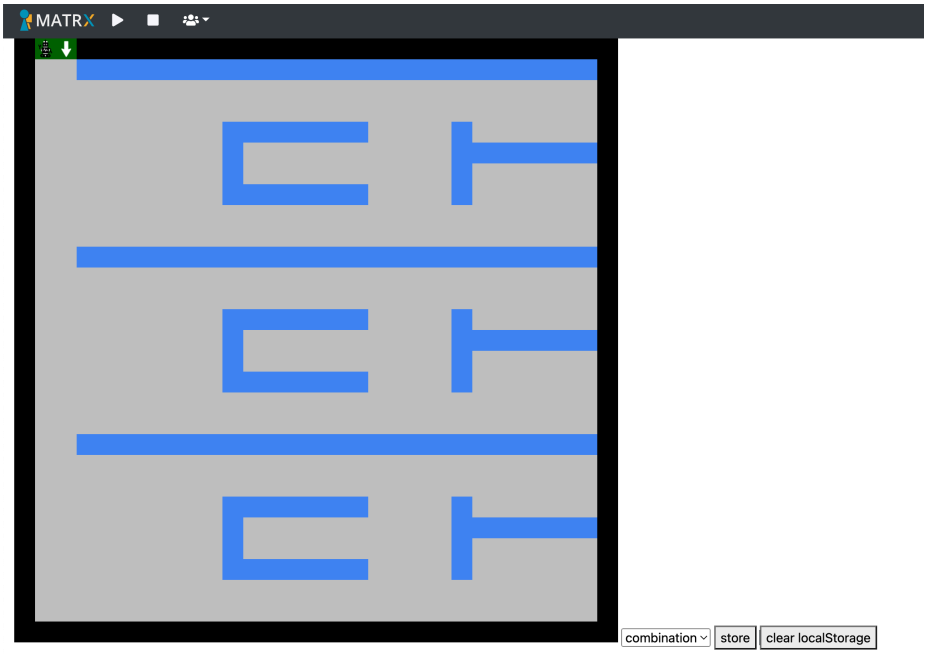


Figure 4.2: God view

4.3 Scenario Settings

The participants played the role of rescue team leader, which was the same as in the expert study. There were three rounds of rescuing. In each round, the robot saved three victims by the priority of moral values. The robot explanations were shown on the interface (see Figure 4.3) and changed depending on the available victims.

4.4 Measurement

We measured the mediating variable (SA) and the dependent variable (subjective experience of MHC) by quantitative analysis. Moreover, a qualitative analysis was conducted to compare different sub-explanations.

4.4.1 Measurement for SA

In section 2.1.3, we mentioned that the simplicity to measure is one of the advantages of the three-layer SA model. We chose SAGAT [33], the most widely used measurement method, as our measurement for SA.

SAGAT should include queries about all operator SA requirements, including level 1 (perception of data), level 2 (comprehension of the meaning), and level 3 (projection of the near future) components. An in-depth cognitive task analysis is required for each domain in which SAGAT is used to determine queries. Combined with our task, we provided some queries on the three levels in Table 4.2. All questions of the questionnaires we used in the experiment study were shown in the appendix C.

4.4.2 Measurement for the Subjective Experience of MHC

We used a Five-points Likert scale to measure the subjective experience of MHC. Van der Waa et al. [97] created a questionnaire to measure participants' control over task performance in a human-agent context. We referred to that and adapted it to our context. The questions included the controllability of the system, the responsibility to the victims, and the robot understandability, which were shown in Table 4.3.

4.4.3 Measurement for Comparing Sub-explanations

We concluded that consequential + counterfactual explanations were more suitable for our scenario than other forms of explanations according to the expert study. However, we wanted to further study the effect and difference between these two sub-explanations. Therefore, we provided some questions for qualitative analysis as shown in Table 4.4, which was based on the statements for measuring the usefulness of explanation types in the study of van der Waa et al. [97].

Perception	Comprehension	Projection
How many victims are there in the danger zone now?	Why do you think the robot will rescue victim XX among them in this situation?	Who do you think the robot will rescue next? And why?
Can you point out which place does victim XX stuck in?	How long does it take for the robot to reach victim XX?	How long do you think it will take for the robot to rescue victim XX after reaching him/her?
Can you describe the information of victim XX?	How long does it take for the robot to rescue victim XX?	How will the robot rescue another victim in this situation?

Table 4.2: SAGAT queries

Measurement on the subjective experience of MHC
It is simple for me to keep an overview of the whole task.
I feel responsible for the well-being of the victims.
I feel all comfortable with all decisions in the task.
The robot's rescue choice matches what I thought.
I can feel control over the system.

Table 4.3: Questions for measuring subjective experience of MHC

For these two sub-explanations, which one:
can be understood better.
helps me to understand the causality better.
shows the robot's understanding of my moral value elicitation better.
help me predict what the robot will do next.

Table 4.4: Questions for sub-explanations

4.5 Participants

For the pilot study, we recruited four participants and tow for each group. For the user study, We recruited 34 participants (17 men, 16 women, and one who preferred not to say) through advertising on the TU Delft campus and personal contacts, 17 for the baseline group and 17 for the conditional group. There were 25 participants (73.5%) who had an age range of 18-24 and nine of them (26.5%) were between 25 and 34. With respect to the education level, three of them obtained some college credits but no degree, 25 of them had already got a bachelor's degree, and six of them had a master's degree. In terms of computer game experience, 13 of the participants played several times a year, ten participants several times a month, three participants several times a week and eight participants played on a daily basis. Detailed participant information can be found in appendix F.

4.6 Hardware and Software

To run the experiment, we used a MacBook Pro 13" laptop and the Human-Agent Teaming Rapid Experimentation (MATRX) software that we mentioned in section 2.1.2. The experiment was compiled and run on a local server. The participants performed the experiment through the interface shown on the Google Chrome browser. Meanwhile, the Situation Awareness questionnaire from Qualtrics and the information survey and control questionnaire from Microsoft Forms were also shown on the Google Chrome browser.

4.7 Procedure

The experiment was conducted in person lasting about 25 minutes and was divided into three parts: the introduction session, the experiment session, and the information and feedback session.

In the introduction session, the participants first filled in the informed consent forms. Then, the instructor introduced some background knowledge about SAR tasks, such as the workflow and organization of SAR. The instructor also presented the scenario that the participants were going to face.

During the experiment session, the participants set their moral values and conducted three rounds of rescuing by supervising rescue robots. They were also asked to fill in the questionnaire about SA after finishing each round of rescuing (one round corresponded to one level of SA). In the third round, the participants had 10 more seconds to check the task information sources and then answered the corresponding SA questions in the projection level.

The third session came after all the victims were rescued. Basic information, such as age and gender, was collected. Moreover, all participants filled out the questionnaire about the subjective experience of MHC. Besides, participants from the conditional group also needed to answer several questions about explanations. We also asked some unstructured questions about control and explanations according to their answers.

4.8 Data Processing

We collected our data from the measurements in section 4.4. In the questionnaire of SAGAT, we had eight questions for the first level, two questions for the second level, and six questions for the third level. We calculated the participants' correct answer rate for each level and used the number of percent as the SA score (range 0-100). So, we got SA score on level 1 (SA score 1), level 2 (SA score 2), level 3 (SA score 3), and the average score on all levels (average SA score).

In measuring the subjective experience of MHC (shown in the appendix C), we had five questions about control for both groups and one more question about explanations for the conditional group. We mapped the level of agreement on the five questions about control to 1-5 ranging from "strongly disagree" to "strongly agree". This way, we could calculate the average score on the subjective experience of MHC (avg. control score) in both groups.

There was no participant who played computer games several times a week in the baseline group. In order to analyze this moderating variable, we added a category called gaming frequency by combining "several times a week" and "daily" into "high frequency" and "several times a month" and "several times a year" into "low frequency". Then, for all moderating variables, we re-coded them by assigning numerical values to the levels of categorical variables, to adapt to further analysis. For example, our gender variable had three categories so we re-coded "prefer not to say" as 0, "female" as 1, and "male" as 2.

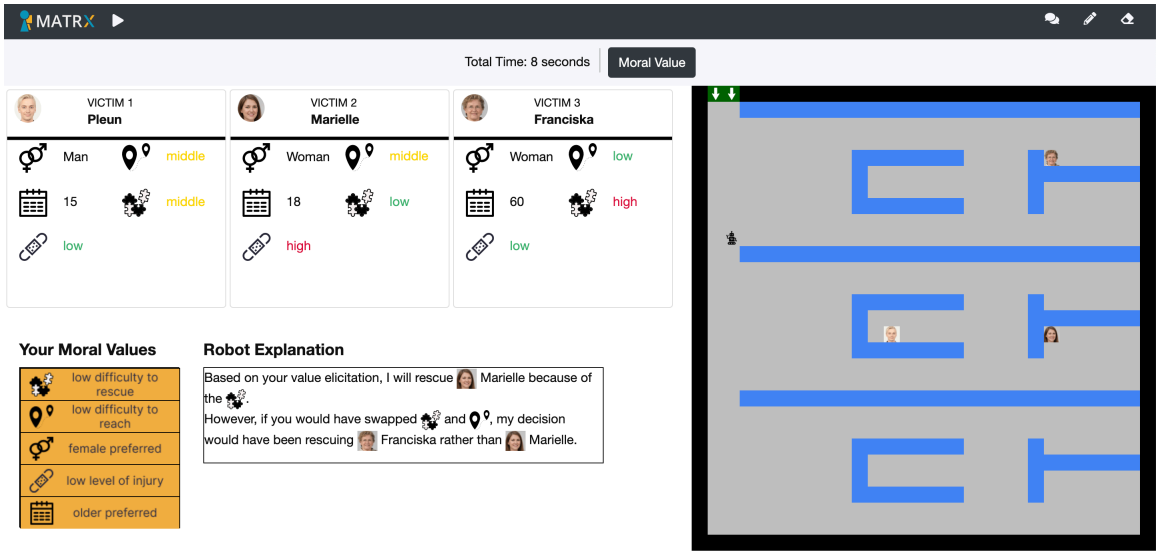


Figure 4.3: Rescue team leader view

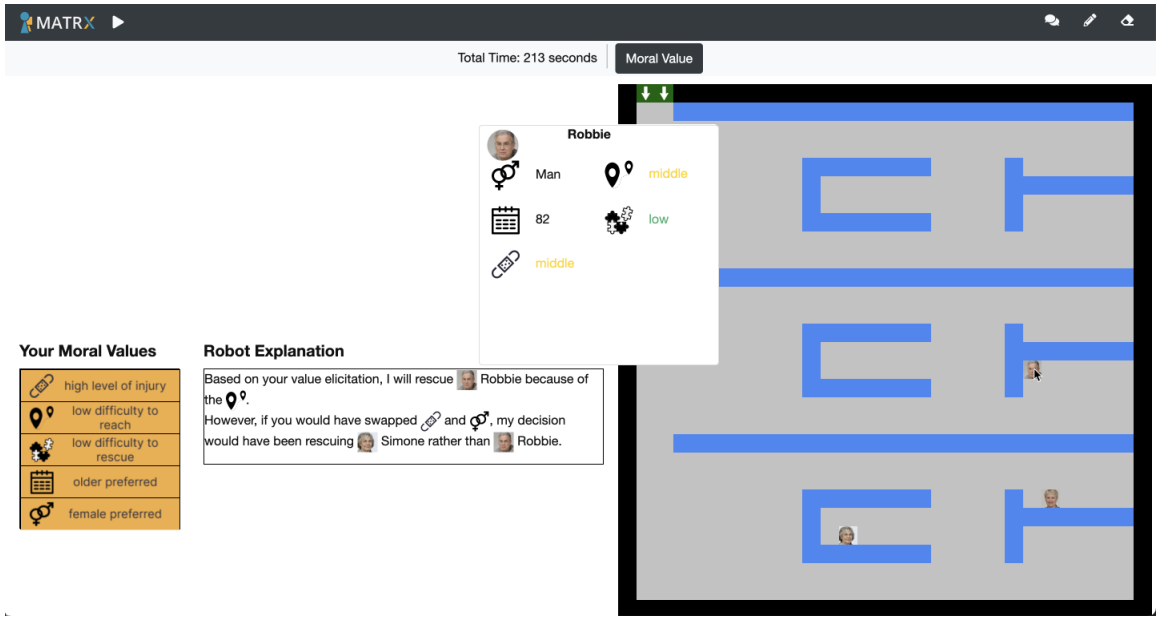


Figure 4.4: information checking by hovering

5

Results

5.1 Moderating Variables

We assigned participants based on balancing the distribution of different moderating variables between the two groups. For each of the above variables, we tried to make the frequency distribution in the two groups the same as possible to remove the influence of the participants' demographics. We used the Kolmogorov-Smirnov two-sample test because the distribution was non-normal. Moreover, since the data followed a non-continuous distribution, a bootstrap version test was performed. The results showed that the frequency distributions of gender ($D(17) = 0.056$, $p = .917$), age ($D(17) = 0.059$, $p = .836$), education level ($D(17) = 0.176$, $p = .368$), and computer game experience ($D(17) = 0.176$, $p = .628$), were not significantly different. Therefore, we excluded the effect of the above moderating variables on the experiment due to their similar distribution in the two groups.

5.2 Effects of Explanations

In this part, we examined the effect of explanations on the different levels of SA and the subjective feeling of control. First, we used the Shapiro-Wilk test to check the normality of all data (shown in table 5.1). The distributions of all data except for the SA score in level2 ($W = 0.765$, $p < .001$) and level3 ($W = 0.915$, $p = .012$) were normal. Next, we used F-tests to check that the variances of two groups on SA score 1 ($F(1,16) = 0.554$, $p = .248$), average SA score ($F(1,16) = 1.407$, $p = 0.503$) and average control score ($F(1,16) = 0.547$, $p = 0.238$) were homogeneous. With the distribution normality and the homogeneity of variances, we used the Two-Sample t-test on the SA score 1, average SA score, and average control score. In contrast, we conducted the Wilcoxon rank-sum test on SA score 2 and SA score 3 to examine the difference between the two groups. The results showed that there was a significant difference in SA score 3 ($W = 85.5$, $p = .036$) between the baseline ($M = 26.47$, $SD = 0.18$) and explanation ($M = 40.20$, $SD = 0.16$) conditions. For average SA score, there was also a significant difference between the baseline group ($M = 30.39$, $SD = 12.32$) and the conditional group ($M = 41.01$, $SD = 14.62$); $t(32) = 2.290$, $p = .029$. However, for the average control score, there was no significant difference between the baseline ($M = 3.12$, $SD = 0.70$) and explanation ($M = 3.10$, $SD = 0.52$) conditions ($t(32) = 0.166$, $p = .869$). The next sections will show further analysis of the relationship between the explanations and scores.

Variables	*W*	*P*
SA score 1	0.962	.277
SA score 2	0.766	<.001 ***
SA score 3	0.915	.012 *
avg SA score	0.972	.507
avg control score	0.947	.098

* <.05, ** <.01, ** <.001.

Table 5.1: Normality check on all scores

In the conditional group, we asked participants if they thought the explanations would help them with their next value elicitation. The majority of them were positive, with 41.2 percent of all participants strongly agreeing and 35.3 percent agreeing with the statement. Only 5.9 percent of people had a negative view.

5.3 Correlation Analysis

We ran a correlation analysis to examine the relationship between SA scores and the average control score. First, we used scatter plots (see Figure 5.1) to check the relationships. We couldn't tell if they were linear relationships at a constant rate. Moreover, because of outliers and the small scale of data, we chose the non-parametric Spearman's rank-order correlation. We found that there was a significant, moderate positive correlation between SA score 3 and average control score ($r_s = 0.42$, p = .015), and between average SA score and average control score ($r_s = 0.50$, p < .001).

5.4 Regression Analysis

Since we found that there was a significant difference between high (M = 3.44, SD = 0.42) and low (M = 2.96, SD = 0.63) frequency gamers on average control score (t(32) = 2.631, p = .014), we used linear regression to investigate whether we could predict average control score and SA scores based on the predictors game experience and explanations. The GVLMA package was used to test the linear model assumptions of normality, heteroscedasticity, linearity, and uncorrelatedness.

We used a multiple linear regression model to predict the average control score. The predictor variables were: SA score 1, score 2, and score 3, game frequency, and explanations. According to the GVLMA, all model assumptions were acceptable. It was shown that the model statistically significantly predicted the average control score (F(5,28) = 3.514, p = .013, adj. $R^2 = .276$). However, only the SA score 3 (p < .001) contributed statistically significantly to the prediction, with a 10 points increase in SA score 3 (range 1-100) being linked with a 0.17 point improvement in the average control score (range 1-5).

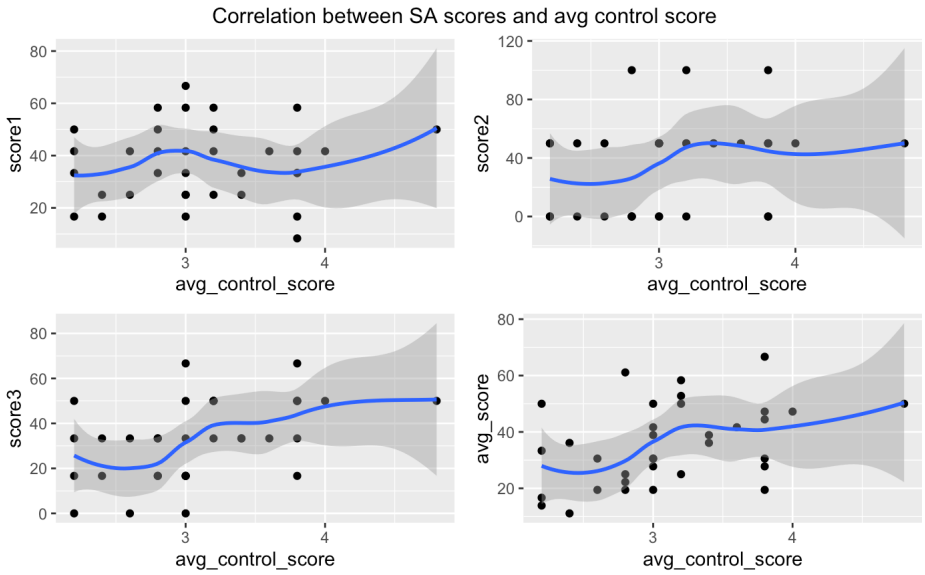


Figure 5.1: Correlation between SA scores and avg control score, the blue line is the fitted smooth curve and the grey part means 95 percent confident interval around the smooth curve.

Then, we used gaming frequency and explanations to predict SA scores. We checked the independence of explanations and game frequency. The chi-squared test showed that they were independent of each other (p = .464). The results showed that SA score 3 ($F(2,31) = 3.396$, p = .046, adj. $R^2 = .127$) and average SA score ($F(2,31) = 3.418$, p = .046, adj. $R^2 = .128$) could be significantly predicted. However, only explanations (p = .039) added statistically significantly to the prediction of the SA score 3, with adding explanations improving 0.13 point on SA score 3, while neither of the predictor variables had a significant contribution to predicting the average SA score.

5.5 Evaluation of Sub-explanations

We mentioned that we used a few statements to measure the usefulness of two different sub-explanations in section 4.4.3. The results showed that in terms of understandability (52.9% vs 47.1%) and predictability (47.1% vs 52.9%), the performances of the consequential and counterfactual explanations were similar, and the participants did not show a preference. However, counterfactual explanations had a better performance in helping people understand the causality (82.4%) and showing the robot's understanding of human moral values (64.7%).

We also did a quantitative analysis on sub-explanations. We checked the

normality of the distribution of statement scores on both explanations with the Shapiro-Wilk test. The results showed that the distributions were non-normal ($p = .682$). Therefore, we used the paired-samples Wilcoxon test to examine the difference between sub-explanations because it was evaluated within subjects. However, it showed that there is no significant difference between the scores of the sub-explanations ($p = .270$).

5.6 Qualitative analysis

We analyzed the answers to the unstructured questions based on grounded theory [100]. We broke answers into excerpts and grouped these excerpts into codes, and then combined similar codes into categories. We could come up with a more reliable theory by combing the text. The results are shown in figure 5.3. We also counted the frequency of descriptive codes with a bar chart as shown in figure 5.2.

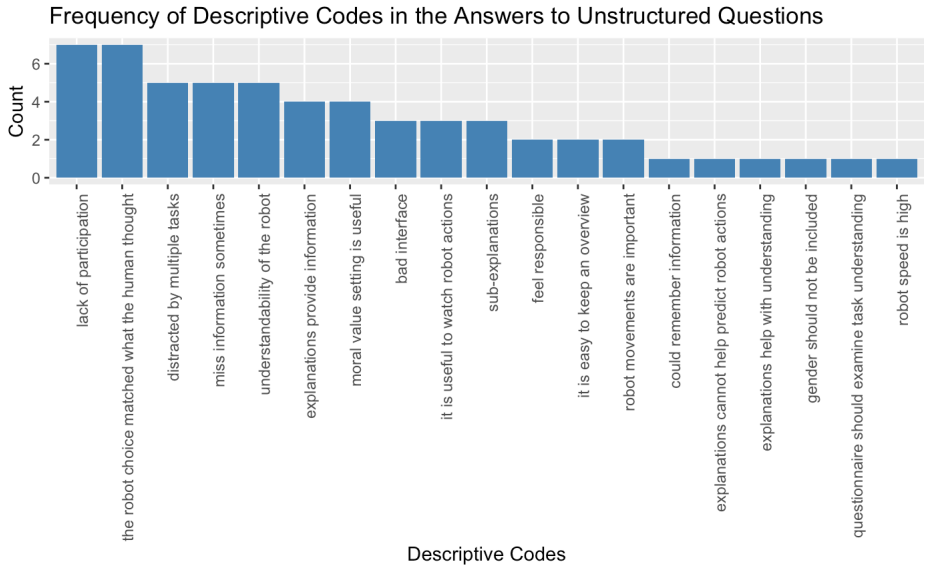


Figure 5.2: Frequency of Descriptive Codes in the Answers to Unstructured Questions

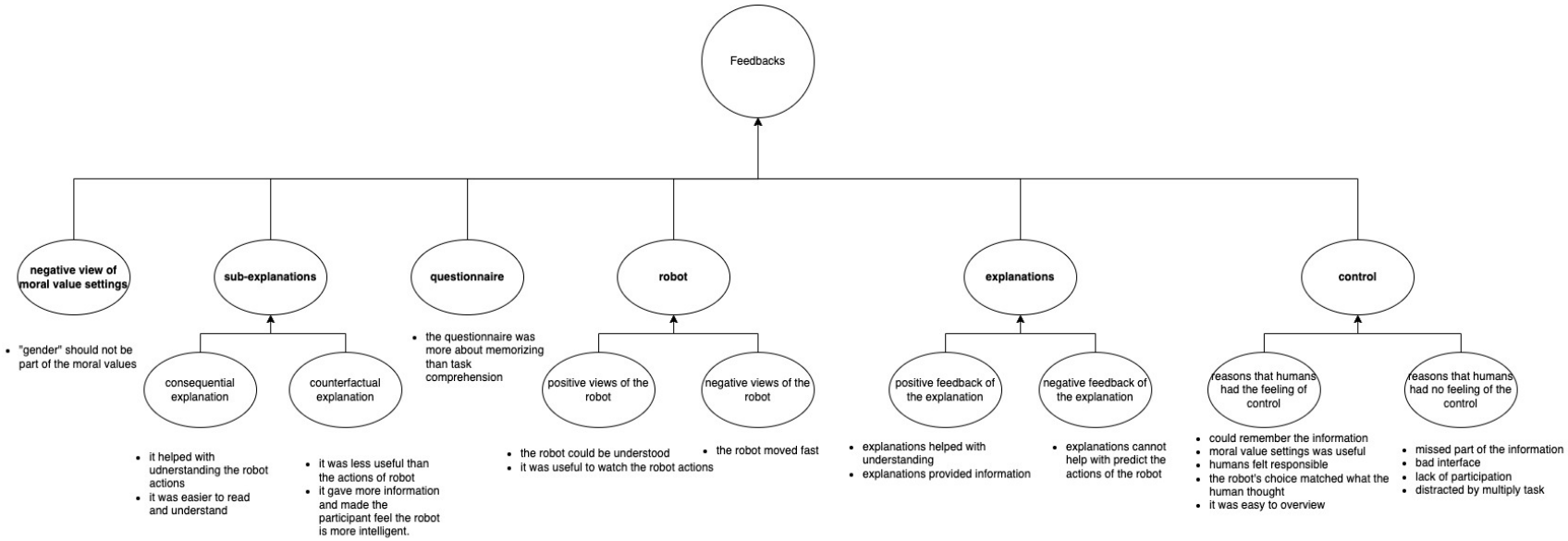


Figure 5.3: Qualitative analysis results in the answers to unstructured questions. The bullet points are codes and the bold texts are categories.

6

Discussion

6.1 Effects of Explanations

According to the results of the effects of explanations, it was shown that the explanations could result in a significantly higher projection level of SA and overall SA, which was in line with what we expected in section 1. The regression analysis also showed that participants who received explanations had more SA in projection level and overall SA during the task.

Both in the correlation and regression analysis, the projection level of SA and overall SA had a strong association with the average control score. It suggests that participants who got more SA tended to feel more in control during the task.

From the qualitative analysis, it can be seen that the participants were positive about the role of explanation in providing information (four participants mentioned it) and understanding the robot (one participant mentioned it). However, there was one participant who thought that the explanations could not help with predicting the actions of the robot.

6.2 Subjective Experience of MHC

Our results showed some positive effects from the explanations. However, the results showed no significant difference in the subjective experience of MHC between the two groups, which suggests that explanations may not affect the feeling of exercised control. This result was inconsistent with our hypothesis, so we made the following possible suggestions.

First, as we mentioned in section 4.1, there was a ceiling effect that almost all participants could achieve high performance on the task. Therefore, we increased the difficulty of the task by adding more victims in each round and speeding up the movement of the robot. However, this increased pace led to another problem: participants sometimes ignored parts of the explanation due to time pressure. As shown in figure 5.2, four participants mentioned that they missed part of information sometimes. We conjectured that the role of explanations, in this case, was not fully reflected, so it might lead to a lack of increase in people's subjective experience of control.

Furthermore, the explanations provided no real control over the operation, such as changing the robot's operation and modifying the participant's own moral values again. This resulted in the participant always being in the position of a supervisor after completing the moral value elicitation and the task started.

Six participants mentioned that they felt a lack of participation in the task. One participant indicated "During the mission, all I could do was to observe the robot's actions, and I could not override its choice of victim."

According to the study of van der Waa et al. [97], none of the experts felt that they had the capacity to control the value elicitation process. The experts stated that they believed the agents acted in accordance with the ethical values evoked, but that this did not result in a sense of being controlled. However, we found some findings that were inconsistent with their conclusions, with some participants feeling a sense of involvement and control in the value elicitation process, and thus scoring high on the sense of control for the entire task. One of the baseline group participants said: "I think the ethical settings section is important and I have found that the bot operation follows my settings. The reason I felt control was that my input in the task, i.e., my morality, got an output that satisfied me, i.e., the correct rescue operation of the robot."

Based on the above discussion, we think that explanations do increase the degree of SA of the task to a certain extent, but they do not affect the experienced control. Combined with the analysis of answers to unstructured questions, we consider that the feeling of exercised control more likely depends on the participants themselves, such as their ability to handle multiple tasks simultaneously (five people mentioned that they were distracted by different tasks but two people said it was easy to keep an overview). Finally, we believe that the subjective feeling of control is more influenced by whether the participants think the moral value elicitation is important and reflected by the robot actions.

6.3 Computer Game Experience

Although the explanations do not affect the feeling of exercised control, we found that one of the moderating variables, computer gaming frequency, has a significant effect on it. The results showed that people who play computer games frequently have a significantly higher subjective feeling of control. We suggest two possible reasons for this. One is that frequent game players have higher cooperative engagement and feelings of cohesion. Ewoldsen et al. found that playing cooperative video games can influence subsequent cooperative behavior [35]. Further, Greitemeyer and Cox raised that cooperative team play boosted emotions of togetherness, which in turn triggered trust, which in turn improved cooperative conduct [43]. So we think that, while working with the robot, players who had more experience with cooperative behavior were able to become more immersed in the cooperation and had more trust in the robot, resulting in a better control experience.

Another potential reason is that frequent game players have relatively better cognitive skills [82]. They can process information well even under time pressure, so the explanations might have been better understood and utilized by them. The full use of explanations could result in a better feeling of overview and exercised control experience, which is in line with the result that the high-frequency gaming group had a significantly higher rating on the question "It is

simple for me to keep an overview of the whole task". Van der Waa et al. [97] reported that participants were not motivated to take over the role of supervisors when they had stress and lacked an overview during the task [97]. Therefore, we suggest that the participants who played computer games frequently could feel better when supervising the task and also process the explanations better.

6.4 Comparison of Sub-explanations

From the qualitative analysis comparing the consequential explanations with the counterfactuals, it was shown that the explanations can help participants in their next moral value elicitation. Although the quantitative analysis didn't show a significant difference, the qualitative results showed that counterfactual explanations were evaluated better in some aspects; e.g., interpreting the causality and showing the robot's understanding of human morality, while in others it was close to the evaluation of consequential explanations. Although this conclusion was reached when some participants of the conditional group sometimes did not read the second part of the explanation, i.e., the counterfactual, we still believe that since we provided complete explanatory content in the questionnaire, this could help participants to recall the task scenarios and reduce the impact of the lack of reading.

The above discussion is also consistent with the analysis of the answers to the unstructured questions. Participants thought the consequential explanations helped understand the robot's actions and were easy to read and understand. Although some participants believed that reading counterfactual explanations was less useful than observing robot actions, it was said that these explanations could provide more information and make the participants feel the robot's intelligence.

6.5 Limitations

We identify a few limitations of our study. First, from the feedback of some participants, it was mentioned that they couldn't finish reading all the explanations due to time pressure. One possible reason was that the explanation itself did contain too much information since counterfactual explanations described what the robot would do in two kinds of situations. Another problem might be that the explanations still contained a lot of text, even if we utilized figures and signs to replace part of it. Therefore, we should use more graphic elements instead of text messages.

Another limitation concerned the experiment interface. We underestimated the effect of the experimental environment interface on the results. One of the participants mentioned, "I need to not only focus on the actions of the robot but also the update of explanations. It took me much time to divert my sight because the explanations part was so far from the robot." This may also affect the participants' incomplete readings. Moreover, some participants claimed that

we should enlarge the figures of the victims since they couldn't remember the correct victim while doing the SA questionnaire.

6.6 Future Work

We identify some possible directions for future work. The design of the explanation was not perfect, neither on the content nor on the interface. More concise explanations instead of a combination of different types of explanations might be worthwhile to use. Moreover, more pictorial and well-placed explanations in the interface might be necessary to help the user to gain more feeling of exercised control.

We are still interested in what will affect the subjective feeling of MHC. Some possible factors can be human trust in the robots and the cognitive skills of the humans. Therefore, more experiments could be designed to investigate the influence.

Van der Waa et al. [97] raised three components for measuring MHC besides the subjective feeling. We could also investigate the link between explanations and the other two components: compliance with ethical guidelines and compliance with moral values. For example, we could use appropriate explanations to make the humans think the agent is more ethical.

7

Conclusion

In this research, we aimed to determine whether explanations could help with improving the subjective experience of Meaningful Human Control (MHC) in delayed control situations. If so, we would like to know which kind of explanations can reach the goal. We chose the Search and Rescue task as our scenario to answer the research question and specified the task settings and workflow. We asked a firefighter as the expert to conduct our expert study, in which we chose the proper explanation types from some previously used ones by other researchers and their combinations. We also improved our experimental design with our expert's advice. After the pilot study, we recruited participants to test our experiment, analyzed the results, and drew the following conclusions.

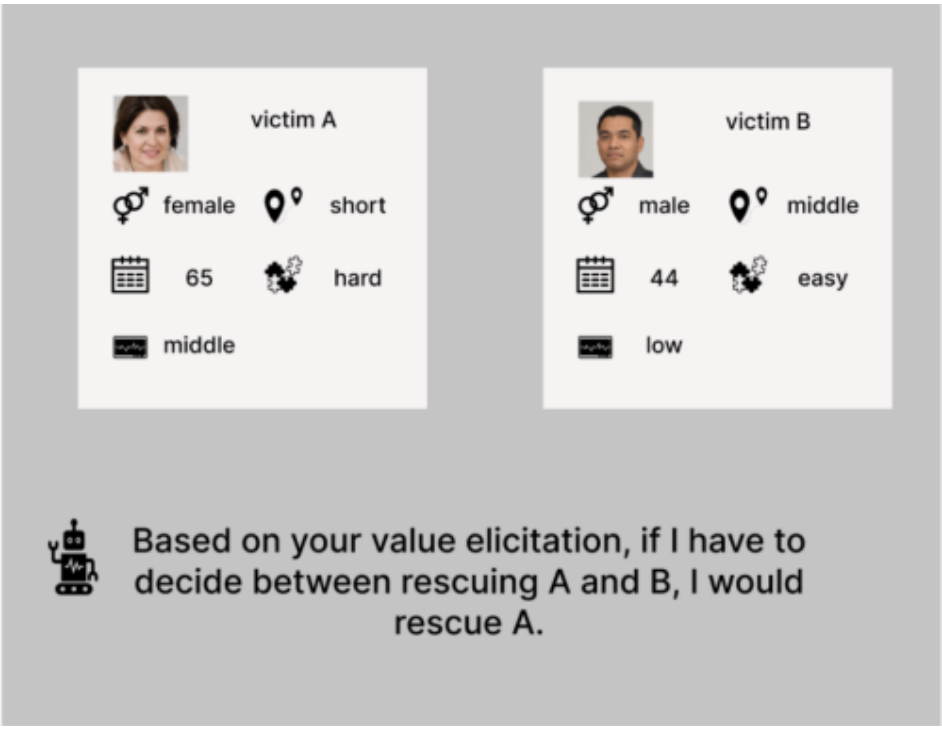
The role of explanations was overall positive as seen in the evaluations of the participants. The explanations could help humans on their next value elicitation and get more SA, especially on the projection level. Moreover, higher SA scores were associated with a better feeling of MHC. However, there was no clear evidence that the subjective experience of MHC could be affected by displaying explanations. Computer game frequency had a significant effect on the subjective feeling of MHC, as higher frequent gaming might result in a higher cooperative engagement and more trust in the robots. We also compared different sub-explanations on their performance, such as understandability and predictability. It was shown that counterfactual explanations seemed to perform better in general.

A

Ranking questions for Explanation in Expert Study

Ranking Questions for Explanations

Here are 5 explanations that could be provided by the robot, rank these explanations in order of relevance to the following statements.



Example A: Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A.

Example B: If you prioritized younger age preferred over gender, my decision would have been rescuing B rather than A.

Example C: Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A. because you prioritize distance the most, and victim A has a closer distance than victim B.

Example D: Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A. For the most important value – distance, A has a short distance.

Example E: Based on your elicitation before, if I have to decide to rescue A or B on the picture, I will rescue A. If you prioritized younger age preferred over gender, my decision would have been rescuing B rather than A.

Q1: This explanation provides the most new information

Example A

Example B

Example C

Example D

Example E

Q2: The explanation provides the most reasons about how to make decisions

Example A

Example B

Example C

Example D

Example E

Q3: I can understand the explanation best

Example A

Example B

Example C

Example D

Example E

Q4: The explanation shows the best robot's understanding of my moral value elicitation

Example A

Example B

Example C

Example D

Example E

Q5: This explanation matches my expectations best

Example A

Example B

Example C

Example D

Example E

Q6: I can predict how the robot will do after arranging it for rescue best

Example A

Example B

Example C

Example D

Example E

Q7: I can feel the most control about the system with the explanation

Example A

Example B

Example C

Example D

Example E

B

SA Questionnaire Before pilot study

info

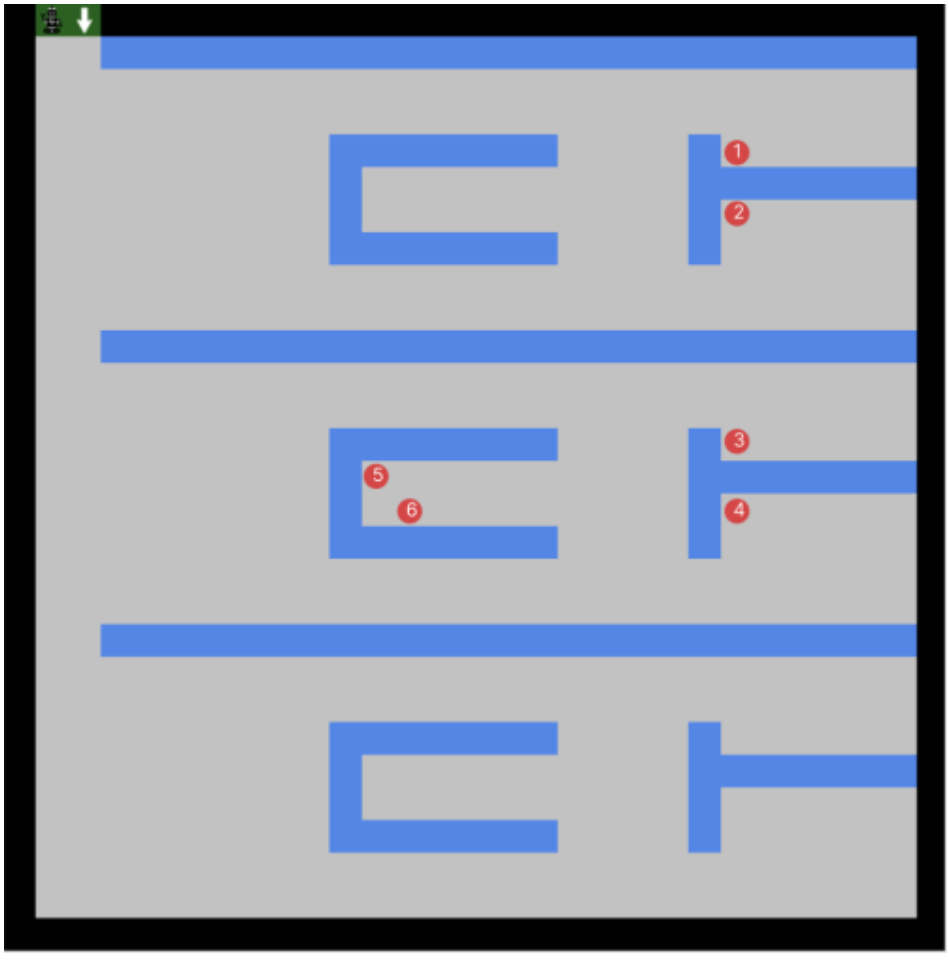
Q0. What is the group number and participant number? (Ask the instructor, eg: 1 2)

First Pause

Q1. How many victims were there in the danger zone?

- 1
- 2
- 3
- 4
- 5
- 6
- 7

Q2. Can you point out which place does Marielle  stuck in?



- 1
- 2
- 3
- 4
- 5
- 6

Q3. What is Pleun's  level of injury?

- low
- middle
- high

Q4. What is Marielle's  difficulty to rescue?

- low
- middle
- high

Q5. What is Franciska's  difficulty to reach?

- low
- middle
- high

Q6. Why do you think the robot will rescue **Pleun**  among them in this situation?

- Because of the gender
- Because of the age
- Because of the level of injury
- Because of the difficulty to reach
- Because of the difficulty to rescue

- [ ] [ ] None of above (please answer why)

Second Pause

Q7. How many seconds does it take for the robot to **reach** this victim?

- [ ] 3-5
- [ ] 6-8
- [ ] 9-11
- [ ] 12-14
- [ ] 15-17
- [ ] 18-20

Q8. How many seconds does it take for the robot to **rescue** this victim?

- [ ] 3-5
- [ ] 6-8
- [ ] 9-11
- [ ] 12-14
- [ ] 15-17
- [ ] 18-20

Third Pause

Q9. Who do you think the robot will rescue next?

- [ ]  Robbie
- [ ]  Lena
- [ ]  Simone

Q10. Why do you think the robot will rescue ${q://QID11/ChoiceGroup/SelectedChoices}?

- [ ] Because of the gender

- ○ Because of the age
- ○ Because of the level of injury
- ○ Because of the difficulty to reach
- ○ Because of the difficulty to rescue
- ○ [ ] None of above (please answer why)

Q11. How many seconds do you think it will take for the robot to reach ${q://QID11/ChoiceGroup/SelectedChoices}? (answer a number)

[ ]

C

SA Questionnaire

info

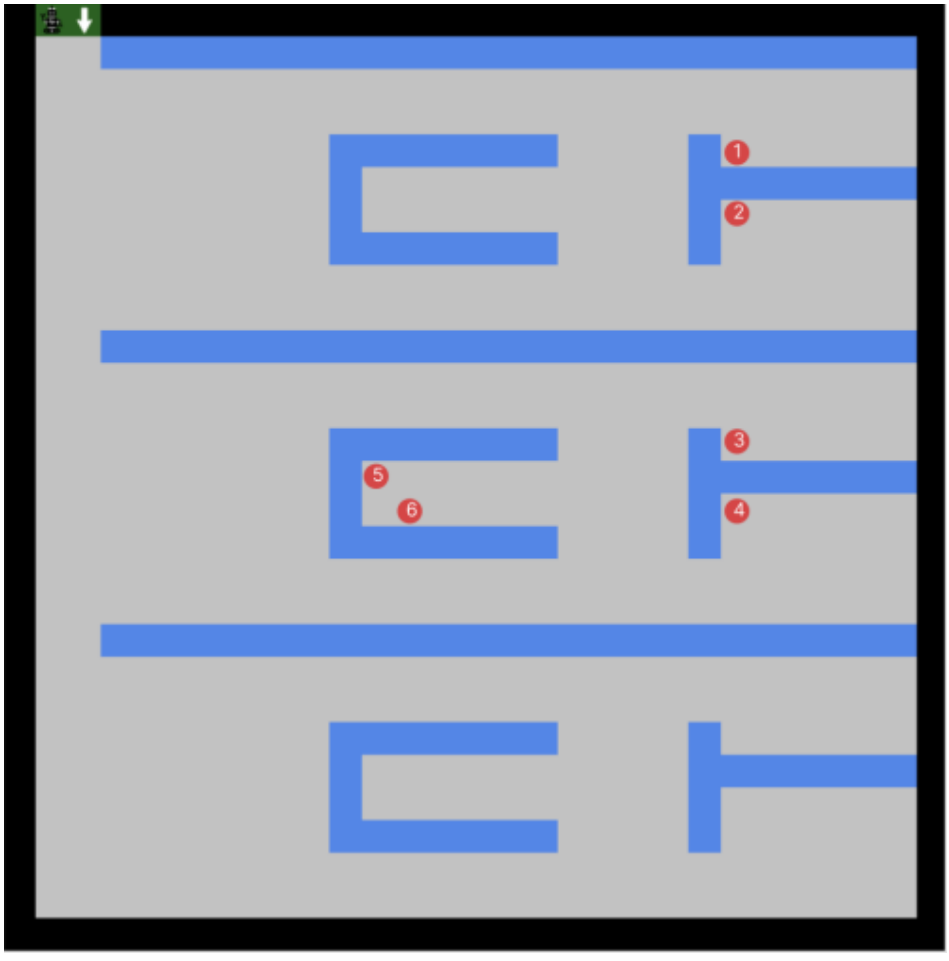
Q0. What is the group number and participant number? (Ask the instructor, eg: 1 2)

First Pause

Q1. How many victims were there in the danger zone?

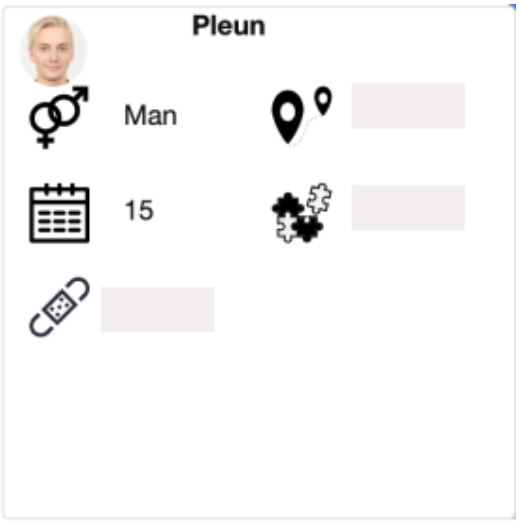
- 1
- 2
- 3
- 4
- 5
- 6
- 7

Q2. Can you point out which place does Marielle  stuck in?



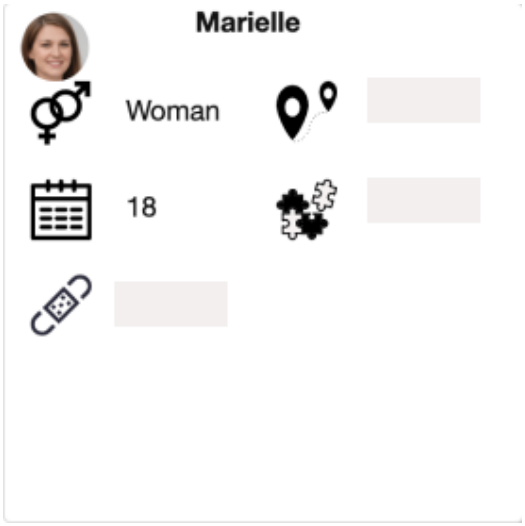
- 1
- 2
- 3
- 4
- 5
- 6

Q3. Could you fill in his level of characteristics?



	low	middle	high

Q4. Could you fill in her level of characteristics?



	low	middle	high
	○	○	○
	○	○	○
	○	○	○

Q5. Could you fill in her level of characteristics?

	low	middle	high
	○	○	○
	○	○	○
	○	○	○

Q6. Why do you think the robot will rescue the **second** victim among them in this situation?

- ○ Because of the gender
- ○ Because of the age
- ○ Because of the level of injury
- ○ Because of the difficulty to reach
- ○ Because of the difficulty to rescue
- ○ [ ] None of above (please answer why)

Second Pause

Q7. How many seconds does it take for the robot to **reach** this victim?

- 3-5
- 6-8
- 9-11
- 12-14
- 15-17
- 18-20

Q8. How many seconds does it take for the robot to **rescue** this victim?

- 3-5
- 6-8
- 9-11
- 12-14
- 15-17
- 18-20

Third Pause

Q9. Who do you think the robot will rescue next?

-  Robbie
-  Lena
-  Simone

Q10. Why do you think the robot will rescue ${q://QID11/ChoiceGroup/SelectedChoices}?

- Because of the gender
- Because of the age
- Because of the level of injury
- Because of the difficulty to reach
- Because of the difficulty to rescue

- ○ [ ] None of above (please answer why)

Q11. How many seconds do you think it will take for the robot to reach, rescue and bring ${q://QID11/ChoiceGroup/SelectedChoices} back? (answer a number)

[ ]

Q12. What would the robot do if Robbie  's level of injury was high rather than middle?

- ○ Robot will save him first
- ○ Robot will save him second
- ○ Robot will save him third

Q13. What would the robot do if Lena  's difficulty to rescue was low rather than high?

- ○ Robot will save her first
- ○ Robot will save her second
- ○ Robot will save her third

Q14. What would the robot do if Simone  's level of injury was high rather than low?

- ○ Robot will save her first
- ○ Robot will save her second
- ○ Robot will save her third

D

Information Survey and Control Questionnaire

Information - for group 1

Information

1. What is your number? (ask the instructor)

2. What is your gender?

- ◯ Woman
- ◯ Man
- ◯ Non-binary
- ◯ Prefer not to say

3. What is your age?

- ◯ Under 18
- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ above 55

4. What is your education level?

- ◯ High school diploma or less
- ◯ Some colledge, no degree
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ PhD

5. What is your experience with computer games?

- ◯ Several times a year
- ◯ Several times a month
- ◯ Several times a week
- ◯ Daily

Experience

6. Read following statements and select which option matches your feeling.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
It is simple for me to keep an overview of the whole task.	◯	◯	◯	◯	◯
I feel responsible for the well-being of the victims.	◯	◯	◯	◯	◯
I feel comfortable with all decisions made in the task.	◯	◯	◯	◯	◯
The robot's rescue choice matches what I thought.	◯	◯	◯	◯	◯
I can feel control over the system.	◯	◯	◯	◯	◯

Information - for group 2

Information

1. What is your number? (ask the instructor)

2. What is your gender?

- ◯ Woman
- ◯ Man
- ◯ Non-binary
- ◯ Prefer not to say

3. What is your age?

- ◯ Under 18
- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ above 55

4. What is your education level?

- ◯ High school diploma or less
- ◯ Some colledge, no degree
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ PhD

5. What is your experience with computer games?

- ◯ Several times a year
- ◯ Several times a month
- ◯ Several times a week
- ◯ Daily

Experience

6. Read following statements and select which option matches your feeling.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
It is simple for me to keep an overview of the whole task.	◯	◯	◯	◯	◯
I feel responsible for the well-being of the victims.	◯	◯	◯	◯	◯
I feel comfortable with all decisions made in the task.	◯	◯	◯	◯	◯
The robot's rescue choice matches what I thought.	◯	◯	◯	◯	◯
I can feel control over the system.	◯	◯	◯	◯	◯
I feel the explanation helped for my next value elicitation.	◯	◯	◯	◯	◯

7. Read following statements and select which sub explanation is more consistent with the statements.

	Based on your value elicitation, if I have to decide rescue between victimA and victimB, I will rescue victimA because of characterist icA.	If you would have swapped characterist icA and charactertis cticB, my decision would have been rescuing victimB rather than victimA.
can be understood better.	◯	◯
helps me to understand the causality better.	◯	◯
shows the robot's understandi ng of my moral value elicitation better.	◯	◯
helps me predict what the robot will do next better.	◯	◯

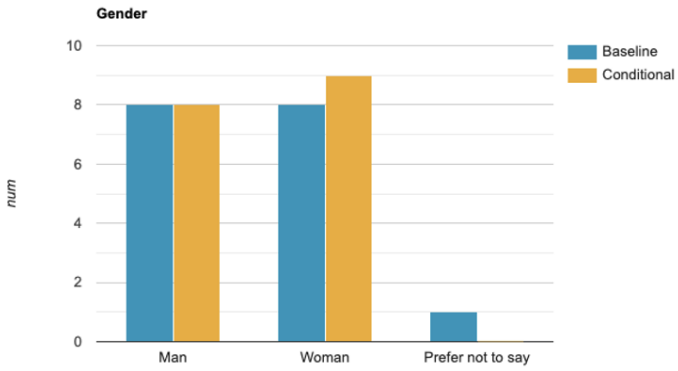
E

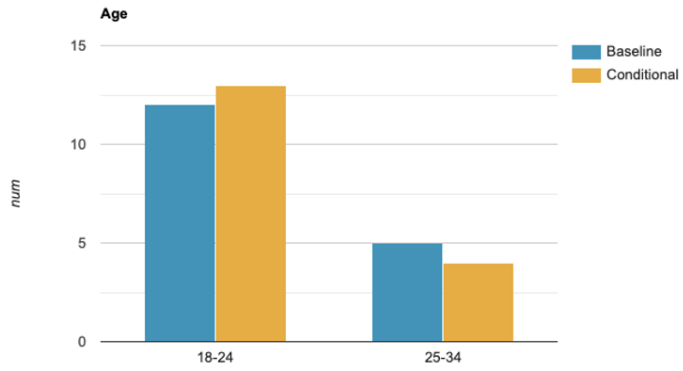
Victim Data

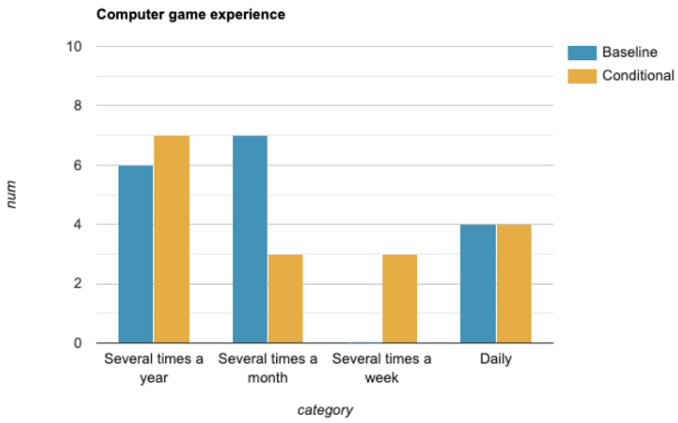
Index	Name	Gender	Age	Difficulty_to_reach	Difficulty_to_rescue	Location	Level_of_injury
0	Pleun	Male	15	middle	middle	12,15	low
1	Marielle	Female	18	middle	low	22,15	high
2	Franciska	Female	60	low	high	22,4	low
3	Manfred	Male	89	low	middle	11,14	middle
4	Aalt	Male	67	low	low	22,21	middle
5	Lucy	Female	60	low	high	25,6	middle
6	Wilhelmus	Male	19	middle	middle	22,16	high
7	Jayden	Male	23	middle	middle	11,15	high
8	Robbie	Male	82	middle	low	22,15	middle
9	Lena	Female	69	high	high	23,22	middle
10	Simone	Female	66	high	low	12,24	low

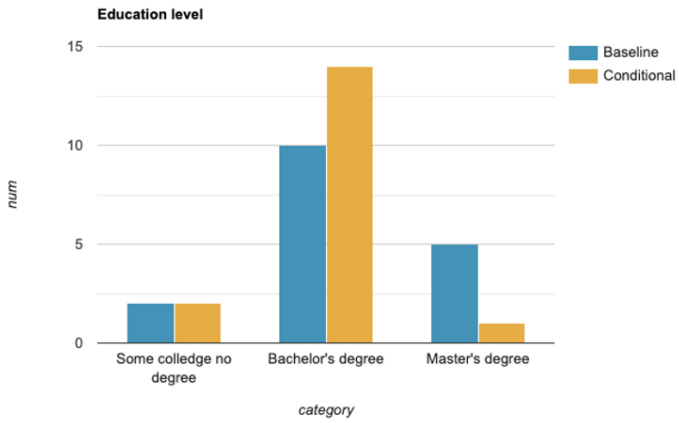
F

Participant Data


Gender
num
10
8
6
4
2
0
Man
Woman
Prefer not to say
Baseline
Conditional


Age
num
15
10
5
0
18-24
25-34
Baseline
Conditional


Computer game experience
num
10
8
6
4
2
0
Several times a year
Several times a month
Several times a week
Daily
category
Baseline
Conditional


Education level
num
15
10
5
0
Some colledge no degree
Bachelor's degree
Master's degree
category
Baseline
Conditional

Bibliography

[1] Allen, C., Smit, I., and Wallach, W. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology 7*, 3 (9 2005), 149–155.

[2] Alley, E. E. 26 Problems of search and rescue in disasters. Tech. rep., 1992.

[3] Amigoni, F., and Schiaffonati, V. Ethics for robots as experimental technologies: Pairing anticipation with exploration to evaluate the social impact of robotics. *IEEE Robotics & Automation Magazine 25*, 1 (2018), 30–36.

[4] Amoroso, D., and Tamburrini, G. Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues. *Current Robotics Reports* (2020), 1–8.

[5] Article36. Key areas for debate on autonomous weapons systems.

[6] Barb Ruppert. Robots to rescue wounded on battlefield, 11 2010.

[7] Battistuzzi, L., Recchiuto, C. T., and Sgorbissa, A. Ethical concerns in rescue robotics: a scoping review, 2021.

[8] Bellaby, R. W. Can AI Weapons Make Ethical Decisions? *Criminal Justice Ethics 40*, 2 (2021), 86–107.

[9] Bogue, R. Disaster relief, and search and rescue robots: the way forward. *Industrial Robot 46*, 2 (5 2019), 181–187.

[10] Bolstad, C. A., and Endsley, M. R. Shared Mental Models and Shared Displays: An Empirical Evaluation of Team Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 43*, 3 (9 1999), 213–217.

[11] Bracken, B., Tobyne, S., Winder, A., Shamsi, N., and Endsley, M. R. Can Situation Awareness Be Measured Physiologically? In *Advances in Neuroergonomics and Cognitive Engineering* (Cham, 2021), H. Ayaz, U. Asgher, and L. Paletta, Eds., Springer International Publishing, pp. 31–38.

[12] Bradshaw, J. M., Feltovich, P., and Johnson, M. Human-Agent Interaction. Tech. rep., 2002.

[13] Brandão, M. Socially Fair Coverage: The Fairness Problem in Coverage Planning and a New Anytime-Fair Method. Tech. rep., 2021.

[14] Brandão, M., Jirotka, M., Webb, H., and Luff, P. Fair navigation planning: A resource for characterizing and designing fairness in mobile robots. *Artificial Intelligence 282* (5 2020).

[15] Bulanda, S. *Ready!: Training the Search and Rescue Dog.* Fox Chapel Publishing, 2012.

[16] Calvert, S. C., Heikoop, D. D., Mecacci, G., and van Arem, B. A human centric framework for the analysis of automated driving systems based on meaningful human control. *Theoretical Issues in Ergonomics Science 21*, 4 (7 2020), 478–506.

[17] Carlsen, H., Johansson, L., Wikman-Svahn, P., and Dreborg, K. H. Co-evolutionary scenarios for creative prototyping of future robot systems for civil protection. *Technological Forecasting and Social Change 84* (5 2014), 93–100.

[18] Casper, J., and Murphy, R. R. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 33*, 3 (6 2003), 367–385.

[19] Cawthorne, D., and Robbins-van Wynsberghe, A. An Ethical Framework for the Design, Development, Implementation, and Assessment of Drones Used in Public Healthcare. *Science and Engineering Ethics 26*, 5 (10 2020), 2867–2891.

[20] Caylin White. How AI is Changing the Future of Digital Marketing, 11 2021.

[21] Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., and Ramos, F. Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics 26*, 2 (2020), 501–532.

[22] Cohen, P., and Levesque, H. J. Teamwork. Tech. Rep. 4, 1991.

[23] Cohen, P. R., Levesque, H. J., and Smith, I. A. On team formation. *Synthese Library* (1997), 87–114.

[24] Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., and Kramer, M. Moral Decision Making Frameworks for Artificial Intelligence. Tech. rep.

[25] Daniël D. Heikoop, Marjan Hagenzieker, Giulio Mecacci, Filippo Santoni de Sio, Simeon Calvert, and Bart van Arem. Meaningful Human Control over Automated Driving Systems. Tech. rep., 2018.

[26] De Bruycker, M., Greco, D., Annino, I., Stazi, M. A., de Ruggiero, N., Triassi, M., de Kettenis, Y. P., and Lechat, M. F. The 1980 earthquake in southern Italy: rescue of trapped victims and mortality. *Bulletin of the World Health Organization 61*, 6 (1983), 1021.

[27] de Sio, F. S., and van den Hoven, J. Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI 5*, FEB (2018).

[28] D.M. McGuigan, B.L. Deam, and D.K. Bull. USAR and the role of the engineer. *NZSEE 2002 Conference* (2002).

[29] Durso, F. T., and Gronlund, S. D. Situation awareness. *Handbook of applied cognition* (1999), 283–314.

[30] Ekelhof, M. Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation. *Global Policy 10*, 3 (9 2019), 343–348.

[31] Elzein, N. The demand for contrastive explanations. *Philosophical Studies 176*, 5 (2019), 1325–1339.

[32] Endsley, M. R. Situation Awareness in Aircraft Systems: Symposium Abstract. *Proceedings of the Human Factors Society Annual Meeting 32*, 2 (10 1988), 96–96.

[33] Endsley, M. R. Measurement of Situation Awareness in Dynamic Systems. Tech. Rep. 1, 1995.

[34] Endsley, M. R. A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. *Human factors 63*, 1 (2021), 124–150.

[35] Ewoldsen, D. R., Eno, C. A., Okdie, B. M., Velez, J. A., Guadagno, R. E., and Decoster, J. Effect of playing violent video games cooperatively or competitively on subsequent cooperative behavior. *Cyberpsychology, Behavior, and Social Networking 15*, 5 (5 2012), 277–280.

[36] Fan, X., and Yen, J. Modeling and simulating human teamwork behaviors using intelligent agents. *Physics of Life Reviews 1*, 3 (12 2004), 173–201.

[37] Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., and Siciliano, B. Autonomy in surgical robots and its meaningful human control. *Paladyn 10*, 1 (2019), 30–43.

[38] Gartner, I. Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form, 1 2019.

[39] Ginsberg ML. counterfactuals. *Artificial intelligence 30*, 1 (1986), 35–79.

[40] Graham-Rowe, D. Cheap drones could replace search-and-rescue helicopters. *New Scientist 207*, 2769 (2010), 20.

[41] Greatbatch, I., Gosling, R. J., and Allen, S. Quantifying search dog effectiveness in a terrestrial search and rescue environment. *Wilderness & environmental medicine 26*, 3 (2015), 327–334.

[42] Greenstreet Berman. Fire and Rescue Service response times: Fire Research Series., 2009.

[43] Greitemeyer, T., and Cox, C. There's no "I" in team: Effects of cooperative video games on cooperative behavior. *European Journal of Social Psychology 43*, 3 (4 2013), 224–228.

[44] Grogan, S., Montréal, P., and Gamache, M. The use of unmanned aerial vehicles and drones in search and rescue operations – a survey. Tech. rep., 2018.

[45] Gunkel, D. J. Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology 22*, 4 (12 2020), 307–320.

[46] Haije, T. Learning Human Intention for Taskable Agents.

[47] Hara, S., Ikeno, K., Soma, T., and Maehara, T. Feature Attribution As Feature Selection.

[48] Harbers, M., Bradshaw, J. M., Johnson, M., Feltovich, P., Van Den Bosch, K., and Meyer, J.-J. Explanation in Human-Agent Teamwork. Tech. rep., 2012.

[49] Harbers, M., de Greeff, J., Kruijff-Korbayová, I., Neerincx, M. A., and Hindriks, K. V. Exploring the ethical landscape of robot-assisted search and rescue. In *A World with Robots*. Springer, 2017, pp. 93–107.

[50] Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., and van Arem, B. Human behaviour with automated driving systems: a quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science 20*, 6 (11 2019), 711–730.

[51] Hoffman, R. R., Ford, K. M., Feltovich, A., Woods, D. D., Feltovich, P. J., and Klein, G. A Rose by Any Other Name…Would Probably Be Given an Acronym. *IEEE Intelligent Systems 17*, 4 (2002), 72–80.

[52] Horowitz, M. C., and Scharre, P. MEANINGFUL HUMAN CONTROL in WEAPON SYSTEMS: A Primer. Tech. rep., 2015.

[53] Inselberg, A. Multidimensional detective. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium* (1997), IEEE, pp. 100–107.

[54] Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., and Goldberg, Y. Contrastive Explanations for Model Interpretability.

[55] Jiang, J., and Karran, A. A Situation Awareness Perspective on Human-Agent Collaboration: Tensions and Opportunities.

[56] Jones, P. Urban Search and Rescue Training in Australia. *Fire Engineers Journal 57* (1997), 11–13.

[57] Kaber, D., Kaber, D. B., and Endsley, M. R. Team situation awareness for process control safety and performance Team Situation Awareness for Process Control Safety and Performance zyx. Tech. rep.

[58] Karaca, Y., Cicek, M., Tatli, O., Sahin, A., Pasli, S., Beser, M. F., and Turedi, S. The potential use of unmanned aircraft systems (drones) in mountain search and rescue operations. *American Journal of Emergency Medicine 36*, 4 (4 2018), 583–588.

[59] Katzenbach, J. R., and Smith, D. K. *The discipline of teams*. Harvard Business Press, 2008.

[60] Koopman, B. O. The theory of search. I. Kinematic bases. *Operations research 4*, 3 (1956), 324–346.

[61] Kring, J., and Mouloua, M. Workload, Situation Awareness, and Teaming Issues for UAV/UCAV Operations Cite this paper Human-Cent ered Design of Unmanned Aerial Vehicles. Tech. rep., 2001.

[62] Kuorikoski, J., and Ylikoski, P. External representations and scientific understanding. *Synthese 192*, 12 (2015), 3817–3837.

[63] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-Velez, F. Human Evaluation of Models Built for Interpretability. Tech. rep., 2019.

[64] Larasati, R., De Liddo, A., and Motta, E. The Effect of Explanation Styles on User's Trust. Tech. rep., 2020.

[65] Lee, Y. H., Jeon, J.-D., and Choi, Y.-C. Air Traffic Controllers' Situation Awareness and Workload under Dynamic Air Traffic Situations. *Transportation Journal 51*, 3 (7 2012), 338–352.

[66] Lenoxt, T., Payne, T., Hahnt, S., Lewist, M., and Sycara, K. MokSAF: How should we support teamwork in human-agent teams? Tech. rep., 1999.

[67] Lewis, D. K. Causal Explanation. 214–240.

[68] Lipton, P. Contrastive Explanation. *Royal Institute of Philosophy Supplement 27* (3 1990), 247–266.

[69] Liu, X., Chi, H.-y., and Xu, X.-d. Seismic emergency organizational structure and technical support platform for 72-hour gold rescue time. In *Proceedings of International Conference on Information Systems for Crisis Response and Management (ISCRAM)* (2011), IEEE, pp. 542–547.

[70] Liu, Y., and Nejat, G. Robotic urban search and rescue: A survey from the control perspective, 11 2013.

[71] Lunde, A., and Tellefsen, C. Patient and rescuer safety: Recommendations for dispatch and prioritization of rescue resources based on a retrospective study of Norwegian avalanche incidents 1996-2017. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 27*, 1 (1 2019).

[72] Lutvica, K., Velagic, J., Kadic, N., Osmic, N., Dzampo, G., and Muminovic, H. Remote path planning and motion control of mobile robot within indoor
}maze environment. In *2014 IEEE International Symposium on Intelligent Control, ISIC
}2014, Juan-les-Pins, France, October 8-10, 2014 (2014), IEEE, pp. 1596–1601.*

[73] Matthias, A. *The responsibility gap: Ascribing responsibility for the actions of learning automata.* Ethics and INformation Technology *(2004), 175–183.*

[74] Mecacci, G., and Santoni de Sio, F. *Meaningful human control as reason-responsiveness: the case of dual-mode vehicles.* Ethics and Information Technology 22*, 2 (6 2020), 103–115.*

[75] Miller, T. *Explanation in artificial intelligence: Insights from the social sciences, 2 2019.*

[76] Mohseni, S., Zarei, N., and Ragan, E. D. *A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems.* ACM Transactions on Interactive Intelligent Systems 11*, 3-4 (12 2021), 1–45.*

[77] Murphy, R. R. *Human-Robot Interaction in Rescue Robotics. Tech. rep., 2004.*

[78] National Search and Rescue Supplement. *National Search and Rescue Plan of the United States. Tech. rep., 2007.*

[79] Neerincx, M. A., van der Waa, J., Kaptein, F., and van Diggelen, J. *Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance. In* Engineering Psychology and Cognitive Ergonomics *(Cham, 2018), D. Harris, Ed., Springer International Publishing, pp. 204–214.*

[80] Nguyen, H. T., Topolsky, N. G., Tarakanov, D. V., and Mokshantsev, A. V. *Multicriteria analysis of firefighter routes in buildings in the case of a fire.* The Journal of Defense Modeling and Simulation *(2020), 1548512920948611.*

[81] Paivio, A. Mental representations: A dual coding approach. *Oxford University Press, 1990.*

[82] Quwaider, M., Alabed, A., and Duwairi, R. *The Impact of Video Games on the Players Behaviors: A Survey.* Procedia Computer Science 151 *(1 2019), 575–582.*

[83] Raitio, R., and Leppävaara, L. *Human Rights in Urban Search and Rescue. Tech. rep., 2016.*

[84] Richardson, H. R., and Stone, L. D. *Operations analysis during the underwater search for Scorpions.* Naval Research Logistics Quarterly 18 *(1971), 141–157.*

[85] Rogers, E., Murphy, R., and Burke, J. *Nsf/darpa study on human-robot interaction.* IEEE Transactions on Systems, Man and Cybernetics, special issue on Human-Robot Interaction *(2004).*

[86] Salas, E. E., and Fiore, S. M. Team cognition: Understanding the factors that drive process and performance. *American Psychological Association, 2004.*

[87] Salmon, P. M., Stanton, N. A., Jenkins, D. P., Walker, G. H., Young, M. S., and Aujla, A. *LNAI 4562 - What Really Is Going on? Review, Critique and Extension of Situation Awareness Theory. Tech. rep., 2007.*

[88] Sarter, N. B., and Woods, D. D. *Situation awareness: A critical but ill-defined phenomenon.* The International Journal of Aviation Psychology 1*, 1 (1991), 45–57.*

[89] Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., and van den Bosch, K. *Human-centered XAI: Developing design patterns for explanations of clinical decision support systems.* International Journal of Human-Computer Studies 154 *(10 2021), 102684.*

[90] Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., and Findlater, L. *No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems *(2020), pp. 1–13.*

[91] Stanton, N. A., Salmon, P. M., Walker, G. H., and Jenkins, D. P. *Is situation awareness all in the mind?* Theoretical Issues in Ergonomics Science 11*, 1-2 (1 2010), 29–40.*

[92] Statheropoulos, M., Agapiou, A., Pallis, G. C., Mikedi, K., Karma, S., Vamvakari, J., Dandoulaki, M., Andritsos, F., and Thomas, C. L. P. *Factors that affect rescue time in urban search and rescue (USAR) operations.* Natural Hazards 75*, 1 (2015), 57–69.*

[93] Sun, X., Zhang, Y., and Chen, J. *High-level smart decision making of a Robot based on ontology in a search and Rescue Scenario.* Future Internet 11*, 11 (11 2019).*

[94] U.S. Coast Guard Research and Development Center. *REVIEW OF SEARCH THEORY: ADVANCES AND APPLICATIONS TO SEARCH AND RESCUE DECISION SUPPORT. Tech. rep., 2001.*

[95] van der Waa, J., and Haije, T. *MATRX Software, 2019.*

[96] van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., and Neerincx, M. *Contrastive Explanations with Local Foil Trees.*

[97] van der Waa, J., Verdult, S., van den Bosch, K., van Diggelen, J., Haije, T., van der Stigchel, B., and Cocu, I. *Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations.* Frontiers in Robotics and AI 8 *(5 2021).*

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

[99] Vidulich, M., Dominguez, C., Vogel, E., and McMillan, G. *Situation awareness: Papers and annotated bibliography. Tech. rep., 1994.*

[100] Walker, D., and Myrick, F. *Grounded theory: An exploration of process and procedure.* Qualitative health research 16*, 4 (2006), 547–559.*

[101] Williams, A., Sebastian, B., and Ben-Tzvi, P. *Review and Analysis of Search, Extraction, Evacuation, and Medical Field Treatment Robots.* Journal of Intelligent and Robotic Systems: Theory and Applications 96*, 3-4 (12 2019), 401–418.*

[102] Yamauchi, B. *PackBot: A Versatile Platform for Military Robotics. Tech. rep., 2004.*

[103] Zeagler, C., Byrne, C., Valentin, G., Freil, L., Kidder, E., Crouch, J., Starner, T., and Jackson, M. M. *Search and rescue: dog and handler collaboration through wearable and mobile interfaces. In* Proceedings of the Third International Conference on Animal-Computer Interaction *(2016), pp. 1–9.*