# Designing Adjustable Work Agreements for Hybrid Intelligent Disaster Management

Advancing Situation Awareness and Task Allocations of First Response Human-Agent Teams

Author:

Marit Leijnse, s1028966, Marit.Leijnse@student.ru.nl

Supervisors:

Prof. Dr. Mark Neerincx[1]
Dr. Serge Thill[2]

[1] Department of Human-Machine Teaming
Expert group of Defence, Safety and Security
TNO

[2] Department of Artificial Intelligence
Faculty of Social Sciences
Radboud University Nijmegen

13-06-2021

**Acknowledgements**

The participation and assistance of so many people made this thesis possible; however, special thanks to the following people:

My supervisors Prof. Dr. Mark Neerincx, and Dr. Serge Thill. Even though we were not able to meet in person due to COVID-19, I believe that we were able to make the most out of it. I would like to thank both of you for the continuous support throughout the whole process.

Beside my supervisors I would like to thank the project manager of ASSISTANCE for the valuable discussions and constructive feedback on my research. I am very grateful for having had the opportunity to be part of the ASSISTANCE project group.

Apart from that, I would like to thank the colleagues from TNO, and the firefighters from the Gezamenlijke Brandweer and Spaansland fire brigade in Enschede who participated in the research for participating. I have learned a lot from you about human-agent teaming in the disaster management domain. Without your help, I would not have been able to complete the thesis.

Furthermore, I would like to thank the fellow TNO interns I met during my time at TNO. The time I could spend with you in the past nine months will not be forgotten.

Finally, I would like to thank my family and friends for their love, care, understanding, and constant encouragement throughout the research period. Your help elevated the thesis to a higher level.

Radboud University, Nijmegen, Marit Leijnse, 09-06-2021

**Abstract.** This research focuses on improving first responders' situation awareness with support from an artificial intelligent agent for decisions with a value-sensitive aspect based on work agreements. To indicate the division of roles, responsibilities, and capabilities of the first responder, a team leader, and an artificial intelligent agent in the decision-making process, three Team Design Patterns were designed, evaluated, and simulated. The patterns describe how work agreements are collaboratively managed to agree on the human-AI collaboration explicitly and how these work agreements are used during disaster management to collaboratively deal with adversarial changes in weather conditions that threaten the first responder's safety. The work agreements addressed in this research indicate who to involve in the interaction, what the notification type should be, and whether task transfer from human to robot should be activated. A formal work agreement model was made to show how work agreements can be managed between more than two actors. The task allocation and information exchange between the first responder, team leader, and AI agent is simulated using the Python Rapid Experimentation software package (MATRX). In total twenty-four participants contributed to this research: eleven human-agent teaming experts from TNO (Artificial Intelligence and Human Factors researchers) and thirteen professional fire fighters from the Gezamenlijke Brandweer (five) and the Spaansland fire brigade in Enschede (eight). Results show that user-tailored notifications and behaviours of the AI agent through work agreements likely improve team performance for value-sensitive decisions significantly.

**Keywords**: Actionable Situation Awareness, Disaster Management, Human-AI Collaboration, Team Design Patterns, Value-Sensitive Decision-Making, Work Agreements

## TABLE OF CONTENTS

## Glossary

| Term (abbreviation) | Description |
|---|---|
| Artificial intelligence agent (AI agent) | An autonomous entity that interacts with the environment to achieve a certain goal. The entity does this by using sensors in the environment to be deriving knowledge from it and act on this knowledge using actuators. |
| AwareneSS tools and tailored training scenarios for increaSing capabiliTies and enhANcing the proteCtion of First RespondErs (ASSISTANCE) | "An international research project commissioned under the Horizon 2020 program in Secure Societies Challenge addressing the SU-DRS02-2018-2019-2020 (Technologies for first responders) topic. The aim of ASSISTANCE is to help and protect FR during the mitigation of large disasters by enhancing their capabilities and skills for facing complex situations to different types of incidents." For more information, see: https://assistance-project.eu/ |
| Commander | A person who directs all fire brigade units in case of mono-disciplinary response and other first response units such as the medical assistance (GHOR) and police. In Dutch this is called the 'Officier van Dienst (OvD)'. |
| Disaster management (DM) | The process of organizing resources and responsibilities of preparedness, response, and recovery of disasters in order to lessen the impact of the disaster. |
| First responder (FR) | A person who is one of the first people to arrive at the location of the disaster to manage the disaster, for example a paramedic, police officer, or fire fighter. In this study, attention is given to the role of fire fighters during disaster management. When we talk about a FR, this person can be seen as an expert in first response and disaster management. |
| Hybrid intelligent system (HIS) | A network of agents with human or Artificial Intelligence. Combining the strengths of human intelligence and Artificial Intelligence can augment and amplify the overall intelligence in within the network. |
| Human-Agent teaming (HAT) | The collaboration between a human agent and an autonomous agent to reach a specific goal. The autonomous agent can be a robotic or a software agent embedded in hardware. This study focuses on the collaboration between a human agent and an artificial intelligent software agent embedded in a mobile device. An expert of human-agent teaming has either knowledge of human factors research or Artificial Intelligence development, which are relevant for human-agent teaming research. |
| Infield fire fighter (infield FF) | A fire fighter who is part of the fire watch and who is sent to the source of the disaster to extinguish fires, to resolve leakages of hazardous substances, to provide technical assistance, and to search and rescue victims. |
| Rapid Experimentation software package (MATRX) | A python software package developed at TNO, which allows users to design human-agent and/or multi-agent team tasks for experimentation or simulation purposes. |
| Team Design Patterns (TDPs) | A methodology that focuses on the requirements of human agents and AI agents for human-agent teaming. Team Design Patterns allow for describing the division of roles, responsibilities, and task |

| | |
|---|---|
| | allocation between a human agent and an AI agent. And the level of involvement of both agents in the teamwork. |
| Team leader | A person who directs a crew of fire fighters that is working at the source of the disaster. In Dutch this is called the 'bevelvoerder'. |
| Work Agreement (WA) | An action guiding statement, which is used to reach a certain goal (norm) that influences the choice of action of agents to promote or demote values. They function as a methodology to reason about (and verify) obligations, permissions, and prohibitions of actions or goals via explicit agreements between a debtor and a creditor. A debtor is an actor that is committed to the execution of the work agreement, and a creditor is an actor to which the commitment to the execution of the work agreement is owed. Work agreements are applicable for any situation in which two agents with different kinds of models for decision-making must cooperate. They act as a bridge in knowledge and understanding. |
| Value-sensitive decision | A decision for which value consideration is integrated in the decision-making process. These decisions are complex as they often involve problems for which several solutions align with the decision-makers values. This can result in value tensions (the situation in which supporting a value challenges another value). Examples of values are security, safety, privacy, loyalty, and teamwork. |
| Mobile Sensing for Safety (MoSeS) | A system developed by TechForFuture in collaboration with the associated Fire Brigade Twente, Thales, and Lucia, which enables fire fighters to operate safer and more efficiently in complex buildings. For more information, see: https://www.brandweer.nl/twente/nieuws-twente/2017/twente-werkt-veiliger-met-moses and https://techforfuture.nl/onderzoek/afgerond/mobile-sensing-safety-moses/ |

# List of figures

# List of text boxes



# List of tables

# 1. Introduction

Industrial incidents are considered to be a major risk by most disaster management (DM) authorities across the European Union (see page 30 of the report of the Directorate-General for European Civil Protection and Humanitarian Aid Operations (ECHO) of the European Commission) [30]. When such a disaster occurs, first responders (fire fighters, paramedics, and police officers) are the first auxiliary troops to arrive on the scene to assist those involved or affected by the event. First responders (FR) are trained to handle the causes (sources) and consequences (effects) of disasters. The tasks of FRs are, amongst others, to warn and evacuate citizens, search for and rescue victims, provide medical aid to injured people, extinguish fires, and resolve leakages and spills of dangerous substances. However, precisely because FRs are among the first dispatched, they enter the disaster scene with limited knowledge of what has happened, what is happening, and what might happen. The information FRs receive and have to process from the dynamic, incomplete, and sometimes uncertain environment continuously impacts their task planning and performance.

The safety region Rotterdam-Rijnmond has the biggest industrial and port area in Europe, which houses 79 businesses that work with hazardous substances each day [1]. These 79 companies together make up 20% of all companies that work with hazardous substances in the Netherlands[2]. The Gezamenlijke Brandweer is trained to help these businesses control hazardous substances and restore safe use of the specific area. To be able to work and make decisions in this demanding environment, the firefighters from the Gezamenlijke Brandweer must be able to assess what has happened, what is happening, and what could happen.

Situation awareness allows FRs to anticipate what is happening and what might happen around them [21]. It gives them the ability to perceive and understand information in the environment to predict what to do next [21]. Suppose a FR has situation awareness of a potential hazard or threat in the environment. In that case, they can perceive, understand, and predict the situation so that they can react more adequately to the situation and avoid being harmed by the potential hazard or threat. Take for example a situation in which a FR is operating in a building in which a fire is burning without an oxygen tank and mouth mask. If the FR perceives the smoke, it can comprehend that something is off and that staying in the building without an oxygen tank and mouth mask will result in, amongst others, toxic air in the lungs from which they might get unwell or even worse unconscious. Based on this comprehension, it can be predicted by the FR what action is best to execute to get out of this toxic air. However, if the FR would not have perceived the smoke in time, it could already have been too late to escape the fire.

To increase the situation awareness of FRs, Artificial Intelligence (AI) systems are increasingly deployed for a wide variety of tasks with the aim to improve safety and performance. However, to improve safety and performance, the technology deployed must complement human capabilities and relieve human actors of demanding tasks [1]. The question, therefore, is how tasks can be allocated between the human agent and the AI agent such that the potential of each of the team members is as optimal as possible and leads to improved team performance?

Neerincx et al. [1] describe the interaction between humans and technology as a (hybrid intelligent) system that encompasses networked Humans, Agents, and Robots in a Teamwork (HART) with joint knowledge-based behaviours [1]. These joint knowledge-based behaviours within a

[1] DCMR Milieudienst Rijnmond, Brzo-bedrijven op kaart, Accessed on: June 7, 2021. [Online]. Available: https://brzoviewer.pzh.nl/

[2] Provincie Zuid-Holland. Risicovolle bedrijven. Accessed on: June 7, 2021. [Online]. Available: https://www.zuid-holland.nl/onderwerpen/veiligheid/risicovolle/

(hybrid intelligent) system are also known as Human-Agent teaming (HAT). To structure HAT research, Team Design Patterns (TDPs) can be used (see [4]-[7]). The use of TDPs during the design phase of (semi-)autonomous systems may help researchers communicate with end-users, AI developers, and Human Factors researchers about the requirements of (joint) knowledge-based behavior in hybrid intelligent systems (HIS) as they describe successful behaviour. The value of TDPs is that they illustrate concisely and intuitively the minimum requirements needed for HAT (roles, responsibilities, and capabilities) [6].

The study of Van der Waa et al. [6] is used by Van Stijn et al. [7] to apply TDPs in the design of FATE for allocation of moral decision-making in human-agent teams. Van Stijn et al. [7] uses TDPs to deal with the design problem of bias mitigation and data sharing for digital assistance for patients and healthcare professionals to prevent, diagnose, treat, and manage diabetes type II. It is the first attempt to utilize the taxonomy of TDPs by Waa et al. [6] in the design of a system for allocation of human-computer teams in morally sensitive situations and to evaluate its usability for a multi-disciplinary research & development group.

## 1.1 Problem statement, motivation, and research objective

Above, I sketched an example in which I highlighted the importance of situation awareness for a fire fighter in case of a fire in which they are not carrying an oxygen tank or wearing a mouth mask. With this example, I illustrated how important it is for a firefighter to perceive smoke to comprehend that the air is toxic and project that the situation should be left before they will lose consciousness. This thought process in which the situation is assessed by following the steps of perception, comprehension, and projection can help the fire fighter decide about the action that should be taken to bring themselves back to safety. However, the situation a firefighter has to deal with can be much more complex than described above.

In the case of DM, such as incident control of hazardous substances, the situation is complex, uncertain, and sometimes incomplete, which means that a FR (e.g., fire fighter) does not always have the ability to perceive the necessary information to comprehend the consequences of the situation and to project the necessary action. It could, for example, be that there is a child in the affected location, which makes the decision to leave the location to bring yourself back to safety difficult (a value-sensitive decision [7]).

Traditionally, the tasks involved in incident control of hazardous substances are performed by humans. However, several fire brigades in different countries, like the Gezamenlijke Brandweer in the Netherlands, are exploring the deployment of unmanned intelligent systems such as Unmanned Aerial Vehicles (AUVs), Unmanned Ground Vehicles (UGVs), and embedded AI agents [3]. For example, a first response team can be supported with technology by means of an AI agent that monitors, supports, and advises each team member during deployment [2], [3]. The AI agent could monitor the environment for the FR and provide early warnings to an FR that possibly is affected by adversarial changes in the environment, such that the FR does not continuously has to monitor the environment themselves while working on their task. Dividing attention between the task and monitoring for threats brings the risk of missing important information when the attention is drawn to the task instead of the environment (or the other way around).

Continuing on the recent TDP studies mentioned above ([4]-[7]), this research project focusses on HAT for decisions with a value-sensitive aspect in the DM domain. Furthermore, the

[3] Brandweer Nederland. Proeftuinen onbemenste inzet van start. April 11, 2018. Accessed: March 18, 2021. [Online]. Available: https://www.brandweer.nl/brandweernederland/nieuws/2018/proeftuinen-onbemenste-inzet-van-start

Socio-Cognitive Engineering (sCE) methodology is used to consider human values throughout the design process and contribute to responsible design, development, and deployment of AI [15]. Using the sCE methodology, researchers pay attention to the requirements and work agreements (WAs) of human-AI teaming by looking into the effect of the assigned roles, capabilities, appearance, and behaviour of the AI on stakeholders.

The research project presented in this thesis is part of the ASSISTANCE project [4]. It investigates the division of roles and responsibilities for the decision-making process in a HIS, constituting of humans (e.g., fire fighters), Unmanned Arial Vehicles (UAVs), Unmanned Ground Vehicles (UGVs), and (embedded) AI agents. The objectives of the work presented in this thesis are 1) to design several TDPs as a generic solution for how an AI agent can support a FR in making decisions with a value-sensitive aspect, and 2) to evaluate and simulate the design solution using input from HAT and DM experts. To simulate the design solution the Python Rapid Experimentation software package (MATRX) was used.

Three TDPs were created using a specific scenario of decision support by an AI agent for decisions with a value-sensitive aspect during incident control of hazardous substances. In this scenario, a fire fighter and an (embedded) AI agent collaborate to resolve a leakage of hazardous substances. The goal is to resolve the leakage without endangering the safety and health of the fire fighter. To reach this goal, WAs are managed at the start of the DM to agree explicitly on the human-AI collaboration during the DM.

WAs provide agents with knowledge and understanding of the behaviour of other agents via explicit agreements on what is obligated, allowed, and prohibited. Therefore, it is assumed that considering what kind of WAs should be established can contribute to reaching the desired level of flexibility in the collaboration between the fire fighter and the AI agent.

A formal work agreement model was created to describe how such explicit agreements can be made for human-AI collaboration for multi-human, multi-robot teams. This model describes how adaptive WAs can be managed between more than two actors.

The research question that is addressed within this research is: "How can an AI agent support a FR with solving a recurrent decision-making problem with a value-sensitive aspect?”. To answer this question the following sub-questions about the design of the scenario and the use of TDPs are addressed:

- Design scenario:
    - What is a representative example of a recurring decision-making problem with a value-sensitive aspect in the ASSISTANCE scenario?
    - What kind of work agreements could help to establish the desired level of flexibility in the human-AI collaboration (e.g., choosing between advice or information provision)?

- Team Design Pattern (TDP):
    - Does the TDP specification provide a clear, concise, complete, and easy-to-understand description of the human-AI decision process for different experts (i.e., first response, human factor, and AI experts)?

[4] For information about the ASSISTANCE project see Section 2.1 or go to the website of the ASSISTANCE project: https://assistance-project.eu/

- Does the TDP specification describe collaborative AI behaviors that are expected to improve team performance with a value-sensitive aspect substantially?

## 1.2 Thesis outline

Chapter 2 outlines the background relevant for this research by describing the context of this project and providing insight into the relevant literature. Chapter 3 describes the design solution for human-agent teaming for decisions with a value-sensitive aspect in the disaster management domain and how this design solution was derived from literature and expert knowledge. Chapter 4 describes how the design solution is evaluated by human-agent teaming and disaster management domain experts. Chapter 5 describes how a knowledge base was made that provides the AI agent with the knowledge needed to be able to support FRs during disaster management with decisions with a value-sensitive aspect. Chapter 6 describes how the design solution is simulated using the Python Rapid Experimentation software package (MATRX). And finally, in chapter 7, the general conclusions and discussion are provided. Within this chapter, an answer is given to the research question and what this means for future research.

# 2. Background

This chapter describes the background relevant for this study. Section 2.1 describes the context of this project. After this, six related topics are discussed, namely: decision-making with a value-sensitive aspect (Section 2.2), the Value-Sensitive Design methodology (Section 2.3), Socio-Cognitive Engineering methodology (Section 2.4), Work agreements (Section 2.5), Team Design Patterns (Section 2.6) and Knowledge representation and reasoning (Section 2.7).

## 2.1 The ASSISTANCE project and TNO's role

ASSISTANCE is an international research project commissioned under the Horizon 2020 program in Secure Societies Challenge addressing the SU-DRS02-2018-2019-2020 (Technologies for FRs) topic. ASSISTANCE stands for AwareneSS tools and tailored training scenarios for increaSing capabiliTies and enhANcing the proteCtion of First RespondErs. The aim of ASSISTANCE is to help and protect FR during the mitigation of large disasters by enhancing their capabilities and skills for facing complex situations during different types of incidents.

In total 19 partners from 8 European countries participate in the ASSISTANCE project. TNO takes part in the ASSISTANCE project with the development of a chemical hazards tool. The aim of developing the chemical hazards tool is to improve the situation awareness of FRs with information on potentially hazardous situations (due to the dispersion of toxic gas) and the effect of these situations on the safety of FRs and the environment (e.g., citizens). Based on this information, the FRs can decide how to approach the situation and mitigate the disaster.

FRs from different organisations will test the ASSISTANCE platform under controlled conditions in three different demonstration pilots of large and complex (natural or man-made) disasters, namely an earthquake, an industrial accident, and a terrorist attack.[5]

To help and protect FR during the mitigation of large disasters, such as the ones described above, with improved situation awareness of potentially hazardous situations and the effects of these situations on the safety of FRs, a better understanding of the factors that influence the decision-making process during the missions (e.g., sensitive values and value tensions) is required. Therefore, below, a closer look has been given to human factors research about situation awareness and value-sensitive decision-making; coordination and distribution of tasks in human-agent teams; and work agreements.

## *2.2 Situation awareness and decision-making*

Situation awareness influences our decision-making process and allows us to anticipate what has happened, what is happening, and what might happen around us [21], which is especially valuable in situations with high information flow where wrong decisions have serious consequences [22], [23]. A high level of situation awareness results in a high degree of knowledge of the situation with respect to input and output of the decision-making process [22], whereas the lack of situation awareness is likely a primary factor in accidents [23], [24]. Such as the phenomenon described in the introduction with the example of a fire fighter in a building on fire without an oxygen tank and mouth mask. In this example, the perceived indicator of smoke could help the firefighter act on bringing himself back to safety.

[5] For more information about the ASSISTANCE project see: https://assistance-project.eu/

Understanding the decision-making process, situation awareness, and the factors that influence the decision-making process and situation awareness (e.g., values and value tensions) is crucial as it can help improve the safety and performance of FRs (firefighters).
In 'Group situation awareness and distributed decision making: From military to civilian application' Wellens [25] describes that situation awareness is vital for mission accomplishment for a range of operational actions during the mission. This chapter clarifies that incomplete or inaccurate situation awareness can result in wrong decisions even by the best trained and most experienced individuals [25]. Ruedy and Schweitzer [26] argue that "many unethical decisions stem from a lack of awareness" (p.1), while high mindfulness leads to a higher likelihood of performing an action that is in line with human values and thus with moral principles in mind.

First response missions can be categorized as potentially high-risk situations characterized by a high information flow in which incomplete or inaccurate situation awareness can result in wrong decisions with severe consequences (safety and health risks) due to the harsh conditions and the involved time pressure. It is often recommended to improve the safety and performance of FRs by complementing the team with an AI agent that monitors the environment using a combination of multiple measures [39], [40]. These measures can be grouped into three different categories: environmental, physiological, and subjective [2]:

- **Environmental measures** describe the local situation and relevant specific circumstances in a timely manner via a diverse range of measures (e.g., weather, wind speed, wind direction, and the dispersion of hazardous substances), such that the firefighter can be supported in his/her decision-making process about the circumstances during his/her mission.
- **Physiological measures** of the fire fighter's health provide the firefighter with relevant information about his personal health/safety situation during the critical mission, such as heart rate, heart rate variability, respiratory rate, and body temperature.
- **Subjective measures** are self-reported performance-based measures by the firefighter. These measures describe how to read the specific local situation at hand and the required actions to take to mitigate the risks involved and improve the performance of the task at hand. This includes asking for help.

Within this research, a closer look is given to supporting FRs with improving their situation awareness with an AI agent that monitors the environment during the missions while the FRs are paying attention to their task. To support FRs with decision-making involving value-sensitive aspects and notify them when such a situation (e.g., a situation of own safety at risk) could occur.

## 2.3 Decision-making with a value-sensitive aspect

Ethical principles, such as the four principles of biomedical ethics: beneficence; non-maleficence; autonomy; and justice from Beauchamp and Childress [10], function as practical guidelines (standards and regulations) that describe values everyone can agree upon [8], [9]. For ethical principles to be action-guiding, it should be known how to apply them in specific situations [10]. However, formalizing ethical principles into practical guidelines is not a straightforward process as value tensions can occur when the ethical principles are applied to specific situations. Therefore, it is important to focus on values and value-tensions to figure out what the ambiguities and gaps are in our understanding of ethical principles [13].

Whittlestone et al. [13] explain the importance and benefits of focusing on values and value-tensions in the field of Artificial Intelligence, noting that value tensions generally occur when it appears necessary to give up one value in order to realize another value. Considering values and thinking of what possible value tensions could occur with the design of AI, we are forced to think about how different values of AI might be interpreted and applied in society. This can help us highlight high-priority areas for which ambiguity and gaps in our understanding of human values occur and direct us to important research directions of the impact of AI on society [13]. As an example in which this leads to improved use of AI, Whittlestone et al. [13] refer to the research of Adel et al. [14] who describe how attempts are made to reduce or even eliminate the trade-off between performance and interpretability for state-of-the-art machine learning systems [13], [14].

There are researchers who specifically focus on moral principles of human-agent teams in the DM domain. For example, within HIS in a search and rescue scenario, there are different types of interaction, such as a human operating the robot from a distance; a human interacting with the robot on sight as a rescue worker; or a human interacting with the robot on sight as a victim. Each of these interactions differs from one another, e.g., the interaction in which the operator controlling the robot has a clear hierarchical structure, whereas victims are somehow dependent on the robot when the robot encounters a victim during a mission. These types of interactions between humans and autonomous systems involve several dilemmas in which value tensions may occur. In this context, Harbers et al. [15] identified the following dilemmas:

- **Safety risks**: "Should SAR robots be employed when they might help saving lives, but their application might also lead to casualties?" (p.10) And "should one develop SAR technology that is intended for peaceful purposes even when it has clear military potential?" (p.10-11)
- **Decreased performance due to the replacement of humans**: "Should one replace infield workers by robots if that leads to suboptimal performance?" (P.11)
- **Loss of relevant information**: "To what extent should information collected by robots be processed to make it more digestible, at the risk of losing or misrepresenting information?"(p.11)
- **False expectations about robot capabilities**: "Should one deploy robots, knowing that this may raise false expectations and runs the risk of degraded performance?" (p.12)
- **Responsibility assignment problems**: "Should one deploy robots that may yield responsibility assignment problems?" (p.12)

These dilemmas and corresponding values and value tensions were gathered via three Value Assessment Workshops with professional firefighters from the Corpo Nazionale Vigili del Fuoco (the Italian national firefighting organization) using the Value Sensitive Design (VSD) methodology [15]. The participants were from different levels in the organization and therefore included both officers and field workers. Furthermore, it is explained that none of the participants worked with robots in their daily work.

The reason mentioned by Habers et al. [15] to explore the ethical landscape surrounding robot-assisted Search and Rescue missions is a lack of attention for responsible development (e.g., requirements) and deployment (e.g., working agreements) of Search and Rescue robots. Harbers et al. claim that it is essential to consider ethics and values for responsible development and deployment. To support this claim, Habers et al. [15] briefly summarize that first-response missions typically are under time pressure and harsh conditions and that these missions can be over prolonged periods of time. They believe that these circumstances bring about specific ethical considerations, such as "what is morally acceptable to ask from rescue workers in terms of mental and physical well-being when lives are at stake" (p.3). Apart from that, they argue that the assigned role, the capabilities, the appearance, and the behaviour of a robot will influence the expectations of the people interacting with the robots. Even more, they point out that as the role robots play in robot-assisted Search and Rescue teams evolves, moral expectations may become heightened, potentially up to levels not achievable by the robots [15].

The value tensions mentioned above: safety risks; a decrease in performance; loss of relevant information; false expectations; and responsibility assignment problem, as identified by Harbers et al. [15] for search and rescue robots, might also occur for embedded AI agents integrated into mobile devices that provide decision-making support. This can, for example, happen when humans rely too much on the support of the AI agent for assessing a situation and deciding about the situation [16]. This illustrates that the goal should not be to maximize the trust in AI, but rather that there should be a balance between reliance and critical thinking about the decisions of the AI to avoid possible negative side effects that stem, for example, from uncertainty in the decision and/or biases. When trust is balanced with critical thinking, humans can benefit from the potentials of AI and, at the same time, avoid the negative effects that stem from overreliance on AI [16]. However, how trust and critical thinking can be balanced is case-dependent. It, therefore, is important to take a close look at values and value tensions that might be affected when the AI agent acts in the way it is designed and developed for. It is important to understand what values and value tensions exist within the human-agent team. This can, for example, be done with the Value-Sensitive Design (VSD) methodology [17].

## 2.4 The Value-Sensitive Design Methodology

In the last few years, the importance of VSD of technology has received an increasing amount of attention [3]. The VSD methodology considers human values throughout the design process of technology within three phases, namely the conceptual, empirical, and technical phase:

- In the **conceptual phase**, the direct and indirect stakeholders of the technology and their values are identified. The direct stakeholders are the ones that will be part of the HIS. They are the humans who interact with the technology to reach a common goal. The indirect stakeholders do not interact with the technology directly, but the technology, to some extent, influences their interests. When the stakeholders are identified, the values of these stakeholders the design will have an impact on will be identified.
- In the **empirical phase**, the stakeholders will be involved in the design process to examine their understanding of and experience with the technology and the implicated values for a specific context.
- In the **technical phase**, specific features of the technology are analysed.

VSD can be seen as a methodology to incorporate ethical considerations into the design. However, it should be noted that the VSD methodology only focusses on the design process and does not give insight into how these values can be integrated with the development process of the technology. To fill this gap Neerincx and Lindenberg [18] created the Socio-Cognitive Engineering methodology, which allows for incorporating the human values identified in the design phase of the technology into the development phase of the technology.

## 2.5 Socio-Cognitive Engineering methodology

The Socio-Cognitive Engineering (sCE) methodology makes it possible for researchers to focus on human values in both the design and the development phase of the technology. The method allows researchers to pay attention to the requirements and WAs of HAT to contribute to responsible design, development, and deployment of AI. This is done by looking into the effect of the assigned roles, capabilities, appearance, and behaviour of the AI on stakeholders.

sCE is developed explicitly for the design of HIS and is based upon the Cognitive Engineering (CE) methodology that originates from the 1980s [19]. The CE methodology aims to improve the performance of computer-supported tasks with increased insight into the cognitive factors of human-computer interaction in an iterative development process. The CE methodology is extended with 'explicit technology input' (p.2) for the sCE methodology. The sCE focuses on the 'specification and generation of ideas' (p.2) to integrate 'explicit made reciprocal effects of technology and human factors' (p.2) [18]. Doing so the sCE methodology combines the components of cognitive engineering, user-centered design, and requirement analysis [20].

The sCE methodology consists of three components, namely the foundation, the specification, and the evaluation. Following these components, the technological aspects and the underlying design rationale behind HIS can be refined, validated, maintained, and reused through the whole design and development process. In the foundation phase, the operational demands, human factors, and technological principles for the envisioned support in the working domain of the human-agent team are defined. Given these specified operational demands, human factors and technological principles objectives (the requirements) are defined for a specific HAT use case (scenarios). For these objectives, claims ("hypothesis") are provided, which could be used in the evaluation phase to test whether the specified functionalities have the proposed effect. The evaluation is done by domain experts, Human Factors experts, and technical experts such as Artificial Intelligence experts. Figure 1 provides an overview of the sCE methodology.

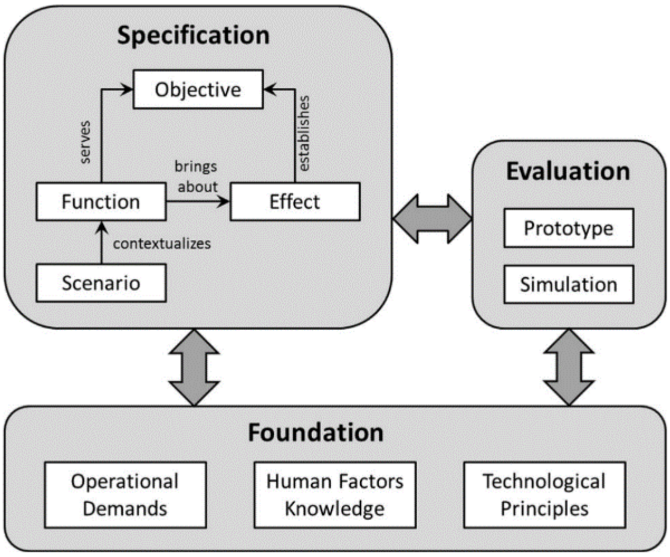



*Figure 1: Socio-Cognitive Engineering methodology [20]*

## 2.6 Work agreements

WAs are action-guiding statements used to reach a certain goal (norms) that influences the choice of action of agents such that it promotes or demotes values [35]. They function as a methodology to reason about (and verify) obligations, permissions, and prohibitions of actions or goals via explicit agreements between a debtor and a creditor [32]. If the new situation derived from the action is better than the old situation, we say that the action promotes that value. On the other hand, if the action leads to a situation that is worse than the old situation the action demotes the value. The foundation of WAs is normative rules in the form of deontic logic [11], [41].

Gutzwiller et al. [33] describe that WAs are applicable for any situation in which two agents with different kinds of decision-making models must cooperate and therefore act as a bridge in knowledge and understanding. In their paper, it is described that the overall intent of WAs within human-agent teams is to provide a means for human agents to indicate preferences concerning how task allocation and execution should be accomplished in collaboration with AI agents. WAs establish rules that allow human agents to predict the AI agent's pattern of operations successfully and increase trust in autonomy [33]. Both the human agent and the AI agent access the WA to guide decisions. WAs affect how team composition is adjusted for individual tasks and can directly affect the quality and quantity of team performance [33]. Furthermore, it is believed by Guztzwiller et al. [33] that WAs increase trust in automation for humans involved in the process as the behaviour of the autonomous systems will be more predictable due to the for humans accessible work agreement model [33].

HAT may vary across situations. In some situations, AI agents can operate without the assistance or control of humans. However, humans need to intervene when the performance degrades or malfunctions happen. This variation in teamwork requires dynamic task allocation, i.e., the ability of the team to shift task allocation depending on the situation at hand and thus a fit-for-purpose collaboration [34].

When working in a team, the agents who are part of the team collaborate to reach a common goal. This common goal makes the agents in the team dependent upon one another as it requires each team member to accomplish their part or a subtask within the teamwork. The role, expertise, and responsibility of a team member influence the distribution of information across various team members, meaning that situations can occur in which one team member possesses relevant information for another team member who needs to be proactively communicated. In essence, the relevant information retrieved by this team member should be shared with the team member it is meant for without explicitly having to request this information. Teamwork can only be effective if the tasks are appropriately coordinated and distributed amongst team members and if every agent is aware of how the common goal will be reached.

Coordination and distribution of tasks within a team strongly depend on the type of team and the context in which it operates, i.e., HAT is influenced by the environment, mission/task, team organization, team dynamics, and communication [11]. As Van Diggelen et al. [11] describes it: “as robots (and agents) become more capable of determining their plans and activities to try and accomplish the team goals, human-robot collaboration gradually moves away from traditional tasking, and towards, e.g., dynamic task allocation, shared initiative, and work agreements” (p.8).

Coordination and distribution of tasks within the team are especially difficult for unstructured, non-routine tasks as the beginning and end of these tasks are not clearly defined [32]. Van Diggelen et al. [11] identified the following types of HAT:

• **Parallel task performance**
Parallel task performance includes situations in which the tasks of individual team members are in coordination with the tasks of other team members. In these situations, the team members can identify, organize, and perform their tasks and responsibilities without the assistance of other team members.

• **Management by exception**
Management by exception includes situations in which team members perform tasks independently; however, when required, ask for help from their supervisor or colleagues—and situations where they are offered help proactively from a team member when they notices the break-down.

• **Training / educating**
Training or educating include situations in which team members strengthen their understanding about the task, the goal, their team members, and the roles within the team using feedback of other team members and monitoring situations. This means that the team members do not perform a task without the supervision or assistance of other team members.

• **Tasking**
Tasking includes situations in which team members are provided a task to perform; however, once the task is done they cannot determine the next task. Tasking is needed when they are unaware of the goal, the surrounding, their team members, and the roles within the team to determine what the next task should be.

Taking the above types of HAT into consideration Van Diggelen et al. [11] address how WAs can help human-agent teams reach *common ground'; 'interpredictability';* and *'directability'* for effective coordination of teamwork. WAs can enable external directability on an agent's behaviour and contribute to maintaining common ground and interpredictability within the team about this behaviour without compromising the team's autonomy (i.e., internal autonomy requirement) [11]. In cases where directability or common ground are not possible, interpredicatability comes into play as it plays a vital role in proactive communication between team members. Interpredictability enables team members to infer to another team member that new information has come available that affects this specific team member [11].
Van Diggelen et al. [11] clarify that common ground is an essential requirement in human-agent teams, especially for situations for which communication is restricted.
Van Diggelen et al. [11] argue that WAs can be used within the design process of human-agent teamwork. It can serve as a methodology to state communicative acts that must be performed for certain circumstances. Examples provided by Van Diggelen et al. [11] of such communicative acts to use WAs for in the design phase are the division of tasks between a human and an agent and the level of human involvement in the teamwork.

## 2.7 Team Design Patterns

One methodology to design human-agent teamwork is TDPs, a methodology that focuses on the requirements of both human agents and AI agents. TDPs are used to structure HAT research and help researchers to communicate requirements of successful (joint) knowledge-based behavior in HIS with a wide range of stakeholders such as end-users, AI developers, and Human Factors researchers (see [4]-[7]). TDPs allow for describing the division of tasks between a human agent and an AI agent. It describes the involvement of both in the teamwork. TDPs can further be used during the design phase of HAT to describe task allocation while simultaneously addressing the moral aspects involved in the interaction between humans and autonomous systems.

Machine ethics research also focuses on the moral components of autonomous systems (such as AI software and unmanned vehicles) in human-agent teams. Machine ethics research aims to integrate ethical rules into autonomous systems such that they act following a normative rule. However, attention is mainly given to the requirements of the Artificial Moral Agents within the human-agent team and thus the role of autonomous systems. What is not focused on in machine ethics research are the human requirements. Therefore, the value of the TDP methodology is that it illustrates concisely and intuitively the minimum requirements needed for HAT (roles, responsibilities, and capabilities) [6] from both the perspective of the human agent and the AI agent.

Van der Waa et al. [6] created a taxonomy that describes this minimal sufficiency needed to express human cognitive components and requirements for human-agent interaction [6]. This taxonomy is used by Van Stijn et al. [7] to apply TDPs in the design of FATE for allocation of moral decision-making in human-agent teams. Van Stijn et al. [7] uses TDPs to deal with the design problem of bias mitigation and data sharing for digital assistance for patients and healthcare professionals to prevent, diagnose, treat, and manage diabetes type II. The research of Van Stijn et al. [7] is the first attempt to utilize the taxonomy of TDPs by Van der Waa et al. [6] in the design of a system for allocation of human-computer teams in morally sensitive situations and to evaluate its usability for a multi-disciplinary research & development group.

Design Patterns originally stem from Alexander [36], who developed a language of design patterns for architecture. The design pattern language was used to describe the effects of the patterns, i.e., the consequences and the relationship of these patterns to other patterns. In the 1990s, pattern languages became a popular tool for software engineering to describe a recurrent problem and the core of a solution for this problem. Gamma et al. [37] created a template to describe the problem space, the solution template, positive and negative consequences, and implementation advice. TDPs also uses a specific template; however, the template used for TDPs is as follows:

- Pattern Name
- Description
- Structure
- Example
- Requirements
- The positive consequences of using the pattern
- The negative consequences of using the pattern

## 2.8 Knowledge representation and reasoning

To enable an AI agent to support humans in specific tasks, such as decision-making, risk assessment, and planning, relevant information from and about the environment is needed. This information can be retrieved from sensors in the environment attached to static objects (such as streetlamps), moving objects (such as AUVs and UGVs), or wearables (such as the suit or helmet of a FR). This sensor information is then structured in a knowledge base. A knowledge base provides an AI agent with semantic knowledge of the world in a structured representation over which the agent can reason [27]. A knowledge base gives an AI agent the ability to inference knowledge as it provides the agent with a certain level of understanding of the agent itself, its environment, and other agents, and thus of the HIS it operates in.

A commonly used methodology to develop a knowledge base is an ontology. An ontology is an object model that represents knowledge in the form of classes, subclasses, and instances. The knowledge is structured within an ontology in the form of relevant concepts and relationships given a specific domain. Ontologies are widely used for research in the information technology domain to exchange knowledge about different HIS [28], [29]. An important advantage of using an ontology to represent knowledge is that both humans and AI agents can interpret it. Another advantage is that there are many tools available for ontologies to perform knowledge reasoning over them. One of these tools is protégé, "a free, open-source ontology editor and framework for building intelligent systems"[6]. Protégé uses the mark-up language Web Ontology Language (OWL) and has an automatic classification reasoner built in.

[6] https://protege.stanford.edu/

# 3. From problem scenario to design solution

This chapter describes the process of transforming a problem scenario of value-sensitive decision-making within the first response domain into a design solution of HAT. Section 3.1 describes the research methodology (the data collection methodology; data characteristics and data analysis process), Section 3.2 describes the findings, and Section 3.3 describes the resulting design solution (i.e., the TDPs that describe the HAT solution for first response teams for decisions with a value-sensitive aspect).

## 3.1 Research methodology

The stakeholders and their associated values and value tensions have been carefully considered to transform the problem scenario into the design solution. The focus was on value tensions that can occur during missions for which first responders have to make a decision and for which agent support might come in handy. Different research techniques have been applied. The literature used to extract the stakeholders and their corresponding values is 'Exploring the Ethical Landscape of Robot-Assisted Search and Rescue' by Harbers et al. [15]. Literature used to create the problem scenario are the Scenario Book External Safety[7], fire brigade industry standards[8], Incident bestrijding gevaarlijke stoffen door de brandweer[9], and Directorate-General for European Civil Protection and Humanitarian Aid Operations (ECHO) (European Commission) [30]. Furthermore, the stakeholders and their corresponding values, the problem scenario, and possible HAT solutions were extracted from a focus group discussion with three HAT experts with knowledge of first response and DM and two semi-structured interviews with a commander and a team leader of the Gezamenlijke Brandweer. The data collection techniques, data characteristics, research procedure, and data analysis techniques for the focus group with HAT experts from TNO are described in Section 3.1.1, whereas that for the semi-structured interviews with DM experts from Gezamenlijke Brandweer is described in Section 3.1.2.

### 3.1.1 Focus group with human-agent teaming experts from TNO

In the literature phase of the research, a focus group was held with three colleagues at TNO with expertise in HAT and DM to discuss the aim of the research and the proposed TDPs for the problem scenario and design scenario as described below in Text box 1 and 2 in Section 3.2.3.

Three colleagues that are all part of the ASSISTANCE project research group of TNO were

---

[7] Scenarioboek Externe Veiligheid. Accessed: September 30, 2020. [Online]. Available: https://www.scenarioboekev.nl/overzichtstabel-scenariokaarten/

[8] Brandweeracademie. (2015, april). Branchestandaarden blijvende vakbekwaamheid. Totaaloverzicht functies manschap a, bevelvoerder en officier van dienst. Accessed: September 10, 2020. [Online]. Available: https://www.ifv.nl/kennisplein/kwaliteit-brandweerpersoneel/publicaties/branchestandaarden-blijvende-vakbekwaamheid

[9] Brandweeracademie. (2014, april). Incident bestrijding gevaarlijke stoffen door de brandweer. Meerjarenvisie. Accessed: September 15, 2020. [Online]. Available: https://issuu.com/brandweernederland/docs/visie_ibgs_definitief_29_april_2014

invited for an online recorded semi-structured focus group session of an hour via Microsoft Teams. At the start of the meeting, the participants were introduced to the study purpose and the goal of the focus group. After this, consent was asked to record the session. The research data of the focus group is processed using pseudonymization techniques to make sure that the personal data of the participants is not traceable to them as a person.

During the focus group a guide was used to structure the meeting. This guide was e-mailed to the participants a day before the meeting. The guide can be found in Appendix A. The session was transcribed, coded, and summarized into key findings. The results of the focus group can be found in Section 3.2.

### 3.1.2 Semi-structured interviews with disaster management experts from Gezamenlijke Brandweer

As mentioned at the start of Section 3.1, the information from the focus group discussion with three colleagues at TNO with expertise in HAT and DM was used in the literature phase of the research in two semi-structured interviews with DM experts, i.e., a commander and team leader of the Gezamenlijke Brandweer. It was chosen to interview both as a commander focuses on the effects of an incident, whereas a team leader focuses on the source of the incident during the mission. This decision was made as it was assumed that the needs and wants in relation to HAT could be different for different functions and therefore taking the perspective from the commander (mitigating the effects of the incident) and a team leader (mitigating the source of the incident) could identify whether this is indeed the case. In principle, the semi-structured interviews functioned as the first verification of the problem scenario and the proposed design solution, such that the design solution matches the needs of FR during DM and could be shared with a larger group of participants for formative evaluation.

A commander and team leader of the Gezamenlijke Brandweer were individually asked for an online 60-minute recorded session in Microsoft Teams for a semi-structured interview. At the start of the meeting, the participants were introduced to the study purpose and the goal of the semi-structured interview. After this, consent was asked to record the session. The research data of the semi-structured interviews are processed using pseudonymization techniques to make sure that the personal data of the participants is not traceable to them as a person. When the recorder was activated, the participants were introduced to the scenario. It was then asked if everything was clear to them or if they had further questions about the scenario before the rounds with questions started. The interview questions from the semi-structured interviews with a commander and team leader of the Gezamenlijke Brandweer can be found in Appendix B. During the rounds of questions, there was room to ask other questions that popped up after a particular answer of the participant. This means that the structures of the interviews with the commander and the team leader are slightly different. However, all the questions in the list were addressed in both interviews. The session was transcribed, coded and key findings were summarized. The results of the semi-structured interviews can be found in Section 3.2.

## 3.2 Results

### *3.2.1 Main findings from the focus group*

In this section, significant findings from the focus group session with three colleagues at TNO with expertise in HAT and DM are described. During this recorded online session in Microsoft Teams, the proposed problem scenario and design solutions were discussed that were designed at that moment. The transcript of the focus group session can be found in Appendix C.

Within the focus group session, we spoke about support for FRs from the AI agent during DM for decisions with a value-sensitive aspect varying from embodied to embedded support. Topics touched during the session were the role of search and rescue during disaster management, trust between human agents and autonomous systems, human control over autonomous systems, and responsibility, roles, and task allocation between human agents and (embodied and embedded) autonomous systems.

Before the focus group session took place I was primarily focused on embodied solutions and did not much touch upon embedded solutions. This was something the participants of the focus group noticed and therefore, the discussion led to embedded support. Examples were given about decision-making support in the form of messages with information about what is inside a building, the dispersion of a gas leakage, and the context the leakage is in. It was questioned by the participants how warnings or predictions of the gas cloud should be shared and with whom in what kind of situations.

Furthermore, attention was given to the distribution of the decision-making. This was identified as one of the points of attention for creating the design solution, e.g., there should be given a close look to who should be involved in the decision-making process and who will be responsible for making the decision. This discussion was in relation to the idea to investigate what WAs should be established for the HAT. It could be that WAs should be established with the team leader. For example, a WA that indicates when a FR is allowed to deviate from procedures to save a victim and thus ignore alert messages from the AI that there is gas inside the building.

Apart from that, relevant research was shared from which concrete values relevant for ASSISTANCE could be extracted (e.g., Harbers et al. [15] and workshops notes from previous research (TRADR)).

### *3.2.2 Main findings from semi-structured interviews*

During two semi-structured interviews, a commander and a team leader of the Gezamenlijke Brandweer went over the problem scenario and proposed design solution, and afterward answered about fifteen questions related to procedures of incident control of hazardous substances, important decision moments during incident control of hazardous substances, risks involved in incident control of hazardous substances, and support of an AI system during incident control of hazardous substances. The transcripts of the semi-structured interviews can be found in Appendices D and E. It depended on the role of the fire fighter I spoke to what was identified as a point of interest for support from the AI agent.

The commander is, for example, responsible for making decisions about the effect of the incident and thus focussed on solutions that could help with better identifying the effects of the incident. In contrast, the team leader focussed on supporting fire fighters sent towards the incident to eliminate its effect and mitigating the danger for them.

However, there was some overlap in the answers given by the commander and the team leader. Both of them agree that the more information there is available about the environment, the better they can decide on safety risks (whether this involves the safety of fire fighters themselves or citizens). Important things that should be measured during the incident that the commander and the team leader identified are the physical parameters of the hazardous substance (e.g., emission, explosion risk, vapor pressure) and measures in the environment (e.g., current weather and weather predictions, and what is in the area of the incident and downwind). Ideas given for situations in which an AI system could help are those where the AI interprets the situation and identifies what the fire fighters should be aware of. I.e., the AI should be able to interpret weather data; the direction and speed of the emission from the hazardous substances; information about the incident itself (e.g., the fire and the leakage and/or spill of hazardous substances), and information about what is in the affected area).

### 3.2.3 From problem scenario to proposed design solution using the input of the focus group and semi-structured interviews.

*Table 1: Stakeholders Value-Sensitive Design analysis*

| Stakeholder | Values |
|---|---|
| **First responder** | Health<br>Safety<br>Performance<br>Autonomy<br>Transparency<br>Privacy |
| **Citizens** | Health<br>Safety<br>Transparency |
| **Human agent-teaming experts (HF researcher or AI developer)** | Health<br>Safety<br>Effectivity of system<br>Satisfaction of stakeholders |

FRs work in complex, dynamic, and uncertain environments, which brings significant risk for their own safety and health and that of citizens affected by the disaster. This problem is related to environmental measures such as sudden changes in wind direction, wind speed, temperature, or the dispersion of hazardous substances. To deal with environmental changes, FRs should have situation awareness. They should constantly monitor the environment to assess what has happened, what is happening, and what might happen to avoid dangerous situations. However, it cannot always be predicted what might happen, and unpredictable environmental changes are very risky. Furthermore, a secondary event could be triggered, which might even bring more risk [30]. The direct and (most important) indirect stakeholders and their corresponding values are listed in Table 1. Within this research, a closer look is taken at the situation in which the FR has to make a decision about his/her own safety due to an adversarial change in the environment. The problem scenario that is addressed is the situation in which a fire fighter has to resolve a leakage of hazardous substances for which there exists the possibility of adversarial changes in the environment. The problem scenario is described in Text Box 1 on the next page.

*‘A fire is burning in the hot zone. The hot zone contains large concentrations of a toxic substance in the air and therefore it is too dangerous to enter this zone without a chemical hazard suit. Apart from that it is known that there are several barrels of ethanol and benzine (UN number 3475, GEVI number 33) leaking in the warm zone close to the hot zone. If the substance in the barrels come into contact with fire in the hot zone it will result in the release of even more toxic gasses in the air than there already are. Based on this information and the information on the ERIC-cards the commander decides that fire brigade personnel need to be send into the hot zone to resolve the leakage. Imagine you are sent into hot zone to contain the leakage of these barrels. You know that the barrels are close to the warm zone, however in an area that cannot be entered without a chemical hazard suit as the air is highly toxic. Apart from that the weather is unpredictable and wind direction and speed might change during the mission, resulting in the fire spreading in your direction instead of in the opposite direction. Changes in weather such as windspeed and wind direction can influence the location of the different zones. There is the possibility that a change in direction of the wind and an increase in windspeed extend the hot zone. As the barrels carry highly flammable substances and are located in the hot zone where a fire is burning as well, the unpredictable weather is a high risk for you. Leaving the barrels in the hot zone can make the situation worse than it already is. The barrels need to be contained to avoid that they come into contact with the fire.‘*

*Text Box 3: The problem scenario of the decision-making process of a first responder with a value-sensitive aspect during disaster management*

To deal with the problem described in Text Box 1 the FR could be supported by a software agent (AI agent) which communicates with FRs via an Android mobile device (tablet) using cloud computing (see the use case in the Text Box 2 below).

*The team leader has decided that you need to contain a leakage of several barrels of ethanol and benzine (UN number 3475, GEVI number 33) within the hot zone in which a fire is burning as well (see the use case described in Text Box 1 in section 4.1.1 for detailed description of the mission). The team leader advices you to bring your mobile device with you on the mission so that the artificial intelligent agent can support you with monitoring the weather and warning you for critical weather changes. Before you start the task you setup a working agreement with the AI agent. The working agreement is set up together with the team leader. With the working agreement it is specified how the AI agent will alert you for the adversarial changes in the environment while you resolve the leakage.*

*Text Box 1: The design scenario of human-agent teaming during disaster management for decisions with a value-sensitive aspect*

Cloud Computing in the context of the research project described in this thesis refers to the remote availability of computer resources (sensor data and code) to support the operations of the AI agent, as described by Kehoe et al. [38]. Using the network, the AI agent can support FRs with decisions with a value-sensitive aspect via adaptive text-based notifications based on specific WAs. The AI agent uses information from sensors in the environment to retrieve information about the weather, hazardous substances, team member's location, and reasons for this information to provide early warnings to FRs in danger. How early warnings are provided to the FR in text-based notifications is based upon predefined WAs. This means that the FR can indicate whether they get information or advice, whether the team leader should be notified, and if task transfer should be enabled (i.e., a robot taking over the task from the FR).

The following core functions for the AI agent are therefore expected to enhance the decision-making process of FRs during DM for decisions with a value-sensitive aspect:

1. Providing assistance in the form of text-based notifications on the tablet to alert the FR in case of adversarial changes in the environment, which is both personal and reliable.
2. Planning and pursuing to joint objectives in a transparent matter that is compliant with the values of FRs during DM.
3. Value-sensitive information sharing, which enables an adequate assessment of values by means of proposals and commitments to WAs.

These core functions for the AI agent are integrated in the design solution using the Team Design Pattern language. In total, three patterns were designed, which are described in the next section.

## 3.3 The Team Design Patterns

Within this research TDPs are used to illustrate the design solution as described in section 3.2. The TDPs illustrate the division of roles and responsibilities between a human agent and an (embedded) AI agent for the decision-making process of decision with a value-sensitive aspect.

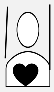
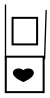
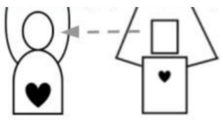
As mentioned in Section 3.1 the TDPs were designed with the help of three human-agent teaming experts, and a commander and team leader of the Gezamenlijke Brandweer. Using their input, it was investigated whether the proposed design solution describe a collaboration between a first responder and AI agent that meets the needs of FR during DM for decisions with a value-sensitive aspect. And thus, whether the design solution describe how a first response team can deal with adversarial environmental changes during missions tailored to the context.

In principle, the TDPs describe the design solution by describing who is doing what (task distribution); how the coordination is structured; and what has been agreed upon and formulated in the working agreements. Using textual and graphical notations, the TDPs describe the objective, the conditions, the involved actors, functional requirements, expected effects (advantages and disadvantages) of HAT for a specific scenario.

In total, three TDPs were created that together form the design solution for decision-making with a value-sensitive aspect during DM. The TDPs describe the collaborative management of WAs and the teamwork of the first responder and AI agent for handling changes in the environment based on these WAs.

To understand the structure of the TDPs see Table 2 on the next page, in which the concepts, type, graphical depiction and an example of the Team Design Pattern Language can be found. The TDPs themselves can be found in Table 3, 4 and 5. The first TDP describes the collaborative management of WAs at the start of the mission (see Table 3), the second TDP describes how the human-agent team handles changing circumstances during the mission while they work on a specific task (see Table 4), and the third TDP describes how the team members react independently on changing circumstances (see Table 5).

*Table 2: Team Design Pattern language concepts, type, graphical depiction and an example*

| **Concept** | **Type** | | **Depiction** | **Example** |
|---|---|---|---|---|
| ***Actor*** | *Human* | |  | *The human task performer, indicated with his/her role or position in the team or organization (e.g., fire fighter).* |
| | *Machine agent* | |  | *(embedded) artificial intelligent agent* |
| ***Work*** | *Direct* | *Moral* | *Make a decision* | *A decision that has implications that might have adversarial consequences (e.g., concerning safety or privacy), for example, deciding between saving a victim versus staying outside the danger zone* |
| | | *Amoral* | *Display information* | *'display relevant information',* |
| | *Indirect* | *Moral* | *Call for an ambulance* | *'alert paramedic that medical care is needed'* |
| | | *Amoral* | *Request action* | *'request the other actor to do something'* |
| ***Work Agreement*** | *Setup* | |  | *A work agreement has collaboratively been established.* |
| | *Specifications* | |  | *Properties of the work agreement are specified.* |
| ***Moral capacity*** | *Human* | |  | *The agent has human or human-like capacity to recognize and make moral decisions* |
| | *Partial* | |  | *The agent has some capacity for recognizing and making moral decisions, but not at a human level* |
| | *None* | | | *The agent has no capacity for recognizing or making a moral decision* |
| ***Asymmetry*** | | |  | *Task division is asymmetric, e.g. 'give support', 'supervise', 'alert'* |

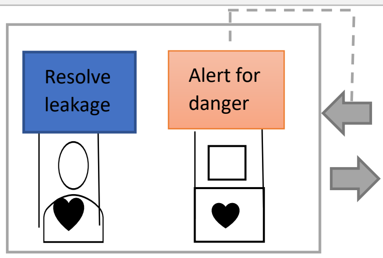
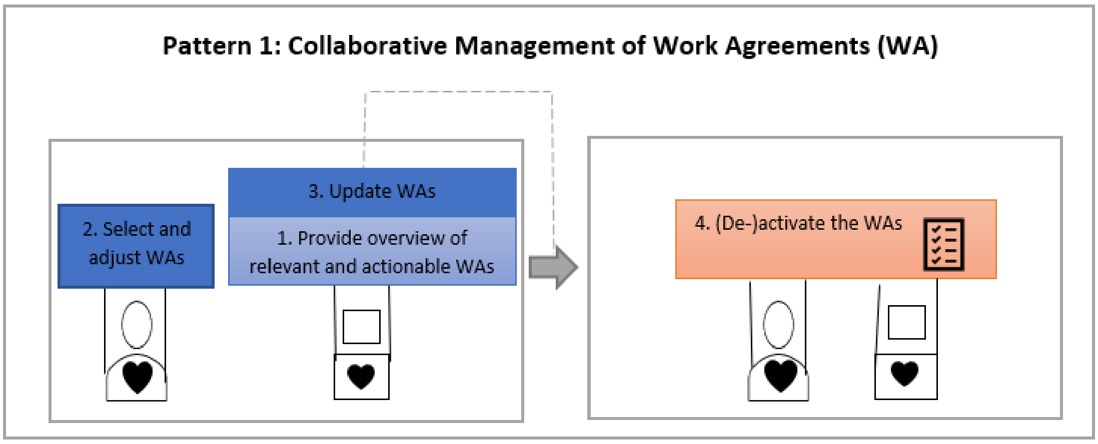
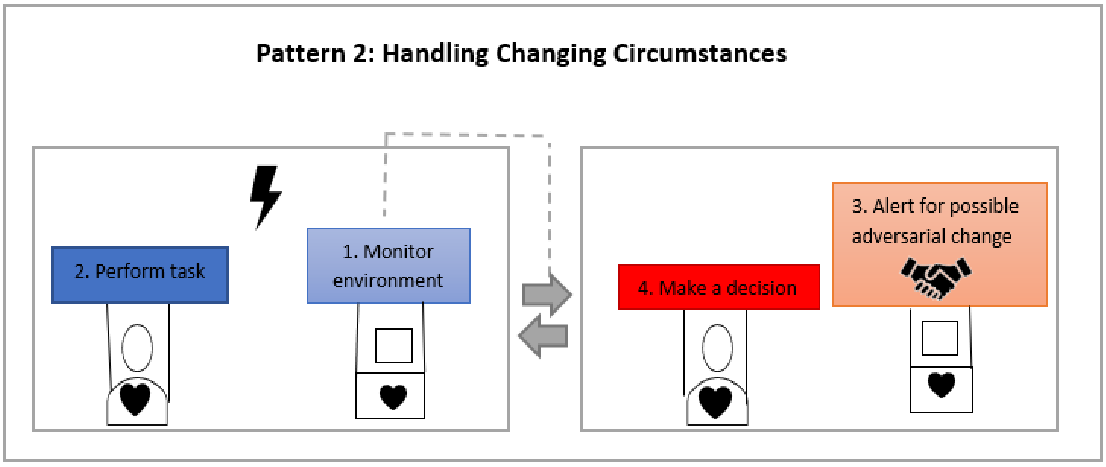
| | | | |
|---|---|---|---|
| ***Transition*** | *Automatic* |  | |
| | *Initiative* | <br> | *If the artificial intelligent agent notices an abnormality, transition is initiated.* |
| | *External influence* |  | *Weather changes.* |

*Table 3: Collaborative Management of Work Agreements Team Design Pattern*

| | |
|---|---|
| **Name** | Collaborative Management of Work Agreements |
| **Description** | At the start of the mission, the team leader assigns tasks to its crew.<br>The AI agent is given a task (monitor environment), and each crew member can indicate how the AI agent should notify them given this task (share dangerous environmental changes). |
| **Structure** | <br> |

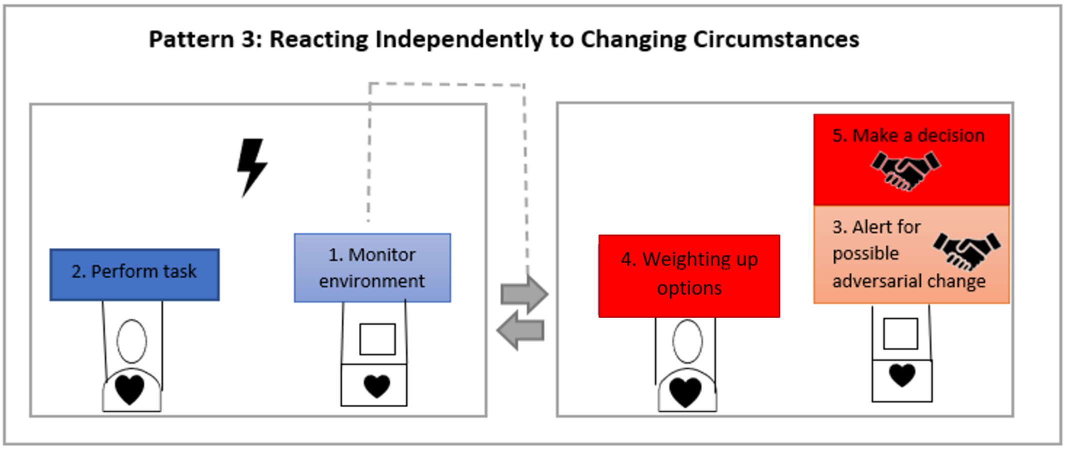
| | |
|---|---|
| **Example** | At the beginning of a mission a team leader assigns tasks to each crew member based on the information they have of the situation. The infield fire fighters (infield FF) are aware of each other's tasks and know that the AI agent will monitor the environment for dangerous changes and doing so support them during the task of resolving the leakage of hazardous substance. This decision has been made by the team leader as the weather is unstable, which has been noticed by observing flapping wind vanes that occasionally change direction. The team leader knows that changes in weather can be dangerous for his crew and therefore decides that the AI agent should monitor the environment (wind direction, wind speed and gas cloud dispersion) for dangerous changes and alert the infield FFs in case of danger based on the preferred interaction type. How this interaction will look like during the mission and thus how the infield FF is alerted for danger will be agreed up at the start of the mission in the form of Work Agreements (WAs). These WAs are collaboratively managed by the infield FF and the AI agent. When the WAs are managed the AI agent provides an overview of the relevant an actionable WAs available to the infield FF. This is a list of 1-n properties. For each property it is described what will be agreed up on. In this case it is established whether the notifications about adversarial changes in the environment should be information-based or advice-based. Furthermore, it is established whether the team leader should be notified about the adversarial change that have consequences for the infield FF or not. The infield FF responds to the AI agent by either selecting the WAs as they are or adjust. After adjustments are made the WAs are updated by the AI agent. The final list with WAs is than activated by the involved parties (I.e., the infield FF and the AI agent). |
| **Requirements** | ➢ The human agent and the AI agent should have the ability to setup and manage work agreements<br>➢ The AI agent should be able to provide interaction style options<br>➢ The AI agent should have the ability to provide information about adversarial changes that have an influence on the task allocation of the actors<br>➢ The AI agent should be able to give advice about adversarial changes that have an influence on the task allocation of the actors<br>➢ The AI agent needs to know who to send the information to<br>➢ The human agent needs to know what interaction style they prefer |
| **Advantages** | + The AI agent can provide a suggestion for the work agreement<br>+ The human agent has the ability to adapt the work agreement<br>+ The human agent is responsible for the final decision<br>+ AI agent does not carry the responsibility for the content of the work agreement as the human agent will make the final decision by accepting the agreement or not |
| **Disadvantages** | – The human agent might be unaware of the preferred interaction style<br>– The interaction possibilities are limited to the amount of pre-installed interaction styles |

*Table 4: Handling Changing Circumstances Team Design Pattern*

| | |
|---|---|
| **Name** | Handling changing circumstances |
| **Description** | The human agent performs his/her assigned task, while the Artificial Intelligent (AI) agent monitors the environment for possible dangerous changes. The AI agent recognizes a dangerous environmental change and initiates a recalculation of the model before the AI agent alerts the human agent. How the AI agent alerts the human agent for danger depends on the working agreement decided upon at the start of the mission, this can either be inform-based or advice-based. When the human agent sees the notification, the human can click on the visualization button to visualize the environmental changes or close the notification instead. Based on the alert and/or the visualization, the human agent can interpret the implications of the environmental changes and decide what to do next. |
| **Structure** |   |
| **Example** | On a first response mission, the agents can communicate about each other's tasks via a tablet they carry. This is useful in situations where the agents who want to communicate work on a different task in different locations or communicate with an AI agent. Communication between team members is especially important if one agent supports the other agent or when information an agent retrieves during his own task is important for another team member to be aware of. During the mission, the human agent performs his task, namely resolving the leakage of a hazardous substance. In the meantime, the AI agent monitors the environment (wind direction, wind speed, and gas cloud dispersion) and scans for adversarial changes. When the AI agent indicates an adversarial change that is harmful for the human agent, an alert will be sent to the human about the possible harmful changes. How the AI agent communicates the risk to the human agent is based on the predetermined interaction style agreed upon at the start of the mission. |
| **Requirements** | ➢ The AI agent needs to be able to monitor the environment (wind direction, wind speed, gas cloud dispersion)<br>➢ The AI agent needs to be able to display information from the environment (wind direction, wind speed, gas cloud dispersion)<br>➢ The AI agent should be able to assess whether the change in the environment is dangerous<br>➢ The AI agent needs to be able to notify the human agent about changes in the environment |

| | |
|---|---|
| | ➢ The AI agent needs to be able to display the changes in an understandable way<br>➢ The human agent has to make the final decision on what to do with the information provided by the AI agent |
| **Advantages** | + The human agent can decide what to do with the information provided by the AI agent<br>+ The human agent is responsible for the final decision<br>+ The human agent can focus on his task<br>+ AI agent is not required to interpret what changes in the environment mean for the task |
| **Disadvantages** | – The human agent might be unable to handle the notification during the task because both hands are needed for the task itself |

*Table 5: Reacting Independently on Changing Circumstances Team Design Pattern*

| | |
|---|---|
| **Name** | *Reacting independently on changing circumstances* |
| **Description** | The human agent performs his/her assigned task, while the Artificial Intelligent (AI) agent monitors the environment for possible dangerous changes. The AI agent recognizes a dangerous environmental change and initiates a recalculation of the model before the AI agent alerts the human agent. How the AI agent alerts the human agent for danger depends on the working agreement decided upon at the start of the mission, this can either be inform-based or advice-based. The human uses the notification to weigh up his/her options on how to react to the situation. Apart from alerting the human about the dangerous environmental change, the AI agent reacts to the changes itself by deciding on the environmental changes. The decision the AI agent takes is sending a robot with the right capabilities to the location where the human is working on the task that has become too dangerous for the human to work on to take over the task from the human. |
| **Structure** |  |

| | |
|---|---|
| **Example** | During the mission, the human agent performs his task, namely resolving the leakage of a hazardous substance. In the meantime, the AI agent monitors the environment (wind direction, wind speed, and gas cloud dispersion) and scans for adversarial changes. When the AI agent indicates an adversarial change that is harmful for the human agent, an alert will be sent to the human about the possible harmful changes. Furthermore, it reacts to the environmental changes by making a decision about it himself too, namely by sending a robot to the location the human is working on the task to take over the task. How the AI agent communicates the risk to the human agent is based on the predetermined interaction style agreed upon at the start of the mission, so is the ability of the system to make the decision to send a robot to the location the human is working on the task to take over the task. |
| **Requirements** | ➢ The AI agent needs to be able to monitor the environment (wind direction, wind speed, gas cloud dispersion)<br>➢ The AI agent needs to be able to display information from the environment (wind direction, wind speed, gas cloud dispersion)<br>➢ The AI agent should be able to assess whether the change in the environment is dangerous<br>➢ The AI agent needs to be able to notify the human agent about changes in the environment<br>➢ The AI agent needs to be able to display the changes in an understandable way<br>➢ The human agent has to make the final decision on what to do with the information provided by the AI agent<br>➢ The AI agent needs to be able to send a robot with the right capabilities to the location of the human agent for a task transfer |
| **Advantages** | + The human agent can decide what to do with the information provided by the AI agent<br>+ The human agent is partially responsible for the final decision<br>+ The human agent can focus on his task<br>+ AI agent is not required to interpret what changes in the environment mean for the task |
| **Disadvantages** | – The human agent might be unable to handle the notification during the task, because both hands are needed for the task itself<br>– The human agent is not aware when and how task transfer is being setup by the AI agent as this is not communicated to the human agent<br>– The human agent has no control over the task transfer once it is enabled |

# 4. Evaluating the design solution

Chapter 3 described how the design of the problem scenario led to the HAT design solution for decision-making with a value-sensitive aspect within the DM domain. The chapter ended with three TDPs that together form the design solution. These TDPs were then used to share the design solution with both HAT experts and end-users such that they could evaluate them.

This evaluation process is described in this chapter. The HAT experts who contributed to this evaluation process are TNO researchers with knowledge of Artificial Intelligence and Human Factors related to HAT. The fire fighters that contributed in this evaluation process are fire fighters from the Gezamenlijke Brandweer in Rotterdam (in short: GB fire fighters) and the Spaansland fire brigade in Enschede (in short: Enschede fire fighters).

It was chosen to include both HAT and first response and DM experts in the design and development process as knowledge from both experts' groups is valuable for assessing the understandability and usability of the design. Fire fighters have knowledge about and expertise in the first response and DM domain and thus in the related decision-making process. Therefore, they can indicate whether the proposed design solutions could work in the circumstances they normally work in, for which they have to make decisions with a value-sensitive aspect. The knowledge and expertise in HAT of the Artificial Intelligence and Human Factors researchers from TNO is valuable as their feedback could help with addressing inconsistencies and errors in the design itself.

The design solutions were qualitatively evaluated in an online evaluation process (online survey on Qualtrics) and within workshops. Both the methodology (does the TDP methodology work for sharing design solutions with different experts (HAT experts and DM experts)?) and context (what do the FRs think of the design solutions?) were evaluated.

Section 4.1 describes the demographical data (the characteristics of the participants within the research), Section 4.2 describes the research methodology (the data collection methodology; data characteristics and data analysis process), section 4.3 describes the findings of the formative evaluation and the workshops, and Section 4.4 describes the conclusions about these findings.

## 4.1 Demographical data of the participants

In total, twenty-four participants participated in this research, of which eleven HAT experts from TNO and thirteen professional fire fighters from the Gezamenlijke Brandweer (5) and the Spaansland fire brigade in Enschede (8) (see Figure 2 on the next page).

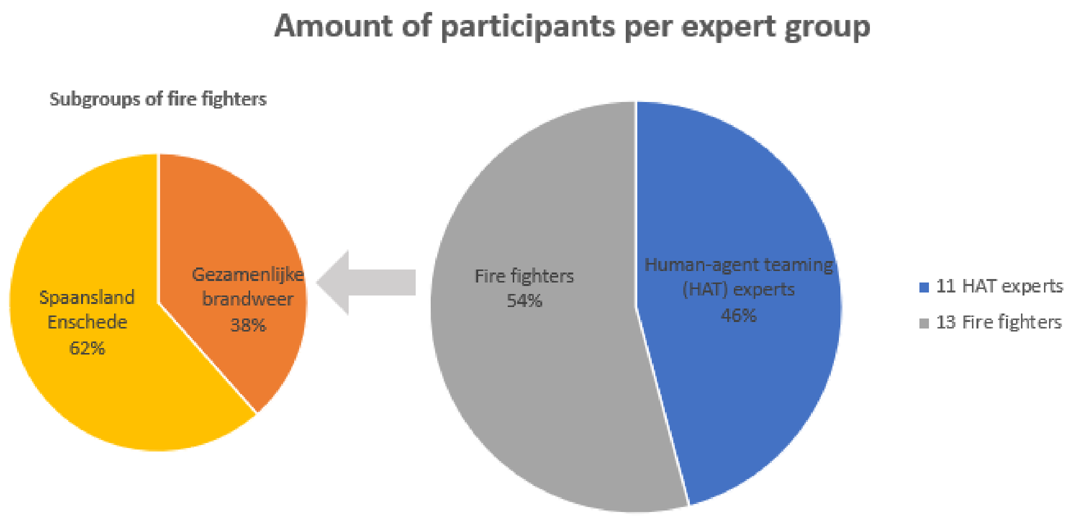



*Figure 2:* Number of participants that participated in this research per expert group

There were male and female participants in the HAT expert group (seven women and five men), whereas all fire fighters who participated in this research are men. Most of them are infield FFs (8); however, also a team leader and commander from the Gezamenlijke Brandweer and the Spaansland fire brigade in Enschede participated. The age of the participants ranges between twenty-five and sixty-four (six of the participant are between twenty-five and thirty-three years old).

At the start of this study, the GB fire fighters were familiar with ASSISTANCE, whereas the Enschede fire fighters had no prior knowledge of ASSISTANCE. Three of the GB fire fighters and two Enschede fire fighters indicated to have a lot of knowledge in hazardous substances. All other fire fighters (8) indicated to have little experience with hazardous substances (see Figure 3).

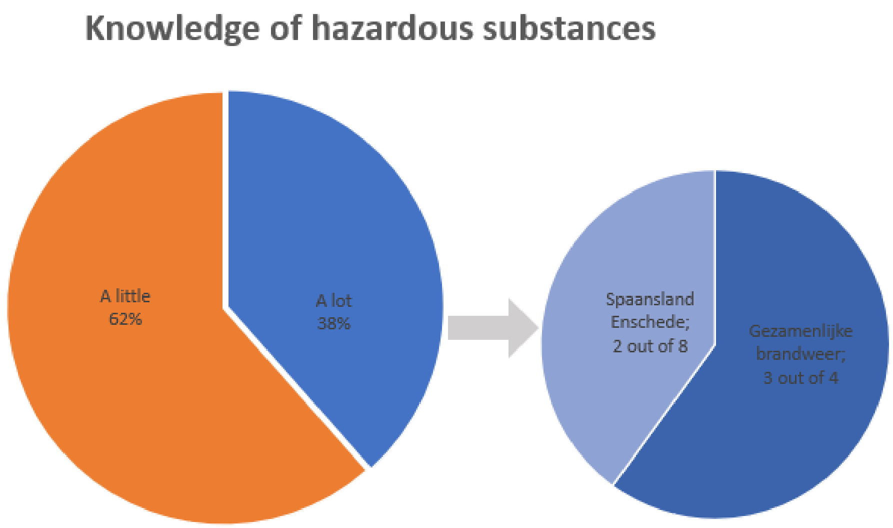



*Figure 3: Indication of the knowledge of fire fighters in hazardous substances*

Seven of the fire fighters indicated to have a lot of experience using tablets for work, and during their free time, the other five indicated to have a little experience. Two fire fighters indicated to have a lot of experience with using UAVs and UGVs during missions, two indicated to have little experience of using UAVs and UGVs, from the other 8 fire fighters it is not known how much experience they have with using UAVs and UGVs as this question was not asked to them.

Three of the HAT experts contributed by taking part in the focus group, all other HAT experts (8) contributed by taking part in the formative evaluation process. One of the fire fighters from the Gezamenlijke Brandweer took part in the semi-structured interviews, the formative evaluation, and workshop. Two GB fire fighters took part in the formative evaluation and the workshop. All other fire fighters contributed by taking part in the semi-structured interviews (a fire fighter from the Gezamenlijke Brandweer), the formative evaluation (a GB fire fighter), or the workshop (8 Enschede fire fighters) (see Figure 4).

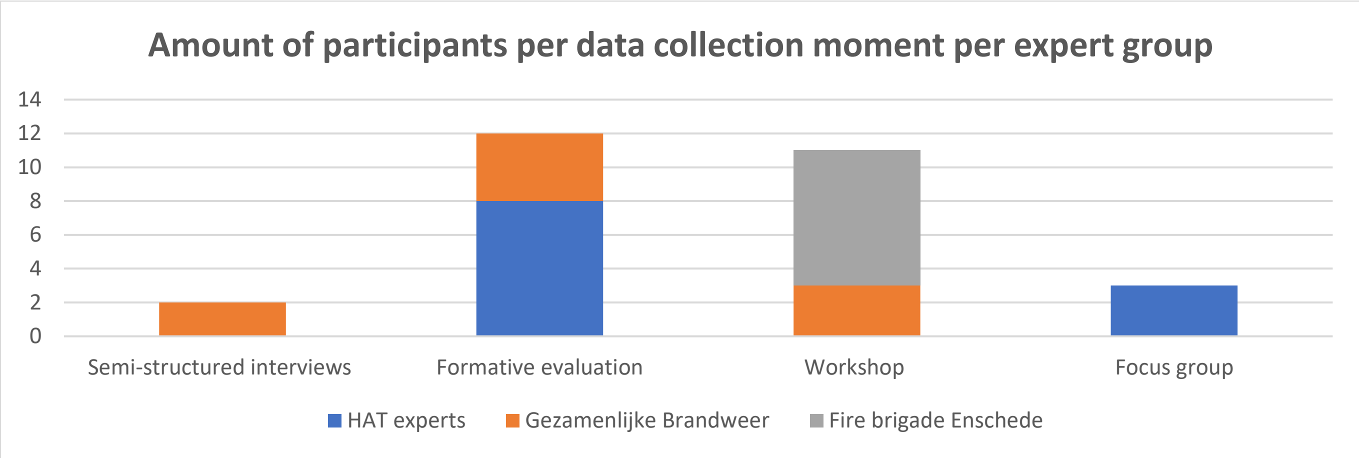



*Figure 4: an overview of the number of participants per data collection moment per expert group*

Of the eleven HAT experts who participated in this study, five have researched in the first response and DM domain of which four have done research within this domain related to HAT. Six of the HAT experts have experience with embedded and embodied agents in relation to research about HAT, whereas five only have experience with embedded agents. Furthermore, 8 of the HAT experts are familiar with TDPs, whereas one stated to have heard of design patterns but did not know much about them, and two stated to have no knowledge of design patterns. For an overview of the experiences of the HAT experts see Figure 5.

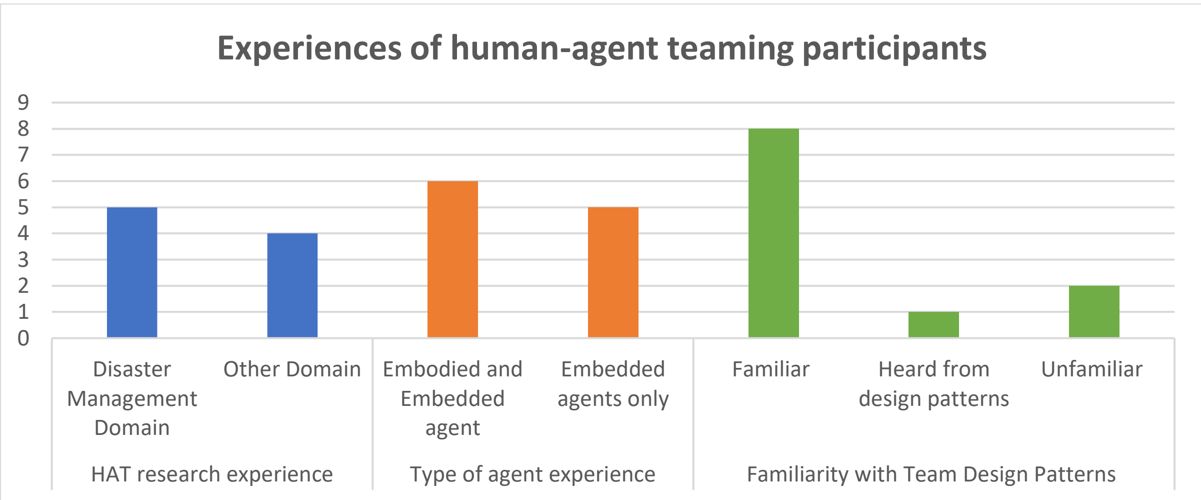



*Figure 5: Overview of the experiences of HAT experts relevant for this research.*

## 4.2 Methodology

An online formative evaluation on Qualtrics and a workshop at the Gezamenlijke Brandweer in Rotterdam and Spaansland fire brigade in Enschede were conducted to evaluate the design solutions. The participants of the formative evaluation process were both HAT experts (AI and HF researchers from TNO) and disaster domain experts (GB fire fighters). This chapter describes the data collection methodology, data characteristics, and data analysis process for the formative evaluation process (Section 4.2.1) and the workshops (Section 4.2.2).

### 4.2.1 Formative evaluation process with human-agent teaming experts and disaster management domain experts

The participants of the formative evaluation process were invited via e-mail to fill in a survey on Qualtrics using an anonymous link. The e-mail explained what research the evaluation process was part of and what was expected from the participants. The participants were asked to fill in a 40-minute survey via Qualtrics in which a case study is addressed to acquire knowledge about the requirements for an AI agent to support a FR in the decision-making process with a value-sensitive aspect. The research data is processed using pseudonymization techniques to make sure that the personal data of the participants is not traceable to them as a person.

` The HAT experts that were asked to participate are researchers from TNO with a background in Artificial Intelligence or Human Factors and experience and/or knowledge in HAT research and/or TDPs. The DM domain experts that were asked are infield FFs from fire brigades that are part of the ASSISTANCE consortium. These are directly the end-users of ASSISTANCE and the ones that will collaborate with the Intelligent System.

The survey form was active for six weeks in total. Within this period, a total of twelve people participated in the research. From these twelve participants, eight are HAT experts, and four are first response and DM domain experts. Of the HAT experts four also indicate to have experience in the first response and DM domain.

When the participants clicked on the link of Qualtrics to start the formative evaluation process, they are introduced to the study purpose and setup before they are asked to fill out the informed consent and general questionnaire. After this, the evaluation process started. Participants were first introduced to the TDPs to show them how the collaboration with the Intelligent System will look like during the evaluation process. After this, the scenario is provided to them in which the participants collaborate with the Intelligent System as proposed by the TDPs during an incident with hazardous substances in the role of an infield FF. The scenario starts with a briefing in which it is shortly described what the situation is the team deals with during the scenario (see Text box 1 and 2 and the description and example of each of the TDPs in Chapter 3). The scenario describes the HAT process during incident control of hazardous substances for decisions with a value-sensitive aspect, i.e., the participant will experience a situation in which a leakage of hazardous substances needs to be resolved before it comes into contact with the fire. At the same time, a potential threat occurs for which a decision needs to be made.

During the scenario, the participants manage WAs with the Intelligent System for the mission, in which a leakage of hazardous materials needs to be resolved before it comes into contact with fire. The participants needed to choose: whether they want to receive information or advice from the AI agent, whether they are the only ones that need to be alerted, or if their team leader also needs to receive the alert, and whether the Intelligent System can send a robot with the right capabilities to the location to take over the task of the human in the case that the human cannot finish the task, because the situation is indicated to be too dangerous for the human agent to do so.

At the end of the scenario, the participants have to make a decision about the alert of the AI agent concerning the adversarial changes in the environment and have to describe why they chose to make such a decision.

The moment the participants were finished with the scenario, they were asked to subjectively evaluate the scenario, the WAs, and the TDPs by rating 7-point scale Likert statements ranging from strongly disagree to strongly agree and indicate why they chose that specific rating. If they did not know how to rate the statement, they could select the option N/A.

The open questions answers and answers to the statements on the Likert scale reflect on the usability of, the representativeness of, and the trust in the current design patterns and propose improvements in more detail. The quality of the data is validated with a sanity check on the overall data and the data of each of the subject matter expert groups to understand the accuracy and completeness of the data.

The scripts of the formative evaluation can be found in Appendices F and G.

### 4.2.2 Workshops with fire fighters of the Gezamenlijke Brandweer and Spaansland fire brigade in Enschede

Four fire fighters from the Gezamenlijke Brandweer and eight fire fighters from the Spaansland fire brigade in Enschede were asked to participate in a 90-minute workshop. The participants of the workshops were invited via e-mail. During these 90-minute workshops, the same use case study was addressed as that of the formative evaluation process (for more information, see Text box 1 and 2 in Chapter 3). The workshop with the Gezamenlijke Brandweer lasted 90 minutes, whereas the workshop with the Spaansland fire brigade lasted 120 minutes. The goal of the workshops was to acquire knowledge about the requirements for an AI agent to support a FR in the decision-making process with a value-sensitive aspect.

Before the workshop started, participants were asked to read the information letter, fill out the informed consent and fill in the general questionnaire. After this, the recorder was started, and the workshop could start. The workshop with the Gezamenlijke Brandweer was recorded. This was also the intention with the workshop at the Spaansland fire brigade in Enschede. Unfortunately, the recording was not saved once the session was over. This means that no transcription could be made of the statements that were made by the fire fighters during the workshop at the Spaansland fire brigade. The information overview I made directly after this Workshop was sent to the fire fighters that participated in the workshop to verify that the information I wrote down was in line with the statements they made during the workshop to avoid possible wrong interpretations of the answers. The research data of the workshops are processed using pseudonymization techniques to make sure that the personal data of the participants is not traceable to them as a person.

The structure of the workshops, the information letter, the informed consent form, and the general questionnaire can be found in Appendices H, I, and J.

## 4.3 Results

HAT experts from TNO (in short: HAT experts) provided their subjective evaluation within the online formative evaluation, the fire fighters from the Spaansland fire brigade in Enschede (in short: Enschede fire fighters) provided their subjective evaluation during a workshop at the fire brigade, and fire fighters from the Gezamenlijke Brandweer (in short: GB fire fighters) provided their subjective evaluation within the online formative evaluation and during a workshop at the fire brigade. The input provided by the HAT experts and fire fighters (in short: HAT experts and FF) are described based on theme, e.g., participants were asked to indicate …

- the representativeness of the use case (Section 4.3.1);
- what WAs could help to reach the desired level of flexibility in the human-AI collaboration (Section 4.3.2);
- whether the TDPs describe the human-AI collaboration in a clear; concise; complete; and easy-to-understand manner that is expected to substantially improve team performance with a value-sensitive aspect (Section 4.3.3).

For each theme, the overall subjective evaluation of all the participants and the differences in subjective evaluation between the different groups of participants is provided. Amongst others, questions were asked about the level of understanding of the participants of the situation and whether they think it is a realistic situation for which FRs need to be prepared. Furthermore, three WAs that might help reach the desired level of flexibility in the human-AI collaboration were provided. Participants were asked if they could indicate if they think these are good WAs and whether these WAs could contribute to trust in the collaboration. Participants were also asked to indicate whether there are other WAs that should be considered.

### 4.3.1 The representativeness of the use case

The use case of this study represents a recurring decision-making problem with a value-sensitive aspect in the ASSISTANCE scenario. It describes how an AI agent supports fire fighters in case of an adversarial change in the environment based on WAs managed at the start of the mission. Within this use case, it is described how the task of a FR (resolving a leakage of hazardous substances from barrels) is interrupted during a mission due to an adversarial change in the environment that threatens the safety of the FR (change of weather conditions). Below we will go over the results collected about this use case. First, we will address the subjective evaluation of all participants (Subsection 4.3.1.1). After that, we will highlight the differences in subjective evaluation between the HAT experts and FF (Subsection 4.3.1.2).

#### 4.3.1.1 The subjective evaluation of all participants on the representativeness of the use case

Overall the participants indicated that the use case represents a recurring decision-making problem with a value-sensitive aspect in the ASSISTANCE scenario. For an indication of this for the formative evaluation, see Figure 6 and the corresponding summary statistics in Table 6 on the next page. However, there were some comments and remarks that should be taken into consideration. For example, the firefighters indicated that the outlined decision-making by the commander and thus the working method of the infield FF deviate from existing procedures. Typically, the procedure for an incident involving a leakage of highly flammable and explosive hazardous substances and a fire is to extinguish the fire before firefighters are deployed within the field to resolve the leakage. It has been indicated, however, that it is possible that a fire may start while working on a leakage of highly flammable and explosive hazardous substances. It, therefore, is possible that this scenario occurs during DM. Furthermore, unexpected situations are not uncommon for FRs.

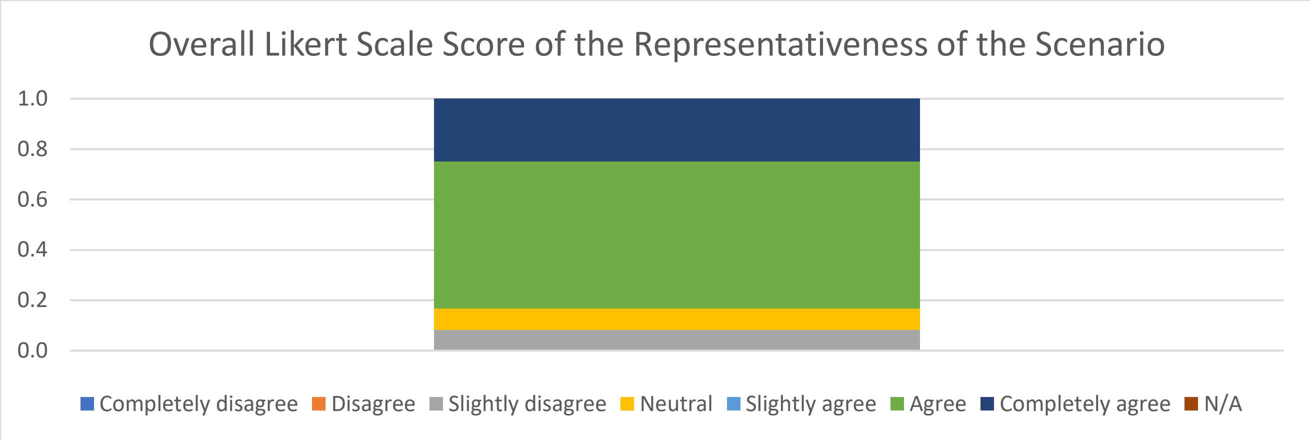



*Figure 6: Overall Likert Scale score of all the participants of the formative evaluation on the representative of the scenario.*

*Table 6: Summary statistics for the data presented in Figure 6.*

| | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| The scenario is a realistic scenario for which we need to be prepared | 3.00 | 7.00 | 6.00 | 1.00 | 1.00 |

Both the HAT experts and firefighters indicated that it is good to notify the FR in case of an adversarial change. This can, for example, be deduced from the following statement of one of the participants from the workshop with the Gezamenlijke Brandweer:

*'In itself it is good for an infield firefighter. We agree on this: if you have this notification then you have to leave. That is possible, if the situation .... But you have to make very clear agreements about it. But now we also say: 112, it is not safe, leave. You can still confirm it with a notification, but that. But that notification ... I think it is crucial, because where most of the information is lost is when I am instructing him and updating the commander, in the meantime the story and the toxic cloud continues. Doing this and that and before I know it … it takes me fifteen minutes, but of course a lot has happened in that fifteen minutes.'*

During the workshop at the Spaansland fire brigade in Enschede, the firefighters further indicated that it might be a good idea to not only focus on the environmental aspects (wind direction, wind speed, dispersion of hazardous substances) but also the health factors of the firefighters. It was suggested to combine the chemical hazard tool with Mobile Sensing for Safety (MoSeS)[10,11], a system capable of tracking where the firefighters are based on GPS data and their health condition. Furthermore, during the workshops with GB and Enschede fire brigade it was emphasized that strategical decisions are not made by infield FFs but by the team leader and higher up in the hierarchy. This means that the consequences of leaving the location before being able to resolve the leakage is considered by the team leader if he knows what to do, else he will ask the commander for advice. This means that it is not the responsibility of the infield FF to adapt the plan if the leakage cannot be resolved without putting him in danger. The firefighters indicated that the higher the position in the hierarchy a firefighter has, the more communication the firefighter will have with the

[10] Brandweer Twente. Twente werkt veiliger met MoSeS. November 3, 2017. Accessed: April 30, 2021. [Online]. Available: https://www.brandweer.nl/twente/nieuws-twente/2017/twente-werkt-veiliger-met-moses

[11] Tech for Future. Mobile Sensing for Safety (MoSeS). Accessed: April 30, 2021. [Online]. Available: https://techforfuture.nl/onderzoek/afgerond/mobile-sensing-safety-moses/

AI system. This means that an infield FF would like to receive a warning notification such that he knows he needs to leave the location, but he will not judge the warning. It is the team leader that will make sense of the warning and decide what to do next. This can, for example, be deduced from the following statement one of the participants gave during the workshop with the Gezamenlijke Brandweer:

*“if he stands there and something happens now, the infield firefighter walks back to a safe place and asks the team leader what should I do now? Literally what happens. And that makes sense because the team leader is the director of the entire incident”*

When we compare the answers of the GB firefighters and the Enschede firefighters, the perspective on the representativeness of the scenario between the GB and Enschede firefighters is similar. Both the GB and Enschede firefighters mentioned that the working procedures described in the scenario for resolving a leakage of highly flammable and explosive substances nearby a burning fire deviate from protocols firefighters follow. However, they agree that the working procedures are understandable for the case study and acknowledge that a fire could start burning while a firefighter is resolving a leakage. Therefore, they agree that it is a realistic scenario for which they need to be prepared.

#### 4.3.1.2 The differences in perspective on the representativeness of the use case between human-agent teaming experts and first response and disaster management experts (firefighters)

In general, the ratings given by firefighters to the statement ‘The scenario is easy to understand’ are more in line with one another than that the ratings of the HAT experts. For an indication of this for the formative evaluation, see Figure 7 and the corresponding summary statistics in Table 7 on the next page. All firefighters agreed with this statement, whereas two out of the nine HAT experts gave a different rating (one slightly agreed and another completely agreed). However, the firefighters are not that much in agreement with one another for the statement ‘It is a realistic scenario for which we need to be prepared’. One of the firefighters slightly disagreed, whereas the others agreed. And also the HAT do not give the same rating. Three of the HAT experts completely agree, four agree, and one is neutral towards the statement.

Reasons given by firefighters for agreeing with the statement that the scenario is easy to understand are amongst others: *“enough data is provided to help you understand what the problem is and how it should be addressed”* and “*Clear explanation and looks familiar”*. The HAT experts do however indicate that the issue lies within the implications and assumptions about prior knowledge: “*Scenario in itself is not difficult; issue is what the implications are, plus what you can assume about prior knowledge about e.g., capabilities of humans and robots”* and “*Basis of scenario and explanation are easy to understand. I would need more background knowledge or information to answer some questions.”*

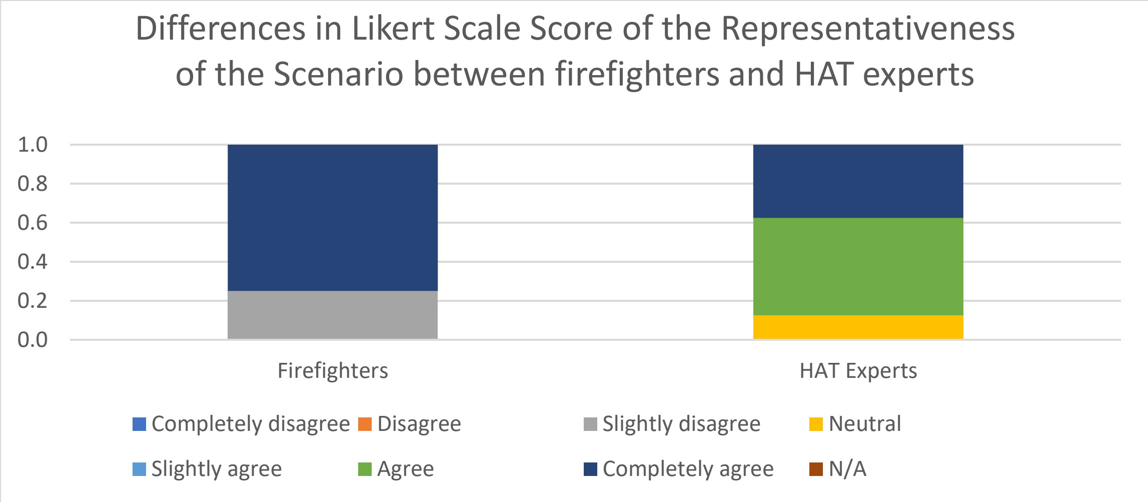



*Figure 7: Differences in Likert Scale Score between firefighters and HAT experts of the representative of the scenario.*

*Table 7: Summary statistics for the data presented in Figure 7.*

| | | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| Firefighters | The scenario is a realistic scenario for which we need to be prepared | 3.00 | 6.00 | 6.00 | 0.00 | 0.00 |
| HAT experts | The scenario is a realistic scenario for which we need to be prepared | 4.00 | 7.00 | 6.13 | 0.93 | 0.86 |

### 4.3.2 Evaluation of the work agreements: the expected useability of the current work agreements and ideas for additional work agreements

Below we will go over the results collected about the current WAs and the ideas provided for new WAs. First, we will address the subjective evaluation of all participants about the current WAs (subsection 4.3.2.1). After that, we will highlight the differences in subjective evaluation between the HAT experts and FF (subsection 4.3.2.2). Finally, we will go over the ideas given by the participants for other WAs that might be relevant (subsection 4.3.2.3.).

#### 4.3.2.1 The subjective evaluation of all participants about the current work agreements

On the question: “what kind of WAs could help for the desired level of flexibility in the human-AI collaboration” the participants answered that the three pre-defined WAs provided to them could support the HAT. These are the WAs that indicate who the notification should be sent to, how the notification should look like, and whether the task transfer from human to robot should be enabled.

Most of the participants indicated that the WAs are good and that they contribute to trust in working together. For an indication of this for the formative evaluation, see Figure 8 and the corresponding summary statistics in Table 8 on the next page. However, it was emphasized that the system cannot sent a robot to the location without a human being aware of it. And that a human should always oversee the robot. The robot thus cannot operate on its own. It was indicated that a system can make its own decisions about sending a robot to the location to take over the task of the human if this has been agreed upon in advance, however, someone must be able to monitor what the system is doing. Meaning the system cannot decide without a person having insight into the decision. The participants therefore were more in line with one another with the WA that indicates who to involve in the interaction and with the WA that indicates how the communication should look like, than the WA that indicates whether the AI agent could send autonomously a robot to the location to take over the task of the human agent. For an indication of this for the formative evaluation, see Figure 8 and the corresponding summary statistics in Table 8 below.

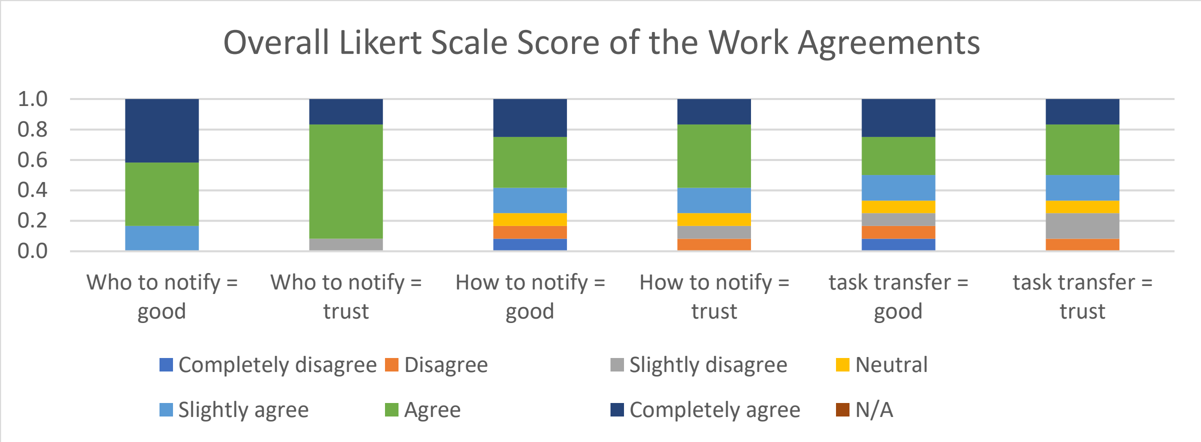



*Figure 8: Overall Likert Scale Score of all the participants of the formative evaluation on the work agreements.*

*Table 8: Summary statistics for the data presented in Figure 8.*

| The Work Agreement (WA) in which it is indicated … is a good WA | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| who should be informed/advised | 5.00 | 7.00 | 6.00 | 1.00 | 1.00 |
| if there should be information or advise about the change | 1.00 | 7.00 | 5.00 | 2.00 | 3.00 |
| if the Intelligent system can send a robot to take over the task | 1.00 | 7.00 | 5.00 | 2.00 | 4.00 |
| The WA in which it is indicated … contributes to trust | Min | Max | Mean | Std | Variance |
| who should be informed/advised | 3.00 | 7.00 | 6.00 | 1.00 | 1.00 |
| if there should be information or advise about the change | 2.00 | 7.00 | 5.00 | 1.00 | 2.00 |
| if the Intelligent system can send a robot to take over the task | 2.00 | 7.00 | 5.00 | 2.00 | 3.00 |

Most of the participants agree that if the safety of the human is threatened, and a robot can take over the task of the human, this should be done. However, the participants doubt the capabilities of robots. This is the main reason given why there should be supervision of a human over the robot and the need for a human agent to oversee the decision of an AI agent sending a robot to the location. This can, for example, be deduced from the following statements of one of the participants from the workshop with the Gezamenlijke Brandweer:

*"People are pretty smart, I say, compared to a robot. You have to tell the robot the same thing we know. Recently, for example, I had an HF leak at BP, one of us had to close a few valves, but yes 1 2 3 4 5 6… how are we going to do that? They all are close to each other and look alike. In such a situation you tell a robot to mark it to be able to turn it off. I think it stops there. Ultimately, the interpretation of the human being is behind it. We determine what to do. It can only be done right once."*

Furthermore, the workshop participants of the Gezamenlijke Brandweer and the fire brigade in Enschede agree that the amount of communication with the AI system depends on the role a firefighter has within the team. This means that the amount of communication with the AI system should be left to the bare minimum for an infield FF, and the higher we go up in the hierarchy, the more communication there can be. Showing the changing situation is only relevant for the team leader or higher in the hierarchy (commander). The infield firefighter is only concerned with the tasks and is not responsible for adjusting the plan. It is important for an infield FF to perform the task safely and if not that he knows that he needs to leave the location. This can, for example, be deduced from the following statements of one of the participants from the workshop with the Gezamenlijke Brandweer:

*"I think your notification should be some sort of standard format where you tell them this is what's going on, these are the readings and you could do this. And then it is done. And then the one who reads it. So that the message is correct for the role that is reading it. If you do, it doesn't matter. You have to leave now is sufficient for him (infield firefighter), but for the commander (team leader) he has to leave now because the measurements are so and so, and for the OvD watch your men leave because the values are now this exclamation mark exclamation sign too high so advice is this and this and that. Layering. If you don't do that and send what the OvD is supposed to know to the troops, they will panic while there may be nothing wrong. "*

Apart from that, it was indicated by the firefighters of the Spaansland fire brigade in Enschede that a notification should only be given if it is proper to give the notification given the location of the firefighter. If the system gives notifications when these are not needed, the system will not be used anymore as it is not trusted anymore. An example given of a situation in which a warning notification is not needed is when an infield FF is working between buildings. Between buildings, the rotation of the wind direction and change in wind speed does not matter for working on the task. The infield firefighter should not be notified; this is distracting instead of supporting.

When we compare the answers of the GB and Enschede firefighters, we can see that the opinions about the current WAs are similar to one another. And thus, that the perspective on the current WAs between the GB and Enschede firefighters is the same. The most important opinion provided by them is that any notification shared with an infield FF should be to the point. It should contain the bare minimum amount of information needed for an infield FF to be alerted about the changing circumstances. They agree with one another that the higher the position of the firefighter in the hierarchy the more communication there can be with the AI agent. Furthermore, they agree with one another that the decisions and actions of the robot should be made under the supervision of a human agent that can provide input or take over the control of the robot when needed.

#### 4.3.2.2 The differences in perspective on the current work agreements between human-agent teaming experts and first response and disaster management experts (firefighters)

In general, compared to the HAT experts, the firefighters who evaluated the WAs gave a slightly higher score on the statements about the WAs contributing to trust. See Figures 9, 10, and 11 and their corresponding summary statistics in Table 9, 10, and 11 on the next page.

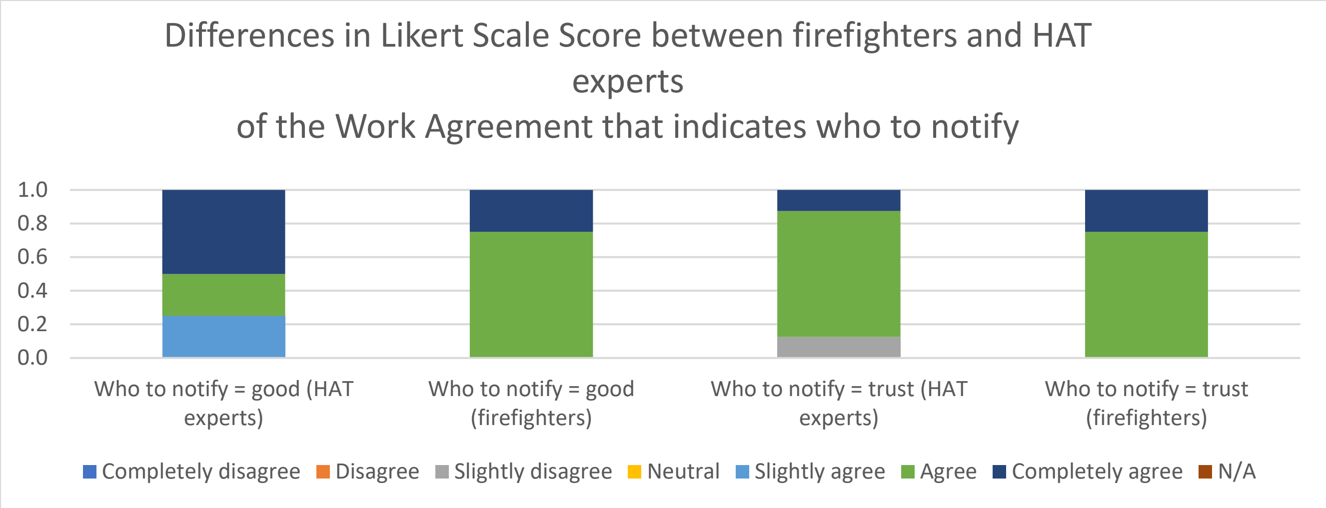



*Figure 9: Differences in Likert Scale Score between firefighters and HAT experts of the Work Agreement that indicates who to notify.*

*Table 9: Summary statistics for the data presented in Figure 9.*

| | The Work Agreement (WA) in which it is indicated ... is a good WA | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | who should be informed/advised | 5.00 | 7.00 | 6.25 | 0.83 | 0.69 |
| Firefighters | who should be informed/advised | 6.00 | 7.00 | 6.25 | 0.43 | 0.19 |
| | The WA in which it is indicated ... contributes to trust | Min | Max | Mean | Std | Variance |
| HAT experts | who should be informed/advised | 3.00 | 7.00 | 5.75 | 1.09 | 1.19 |
| Firefighters | who should be informed/advised | 6.00 | 7.00 | 6.25 | 0.43 | 0.19 |

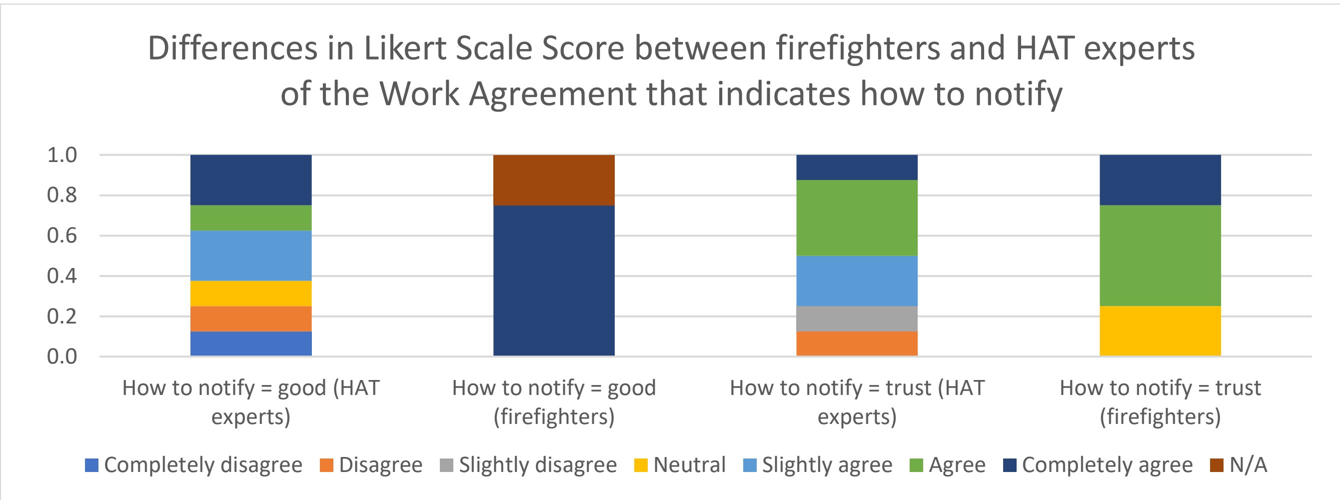



*Figure 10: Differences in Likert Scale Score between firefighters and HAT experts of the Work Agreement that indicates how to notify.*

*Table 10: Summary statistics for the data presented in Figure 10.*

| | The Work Agreement (WA) in which it is indicated … is a good WA | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | if there should be information or advise about the change | 1.00 | 7.00 | 4.63 | 2.06 | 4.23 |
| Firefighters | who should be informed/advised | 6.00 | 7.00 | 6.25 | 0.43 | 0.19 |
| | **The WA in which it is indicated … contributes to trust** | **Min** | **Max** | **Mean** | **Std** | **Variance** |
| HAT experts | if there should be information or advise about the change | 3.00 | 7.00 | 5.00 | 1.58 | 2.50 |
| Firefighters | who should be informed/advised | 4.00 | 7.00 | 5.75 | 1.09 | 1.19 |

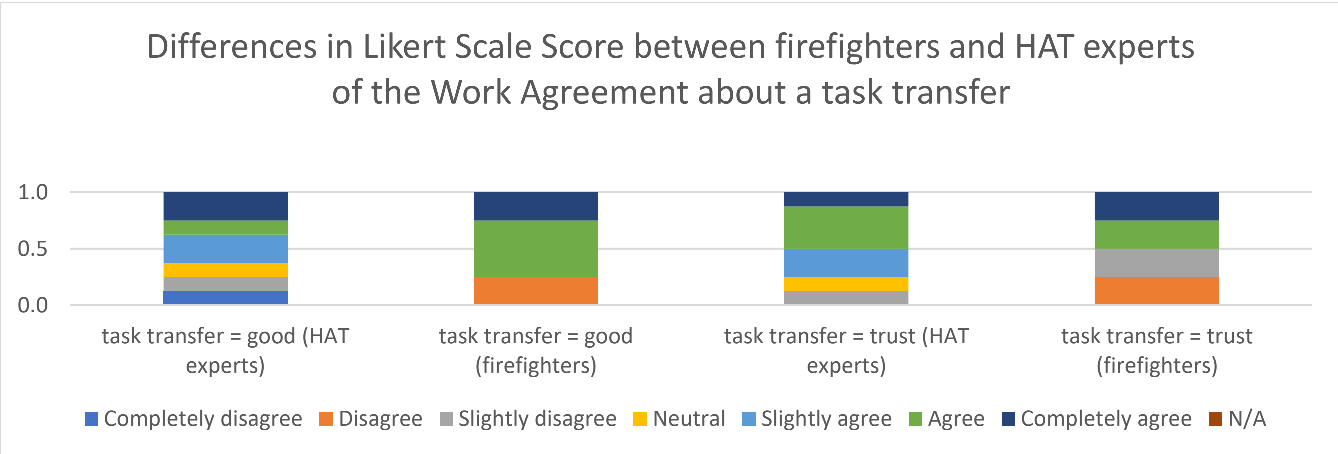



*Figure 11: Differences in Likert Scale Score between firefighters and HAT experts of the Work Agreement about a task transfer*

*Table 11: Summary statistics for the data presented in Figure 11.*

| | The Work Agreement (WA) in which it is indicated … is a good WA | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | if the Intelligent system can send a robot to take over the task | 1.00 | 7.00 | 4.75 | 1.92 | 3.69 |
| Firefighters | if the Intelligent system can send a robot to take over the task | 2.00 | 7.00 | 5.25 | 1.93 | 3.69 |
| | **The WA in which it is indicated … contributes to trust** | **Min** | **Max** | **Mean** | **Std** | **Variance** |
| HAT experts | if the Intelligent system can send a robot to take over the task | 3.00 | 7.00 | 5.25 | 1.20 | 1.44 |
| Firefighters | if the Intelligent system can send a robot to take over the task | 2.00 | 7.00 | 4.50 | 2.06 | 4.25 |

Within the formative evaluation process, the participants (8 HAT experts, 4 GB firefighters) went over the use case, i.e., the scenario in which the participants had to decide about what WAs to manage and how to deal with adversarial changes in the environment (changes in weather condition) with the help of the AI agent. The participants were asked to decide whether to keep the three WAs as they are or to adapt them. Overall, the results illustrate that most of the participants (nine out of twelve) like the WAs as they are; however, three participants choose to change at least one of the WAs. The two HAT experts and the firefighter that decided to change the WAs disabled the task transfer option. These participants indicated that the system cannot make this decision independently if not all the conditions under which a robot can operate can be made clear.

Furthermore, one of the HAT experts changed the notification type from advice to information. This participant later indicated to prefer receiving information and advice rather than choosing between information and advice.

The question whether it is good to allow the system to send a robot to the location was answered by seven of the participants as good, whereas five did not allow the system to make this decision.

When the same participants had to rate the WAs, they slightly agree that deciding about what the notification should look like and task transfer are good WAs that contribute to trust and agree that the WA about who should be notified is a good WA that contributes to trust. For each of the WAs an in-depth analysis will now be provided about the ratings the participants gave:

- In general, the participants agree with the statement '*I think the work agreement which specifies who should be informed/advised about changing circumstances is a good work agreement*'. See Figure 12 and the corresponding summary statistics in Table 12. Ratings vary from slightly agree to completely agree with one exception. Five of the participants completely agreed, five agreed, and two slightly agreed with the statement. The two that slightly agreed are HAT experts. Only one participant (a HAT expert) slightly disagreed with the statement that this work agreement contributes to trust in collaboration with the AI agent.

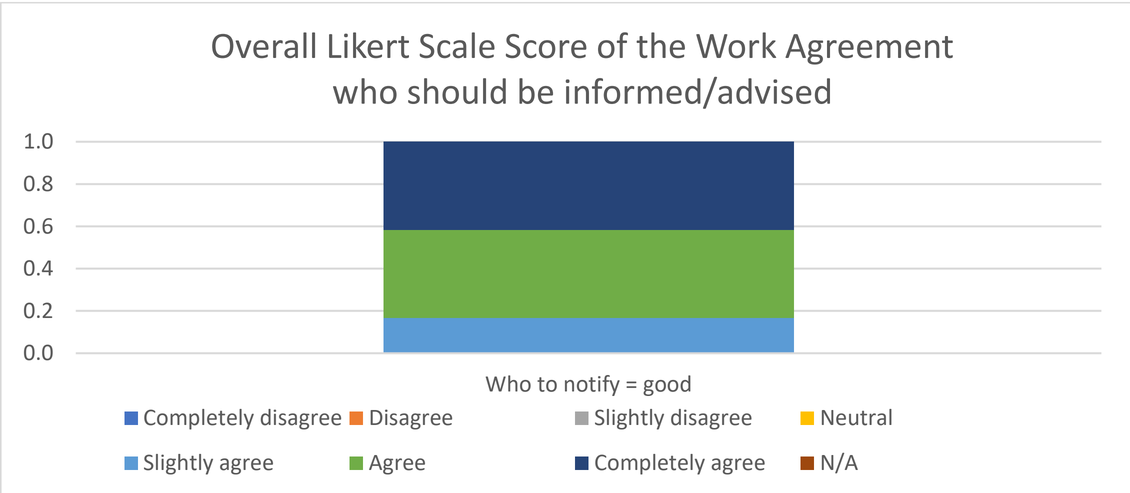



*Figure 12: Overall Likert Scale Score of all the participants of the formative evaluation of the work agreements who should be informed/advised.*

*Table 12: Summary statistics for the data presented in Figure 12.*

| The Work Agreement (WA) in which it is indicated … is a good WA | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| who should be informed/advised | 5.00 | 7.00 | 6.00 | 1.00 | 1.00 |

- The statement '*I think the work agreement which specifies if informed or advised should be given about the changing circumstances is a good work agreement'* was slightly differently rated by HAT experts and firefighters. See Figure 13 and the corresponding summary statistics in Table 13. Generally, the participants slightly agree with the statement, with ratings varying between completely disagree to completely agree. Three of the firefighters agreed and one completely agreed with the statements, whereas the HAT experts gave a rating from completely disagree to completely agree. Even though the ratings of the HAT expert vary a lot about whether this is a good working agreement six of the HAT experts do think that this work agreement could contribute to trust in the collaboration with the AI agent (two slightly agreed, three agreed, one completely agreed). Three of the firefighters agree as well, the other firefighter rated the statement with neutral. This means that the participants slightly agree that the work agreement contributes to trust in the collaboration with the AI agent.

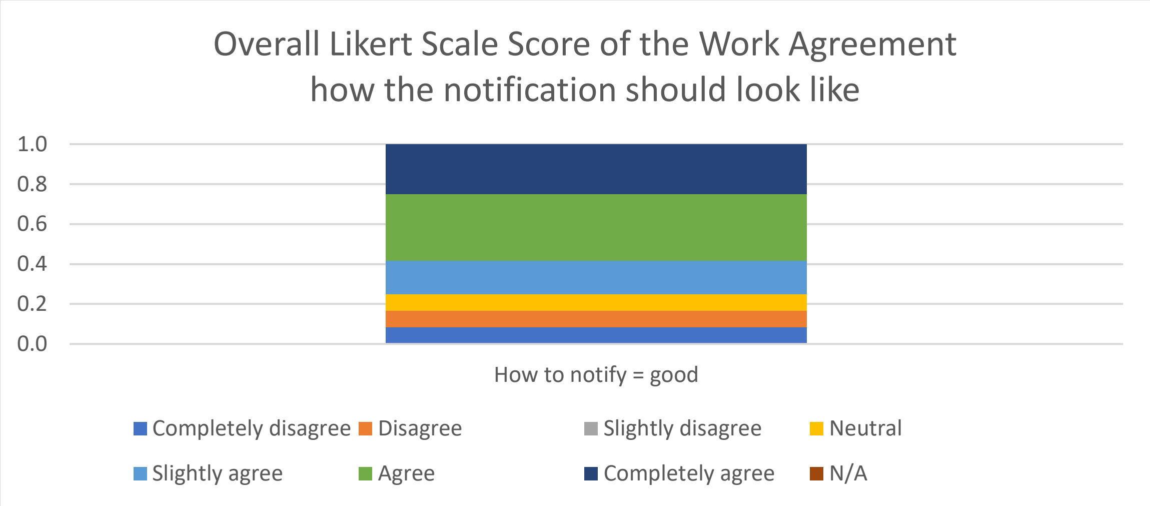



*Figure 13: Overall Likert Scale Score of all the participants of the formative evaluation of the work agreement how the notification should look like.*

*Table 13: Summary statistics for the data presented in Figure 13.*

| The Work Agreement (WA) in which it is indicated … is a good WA | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| if there should be information or advise about the change | 1.00 | 7.00 | 5.00 | 2.00 | 3.00 |

- In general, the participants slightly agree with the statement '*I think the work agreement which specifies if the intelligent system can send a robot with the right capabilities to the location to take over the task of the human agent is a good work agreement*'. See Figure 14 and the corresponding summary statistics in Table 14. Eight of the participants agreed with the statement, one HAT expert completely disagreed, one firefighter disagreed, one HAT expert slightly disagreed, and one HAT expert was neutral about the statement. Also, the ratings for the statement '*I think the work agreement which specifies if the intelligent system can send a robot with the right capabilities to the location to take over the task of the human contributes to trust in the collaboration*' were quite different from one another. Generally, the participants slightly agree that the work agreement contributes to trust in collaboration with the AI agent. One firefighter and one HAT expert completely agreed, three HAT experts agreed, one HAT expert slightly agreed, one HAT expert rated neutral, one firefighter and HAT expert slightly disagreed, and one firefighter and HAT expert disagreed.

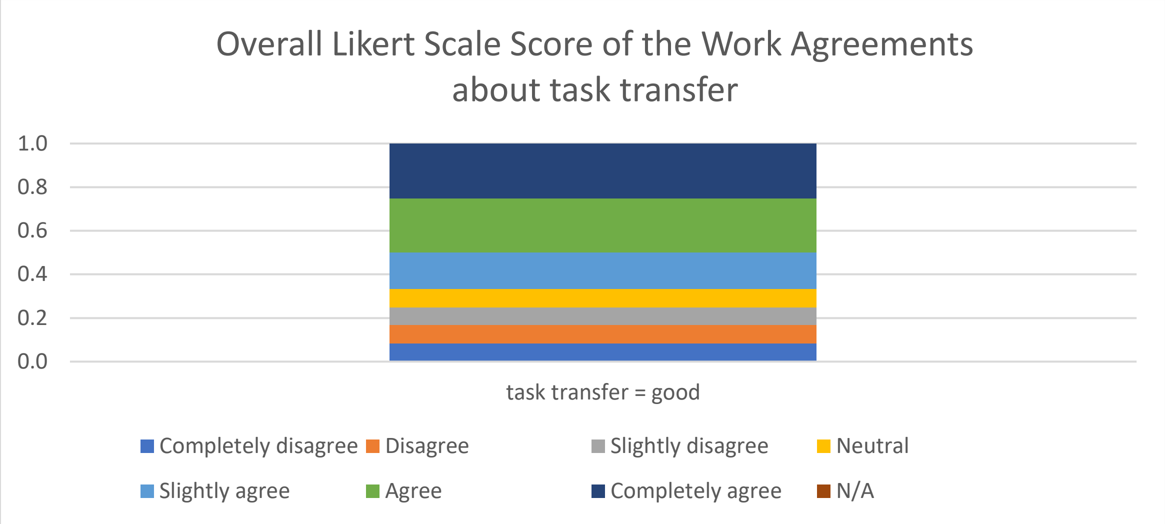



*Figure 14: Overall Likert Scale Score of all the participants of the formative evaluation of the work agreement about task transfer from human to robot.*

`

*Table 14: Summary statistics for the data presented in Figure 14.*

| The Work Agreement (WA) in which it is indicated … is a good WA | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| if the Intelligent system can send a robot to take over the task | 1.00 | 7.00 | 5.00 | 2.00 | 4.00 |

#### 4.3.2.3 Ideas provided by the participants for additional work agreements that should be considered

The participants also indicated additional configurations for WAs or additional WAs that could be considered, namely configurations for WAs or additional WAs that give the ability to…

- not only decide to interrupt the task based on weather changes but also based on the size and temperature of the fire and/or the health status of the firefighter;
- have more nuance in the options to choose from rather than only being able to enable or disable the WA, for example, for the task transfer WA;
- choose auditive notifications and voice-interaction instead of only visual communication with the AI agent;
- watch along with the firefighter via body cams;
- being able to supervise the robot with camera vision;
- capture important information (keywords) from mutual conversations between infield FFs during missions that the team leader needs to be aware of or for further notice in the emergency center and on the press monitor.

### 4.3.3. Evaluation of whether the TDP specification describes the human-AI decision process in a clear, concise, complete and easy-to-understand manner

#### 4.3.3.1 Overall subjective evaluation of whether the TDP specification describes the human-AI decision process in a clear, concise, complete and easy-to-understand manner

Most participants indicated that the TDP specification describes the human-AI decision process in a clear, concise, complete, and easy-to-understand manner. For each of the TDPs, the overall opinion of the participants is described below, followed by an in-depth description of the differences between HAT experts and firefighters in Section 4.3.3.2.

The participants slightly agree that the TDP for 'collaborative management of the working agreements' is easy to understand and agree that the images and textual information of the pattern complement each other well. See Figure 15 and the corresponding summary statistics in Table 15.

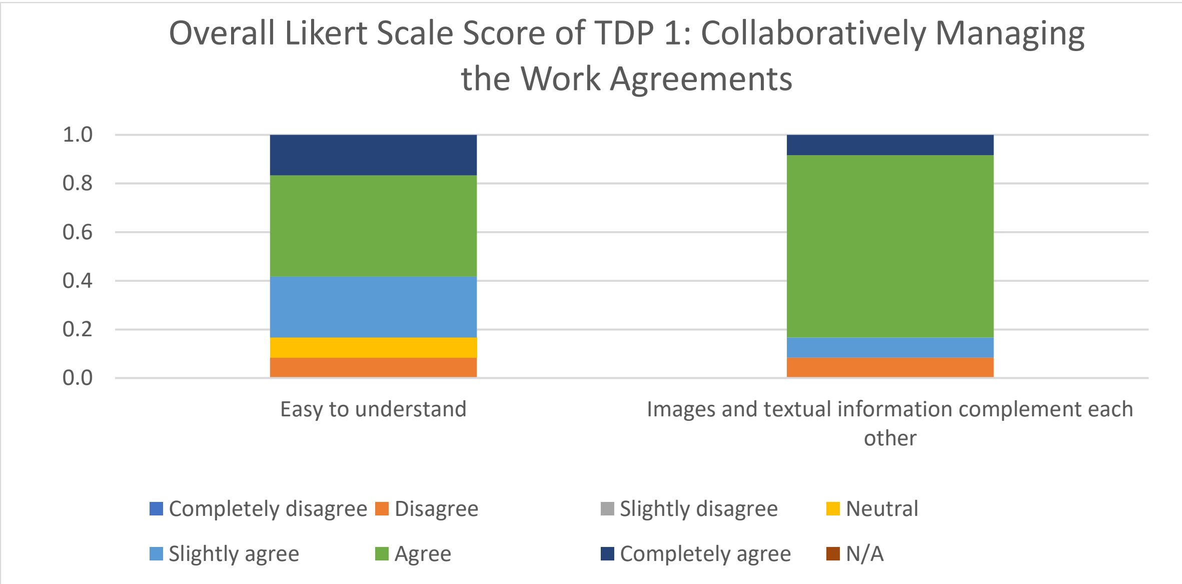



*Figure 15: Overall Likert Scale Score of all the participants of the formative evaluation of TDP 1: Collaboratively Managing the Work Agreements.*

*Table 15: Summary statistics for the data presented in Figure 15.*

| TDP 1: Collaboratively Managing the Work Agreements | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| The proposed design solution is easy to understand | 2.00 | 7.00 | 5.00 | 1.00 | 2.00 |
| The images and textual information of the pattern complement each other | 2.00 | 7.00 | 6.00 | 1.00 | 1.00 |

Furthermore, the participants agree that the TDP for 'dealing with changing circumstances' is easy to understand and agree that the images and textual information of the pattern complement each other well. See Figure 16 and the corresponding summary statistics in Table 16.

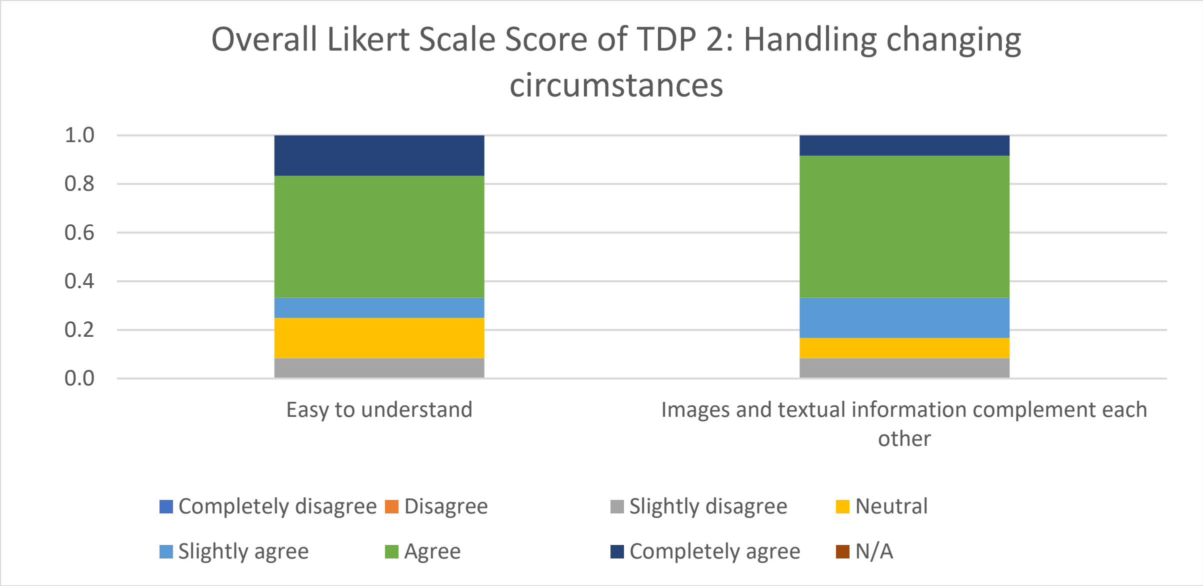



*Figure 16: Overall Likert Scale Score of all the participants of the formative evaluation of TDP 2: Handling changing circumstances.*

*Table 16: Summary statistics for the data presented in Figure 16.*

| TDP 2: Handling changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| The proposed design solution is easy to understand | 3.00 | 7.00 | 6.00 | 1.00 | 1.00 |
| The images and textual information of the pattern complement each other | 3.00 | 7.00 | 6.00 | 1.00 | 1.00 |

Finally, the participants slightly agree that the TDP for 'independently reacting to changing circumstances' is easy to understand and agree that the images and textual information of the pattern complement each other well. See Figure 17 and the corresponding summary statistics in Table 17.

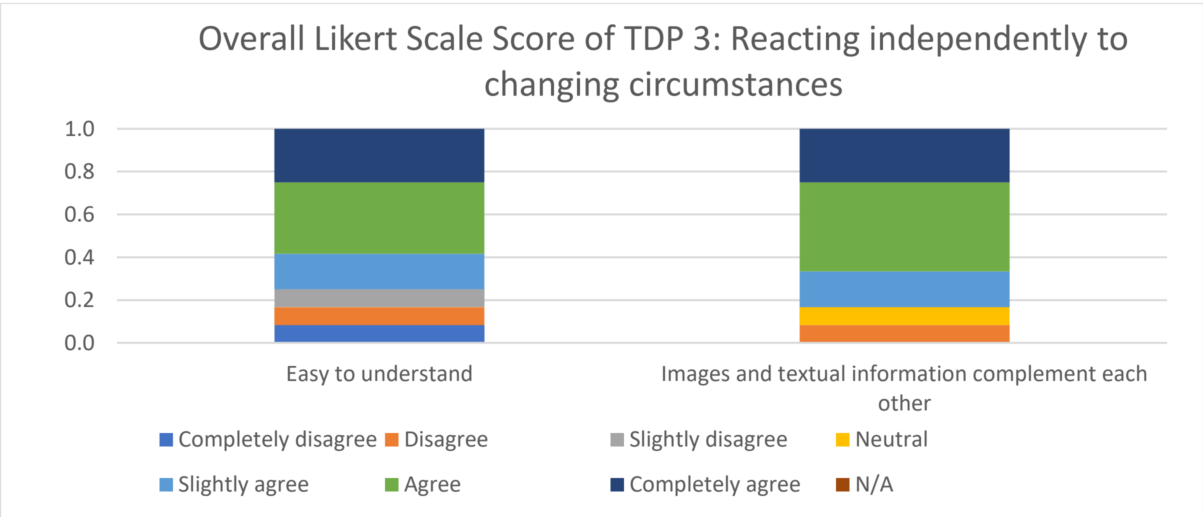



*Figure 17: Overall Likert Scale Score of all the participants of the formative evaluation of TDP 3: Reacting independently to changing circumstances.*

Table 17: Summary statistics for the data presented in Figure 17.

| TDP 3: Reacting independently to changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| The proposed design solution is easy to understand | 1.00 | 7.00 | 5.00 | 2.00 | 4.00 |
| The images and textual information of the pattern complement each other | 2.00 | 7.00 | 6.00 | 1.00 | 2.00 |

#### 4.3.3.2 The differences in perspective between human-agent teaming experts and firefighters about the TDP specification being clear, concise, complete and easy-to-understand

The ratings of the firefighters for the statement that the TDP specification describe the human-AI decision process in a clear, concise, complete, and easy-to-understand manner are more in line with one another than the ratings of the HAT experts. Furthermore, firefighters agree more with the statement than the HAT experts do.

The ratings for 'the collaborative management of work agreements' TDP vary from disagree to completely agree. For the statement that "*the proposed design pattern of collaborative management of the working agreements is easy to understand*", this means that one disagrees, one is neutral, three slightly agree, five agree, and two completely agree. See Figure 18 and the corresponding summary statistics in Table 18.

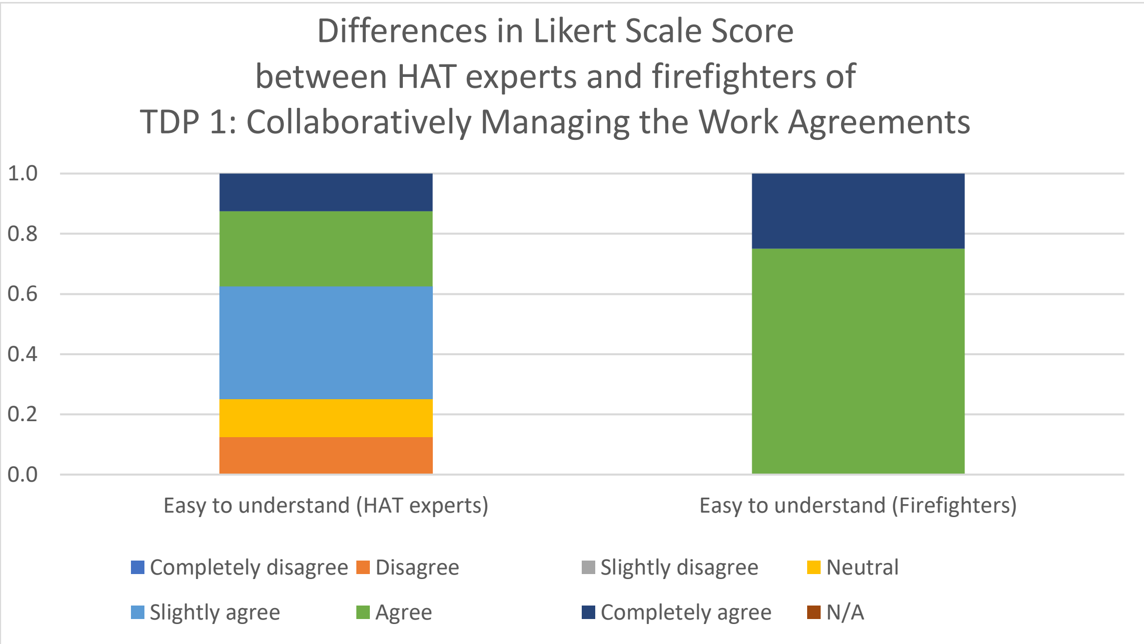



Figure 18: Differences in Likert Scale Score between HAT experts and firefighters on whether TDP 1: Collaboratively Managing the Work Agreements is easy to understand.

Table 18: Summary statistics for the data presented in Figure 18.

| | TDP 1: Collaboratively Managing the Work Agreements | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | The proposed design solution is easy to understand | 2.00 | 7.00 | 5.00 | 1.41 | 2.00 |
| Firefighters | The proposed design solution is easy to understand | 6.00 | 7.00 | 6.25 | 0.43 | 0.19 |

Some of the HAT experts did have questions about the task transition, especially from step 3 to 4, such as the following participant:

‘I miss the interpretation of the context in which the two tasks take place, or is this irrelevant? *e.g., The left block in which the work agreements are made takes place before the start of the work, and the right block '(de) activate work agreements during the execution of the work, right? Maybe not, but the task transit is not clear to me. - I can immediately imagine tasks 1-3, but task 4 (de-) activating the work requires more explanation, which human value plays a role, how when does this happen? Or will that be discussed in a later pattern? - Why do human value not play a role in Task 1-3, but in Task 4. In Task 1-3, the conditions that determine task 4 are included?*’

The scores for the statement that the images and textual information of ‘the collaborative management of the working agreements’ TDP complement each other well are that: one disagrees, one slightly agrees, nine agree, and one completely agrees. See Figure 19 and the corresponding summary statistics in Table 19.

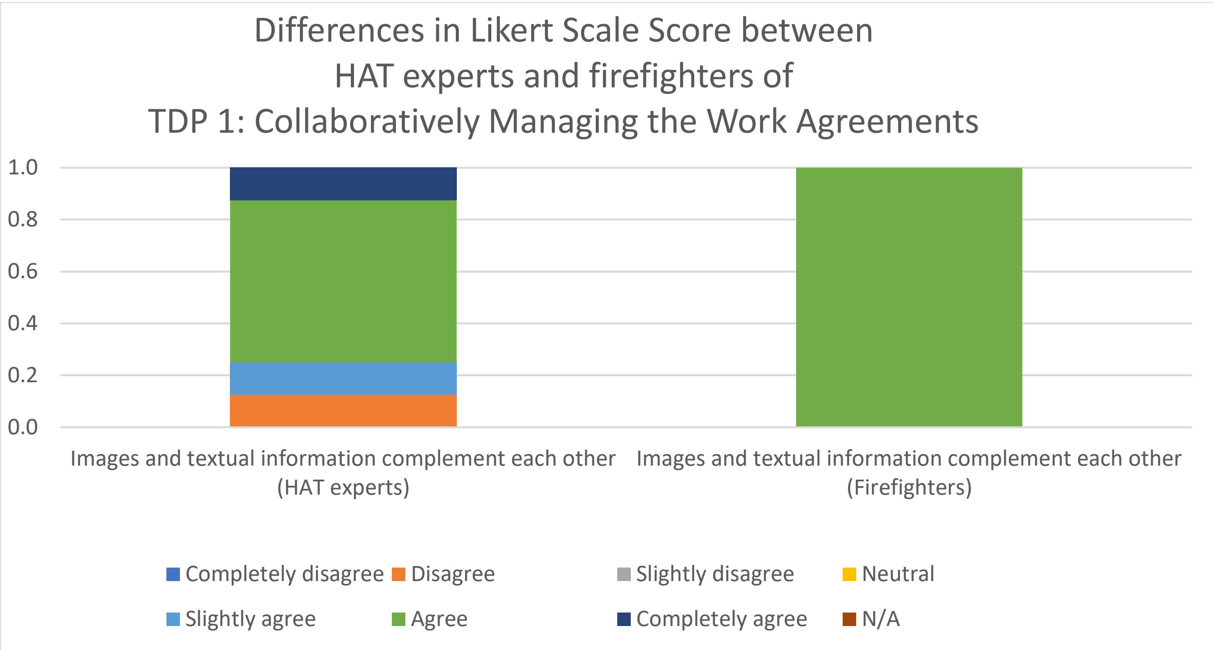



*Figure 19: Differences in Likert Scale Score between HAT experts and firefighters on whether the images and textual information of TDP 1: Collaboratively Managing the Work Agreements complement each other well.*

*Table 19: Summary statistics for the data presented in Figure 19.*

| | TDP 1: Collaboratively Managing the Work Agreements | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | The images and textual information of the pattern complement each other | 1.00 | 8.00 | 5.63 | 1.93 | 3.73 |
| Firefighters | The images and textual information of the pattern complement each other | 6.00 | 6.00 | 6.00 | 0.00 | 0.00 |

One of the firefighters pointed out that it should not be too much text and three HAT experts pointed out that the legend is necessary and can help for understanding.

The ratings of ‘the dealing with changing circumstances’ TDP vary from slightly disagree to completely agree. For the statement that the proposed design pattern of dealing with changing circumstances is easy to understand, this means that one slightly disagrees, two are neutral, one slightly agrees, six agree, and two completely agree. For the statement that the images and textual information of 'the dealing with changing circumstances' TDP complement each other well this means that one slightly disagrees, one is neutral, two slightly agree, seven agree and one completely agrees. See Figure 20 and the corresponding summary statistics in Table 20.

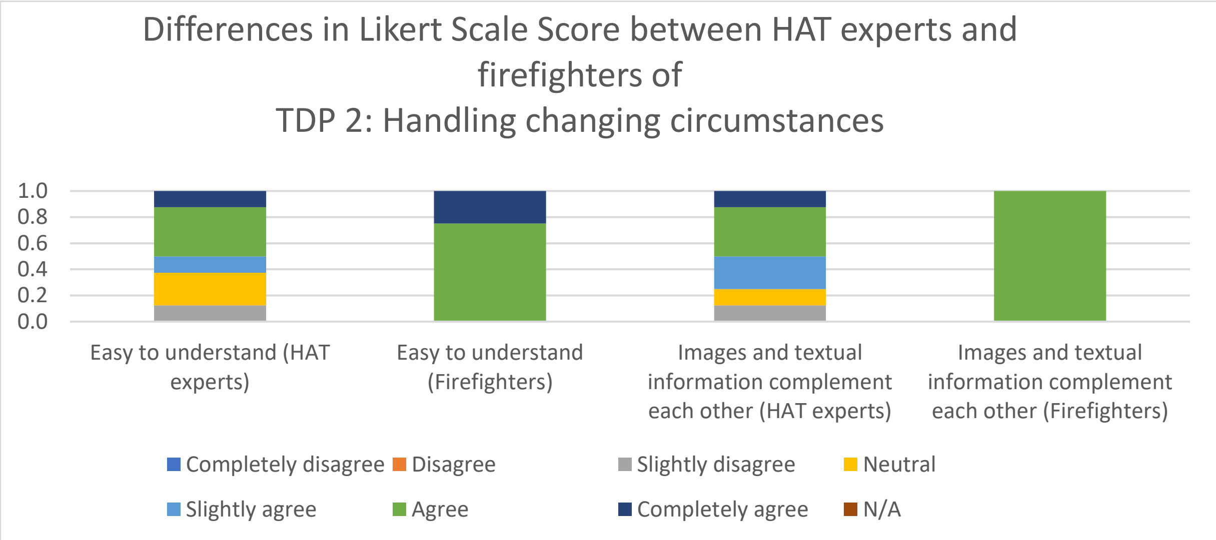



*Figure 20: Differences in Likert Scale Score between HAT experts and firefighters on whether information of TDP 2: Handling changing circumstances is easy to understand, and the images and textual information complement each other well.*

*Table 20: Summary statistics for the data presented in Figure 20.*

| | TDP 2: Handling changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | The proposed design solution is easy to understand | 3.00 | 7.00 | 5.13 | 1.27 | 1.61 |
| Firefighters | The proposed design solution is easy to understand | 6.00 | 7.00 | 6.25 | 0.43 | 0.19 |
| HAT experts | The images and textual information of the pattern complement each other | 3.00 | 7.00 | 5.25 | 1.20 | 1.44 |
| Firefighters | The images and textual information of the pattern complement each other | 6.00 | 6.00 | 6.00 | 0.00 | 0.00 |

Some of the HAT experts did have questions about the order of tasks as some tasks are parallel while others are sequential such as described by the following participant:

*'task 1 and 2 seem like parallel tasks to me, can this also be visualized? Numbering makes it seem sequential.'*

The ratings for the TDP 'reacting to changing circumstances' vary from completely disagree to completely agree. For the statement that the proposed design pattern of 'independently reacting to changing circumstances' is easy to understand this means that one completely disagrees, one disagrees, one is neutral, two slightly agree, four agree and three completely agree. For the statement that the images and textual information of the pattern complement each other well this means that one disagrees, one is neutral, two slightly agree, five agree and three completely agree. See Figure 21 and the corresponding summary statistics in Table 21.

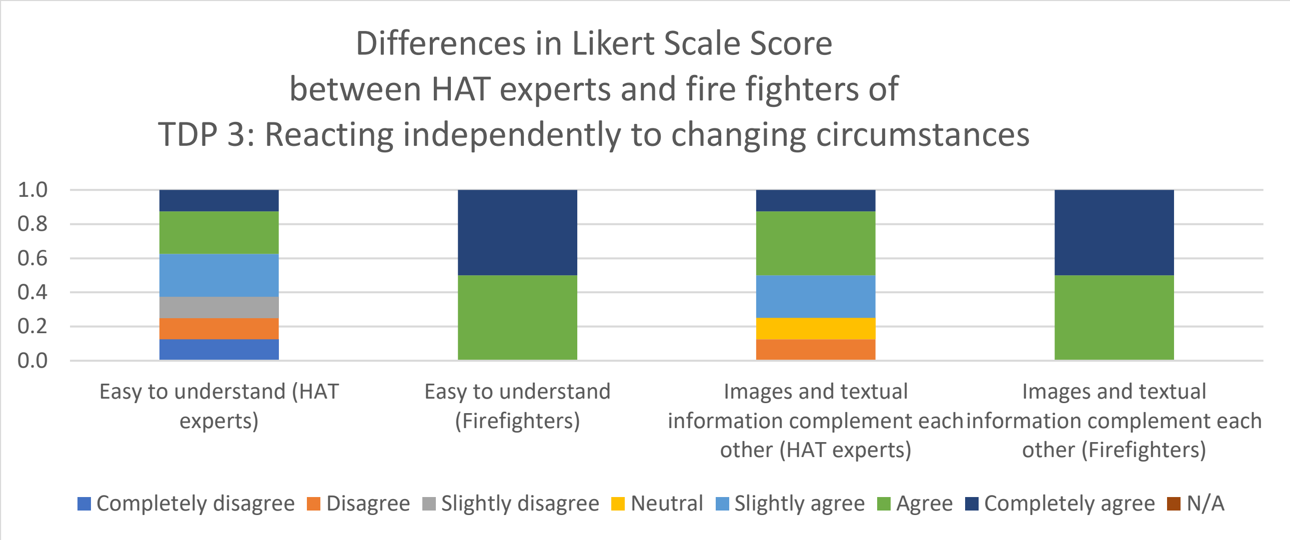



*Figure 21: Differences in Likert Scale Score between HAT experts and firefighters on whether of TDP 3: Reacting independently to changing circumstances is easy to understand, and the images and textual information complement each other well.*

*Table 21: Summary statistics for the data presented in Figure 21.*

| | TDP 3: Reacting Independently to changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | The proposed design solution is easy to understand | 1.00 | 7.00 | 4.38 | 2.00 | 3.98 |
| Firefighters | The proposed design solution is easy to understand | 6.00 | 7.00 | 6.50 | 0.50 | 0.25 |
| HAT experts | The images and textual information of the pattern complement each other | 2.00 | 7.00 | 5.13 | 1.45 | 2.11 |
| Firefighters | The images and textual information of the pattern complement each other | 6.00 | 7.00 | 6.50 | 0.50 | 0.25 |

Some of the HAT experts did have questions about task four and five as described by the following participant:

*‘ratio between 4 and 5 is difficult, does 5 overrules 4? is it not better to inform the human of the choice made?’*

### 4.3.4. Evaluation of whether the TDP specification describe collaborative AI behaviors that are expected to improve team performance with a value-sensitive aspect substantially

#### 4.3.4.1 Overall subjective evaluation whether the TDP specification describe collaborative AI behaviors that are expected to improve team performance with a value-sensitive aspect substantially

Overall, the participants agree that the described collaborative AI behaviors are expected to improve team performance with a value-sensitive aspect substantially. Most of the participants indicated that the TDP specification describes the collaborative AI behaviors expected to improve team performance with a value-sensitive aspect substantially. For each of the TDPs, the overall opinion of the participants is described below, followed by an in-depth description of the differences between HAT experts and fighters in Section 4.3.4.2.

The participants agree that the implementation of collaborative management of the working agreements results in good decision-making. And that this TDP can be applied in other situations involving similar choices. See Figure 22 and the corresponding summary statistics in Table 22.

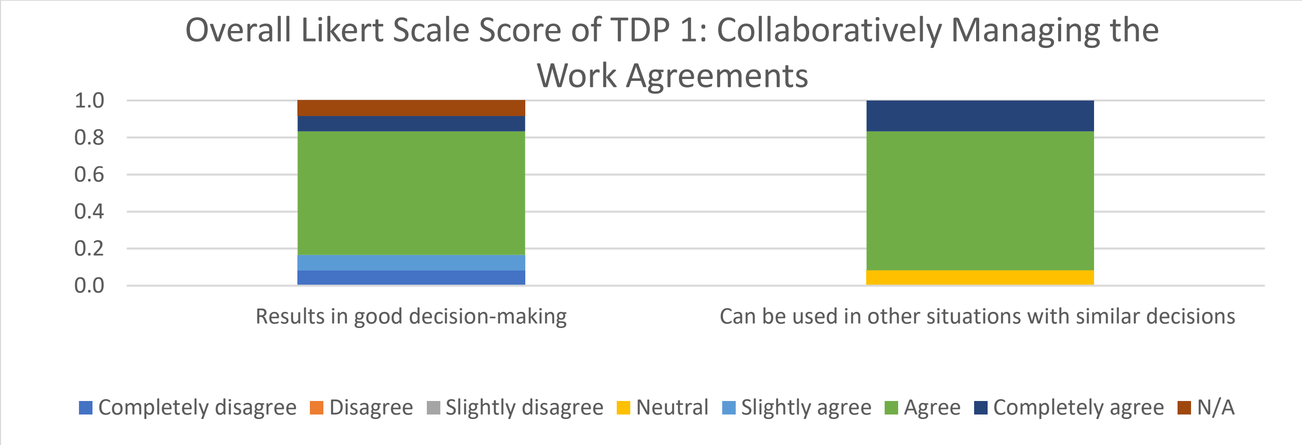



*Figure 22: Overall Likert Scale Score of all the participants of the formative evaluation on whether TDP 1: Collaboratively Managing the Work Agreements results in good decision-making and can be used in situations in which similar decisions need to be made.*

*Table 22: Summary statistics for the data presented in Figure 22.*

| TDP 1: Collaboratively Managing the Work Agreements | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| The implementation of the design pattern will result in good decision-making | 1.00 | 8.00 | 6.00 | 2.00 | 3.00 |
| The implementation of the design pattern can be used in other situations with similar decisions | 4.00 | 7.00 | 6.00 | 1.00 | 1.00 |

Furthermore, the participants slightly agree that the implementation of ‘dealing with changing circumstances’ results in good decision-making and agree that this team design pattern can be applied in other situations involving similar choices. See Figure 23 and the corresponding summary statistics in Table 23.

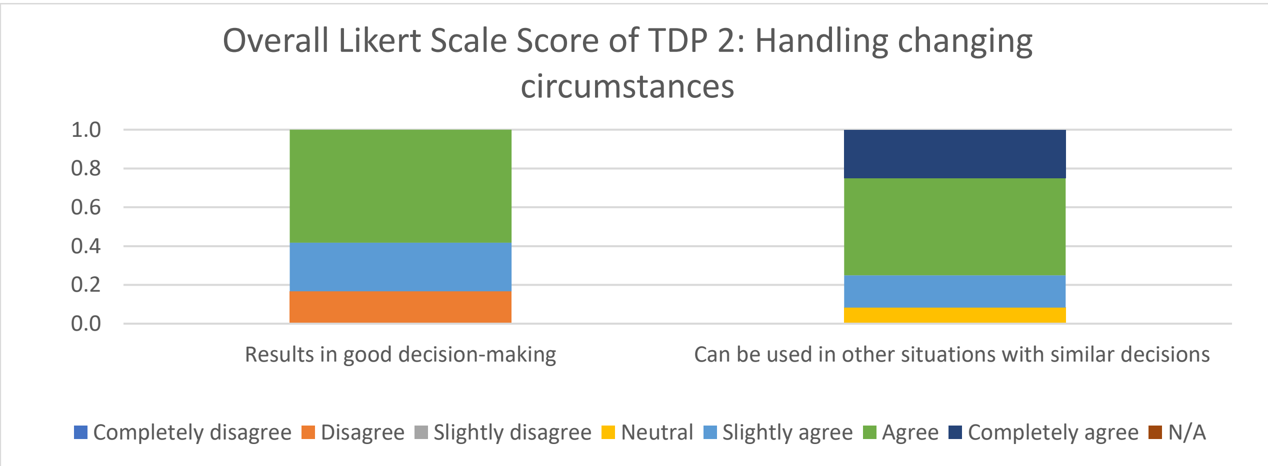



*Figure 23: Overall Likert Scale Score of all the participants of the formative evaluation on whether TDP 2: Handling changing circumstances results in good decision-making and can be used in situations in which similar decisions need to be made.*

Table 23: Summary statistics for the data presented in Figure 23.

| TDP 2: Handling changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| The implementation of the design pattern will result in good decision-making | 2.00 | 6.00 | 5.00 | 1.00 | 2.00 |
| The implementation of the design pattern can be used in other situations with similar decisions | 4.00 | 7.00 | 6.00 | 1.00 | 1.00 |

Finally, the participants slightly agree that the implementation independently reacting to changing circumstances results in good decision making and slightly agree that this team design pattern can be applied in other situations involving similar choices. See Figure 24 and the corresponding summary statistics in Table 24.

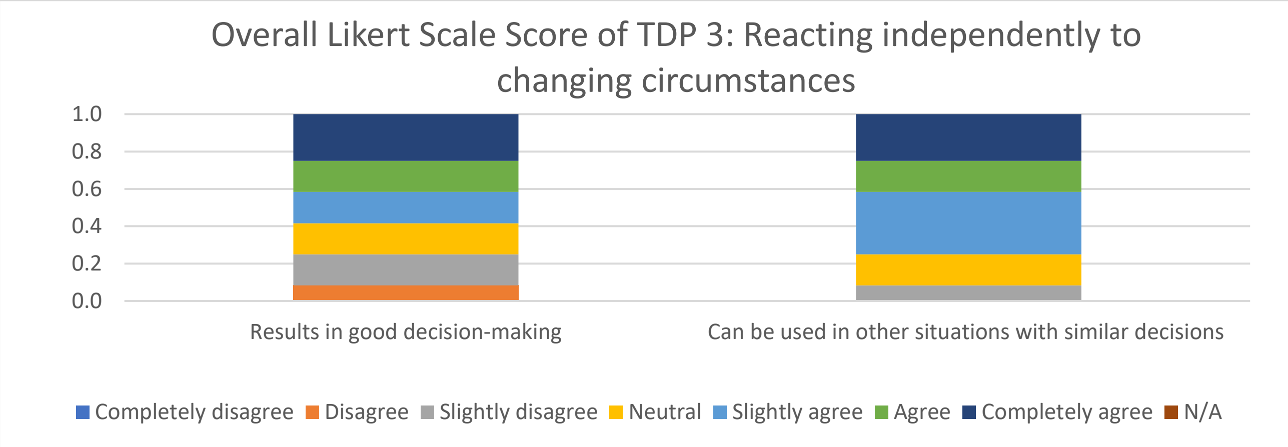



Figure 24: Overall Likert Scale Score of all the participants of the formative evaluation on whether TDP 3: reacting independently to changing circumstances results in good decision-making and can be used in situations in which similar decisions need to be made.

Table 24: Summary statistics for the data presented in Figure 24.

| TDP 3: Reacting independently to changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|
| The implementation of the design pattern will result in good decision-making | 2.00 | 7.00 | 5.00 | 2.00 | 3.00 |
| The implementation of the design pattern can be used in other situations with similar decisions | 3.00 | 7.00 | 5.00 | 1.00 | 2.00 |

#### 4.3.4.2 The differences in perspective between human-agent teaming experts and firefighters regarding the effect of the AI behaviours on team performance for value-sensitive decisions

The ratings of the firefighters are more in line with one another than the ratings of the HAT experts (see figures 25, 26, and 27 below). Compared to HAT experts, firefighters agree more with the statement that TDP specifications describe collaborative AI behaviors expected to improve team performance with a value-sensitive aspect substantially.

The ratings towards the statement that the TDP about collaborative management of WAs results in good decision-making vary from completely disagree to completely agree, e.g., one participant completely disagrees, one slightly agrees, eight agree, one completely agrees, and one filled in N/A. See Figure 25 and the corresponding summary statistics in Table 25 below.

` The ratings towards the statement that the TDP about collaborative management of WAs can be applied to other situations involving similar choices vary from neutral to completely agree, e.g. one participant is neutral about it, nine agree, and two completely agree. See Figure 25 and the corresponding summary statistics in Table 25 below.

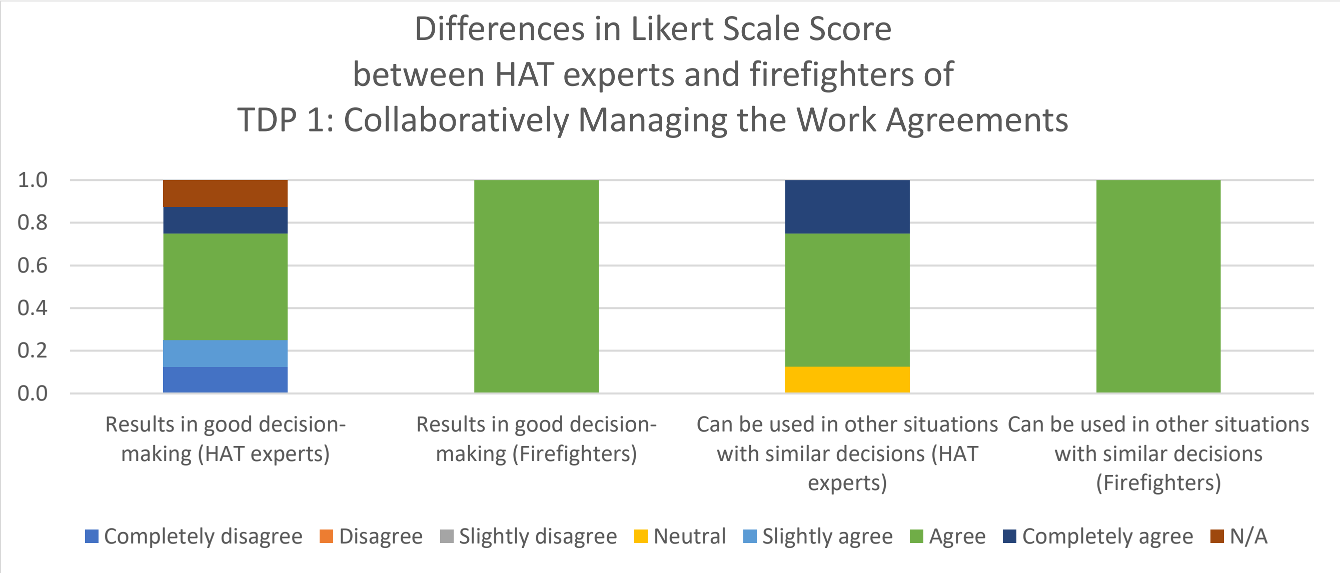



*Figure 25: Differences in Likert Scale Score between firefighters and HAT experts on whether TDP 1: collaborative management of work agreements results in good decision-making and can be used in situations in which similar decisions need to be made.*

*Table 25: Summary statistics for the data presented in Figure 25.*

| | TDP 1: Collaboratively Managing changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | The implementation of the design pattern will result in good decision-making | 1.00 | 8.00 | 5.63 | 1.93 | 3.73 |
| Firefighters | The implementation of the design pattern will result in good decision-making | 6.00 | 6.00 | 6.00 | 0.00 | 0.00 |
| HAT experts | The implementation of the design pattern can be used in other situations with similar decisions | 4.00 | 7.00 | 6.00 | 0.87 | 0.75 |
| Firefighters | The implementation of the design pattern can be used in other situations with similar decisions | 6.00 | 6.00 | 6.00 | 0.00 | 0.00 |

The ratings towards the statement that the TDP about handling with changing circumstances results in good decision-making vary from disagree to agree, e.g., two participants disagree, three slightly agree, seven agree. See Figure 26 and the corresponding summary statistics in Table 26.

The ratings towards the statement that the TDP about handling changing circumstances can be applied to other situations involving similar choices vary from neutral to completely agree, e.g., one participant is neutral about it, two slightly agree, six agree, and three completely agree. See Figure 26 and the corresponding summary statistics in Table 26.

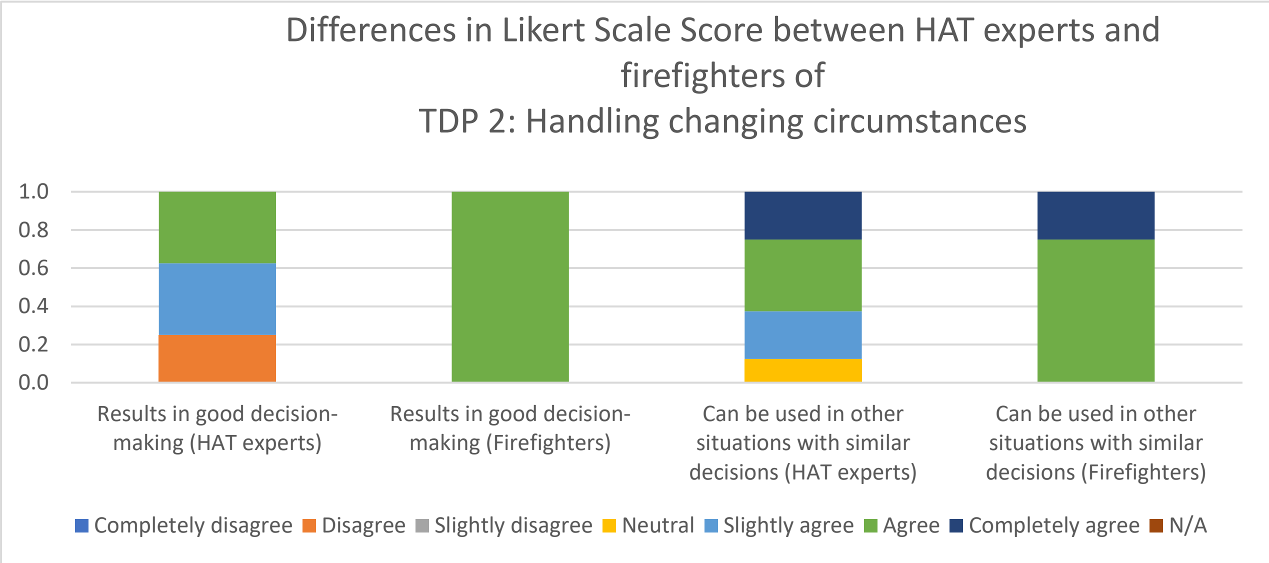



*Figure 26: Differences in Likert Scale Score between firefighters and HAT experts on whether TDP 2: handling changing circumstances results in good decision-making and can be used in situations in which similar decisions need to be made.*

*Table 26: Summary statistics for the data presented in Figure 26.*

| | TDP 2: Handling changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | The implementation of the design pattern will result in good decision-making | 2.00 | 6.00 | 4.63 | 1.58 | 2.48 |
| Firefighters | The implementation of the design pattern will result in good decision-making | 6.00 | 6.00 | 6.00 | 0.00 | 0.00 |
| HAT experts | The implementation of the design pattern can be used in other situations with similar decisions | 4.00 | 7.00 | 5.75 | 0.97 | 0.94 |
| Firefighters | The implementation of the design pattern can be used in other situations with similar decisions | 6.00 | 7.00 | 6.25 | 0.43 | 0.19 |

Some of the HAT experts did have questions about the consequence of stopping the task to be able to see the notification, such as the following participant:

*"In order to see the notification and advice from the intelligent system, the user must abort his physical task without knowing how important the notification is. This requires a considerable task switching from the user. Would it be possible, for example, to give the user an indication of the notification and the seriousness via voice, or to have a large part of the interaction take place via audio / voice so that he only has to switch to the screen for more visual information that can help make the decision. This may also be included in the working agreements, including the option to choose other settings based on the severity of the report."*

The ratings towards the statement that the TDP about independently reacting to changing circumstances results in good decision-making vary from disagree to agree, e.g., one participant disagrees, two slightly agree, two are neutral, two slightly agree, two agree, and three completely agree. See Figure 27 and the corresponding summary statistics in Table 27 below. One of the firefighters pointed out as a reason to agree with this statement that the decision-making is based on agreements. One of the HAT experts described that they imagine that this depends on the task.

The ratings towards the statement that the TDP about independently reacting to changing circumstances can be applied to other situations involving similar choices vary from slightly disagree to completely agree, e.g., one participant slightly disagrees, two are neutral about it, four slightly agree, two agree, and three completely agree. See Figure 27 and the corresponding summary statistics in Table 27.

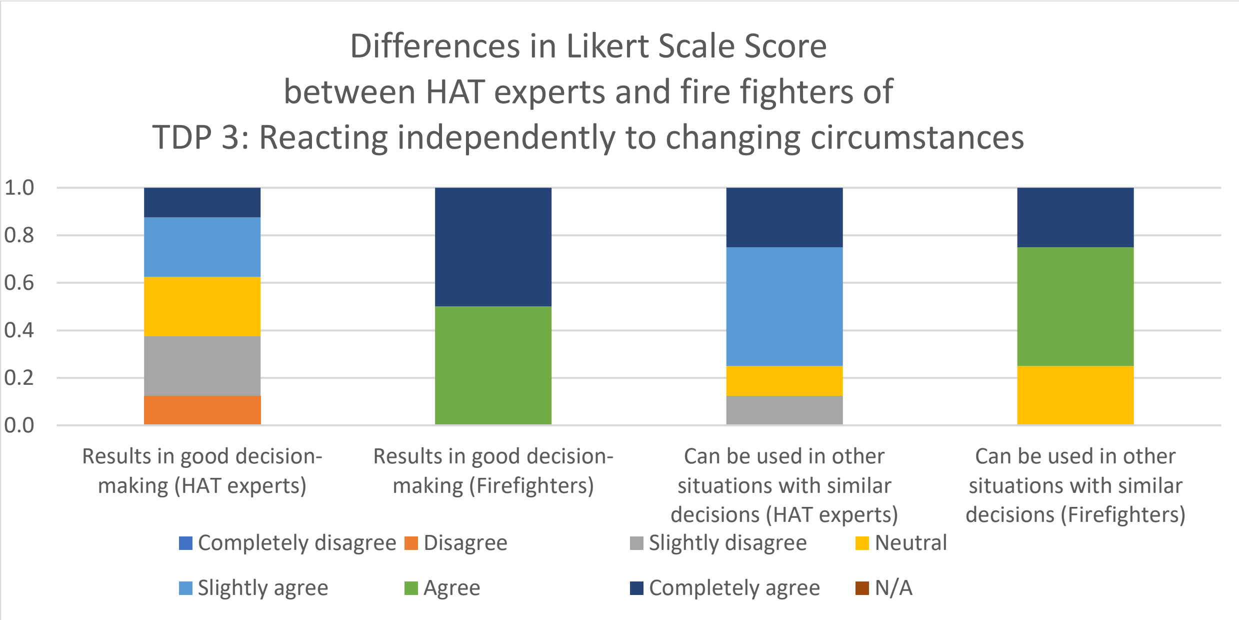



*Figure 27: Differences in Likert Scale Score between firefighters and HAT experts on whether TDP 3: reacting independently to changing circumstances results in good decision-making and can be used in situations in which similar decisions need to be made.*

*Table 27: Summary statistics for the data presented in Figure 27.*

| | TDP 3: Reacting independently to changing circumstances | Min | Max | Mean | Std | Variance |
|---|---|---|---|---|---|---|
| HAT experts | The implementation of the design pattern will result in good decision-making | 2.00 | 7.00 | 4.13 | 1.45 | 2.11 |
| Firefighters | The implementation of the design pattern will result in good decision-making | 6.00 | 7.00 | 6.50 | 0.50 | 0.25 |
| HAT experts | The implementation of the design pattern can be used in other situations with similar decisions | 3.00 | 7.00 | 5.13 | 1.27 | 1.61 |
| Firefighters | The implementation of the design pattern can be used in other situations with similar decisions | 4.00 | 7.00 | 5.75 | 1.09 | 1.19 |

## 4.4. Conclusions

This section summarizes the results of the evaluation of the design solution. First, a summary is given of the subjective evaluation of the participants on the use case. After that, the subjective evaluation of the Was and suggestions for other WAs are provided. Finally, a closer look is given to the subjective evaluations of the TDPs.

- **Conclusions about the use case:** The use case, which represents how an AI can support firefighters during the mission in case of an adversarial change in the environment with a notification alert based on WAs managed before the mission, has been identified as representative for a recurring decision-making problem with a value-sensitive aspect in the ASSISTANCE scenario. It is, however, emphasized that it is important to take the responsibility hierarchy into considerations as the consequences of not being able to resolve a leakage is not considered by the infield FF who is working on the leakage, but the team leader (or the commander if the team leader does not know what to do with the situation). The firefighters indicated that the higher the position in the hierarchy a firefighter has, the more communication the firefighter will have with the AI system. This means that an infield FF would like to receive a warning notification such that he knows he needs to leave the location, but he will not make sense of the warning. It is the team leader who will make sense of the warning and decide what to do next. Furthermore, they indicated that it is important to monitor the environment and the health of the firefighters. It is suggested that the AI system should monitor the health of the firefighters in addition to monitoring the environment. As the health of a firefighter influence the decision to continue working on the task as well.

- **Conclusions about the work agreements:** All three working agreements that were provided to the participants were identified as good working agreements. Some comments for improvement were made.

  Suggestions were given for auditive notifications and voice-interaction options. The FR does not have to interrupt his task to communicate with the AI agent when these options are available for them (which the FR has to do when the communication with the AI agent is text-based).

  Furthermore, it was emphasized that the system cannot send a robot to the location without a human being aware of it and that a human should oversee the robot. It was indicated that, if it has been agreed upon in advance, the system can decide to send a robot to the location to take over the task of the human. However, someone must be able to watch what the system is doing, and thus the system cannot decide without a person having insight into the decision.

  The results of the formative evaluation illustrate furthermore that most of the participants (nine out of twelve) like the WAs as they are (e.g., an advice-based notification is sent to the infield FF and the team leader, and a robot with the right capabilities is sent to the location in the case that a human cannot finish the task because of danger). Only three participants choose to change the WAs. The two HAT experts and the firefighter who decided to change the WAs disabled the task transfer option. These participants indicated that the system cannot make this decision on its own as long as all the conditions under which a robot can operate are not clear. Furthermore, one of the HAT experts changed the notification type from advice to information. This participant did, however, prefer to receive both information and advice from the AI agent.

  The following four suggestions were given to add to the WAs:

  - the ability to choose to not only take environmental factors into considerations for the alert notification but also the health status of the firefighter;
  - options for auditive notifications and voice-interaction instead of visual communication;
  - an option to be able to watch along with the firefighter and the robot via body cams;
  - an option to capture important information (keywords) from mutual conversations between infield FF during missions.

  Furthermore, it was suggested to have more nuance in the options than yes and no.

- **Conclusions about the Team Design Patterns:** Overall, it is to some extent considered that the TDPs describe the human-AI collaboration in a clear; concise; complete and easy-to-understand manner, however the answers of the firefighters are more in line with one another than the answers of the HAT experts. Some of the HAT experts did have questions about the task transition in the work agreement TDP from step 3 to 4, the order of tasks in the interrupting to changing circumstances TDP as some tasks are parallel (task 1 and 2) while others are sequential, and the ratio between task 4 and 5 in the reacting independently to changing circumstances.

  Furthermore, the participants slightly agree that the AI's described behaviours are expected to improve team performance with a value-sensitive aspect substantially. The participants agree that the implementation of collaborative management of the working agreements results in good decision-making. And that this team design pattern can be applied in other situations involving similar choices. Furthermore, the participants slightly agree that the implementation dealing with changing circumstances results in good decision making and agree that this team design pattern can be applied in other situations involving similar choices. Finally, the participants slightly agree that the implementation independently reacting to changing circumstances results in good decision making and slightly agree that this team design pattern can be applied in other situations involving similar choices.

# 5. Knowledge representation and reasoning

Chapter 4 described how the design solution is evaluated by HAT experts from TNO and DM experts (firefighters from the Gezamenlijke Brandweer and Spaansland fire brigade in Enschede). This chapter describes the knowledge base that provides information about the domain and how to manage WAs. The knowledge base provides the AI agent with the ability to inference knowledge from itself; the environment; and other agents (i.e., the HIS it operates in), such that it can alert FRs in case of adversarial changes in the environment given predefined WAs with the FRs. I.e., the knowledge base provides the AI with information that he can use to support FRs with decisions with a value-sensitive aspect. Section 5.1 describes the methodology used to make the knowledge base (e.g., the ontologies and the formalization). Section 5.2 provides the graphical descriptions of the ontologies (domain ontology and work agreement ontology) and the formalization of the WAs. Section 5.3 provides a conclusion about the knowledge representation and reasoning.

## 5.1 Methodology

### *5.1.1 Ontologies*

This section describes the process that led to the knowledge base of HAT for decisions with a value-sensitive aspect. The knowledge base contains two independent models that capture relevant knowledge of human-AI teaming for decisions with a value-sensitive aspect. The first model is a domain ontology called ASSISTANCE, and the second model is the work agreement ontology and its corresponding formalization. The formalization methodology is described in Section 5.1.2. The ontologies are made in Protégé, and the graphical notations of the ontologies are created using the OntoGraph plugin within Protégé. Protégé uses the Web Ontology Language (OWL), which allows for easy sharing such that the models can be re-used in other HAT research.

### *5.1.2 Formalization*

Singh [12] addresses social commitments for multiagent systems with an emphasis on the interplay of commitments and social structures. The formalism of Singh [12] is written in a formal language that is loosely based on predicate logic and states constraints on how the various predicates and operators relate to one another. Mioch et al. [32] use the formalization of Singh [12] as the base for their formalism which addresses the key concepts of WAs and how WAs are established between a creditor and a debtor. Within computational Logic and Set theory relationship elements ($\in$, $=$), logical operators ($\wedge$, $\vee$, $\neg$, $\rightarrow$, $\leftrightarrow$), set operators ($\cap$, $\cup$, $\subseteq$, $\emptyset$) and quantifiers ($\forall$, $\exists$) are used to provide structure and order to outline the atomic formulas that express the object in the real world, in this case, WAs for more than two actors. Table 28 provides an overview of the relationship elements, logical operators, and quantifiers of the Logic and Set Theory and their meaning.

Following the abstract representation of commitments as proposed by Singh [12] for commitments within a context group and the formalism of Mioch et al. [32] for WAs, this study defines a formalization of WAs between two or more actors. The formalism can be seen as an extension of the formalism of Mioch et al. [32] as it adds the possibility of WA for more than two actors based on the formalism of Singh [12].

*Table 28: Logic and Set Theory Legend*

| | | |
|---|---|---|
| **Relationship elements** | ∈ | Is an element of |
| | = | Is equivalent to |
| | := | Is defined by |
| **Logical operators** | ∧ | Conjunction |
| | ∨ | Disjunction |
| | ¬ | Negation |
| | → | Conditional |
| | ↔ | Biconditional |
| **Set operators** | ∩ | Intersection |
| | ∪ | Union |
| | ⊆ | Subset |
| | ∅ | Empty set |
| **Quantifiers** | ∀ | Universal |
| | ∃ | Existential |

## 5.2 Results

### 5.2.1 Domain ontology

The ASSISTANCE domain ontology describes the characteristics of DM, the decision-making process (e.g., communication, task allocation, and goals), the direct stakeholders (end-users), and the environment. The ASSISTANCE domain ontology consists of three branches: ActorDomain, CommunicationDomain, and MissionDomain. Each of these three branches is depicted in Figures 28, 29, and 30. For an enlarged view of each of these branches of the domain ontology, see Appendices K, L, and M.

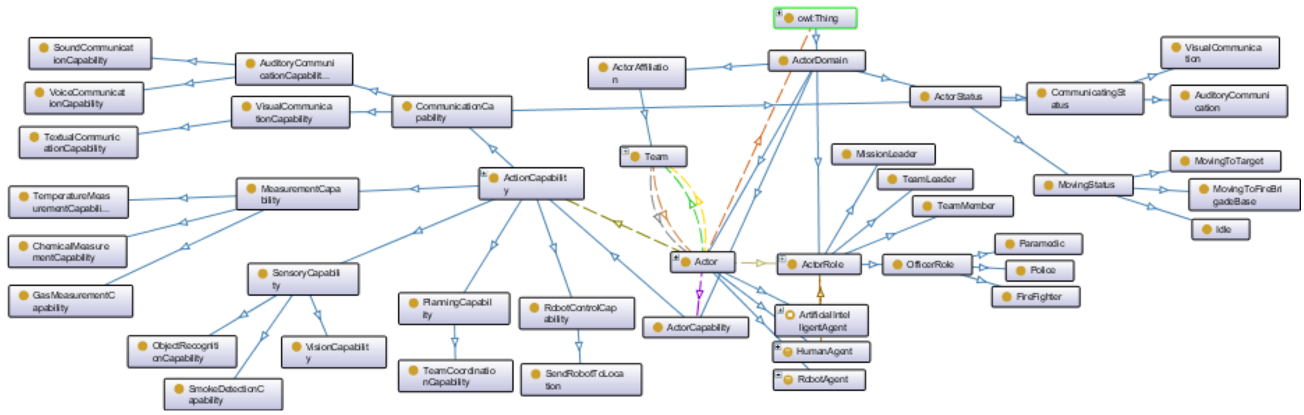



*Figure 28: The ActorDomain branch of the assistance domain ontology*

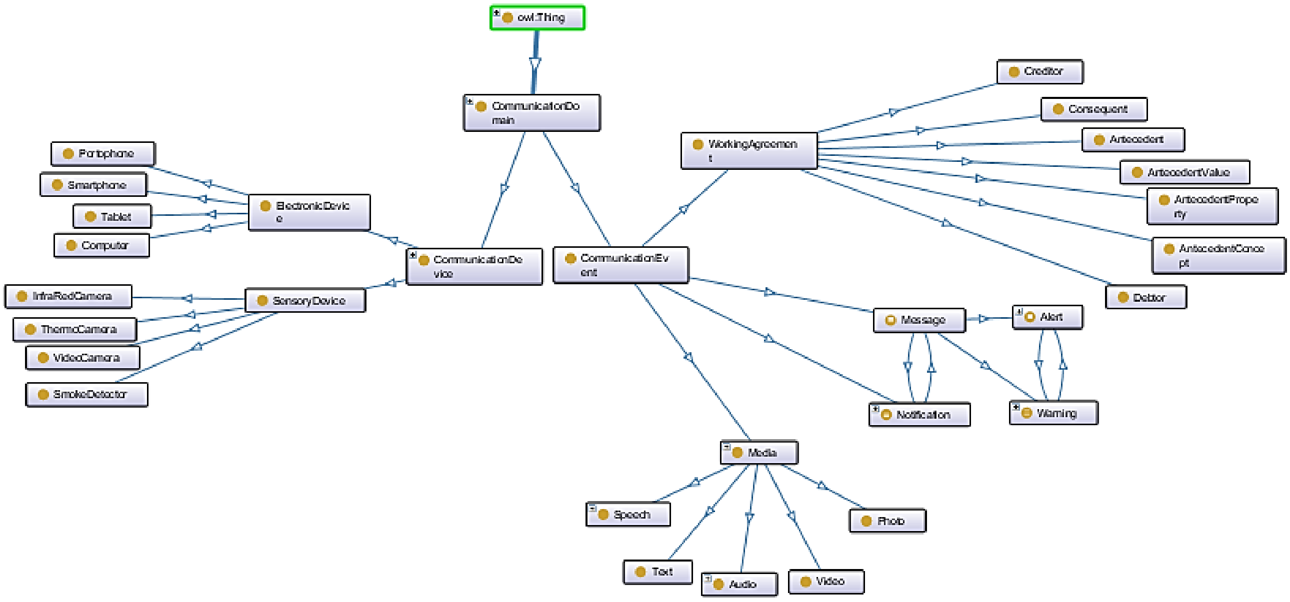



*Figure 29: The CommunicationDomain branch of the assistance domain ontology*

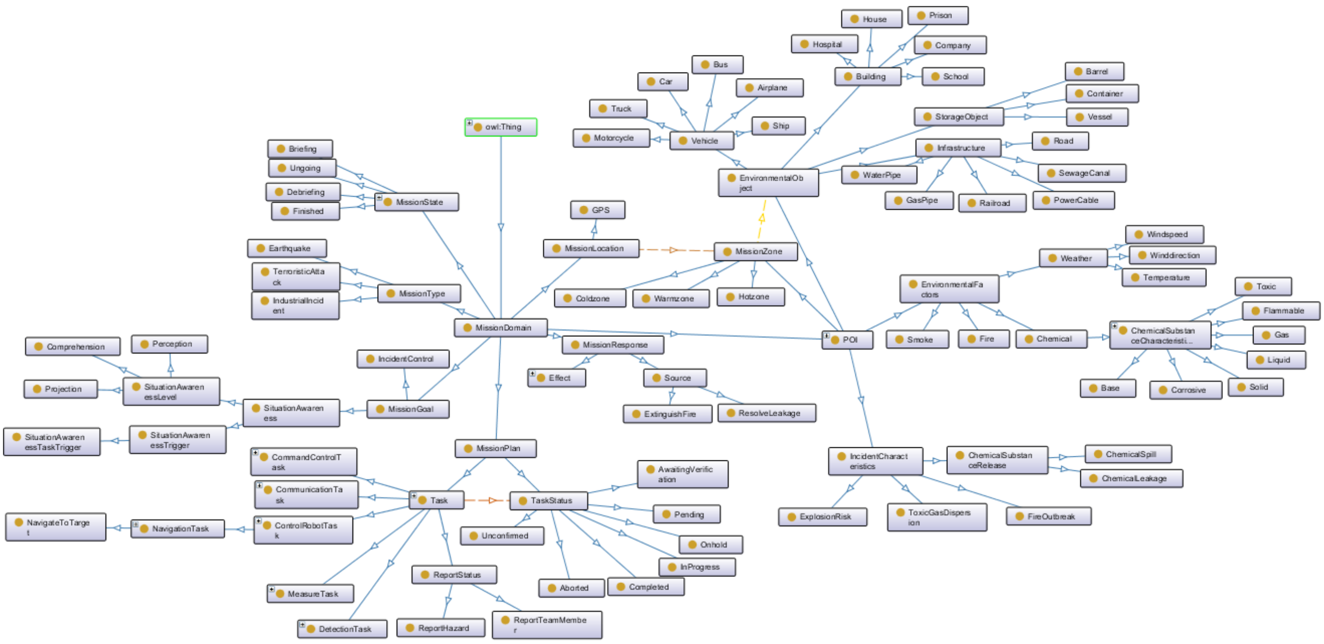



*Figure 30: The MissionDomain branch of the assistance domain ontology*

### 5.2.2 Work agreement ontology and formalization

How WAs are established is described in the work agreement ontology, which is connected to the domain ontology via the CommunicationDomain branch. The graphical description of the work agreement ontology is depicted below in Figure 31.

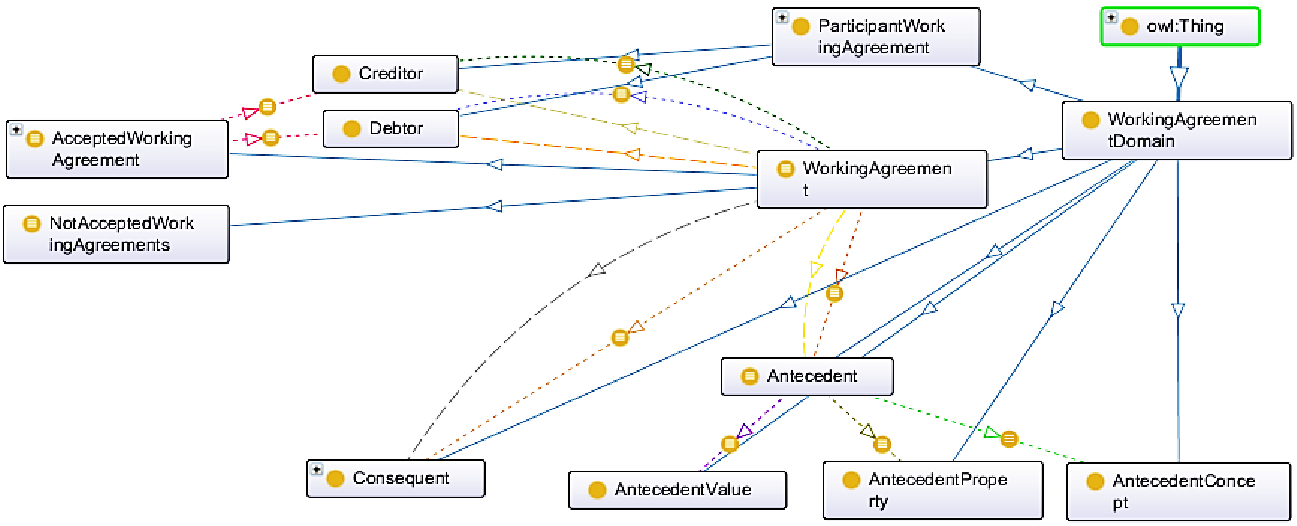



*Figure 31: The work agreement ontology*

The work agreement model specifies the underlying concepts and relations of WAs between two or more actors. To specify these, the paper of Singh [12] and Mioch et al. [32] were used. This study focuses on data sharing (text-based notifications) and enabling/disabling task transfer specifically.

In principle, a WA is a commitment between a debtor towards a creditor for a given condition; this is defined by Mioch et al. [32] using the concepts: *creditor, debtor, Antecedent, consequent, L, E*. A debtor is an actor that is committed to the execution of the WA; A creditor is an actor to which the commitment to the execution of the WA is owed; the antecedent is a condition that needs to be valid to activate the WA; a consequent is a task or goal state, i.e., the norm (obligation, permission, prohibition); L stands for lifespan and indicates the timespan for which the WA is applicable; and E is the status that indicates for each of the involved actors if the WA is accepted or declined. The use of each of these concepts is explained underneath definition 1, which defines the WA as a commitment between a debtor towards a creditor for a given condition, such as is defined by Mioch et al. [32].

We can also take the context group and the relations within this context group into consideration. WAs are made in a social context and therefore intuitively include the norms that apply in the group in which the WAs are established. This group can be referred to as the context group and includes not only the agents that commit to WAs but also agents that do not commit to WAs but nevertheless are affected by these WAs. When the context group is taken into consideration attention will not only be given to the effect of WAs towards actors that commit on a WA, but also to the actors that do not commit but are nevertheless within the context. The formalism which includes the context group, is described in definition 2 and is based on the formalism of Singh [12].

If we consider the context group, we can clarify whether the actor is committed with a WA with specific actors, which only applies to these specific actors within the context group (this is described in definition 2d) or a WA with a specific actor which is applicable to all actors of the context group that are affected by the WA (this is described in definition 2e). We will now illustrate what this difference between these two commitments means: Let us say the context group consists of six agents in total, namely a team leader (l), four infield FF (f1, f2, f3, f4), and an intelligent system (s). For this context group, two WAs are established, namely general team WA and sub-team WA. The sub-team WA is a commitment applicable towards specific actors of the context group, whereas general team WA is a commitment that is applicable towards the whole context group. Let us say that a sub-team WA is established that indicates that an Intelligent system (s) commits towards infield FF (f1) and the team leader (l) to notify them in case infield FF (f1) is in danger. In this case, the intelligent system (s) is the debtor, whereas the infield FF (f1) and the team leader (l) are the creditors. For this WA, the Intelligent system (s) has a commitment that is applicable towards infield fighter (f1) and the team leader (l), but not towards any of the other members in the context group (this is described with definition 2d). Both the infield fighter (f1) and the team leader (l) know that they are a creditor and if one of them does not accept the commitment the WA needs to be revised.

For the use case in this research, three WAs are defined, which allows the AI agent to send an alert notification and to setup a task transfer from human to robot. These WAs indicate who the AI agent should notify in case of an adversarial change in weather conditions, how the notification should look like, and whether task transfer should be enabled or not. These WAs can be seen as sub-team WAs as the commitments made with these WAs are between the AI agent and specific team members, namely the infield FF the AI agent is supporting and if agreed upon with the work agreements the team leader.

However, it could be that there also is a general team WA established apart from a sub-team WA. In this type of Working Agreement, the Intelligent system (s) commits towards the team leader (l) to notify all actors in the context group that are in danger. This means that the Intelligent system (s) commits towards the team leader (l); however, each member that is part of the context group will be affected by this commitment. The intelligent system (s) is the debtor, and the team leader (l) is the creditor; however, the commitment the Intelligent System (s) makes applies to all actors in the context group (G) (this is described with definition 2e).

DEFINITIONS:

1. A work agreement from a set of actors towards another set of actors without taking the context group in consideration is denoted as: *WA(creditor := {f1, l}, debtor := {s}, Antecedent := p* $\in$ *P, consequent, L,E),* and can be described *as* a commitment between debtor(s) and creditor(s) given an antecedent, consequent, Lifespan (L), and status (E).

   In this relation the following is defined:

   a) Debtor is a non-empty set of actors that is committed to the execution of the work agreement. This list of actors can be written as follow: *debtor := {a1; a2; : : : ; an}* $\in$ *A.* Where A is a list of actors that are debtor and/or creditor. Within the example it is defined as a list of one actor, namely intelligent system (s); and thus as debtor := {s}.

   b) Creditor is a non-empty set of actors to which the commitment to the execution of the work agreement is owed. This list of actors can be written as follows: *creditor := {b1; b2; : : : ; bn}* $\in$ *A.* Where A is a list of actors that are debtor and/or creditor. Within the example it is defined as a list of two actors, namely infield FF (f1) and team leader (l); and thus as creditor := {f1, l}.

   c) Antecedent is a condition that needs to be valid to activate a work agreement. This is a predicate (p) and can be written as follow: *Antecedent := p* $\in$ *P*

   d) Consequent is a task or goal state, i.e., the norm (obligation, permission, prohibition) and thus a proposition committed to by the debtor to execute once the antecedent is valid. The consequent can be written as follow: *consequent := NORM (x* $\in$ *{K* $\cup$ *P}).* Within this relation, the NORM can be obligated, permitted, or forbidden.

*NORM(k $\in$ K)* implies that the actor is either obliged, permitted, or forbidden to perform task (k), whereas *NORM (p$\in$ P)* implies that the actor is either obliged, permitted, or forbidden to see to it that predicate (p) is true.

e) Lifespan indicates the timespan for which the work agreement is applicable, if none is described the work agreement is generally applicable. The Lifespan can be written as follow: *Lifespan:= BeAf (t$\in$ T)* or *BeAf (p$\in$ P).* Within this relation, BeAF indicates whether the work agreement is activated before or after a specific time (T) or a specific set of predicates (P).

f) E indicates the status of acceptance or decline of the work agreement for each of the involved actors. E is a Boolean and can be written as follow: *E := $\forall$x $\in$ {debtor; creditor}, ex := true V false*

2. A work agreement from a set of actors towards another set of actors for which the context group is taken into consideration is denoted as: WA*(debtor := {s}, creditor := {f1, l}, G := {l, f1, f2, f3, f4, s}, p)* and can be described as a work agreement for a given proposition (p) for which the debtor(s) commits to towards the creditor(s) within the context group (G). With this type of work agreements, the context group is taken into consideration as it can be defined whether the debtor commits towards specific actors within the context group after a specific request of these same actors, such as described in definition 2b or whether the commitment applies towards everyone in the context group (G) such as described in definition 2c.

   In this relation the following is defined:

   a) The context group (G) is a non-empty set of actors that are defined within the context. This list can be written as follow: G:= {g1, g2, …, gn}. The set includes both the actors that commit to a work agreement and the ones that do not. This means that an actor can be a debtor and a creditor or none of the both (e.g., infield FF (f1):= Creditor $\cap$ Debtor $\cup$ Other). Within the example, it is defined as a list of six actors, namely a team leader (l), four infield FF (f1, f2, f3, f4), and an intelligent system (s); and thus as G:= {l, f1, f2, f3, f4, s}.

   b) Debtor is a non-empty set of actors committed to the execution of the work agreement. This list of actors can be written as follow: *debtor := {a1; a2; : : : ; an} $\in$ A, where A $\in$ G.* Within the example it is defined as a list of one actor, namely intelligent system (s); and thus as debtor := {s}.

   c) Creditor is a non-empty set of actors to which the commitment to the execution of the work agreement is owed. This list of actors can be written as follow: *creditor := {b1; b2; : : : ; bn} $\in$ A, where A $\in$ G.* Within the example it is defined as a list of two actors, namely infield FF (f1) and team leader (l); and thus as creditor := {f1, l}.

   d) A work agreement: WA *(debtor, creditor, G, p*)) describes a commitment to proposition (p) of at least one debtor (intelligent system (s)) towards at least one creditor (infield FF (f1) and team leader (l)) in a context group (G). The commitment is specific for several actors of the context group if the commitment is created by the debtor(s): *create (debtor, C (debtor, creditor, G, p))*, on request by the creditor(s): *request (creditor,* *debtor, p)* after request of the creditor(s): *request (creditor, debtor, p).* This commitment of s to specific actors of the context group (f1 and l) can be written as: $\forall$*creditor, G , p : creditor* $\in$ *G* → *C(debtor, creditor, G, q)), where q* $\triangleq$ *(request(creditor, debtor, p)* → *create(debtor, C(debtor, creditor, G, p))*

e) A work agreement: WA*(debtor, creditor, G, p)*) describes a commitment to proposition (p) of at least one debtor (intelligent system (s)) towards at least one creditor (infield FF (f1) and team leader (l)) in a context group (G). The commitment can be applicable towards the whole context group, which is written down as: *C (debtor, G, G, (*$\forall$*creditor , p: creditor* $\in$ *G* → *C(debtor, creditor, G, q)))*

## 5.3 Conclusions

As was described in Section 2.8, to be able to support humans in specific tasks, such as decision-making; risk assessment; and planning, an AI agent needs information from sensors of static, moving or wearable objects in the environment. This sensor data provides the agent with a certain level of understanding of the agent itself, its environment, and other agents, and thus of the HIS it operates in.

Within this research, the knowledge that is provided to the AI agent is knowledge about the domain (i.e., DM) and about managing and working with WAs (i.e., as represented in section 5.2.2) that together form the knowledge base. The knowledge base provides the AI agent with the necessary information to derive knowledge about the environmental factors that may influence a FR's decision-making about his/her safety so that they can alert the FR in accordance with the WAs.

The TRADR ontology (see [32]), a knowledge base for search and rescue missions during DM, functioned as the source of inspiration for the knowledge base in this research. However, within this research, there is investigated how WAs can be managed between more than two actors. To formalize this the research of Singh (1999) [12] and Mioch et al. (2018) [32] is used. This allows us to formalize WAs from a set of actors towards another set of actors with and without considering the context group. This research evaluated how this would work specifically for WAs related to data sharing (text-based notifications) and enabling/disabling a task transfer from human to robot. Future research could take other WAs into consideration, such as the one suggested by the participants of this research (see Section 4.3.2.3).

# 6. Simulation-based evaluation in MATRX

Chapter 5 described the knowledge base that provides the AI agent with the ability to inference knowledge from itself; the environment and other agents (i.e., the HIS it operates in), such that it can alert FRs in case of adversarial changes in the environment given predefined WAs with the FRs. This chapter describes how the HAT is simulated in MATRX to evaluate the support for the FR simulation-based. This simulation simulates the communication and task allocation of the agents, i.e., it is demonstrated how the WAs are collaboratively managed by the FR, team leader, and AI agent before managing the disaster, and how the agents collaborate with one another based on these WAs during the DM. The WAs provide the agents with the necessary information about what kind of notification is sent by the AI agent and to who, and whether the AI agent can send a robot to the location to take over the task of the human when the human cannot finish the task due to an adversarial change in weather conditions. Section 6.1 describes the methodology used to make the simulation (i.e., MATRX), Section 6.2 provides graphical descriptions of the simulation and a video-based demonstration, and Section 6.3 provides a conclusion about the simulated-based evaluation.

## 6.1 Methodology

The simulation was created using the Python Rapid Experimentation software package (MATRX)[12]. With this package, human-agent and/or multi-agent team tasks can be designed for experimentation or simulation purposes. MATRX allows a user to define a world containing objects, agents, actions, and simulated goals and create a scenario to simulate desired events within this world. Within this scenario, agents then mutate the world through performing actions, i.e., when an agent decides on an action, it will communicate this in the world. The agents' reaction to the world is whether the action succeeded or not and a potential reason for this. Agents can be controlled by human users or can be programmed to act autonomously. Agents communicate with one another by sending and receiving messages about their actions within the world or actions they wish other agents to do in the world. MATRX is open source and can be found on GitHub[13].

Within this project, MATRX was used to simulate the communication and task allocation of the FR, AI agent, and team leader for managing WAs and dealing with task interruption based on these WAs. As was already described in the introduction of this chapter, what is demonstrated in MATRX is how the FR, team leader, and AI agent collaboratively manages the WAs before the disaster is managed. And how the agents collaborate based on these WAs during the DM. The WAs provide the agents with the necessary information about what kind of notification is sent by the AI agent, to who, and whether the AI agent can send a robot to the location to take over the task of the human when the human cannot finish the task due to an adversarial change in weather conditions.

The world, visualized on the left in MATRX represents the disaster site in which the three agents operate. On the right in MATRX, the communication between the agents is visualized. See figure 32 on the next page for a representation of the MATRX environment.

[12] https://matrx-software.com/

[13] https://github.com/matrx-software/matrx

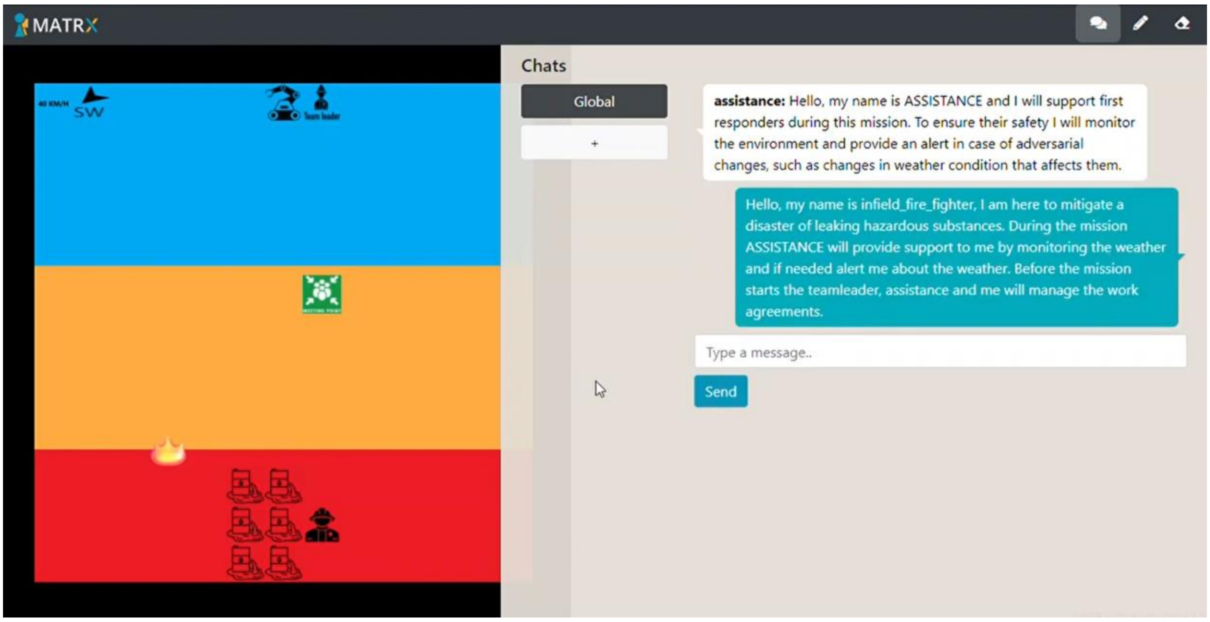



*Figure 32: Visualization of the MATRX environment with on the left the world and on the right the chat.*

The disaster site is represented with the three zones (cold zone, warm zone, and hot zone). FRs use these three zones to divide the disaster area into a safe area to coordinate the incident response from, possible unsafe area, and the affected area of the incident. The blue-colored area represents the cold zone, which is the installation location and a safe area to be in without special equipment. The orange-colored area represents the warm zone which is the possibly unsafe area between the final line-up and the preliminary line-up. The final line-up indicates the transition from cold zone to warm zone (i.e., transition from safe zone to possible unsafe zone), and preliminary line-up indicates the transition from warm zone to hot zone (i.e., transition from possible unsafe zone to affected area). The warm zone is polluted by the deployment of the fire brigade, for example, because a victim who has been in contact with hazardous substances is brought to this zone. All contaminated materials and resources remain in the warm zone. The red-colored area represents the hot zone, which is the source area that is contaminated with hazardous substances as a result of the incident.

The disaster site contains a burning fire and 6 leaking barrels within the hot zone and a meeting point in the warm zone to which the FR who is working on resolving the leaking barrels is expected to go to in case of danger.

In the world, agents interact with the environment; i.e., the team leader oversees and communicates with his team from the cold zone, the FR is resolving the leaking barrels in the hot zone, the AI agent monitors the weather to indicate whether the wind direction and wind speed have changed (the wind direction is represented with the arrow and letter and the wind speed is provided in km/h at the top left corner of the world), and a UGV stands ready next to the team leader to take over the task of the FR once the AI agent gives him that order.

The communication between the team leader; FR; and AI agent is provided in the chat menu at the right. Within this chat menu, it is visualized how the WAs between the team leader, FR, and AI agent are managed at the start of the mission; and how based on these WAs, the AI agent alerts the FR and the team leader about the changing weather conditions and sends a robot to the location of the FR to take over the task of the FR. The visualizations of this simulation (including a video demonstration) are represented in the next section (Section 6.2).

## 6.2 Results

This section provides the visualisations of the above-described simulation (i.e., the communication and task allocation of the FR, AI agent, and team leader for managing WAs and dealing with adversarial changes in the environment).

Figure 33 below illustrates the communication between the FR and AI agent while managing the WAs. Furthermore, the world is illustrated before the adversarial change of weather conditions, e.g., the wind speed is moderate (40 km/h), and the wind direction is southwest, which is not towards the barrels.

Figure 34 on the next page illustrates the communication between the AI agent and the FR at the moment the AI agent registered the adversarial change of the weather condition, i.e., the wind speed changed from moderate (40 km/h) to strong (60 km/h) and the wind direction has now turned more towards the barrels (e.g., facing south). The AI agent alerted the FR of this change and communicated that a robot is sent to the location. Furthermore, it can be seen in the world that the FR has resolved three barrels; however, that three barrels are still leaking. Furthermore, it can be seen that the task transfer occurred, meaning that the FR has left the location, and the robot is now at the location to resolve the remaining leaking barrels.

The simulation can be viewed in the following video: [https://youtu.be/_Iyc7e9Y0dw](https://youtu.be/_Iyc7e9Y0dw). It is recommended to turn on the audio because the video contains an auditory explanation of what is happening during the simulation.

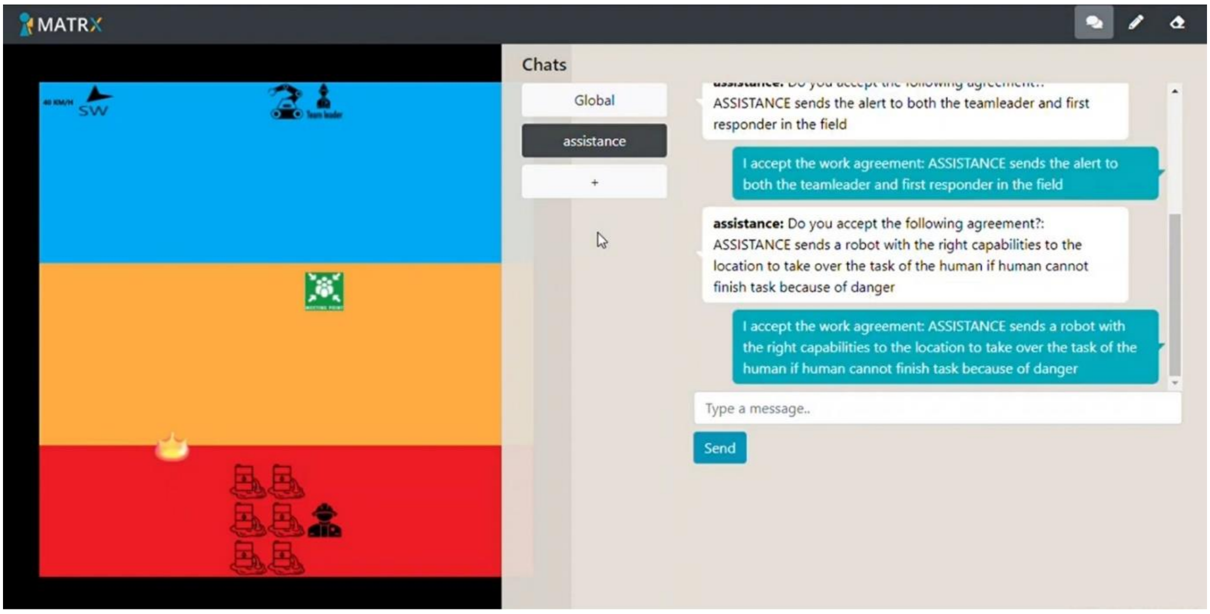



*Figure 33: Visualization of managing work agreements and the situation of the disaster management before the adversarial change in weather conditions.*

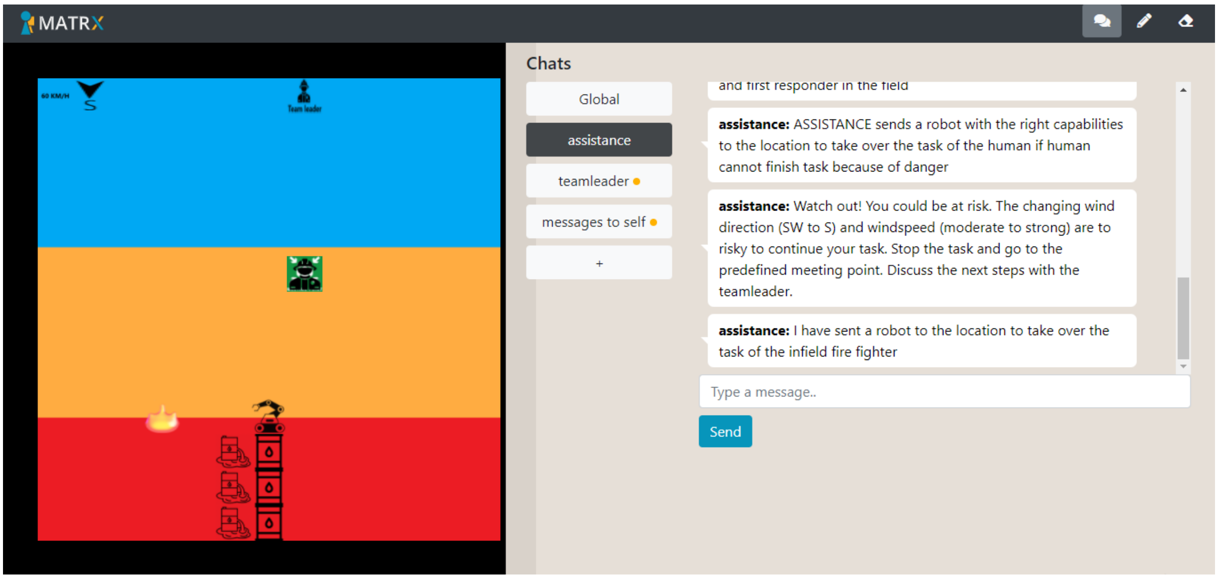


*Figure 34: Visualization of managing work agreements and the situation of the disaster management before the adversarial change in weather conditions.*

## 6.3 Conclusions

In the previous section, it was shown what the MATRX simulation looks like. This simulation visualises the communication and task allocation of the FR; AI agent; and team leader for managing WAs and dealing with task interruption based on these WAs. With on the left side the world in which the agents operate in (i.e., the disaster site), and on the right side the chat in which the agents communicate with one another about the WAs for the DM (before the DM starts) and about the adversarial changes in weather condition (during the DM).

Within this simulation it has been demonstrated how the AI agent can communicate with the FR and team leader about managing a set of working agreements at the start of the DM that will be used during the DM. These working agreements provide the AI agent the ability to understand during the DM what kind of notification to send and to who, and whether it can send a robot to the location to take over the task of the human in case the human cannot finish the task because of danger or not. Furthermore, it has been demonstrated how these WAs are then used by the AI agent to communicate with the FR and team leader about adversarial changes and to decide about dealing with the adversarial changes.

The simulation can be used to evaluate the WAs and task allocation with FRs. Currently the responses of the FR (and team leader) are programmed; however, future research could integrate a functionality to check whether a human user chooses to accept an agreement or not. This can, for example, be done by adding an accept and decline button underneath the chat. Furthermore, future research could add more WAs to the simulation. Apart from that, the work agreement ontology described in Chapter 5 could be integrated in the agents' knowledge base. Currently, the WAs are integrated following a rule-based methodology without considering learning abilities, which means that both the knowledge and the agents' behavior is fixed. To test if the model brings forward the predicted actions in different scenarios and to test that it is accurate for communicating, instantiating, and de-activating the WAs, the knowledge of the work agreement ontology should be implemented.

# 7. General conclusions and discussion

This chapter provides an answer to the research question: "How can an Artificial Intelligent agent support a FR with solving a recurrent decision-making problem with a value-sensitive aspect?”. Furthermore, it is discussed what the results mean for future research. Section 7.1 describes the main outcome and further insights. Section 7.2 provides insight into the significance of this research. Section 7.3 describes the limitations of this research. And Section 7.4 describes the recommendations for future research.

## 7.1 Main outcome and further insights

Within this research, it has been investigated how an AI agent can support a FR with solving a recurrent decision-making problem with a value-sensitive aspect through …

1) the design and evaluation of TDPs that illustrate the collaborative management of WAs in HIS for DM and the use of these WAs during the DM to make a decision about the safety of the FR;
2) the development of a formal work agreement model for multi-human, multi-robot teams that describe how adaptive WAs can be managed between more than two actors;
3) a simulation of the task allocation and information exchange as described in the TDPs using the Rapid Experimentation software python software package (MATRX), which can be extended or transformed into a user experiment.

The WAs describe the communicative acts that the AI agent must perform to be able to support a FR with making a decision with a value-sensitive aspect (e.g., dealing with adversarial changes in weather conditions that can be a risk for the safety of the FR). Within this research, HAT and DM experts evaluated three working agreements, namely WAs that describe

- who should receive an alert notification from the AI agent, e.g., only the FR who is affected by the changing weather condition or the team leader as well;
- how the alert notification should look like, e.g., information-based or advice-based;
- whether task transfer from human to robot should be enabled, and thus if the AI agent can send a robot to the location when the FR has to leave the location due to the adversarial changes in weather conditions.

Both HAT experts and DM experts have indicated these WAs to be a valuable method for establishing a bridge of knowledge and understanding of the behaviour of the AI agent in the human-AI collaboration for decisions with a value-sensitive aspect during DM. It has been indicated that these WAs contribute to establishing the desired level of flexibility in the human-AI collaboration for decisions with a value-sensitive aspect. Moreover, they are a good starting point for tailoring the communication and task allocation between the AI agent and a FR during DM. However, the participants did have suggestions for improving the interactions. It was, for example, emphasized that a human agent should be aware of the AI agent initiating a task transfer and that furthermore, the human agent should have the ability to oversee the task transfer from human to robot. This means that decisions and actions of the robot should be under the supervision of a human agent that can provide input or take over the control whenever needed. It was also suggested to have auditive communication options instead of textual communication options, such that the FR does not have to interrupt his/her task to be able to receive a warning from the AI agent about adversarial changes in weather conditions.

The results show furthermore that it is important to consider the responsibility hierarchy of a first response team during DM. The higher the position in the hierarchy a FR has the more communication there is expected to be with the AI system. An infield FF likes to receive a warning notification to understand when a situation is not safe; however, they will not decide what to do next. This is the responsibility of the team leader. Therefore, the team leader should receive information as to why the notification has been sent to the infield firefighter to make a new strategy. Any notification shared with an infield FF should be to the point and should contain the bare minimum amount of information needed for an infield FF to be alerted about the changing circumstances. Furthermore, it has been indicated that it is not only important to monitor the environment, but also the health of the firefighters as both environmental factors and physiological measures of the health of the firefighter influence the decision to continue working on the task or to leave the location instead because of safety risks.

The main outcome of this research, therefore, is that an AI agent can support a FR in his/her decision-making process in situations with a value-sensitive aspect (security) by means of an alert notification about the situation and adaptation to the situation with the help of WAs that determine who receives the notification, what the notification looks like for the FR, and whether a robot (with the right capabilities and under supervision of a human agent) can be sent to the location to take over the task from the FR if the FR cannot complete the task himself due to the changing circumstance. Apart from subjective evaluations from HAT experts and Dutch firefighters, the support of the AI agent for decisions with a value-sensitive aspect with the help of WAs has been demonstrated with a formal work agreement model for multi-human, multi-robot teams that describe how these adaptive WAs can be managed between more than two actors, and a MATRX simulation of the task allocation and information exchange between the different agents in relation to these working agreements.

## 7.2 Significance

To manage a disaster means being able to mitigate uncertainties a disaster brings to society and the lives of individuals involved in it. Value-sensitive decision-making brings uncertainty as these are complex decisions for which there is not a single correct solution. Thinking about your own safety and the safety of others while mitigating a disaster is such a decision. Safety is not a given during DM and when a FR has to divide his attention between his/her task and monitoring the environment for potential threats risks exist for missing important information when attention is drawn to the task when attention should have been drawn to the environment (or the other way around). Supporting FRs during DM by improving their situation awareness with an AI agent that monitors the environment and notifies the FR in case of adversarial changes in weather conditions using WAs this risk could be minimized. WAs allow a FR to understand and know when and how to expect support as it tailors the support of the AI to the needs of the FR. Minimizing the risk for FRs during DM could improve their ability to mitigate the disaster as the attention that would typically be given to think about their safety could now be used to perform the task.

Furthermore, the design solution of user-tailored notifications through working agreements for two or more than two actors is generalizable to situations other than DM in which value-sensitive decisions have to be made. The formal work agreement model for multi-human, multi-robot teams to manage WAs for more than two actors is generally helpful for HAT research. It sheds light on the operational patterns of autonomous actors and thus forms a bridge for knowledge and understanding about their behavior. This information can then be used to determine whether this behavior promotes values we wish to be promoted.

## 7.3 Limitations

As was described in Section 7.1 within this research, attention has only be given to textual communication with the AI agent. This has been marked as a limitation by participants as the FRs can only communicate with the AI agent if they interrupt their task. It, therefore, is suggested to include options for auditive communication.

The second limitation is that the DM experts are all Dutch firefighters, while the goal is to deploy ASSISTANCE on a European scale. For this research project, the DM experts that are considered are firefighters from the Gezamenlijke Brandweer in Rotterdam and from the Spaansland fire brigade in Enschede (these are both fire brigades in the Netherlands). These Dutch firefighters have noted that the firefighters in Belgium and Germany (countries on the borders of the Netherlands) operate differently than the Dutch fire brigade does. It could be that the different ways of working of these other fire brigades require other WAs that are not addressed in this research. Furthermore, the solution of ASSISTANCE is directed towards all types of FRs that are part of the DM team. Therefore, it is important to involve firefighters and other first response teams such as police and paramedics. To improve the design solution future research could include other first response teams (police and paramedics) and firefighters from outside the Netherlands to investigate if different ways of working indeed require different WAs.

Apart from widening the perspective on the design solution by involving more DM experts with a different role or from a different country in Europe, it is suggested by the firefighters that participated in this research that the AI agent should not only provide alert notifications based on environmental factors such as adversarial changes in weather conditions but also based on adversarial changes in physiological factors (i.e., the health condition of the firefighter). The safety of the firefighter is not only influenced by environmental conditions but also by their physiological state. Therefore, the AI agent will not have the full picture of the situation if it only considers the environmental factors. Better support could be provided if adversarial changes in health conditions are considered as well. Future research could investigate how to include this aspect.

Furthermore, it should be investigated how the knowledge base can be integrated in the MATRX simulation. The MATRX simulation is currently not connected to the knowledge base (which includes information on the DM domain and managing WAs for two or more than two actors). However, instead, communication and task allocation in relation to managing and working with WAs in a hybrid intelligent team is implemented using a rule-based method (i.e., the knowledge is fixed, and the AI agent has no learning ability for behavior other than defined in the rules). To transform the simulation into a user experiment the reasoning behind managing and working with the WAs should be added. If this reasoning is added the interaction between the different agents would be depended upon what the human controller would choose, which could indicate (if a large enough group of participants would participate) what the overall attitude is towards the communication and task allocation with the AI agent. This information could then be used to integrate the work agreements model in the ASSISTANCE platform such that the functionality could be tested in the real world within the pilot tests. Currently, the design solutions are only evaluated using descriptive and graphical methods in which participants read about and see pictures of the scenario and have to imagine to be in a situation in which they collaborate with an AI agent to make a decision about their safety. It could be that the behavior and opinions of the participants are different when they would be able to interact with the AI agent and that this has an influence on their expectations about the AI agent. A pilot test in the real world could investigate this.

## 7.4 Future research

Future research could investigate how to transform from text-based notifications to auditive notifications. Furthermore, it could be tested whether auditive notifications are better than text-based notifications or that maybe a combination of the two would be the best.

Another thing that could be investigated in future research is how to include monitoring the health conditions of the FR in the design. During the workshop at the Spaansland fire brigade in Enschede it was suggested to combine the approach of this research with MoSeS. Future research could investigate whether the integration of ASSISTANCE and MoSeS is achievable. If this is achievable, the AI agent could base the notification not only on adversarial changes in weather conditions but also on adversarial changes in health conditions of the FR.

Furthermore, I would like to give some suggestions for scaling up. As was already explained in Section 7.2 within this research focus has only been given to the role of firefighters during DM, and the firefighters that have been included in the research are all Dutch firefighters. It might be a good idea to investigated how to involve other DM experts as well, e.g., FRs with a different role during DM (e.g., police or paramedics), or firefighters from fire brigades outside the Netherlands. It could, for example, be that first response teams with different roles and firefighters from outside the Netherlands operate differently during DM. These different ways of working might require other WAs that are not considered within this research. Including other DM experts, therefore, could improve the design solution.

As described in Section 7.2 these DM experts can, for example, participate in a user experiment in MATRX in which they experience the task allocation and information sharing for the existing WAs. After the experiments, an open question could be asked to the participants that ask them to indicate any other working agreements they would like to set up for the situation they experienced during the experiment. A user experiment in MATRX is a good starting point for evaluating the existing scenario or different scenarios with value-sensitive decisions with other DM experts as it can easily be done online. Suppose the user experiment in MATRX shows that managing the WAs with the existing WAs works as intended for all types of FRs. In that case, this functionality could then be integrated into the ASSISTANCE platform. Integrating the functionality of managing the WAs in the ASSISTANCE platform allows the design solution to be tested in real-world situations during the pilot tests.

# 8. References

[1] M. Neerincx, J. van Diggelen, & L. van Breda, "Interaction design patterns for adaptive human-agent-robot teamwork in high-risk domains," In International conference on engineering psychology and cognitive ergonomics Springer, Cham, (pp. 211-220), July 2016.

[2] M. Harbers, & M. Neerincx, "Value sensitive design of a virtual assistant for workload harmonization in teams," Cognition, Technology & Work, 19(2-3), pp. 329-343, 2017.

[3] M. Harbers, & M. Neerincx, "Value sensitive design of automated workload distribution support for traffic control teams," In International Conference on Engineering Psychology and Cognitive Ergonomic Springer, Cham, pp. 12-21, June, 2014.

[4] J. Van Diggelen, M. Neerincx, M. Peeters, & J. M. Schraagen, "Developing effective and resilient human-agent teamwork using team design patterns," IEEE intelligent systems, 34(2), pp. 15-24, 2018.

[5] J. van Diggelen, & M. Johnson, "Team Design Patterns". In Proceedings of the 7th International Conference on Human-Agent Interaction, pp. 118-126, September 2019.

[6] J. van der Waa, J. van Diggelen, L. C. Siebert, M. Neerincx, C. Jonker, "Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach,". In: Harris D., Li WC. (eds) Engineering Psychology and Cognitive Ergonomics. Cognition and Design. HCII 2020. Lecture Notes in Computer Science, vol 12187. Springer, Cham. 2020, https://doi.org/10.1007/978-3-030-49183-3_16

[7] J. van Stijn, M. Neerincx, A. ten Teije, & S. Vethman, "Team Design Patterns for Moral Decisions in Hybrid Intelligent Systems: A Case Study of Bias Mitigation", 2021.

[8] R. Gillon, "Defending the four principles approach as a good basis for good medical practice and therefore for good medical ethics". J Med Ethics. 2015;41:111–116, 2015, doi: 10.1136/medethics-2014-102282.

[9] A. Winfield, and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180085, 2018.

[10] T. Beauchamp, and J. Childress, "Principles of biomedical ethics". Oxford University Press, USA, 2001.

[11] J. van Diggelen, J. Barnhoorn, M. Peeters, W. van Staal, M. Stolk, B. van der Vecht, & J. M. Schraagen, "Pluggable social artificial intelligence for enabling human-agent teaming". arXiv preprint arXiv:1909.04492, 2019.

[12] M. Singh, "An ontology for commitments in multiagent systems. Artificial intelligence and law". 7(1), pp. 97-113. 1999.

[13] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, pp. 195–200, 2019. https://doi.org/10.1145/3306618.3314289

[14] T. Adel, Z. Ghahramani, and A. Weller, “Discovering interpretable representations for both deep generative and discriminative models,” In International Conference on Machine Learning. pp. 50–59. 2018.

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

[16] J. Lee, K. See, “Trust in automation: designing for appropriate reliance,” Hum Factor 46(1): pp. 50–80, 2004

[17] B. Friedman, “Value-sensitive design,” interactions, 3(6), pp. 16-23, 1996.

[18] M. Neerincx, & J. Lindenberg, “Situated cognitive engineering for complex task environments, ” Aldershot: Ashgate Publishing Limited. 2008.

[19] M. Neerincx, G. Te Brake, J. Van de Ven, H. Arciszewski, T. De Greef, & J. Lindenberg, “Situated cognitive engineering: Developing adaptive track handling support for naval command and control centers”. In HCP-2008-Third Internation Conference on Human-Centered Processes, pp. 3-20, January 2008.

[20] M. Neerincx, W. van Vught, O. Blanson Henkemans, E. Oleari, J. Broekens, R. Peters, & B. Bierman Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management. Frontiers in Robotics and AI, 6, 118. 2019.

[21] M. Endsley, “Design and evaluation for situation awareness enhancement, ”. In Proceedings of the Human Factors Society annual meeting, Vol. 32, No. 2. Sage CA: Los Angeles, CA: SAGE Publications, pp. 97-101, October 1988.

[22] G. Vlahakis, D. Apostolou, & E. Kopanaki, “Enabling situation awareness with supply chain event management,” Expert Systems with Applications, 93, pp. 86-103. 2018.

[23] M. Tolcott, F. Marvin, and P. Lehner. “Expert decision making in evolving situations,” IEEE Transactions on Systems, Man and Cybernetics 19(3): pp. 606-615, 1989.

[24] M. Naderpour, J. Lu, & G. Zhang, “An intelligent situation awareness support system for safety-critical environments,” Decision Support Systems, 59, pp. 325-340, 2014.

[25] A. Wellens, “Group situation awareness and distributed decision making: From military to civilian applications,” Individual and group decision making: Current issues, pp. 267-291, 1993.

[26] N. Ruedy, M. Schweitzer. “In the Moment: The Effect of Mindfulness on Ethical Decision Making,” ournal of Business Ethics, 95(1), pp. 73–87, 2010, https://doi.org/10.1007/s10551-011-0796-y

[27] A. Atyabi, S. MahmoudZadeh, and S. Nefti-Meziani, “Current advancements on autonomous mission planning and management systems: An AUV and UAV perspective,” Annual Reviews in Control, 2018.

[28] C. Schlenoff, S. Foufou, and S. Balakirsky, “Performance evaluation of robotic knowledge representation (PERK),” in PerMIS, 2012.

[29] F. Maurelli, M. Carreras, J. Salvi, D. Lane, K. Kyriakopoulos, G. Karras, M. Fox, D. Long, P. Kormushev, and D. Caldwell (2016) “The PANDORA project: A success story in AUV autonomy,” in OCEANS 2016 - Shanghai, pp. 1–8, April 2016.

[30] Directorate-General for European Civil Protection and Humanitarian Aid Operations (ECHO) (European Commission), “Overview of natural and man-made disaster risks the European, Natural disasters,” Union may face, December 2017, https://doi.org/10.2795/861482

[31] G. A. Klein, R. Calderwood and D. MacGregor, "Critical decision method for eliciting knowledge," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, no. 3, pp. 462-472, May-June 1989, doi:10.1109/21.31053.

[32] T. Mioch, M. Peeters, & M. Neerincx, “Improving adaptive human-robot cooperation through work agreements,” In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 1105-1110, August 2018.

[33] R. Gutzwiller, S. Espinosa, C. Kenny, & D. Lange, “A design pattern for working agreements in human-autonomy teaming,” In International Conference on Applied Human Factors and Ergonomics Springer, Cham. pp. 12-24, July 2017.

[34] M. Goodrich, T. McLain, J. Anderson, J. Sun, & J. Crandall, “Managing autonomy in robot teams: observations from four experiments,” In 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 25-32, March 2007.

[35] [authors removed], “user-centered social commitment model for location sharing applications in the family life domain International Journal of Agent-Oriented Software Engineering,” Vol. x, No. x, pp.xxx–xxx.

[36] C. Alexander, S. Ishikawa, M. Silverstein, “A Pattern Language: Towns, Buildings, Construction,” Oxford University Press, New York, 1977.

[37] E. Gamma, R. Helm, R. Johnson, J. Vlissides, “Design Patterns: Elements of Reusable Object-Oriented Software,” Addison-Wesley, Reading, 1995.

[38] B. Kehoe, S. Patil, P. Abbeel, & K. Goldberg, A survey of research on cloud robotics and automation. IEEE Transactions on automation science and engineering, 12(2), pp. 398-409, 2015.

[39] E. Galy, M. Cariou, C. Mélan, “What is the relationship between mental workload factors and cognitive load types?,” International journal of psychophysiology, 83(3), pp. 269-275, 2012.

[40] F. Eggemeier & G. Wilson, “Performance-based and subjective assessment of workload in multi-task environments,” Multiple-task performance, pp. 217-278. 1991.

[41] R. Wieringa, & J. Meyer. “Applications of deontic logic in computer science: A concise overview,” Deontic Logic in Computer Science, 17–40. 1993.