TRANSPARENT AND EXPLAINABLE AGENTS FOR HUMAN-AGENT TEAMING

RUBEN VERHAGEN

Transparent and Explainable Agents for Human-Agent Teaming

Transparent and Explainable Agents for Human-Agent Teaming

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. H. Bijl,
chair of the Board for Doctorates,
to be defended publicly on
Thursday 2 April 2026 at 15:00 o'clock

by

Ruben Sebastiaan VERHAGEN

This dissertation has been approved by the promotor and the copromotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. M.A. Neerincx,	Delft University of Technology, *promotor*
Dr. M.L. Tielman,	Delft University of Technology, *copromotor*
Independent members:	
Prof. dr. ir. D.A. Abbink,	Delft University of Technology
Dr. J.B. Lyons,	Air Force Research Laboratory, United States
Prof. dr. ir. J.F.M. Masthoff,	Utrecht University
Prof. dr. F. Santoni de Sio,	Eindhoven University of Technology
Prof. dr. G.C.H.E. de Croon,	Delft University of Technology, *reserve member*

SIKS Dissertation Series No. 2026-22.
The work in the thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Explainable AI, Human-Agent Teaming, Meaningful Human Control

Printed by: Proefschriftspecialist | https://www.proefschriftspecialist.nl/

Cover: Asa Marjew

Style: TU Delft House Style, with modifications by Moritz Beller

ISBN 978-94-6518-278-0

An electronic version of this dissertation is available at http://repository.tudelft.nl/.

Contents

Summary

Human-agent teaming in high-stakes domains is already contributing positively to society, yet these AI agents are often still tools directly controlled by humans. Becoming teammates requires agents to be more autonomous and interdependent, two factors that determine what humans need to know about agents. Agent transparency and explanations can provide this necessary knowledge for effective and responsible collaboration. However, we lack an understanding of what agents should disclose and clarify across interdependencies and autonomy levels. Accordingly, this thesis examines how to design transparent and explainable agents that foster effective and responsible human-agent teaming.

We first develop a conceptual framework that distinguishes agent transparency (disclosing information) from explainability (clarifying that information) and relates these concepts to interpretability and understandability, resolving common ambiguities. Using simulation environments, we then demonstrate that interdependence influences how transparency and explanations impact human-agent teaming processes, underscoring its importance in studies on transparent and explainable agents. Next, we examine the trust calibration process across interdependencies. We find first evidence that interdependence relationships influence trust calibration in human-agent teams, suggesting that engaging in joint actions facilitates more accurate trust calibration.

To support responsible human-agent teaming, we develop an evaluation method for meaningful human control based on expert knowledge, operationalizing traceability through objective and subjective indicators and eliciting reasons underlying outcomes. We apply this method to study agent autonomy and explanations in morally sensitive situations. The findings suggest that people prefer more involvement over greater agent autonomy and that they take on greater moral responsibility when agents explain potential consequences. These insights are crucial for designing agents that enhance human moral awareness and human-agent teaming in morally sensitive situations.

Translating these insights to practice, we design TEAMS (Transparent and Explainable Autonomy for Mapping and Searching). This human-robot collaboration system for firefighting moves beyond teleoperation by proposing and explaining intermediate navigation destinations while autonomously navigating towards them. This system is grounded in expert firefighting knowledge and can address the challenge of camera-based teleoperation in low-visibility conditions. We highlight the importance of training, iterative and human-centered refinements, and software optimization to further enhance the system.

Finally, we synthesize a research agenda with taxonomies and guidelines, team design patterns, modular testbeds, and study templates to advance the field. Taken together, this thesis offers a path from concept to practice: a conceptual framework, studies in simulation environments, an evaluation method for meaningful human control, and TEAMS in a practically grounded setting, complemented by a research agenda. By doing so, this thesis supports the design of transparent and explainable AI agents that foster effective and responsible human-agent teaming.

Samenvatting

Samenwerking tussen mensen en kunstmatig intelligente agenten in domeinen waar veel op het spel staat levert al een positieve bijdrage aan de samenleving, ondanks dat deze agenten vaak nog hulpmiddelen zijn die rechtstreeks door mensen worden aangestuurd. Om volwaardige teamgenoten te worden moeten agenten autonomer en onderling afhankelijker worden, twee factoren die bepalen wat mensen over agenten moeten weten. Transparantie en uitleg van agenten kan deze noodzakelijke kennis bieden voor een effectieve en verantwoorde samenwerking. Er ontbreekt echter nog begrip van wat agenten precies zouden moeten onthullen en toelichten binnen verschillende vormen van onderlinge afhankelijkheid en niveaus van autonomie. Dit proefschrift onderzoekt daarom hoe transparante en uitlegbare agenten kunnen worden ontworpen die effectieve en verantwoorde mens-agentsamenwerking bevorderen.

Allereerst ontwikkelen we een conceptueel kader dat agent transparantie (het onthullen van informatie) onderscheidt van uitleg (het toelichten van die informatie) en deze begrippen relateert aan interpreteerbaarheid en begrijpelijkheid, waarmee veelvoorkomende onduidelijkheden worden opgelost. Vervolgens tonen we in gesimuleerde omgevingen aan dat onderlinge afhankelijkheid invloed heeft op de manier waarop transparantie en uitleg van agenten de processen binnen mens-agent teams beïnvloeden, wat het belang van deze factor onderstreept in onderzoek naar transparante en uitlegbare agenten. Daarna onderzoeken we het proces van vertrouwenskalibratie binnen verschillende vormen van onderlinge afhankelijkheid. We vinden eerste aanwijzingen dat deze afhankelijkheidsrelaties de vertrouwenskalibratie in mens-agent teams beïnvloeden, suggererend dat gezamenlijke acties een meer nauwkeurige vertrouwenskalibratie mogelijk maken.

Om verantwoorde mens-agentsamenwerking te ondersteunen, ontwikkelen we een evaluatiemethode voor betekenisvolle menselijke controle op basis van deskundige kennis. Deze methode operationaliseert traceerbaarheid via objectieve en subjectieve indicatoren en achterhaalt de redenen achter uitkomsten. We passen deze methode toe om autonomie en uitleg van agenten te bestuderen in moreel gevoelige situaties. De resultaten suggereren dat mensen meer betrokkenheid verkiezen boven meer autonome agenten, en dat zij meer morele verantwoordelijkheid nemen wanneer agenten potentiële consequenties toelichten. Deze inzichten zijn essentieel voor het ontwerpen van agenten die het moreel bewustzijn van mensen versterken en samenwerking in moreel gevoelige situaties verbeteren.

Om deze inzichten naar de praktijk te vertalen, ontwerpen we TEAMS (Transparent and Explainable Autonomy for Mapping and Searching). Dit systeem voor mens-robotsamenwerking tijdens brandbestrijding gaat verder dan tele-operatie door tussentijdse navigatiebestemmingen voor te stellen en uit te leggen, terwijl het autonoom naar deze bestemmingen navigeert. Het systeem is gebaseerd op deskundige kennis uit de brandweerpraktijk en kan het probleem van cameragebaseerde tele-operatie onder omstandigheden met beperkte zichtbaarheid aanpakken. We benadrukken het belang van training, iteratieve en mensgerichte verfijning, en software-optimalisatie om het systeem verder te verbeteren.

Tot slot synthetiseren we een onderzoeksagenda met taxonomieën en richtlijnen, teamontwerppatronen, modulaire testomgevingen, en studietemplates om het vakgebied verder te brengen. Gezamenlijk biedt dit proefschrift een traject van concept tot praktijk: een conceptueel kader, studies in gesimuleerde omgevingen, een evaluatiemethode voor betekenisvolle menselijke controle, en TEAMS in een praktijkgerichte setting, aangevuld met een onderzoeksagenda. Op deze manier ondersteunt dit proefschrift het ontwerp van transparante en uitlegbare AI-agenten die effectieve en verantwoorde mens-agentsamenwerking bevorderen.

1

Introduction

1.1 Problem and Motivation

On May 13, 2025, the Limburg-Noord Fire Brigade successfully fought a big fire at a large retail chain of 1000 m^2. For the first time ever, they deployed their teleoperated exploration and extinguishing robot. The robot immediately proved its value by first extinguishing from outside and then entering the building to extinguish in places too dangerous for firefighters because of hazardous substances. Despite this success, the Fire Brigade's future vision extends beyond teleoperating the robot, as this becomes increasingly challenging in low-visibility conditions and adds another burden to their already substantial workload.

Ideally, firefighters would collaborate with a more autonomous robot. This shifts the robot's role from that of a tool to that of a teammate. However, this new role introduces challenges in shaping and defining the human-agent collaboration. Firefighters may need to rely on the robot for physical work inside too dangerous buildings, such as extinguishing fires and rescuing victims. In contrast, it is not recommended that the robot be responsible for deciding which victims can be rescued. Instead, firefighters may be the performers of such cognitive and morally sensitive work, while being supported by the robot. Other tasks may require the robot to provide relevant sensor data to the firefighters, who then make decisions such as selecting navigation goals. These interdependence relationships between the robot and firefighters result in various observability, predictability, and directability requirements [99]. Robot transparency and explanations are essential for supporting these requirements and fostering effective, responsible human-robot teaming [208].

This ideal future situation illustrates how humans and artificial intelligence (AI) agents can augment each other and achieve outcomes that would otherwise be impossible. Humans are often limited in what they can achieve together due to capacity or safety constraints, especially in high-stakes domains characterized by uncertainty and high workload, such as emergency response, healthcare, and defense. Teaming up with AI agents has enabled us to start addressing these limitations, even though most state-of-the-art AI agents are often directly controlled by humans (e.g., through teleoperation).

To achieve this ideal, human-agent teaming in these high-stakes domains must be effective and responsible. Effective teamwork requires high-quality outcomes and well-performing teams. Responsible human-agent teaming means aligning behavior and outcomes with human values, legal standards, and ethical principles to ensure meaningful human control [30]. Shared mental models of task and team knowledge between humans and agents are crucial to achieve these goals [28, 45]. Developing and maintaining such models requires behavioral transparency and explanations from both humans and agents. However, what exactly humans need to know about agents during human-agent teaming in high-stakes domains depends strongly on the agents' level of autonomy and the interdependencies between both team members. These two factors result in different transparency and explanation requirements for effectively and responsibly executing tasks [99, 101].

Concretely, interdependencies in human-agent teams determine the relationships, coordination, and information necessary to collaborate [99]. For example, some tasks require joint execution, others can be performed independently but benefit from collaboration, and some may be restricted to specific team members. Successfully managing these interdependencies requires humans and agents to observe, predict, and direct each other's behavior [99, 101]. Interdependence is often strongly linked with autonomy. Higher levels of autonomy allow agents to perform distinct tasks and collaborate interdependently with humans to achieve shared goals [157]. At the same time, higher levels of autonomy can also enable agents to act and decide more independently rather than collaborate interdependently with humans. Together, interdependence and autonomy thus strongly shape what agents should disclose and clarify to be observable, predictable, and directable for humans. The ongoing, rapid developments in robotics and AI will ultimately make agents proper teammates able to meet these observability, predictability, and directability requirements.

As interdependence and autonomy shape the information necessary to support these requirements [208, 224], they also moderate the effects of agent transparency and explanations on team processes such as trust. The field that studies what explanations agents should provide, how they should do so, and what the effects are on humans, is called Explainable AI (XAI). The motivation behind XAI in general is that AI agents should be able to disclose and clarify their observations, knowledge, actions, and goals. This communication should enhance human understanding of the agents and increase awareness of their responsibility for outcomes resulting from agent behavior [30]. Otherwise, humans will struggle to trust agent behavior appropriately, as they lack accurate mental models of their capabilities and limitations [133, 145]. Additionally, humans will struggle to determine whether and when to override agent behavior that does not align with human values, legal standards, and ethical principles [30]. Explainable AI is a field that studies all types of human-agent interaction, but it is particularly relevant to the teamwork vision sketched before.

Autonomy and interdependence shape information requirements for effective and responsible human-agent teaming, making both factors crucial to consider when designing and evaluating transparent and explainable agents. However, we currently lack an empirically grounded understanding of what agents should disclose and clarify to human teammates to foster effective and responsible teaming, particularly across various interdependencies and autonomy levels. This motivates the following main research question of this dissertation:

> How should we design transparent and explainable agents that foster effective and responsible human-agent teaming across interdependencies and autonomy levels?

1.2 Transparent and Explainable Agents in Teams

Before confidently implementing transparent and explainable agents for real-world human-agent teams, we must study how such agents impact collaboration and outcomes. This first requires an understanding of the most relevant concepts and their relationships, the current state of the art, and the most significant challenges.

1.2.1 Human-Agent Teaming in Complex Environments

Human-agent teaming can be defined as at least one human and AI agent collaborating towards a common goal [99]. This agent can range from a virtual decision support system to a physical robot performing actions. Human-agent teams can be involved in cognitive and/ or physical work [217]. Cognitive work includes mental or information processing activities, such as firefighters monitoring camera images from their exploration and extinguishing robot. In contrast, physical work includes manipulating tangible objects in the world, such as firefighting robots pushing aside rubble. Our firefighting examples perfectly illustrate how both humans and AI agents are often limited in what outcomes they can achieve independently, especially in high-stakes domains with demanding and dynamic environments. Human-agent teams aim to achieve these otherwise impossible outcomes by combining the unique strengths of humans and agents, augmenting each other's capabilities [3].

Humans and agents usually have explicit roles during this collaboration, such as supervisor, performer, or supporter [99]. Higher levels of autonomy can enable agents to take on unique roles or sets of tasks and work interdependently with human team members to achieve shared goals [157]. Agent autonomy is typically classified into ten levels, ranging from no autonomy to full autonomy, each defining the agent's roles, capabilities, and degree of self-government and self-directed behavior [159]. Team design patterns can be used to shape and define human and agent roles by expressing collaboration forms with various team properties [217, 218]. These patterns outline how humans and agents collaborate and communicate, the requirements that enable such interactions, and the advantages and limitations of their use.

Teamwork differs from regular interaction because it involves interdependencies between the activities and outcomes of humans and agents able to execute actions independently and proactively [131, 157]. These interdependencies reflect how humans and agents mutually depend on each other during teamwork, shaping coordination, communication, and collaboration requirements [101]. They can result from task structures such as pooled, sequential, reciprocal, or team relationships [173]. These task interdependencies form a hierarchy that represents increasing needs for coordination to manage dependencies. For example, pooled interdependence involves independent task execution without interaction, whereas sequential interdependence involves task execution where team members wait for others to complete their task first. In contrast, reciprocal interdependence involves team members taking turns to partially complete tasks, while team interdependence involves concurrent task execution or even joint actions.

Unfortunately, these task interdependencies fail to capture all the dynamics of human-agent collaboration. Addressing these dynamics also requires understanding the interdependencies between team members during joint actions. Such actions can result in required or opportunistic interdependencies [99]. Required interdependencies arise from a lack of necessary knowledge, skills, abilities, or resources to competently execute actions independently. In such cases, joint action execution is the only option. In contrast, opportunistic interdependence is optional and arises from recognizing opportunities to be more effective by working jointly.

These interdependencies will become increasingly significant as AI agents take on more autonomous roles as teammates. Moreover, they result in different observability, predictability, and directability requirements during collaboration [99]. Human-agent teams need mechanisms that can support these requirements and successfully manage interdependencies. Otherwise, human-agent teaming will not be effective and responsible.

1.2.2 Effective and Responsible Human-Agent Teaming

Effective human-agent teaming primarily focuses on high-quality team outcomes and well-performing teams. One general requirement for successful human-agent teaming is that both team members have accurate mental models of each other's capabilities, limitations, and knowledge [107, 175]. This way, they can appropriately understand and trust each other [13]. Interdependence relationships not only shape human-agent collaboration and create different information requirements, they can also facilitate human assessment of agent trustworthiness [97]. Such assessment is crucial for accurate trust calibration towards appropriate trust [141]. The trust calibration process involves adjusting trust over time and through repeated interactions, in response to changes in agent reliability and trustworthiness [141, 156]. This process ideally results in humans' trust matching agents' actual trustworthiness (i.e., appropriate trust) [141, 182]. Fostering such appropriate trust is crucial because a lack thereof can cause over- or under-trusting AI agents, potentially resulting in detrimental outcomes [118, 158].

Facilitating effective teamwork requires not only mutual understanding to enable appropriate trust, but also sufficient situation awareness during the task [61]. Situation awareness involves perceiving elements in situations, comprehending their meaning, and projecting their future status [61]. These are critical prerequisites for human decision-making and action execution, and thus for effective teamwork [60]. At the same time, human workload should be adequately balanced to avoid negative effects on situation awareness [60].

High-quality outcomes and well-performing teams are not the only desired outcomes of human-agent teams. Responsible human-agent teaming is also essential, especially in morally sensitive domains such as emergency response, healthcare, and defense [178]. Such teamwork does not necessarily strive to optimize performance but instead prioritizes aligning agent behavior and outcomes with human values, legal standards, and ethical principles [30]. Discussions on responsible AI and teaming accelerated in response to autonomous weapon systems and related questions about accountability and responsibility. These discussions resulted in a new concept called meaningful human control [179].

As agents take on more autonomous roles, responsible human-agent teaming becomes even more important. Higher levels of autonomy allow agents to perform distinct tasks and collaborate interdependently with humans to achieve shared goals [157]. However, these same capabilities also increase the need for and importance of meaningful human control to ensure responsible human-agent teaming. Meaningful human control requires that humans ultimately remain in control of and be morally responsible for the behavior of AI agents [179]. This becomes particularly challenging with increasingly autonomous agents that deal with morally sensitive situations in which people's welfare, rights, and values may be affected. It is crucial that humans can be held accountable for outcomes resulting from agent behavior, especially in those situations [179, 214]. Otherwise, gaps in culpability, moral and public accountability, and active responsibility may arise [178]. Therefore, meaningful human control is increasingly imposed as a requirement for AI agents [214].

AI agents should meet tracking and tracing requirements to ensure meaningful human control [179]. Tracking requires agents to adapt to the moral reasons of humans, who are then considered in control of and morally responsible for these agents. These reasons have been ordered based on their proximity and complexity in influencing agent behavior. More proximal reasons, such as intentions, are argued to be simpler and closer in time to agent behavior than more distal reasons, such as values [139]. Furthermore, tracing requires at least one human involved in the design or interaction with AI agents to understand their moral and technical behavior properly [179]. This way, agent behavior should always be traceable to their designers, deployers, or users [179]. Since these requirements are relatively abstract, more actionable solutions to ensure meaningful human control have also been proposed [30]. These include team design patterns to shape meaningful human control [214, 219], value sensitive design to respect norms and values [75], machine ethics to implement artificial moral agents [6], explainable AI to achieve human moral awareness [30, 214], and variable autonomy to allow human control and responsibility [144].

1.2.3 Agent Transparency and Explanations

Effective and responsible human-agent teaming requires successfully managing interdependencies and ensuring meaningful human control. This is only possible when both team members have accurate mental models of each other's capabilities, limitations, and knowledge [107, 175]. Building these accurate mental models requires behavioral transparency and explanations, which, unfortunately, do not come naturally to AI agents [107].

One reason why such explanations are crucial for human-agent teaming is that otherwise humans attribute agent behavior by assigning inappropriate mental states that explain the behavior [132, 133, 145]. Without explanations, these mental states can involve incorrect beliefs, goals, emotions, and intentions. In contrast, explaining the reasons underlying agents' behavior helps humans to assign the correct mental states to their behavior. This will help humans better understand the capabilities and limitations of AI agents, which can enhance human-agent teaming [9].

Explainable AI is a field that develops methods for such behavioral transparency and explanations. These methods aim to make AI agents better understandable by explaining their perceptions and behavior, ideally fostering appropriate trust and enhancing human-agent interaction or collaboration [9, 99, 101, 115]. The field is generally divided between data-driven/perceptual and goal-driven/cognitive explainable AI [9, 149]. Data-driven/perceptual explainable AI involves explaining and understanding the decisions and inner workings of machine learning algorithms, given certain input data [9, 79]. In contrast, goal-driven/cognitive explainable AI involves explaining and understanding the actions, decisions, and underlying reasons of goal-driven AI agents [9, 115]. Explanations are often characterized as global if they concern general system behavior or local if they concern specific decisions or actions [56, 79].

Explanation methods can be divided into generation, communication, and reception phases [149]. The generation phase involves extracting explanations from the underlying models used by AI agents, such as which features influence their behavior. The communication phase encompasses the content and form of explanations, such as textual, visual, or hybrid explanations [197]. The reception phase involves empirical research on explanation effectiveness, which is something fewer than 1% of explainable AI studies do [195]. Moreover, these studies often conduct human-grounded evaluations with laypeople as participants, rather than application-grounded evaluations with a representative population sample that fits the context [56].

Nowadays, some of the most common explanation types include feature attributions, confidence explanations, and contrastive explanations [210]. Feature attributions clarify which relevant situational features influenced AI agent behavior and are useful to enhance agent predictability and identify biases that require adjustments [1, 214]. Confidence explanations clarify how certain agents are that their decisions are correct and are helpful to decide whether to trust agents [212, 214]. Contrastive explanations clarify why agents made certain decisions instead of others and are useful to enhance agent predictability and human understanding of their reasoning [145, 214].

Several studies have already investigated how agent transparency and explanations influence the prerequisites for effective human-agent teamwork. Although increasingly transparent and explainable agents often enhance situation awareness, trust, and performance, they can also negatively impact workload [34, 142, 157, 187]. Commonly identified challenges and goals include designing agents that can adapt their transparency and explanations based on both user and context [157]. This could be achieved by modelling both user and context and using that model to adapt the transparency and explanations generated and/or communicated by the agent [9, 145]. Before developing such adaptive agents, we need to know under what collaboration conditions agent transparency and explanations are beneficial or detrimental. These insights remain limited because agent transparency and explanations are often studied without considering the dynamic nature of teamwork arising from interdependencies between humans and AI agents.

Research on transparent and explainable agents for meaningful human control and responsible human-agent teaming is also limited. The tracking and tracing requirements of meaningful human control essentially tell us that agent behavior and human understanding of that behavior determine meaningful human control. Agent transparency and explanations can facilitate this behavioral understanding required for humans to exercise control properly [214]. Rather than shifting accountability to agents, such transparency and explanations should foster human moral awareness by meeting the epistemic condition of moral responsibility [16, 125, 172].

Meaningful human control requires humans to be aware and able to act upon this responsibility [30]. How human-agent collaboration is designed and shaped can partially achieve this, for example, by giving humans the authority always to intervene and override agent behavior. However, this is not sufficient to ensure that human control is meaningful. Agents should also be transparent about their behavior and provide explanations for it. Such communication should enhance human understanding of the agents and increase awareness of their responsibility for outcomes resulting from agent behavior [30]. However, we currently lack a clear understanding of what agents should disclose and clarify to ensure meaningful human control across various levels of agent autonomy.

The relationships between transparency, explanations, interdependence, autonomy, and human-agent teaming measures studied in this thesis, can be shown in a conceptual model (Figure 1.1). This model illustrates how interdependencies and autonomy levels shape human and agent behavior during collaboration. It also highlights how they moderate the influence of agent transparency and explanations on key effective and responsible human-agent teaming measures such as performance, workload, situation awareness, trust, and meaningful human control.

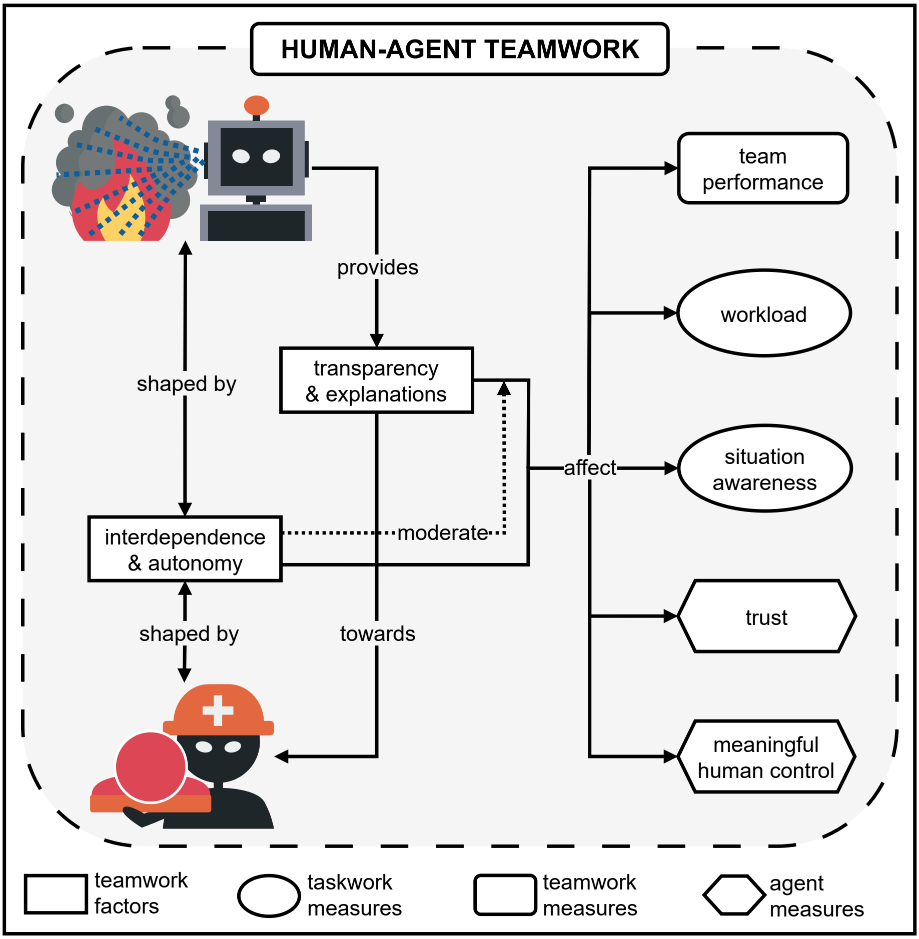


Figure 1.1: Overall conceptual model of this thesis.

1.3 Research Approach

This thesis consists of both analytical and empirical studies on human-agent[1] teaming. We conducted the analytical studies to identify and define our key concepts and their relationships. This includes the development of a conceptual framework that defines and relates agent transparency and explainability, grounded in explainable AI and social science literature. We also operationalize the concept of meaningful human control based on ethical, legal, and technical literature. A qualitative focus group study with experts further informed this work. The final analytical study provides a research agenda for the human-agent teaming research community, grounded in the lessons learned and challenges identified throughout the thesis. Moreover, it is based on the organization and outputs of a Lorentz Center workshop on research environments for human-agent teaming. During this workshop, we discussed and collaborated towards addressing frequently encountered challenges in human-agent teaming research. We summarized the workshop outputs into a research agenda with concrete goals and deliverables for the community.

These analytical studies informed our empirical work by defining the independent and dependent variables. We used simulation environments for our more fundamental research questions on the effects of interdependence, autonomy, and explanations on team processes such as trust. Such environments are good to control and can accommodate user studies efficiently. We simulated emergency response tasks in these environments, where human subjects collaborated with rule-based AI agents to search and rescue victims. Humans and agents collaborated in a two-dimensional grid environment and communicated through a chat window. These environments enabled us to rapidly implement, manipulate, and evaluate interdependencies and agent autonomy, transparency, and explanations. For example, by adding obstacles that could only be removed collaboratively, or by making agents communicate more reasoning information. This way, these environments facilitated (1) studying transparent and explainable agents during human-agent teaming and (2) obtaining actionable insights required to implement such agents in real-world scenarios.

To take a step towards applying the envisioned use case illustrated at the start of this thesis, we conducted our final empirical study with a physical robot in a realistic, practically grounded setting, inspired by real-world firefighting scenarios. This also contributed to the ecological validity of the study and its findings [8]. Co-designed with the Rotterdam Fire Brigade, we developed a human-robot collaboration system and compared it with teleoperation. We simulated firefighting tasks to map environments and find victims.

We primarily assessed our empirical studies with laypeople to evaluate more general human-agent teaming processes, obtaining insights that can later be validated with domain experts. A second practical driver of this approach is that recruiting sufficient domain experts is challenging [56, 195]. However, we involved domain experts in the design and evaluation of our simulation environments discussed above. We also recruited one domain expert to compare our human-robot collaboration system with teleoperation based on

[1] We use *(AI) agent* as an umbrella term for the artificial teammates studied in this dissertation. Chapter-specific terms (human-*agent*/*AI*/*machine*/*robot* teaming) follow the original papers. Unless otherwise stated, read them as referring to the same construct. The only systematic distinction is embodiment: *robot* denotes an embodied (AI) agent; *agent/AI* denotes a software (disembodied) instantiation. By *machine* we mean the broader human-machine interaction umbrella that encompasses both software agents and robots; we use it when the type of machine is not yet specified.

video-recorded interactions.

We employed quantitative methods to evaluate the empirical studies in simulation environments. Measures included subjective self-report scales (e.g., trust and workload), objective probe-based checks (e.g., situation awareness queries), and behavioral performance metrics (e.g., task completion and time). While these quantitative methods can provide generalizable results, they fail to expose the in-depth explanations for perceptions and behavior that qualitative methods can. Mixed-methods research combines the complementary strengths of both approaches, pairing generalizable effects with nuanced experiential insights [198]. Therefore, we evaluated our final empirical study with the physical robot in a mixed-methods study. This study combined high-fidelity, in-person evaluations by laypeople with medium-fidelity, video-based evaluations by both laypeople and a domain expert. The in-person interactions were assessed qualitatively through semi-structured interviews that focused on human-agent teaming processes such as trust and workload. The video-based evaluations by laypeople were quantitatively assessed using self-report scales on the same team processes. Finally, the domain expert qualitatively evaluated the videos of our human-robot collaboration system and teleoperation, assessing their challenges, advantages, disadvantages, and applicability.

1.4 Research Questions and Thesis Structure

This thesis consists of several research sub-questions to answer the main research question. Each thesis chapter answers a separate sub-question. Figure 1.2 presents an overview of the thesis chapters, organized by research type, concepts, and dependent and independent variables.

1.4.1 Defining Agent Transparency and Explainability

There is an extensive amount and variety of research and methods on transparent and explainable agents. Significant progress has been made in extracting various explanation types from AI agents, different ways to communicate these, and insights into their effectiveness [157]. However, the explainable AI community still lacks clear definitions of and relationships between its key concepts transparency, explainability, interpretability, and understandability. These concepts are often used interchangeably or within each other's definitions. The resulting ambiguity makes it challenging to comprehend research on these concepts and should first be resolved before we can investigate, manipulate, or implement them. Therefore, our first research sub-question is:

> How should we define and relate agent transparency, explainability, interpretability, and understandability? (Chapter 2)

To address this question, Chapter 2 develops a conceptual framework that offers clear and distinct definitions of transparency, explainability, interpretability, and understandability. This framework also clarifies how these concepts relate to one another, resolving common ambiguities in the explainable AI literature.

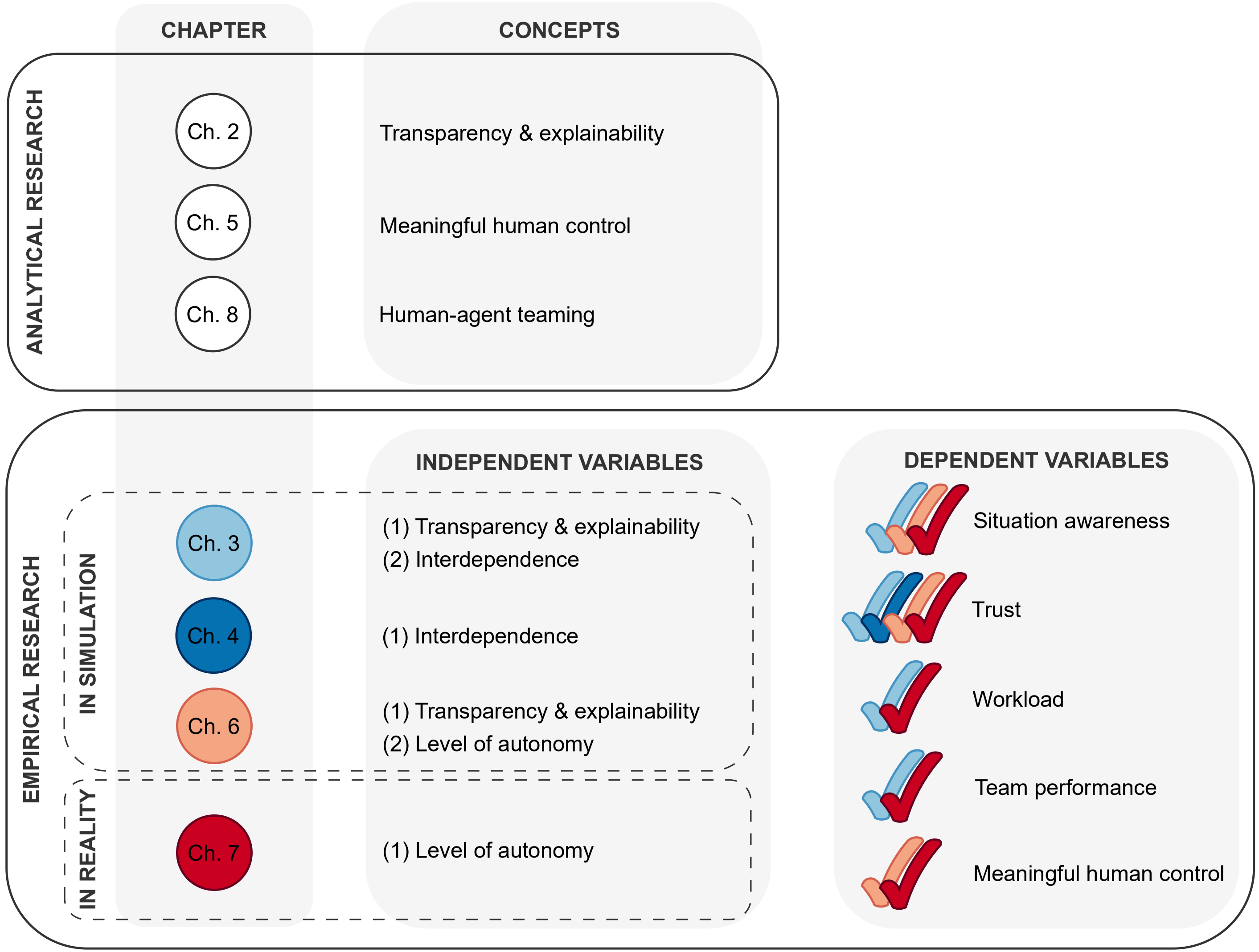


Figure 1.2: Overview and structure of the dissertation chapters, grouped based on research type, concepts, and dependent and independent variables.

1.4.2 Transparency and Explanations across Interdependence

Interdependence relationships between humans and AI agents will become increasingly important as agents take on more autonomous roles as teammates [101]. These relationships result in different observability, predictability, and directability requirements for effective human-agent teaming [99]. Agent transparency and explanations can support these requirements, especially by enhancing the observability and predictability of their behavior. However, it remains unclear how different levels of interdependence influence the effectiveness of agent transparency and explanations in fostering effective human-agent teaming. Therefore, the second research sub-question is:

> How do interdependence and agents' transparency and explanations influence effective human-agent teaming, individually and interactively? (Chapter 3)

To address this question, Chapter 3 presents a user study with laypeople on the effects of transparent and explainable agents across interdependence. This study provides empirical insights into the human-agent teaming conditions under which agent transparency and explanations are either beneficial or detrimental.

1.4.3 Interdependence and Trust Calibration

Appropriate human trust in agents is crucial during collaboration to avoid the disuse and misuse of agents [118]. Several approaches, such as confidence explanations, uncertainty communication, and trustworthiness cues, can foster appropriate trust [141]. Interdependence relationships can give rise to these approaches by requiring mechanisms that support observability, predictability, and directability. By doing so, these mechanisms can facilitate active and continuous trust exploration between humans and agents. This exploration should ensure that humans' assessments of agents' trustworthiness are appropriate for achieving optimal outcomes [97]. However, the required coordination and mutual dependencies can vary between interdependence relationships, such as independent task execution or optional collaboration. It currently remains unclear how these different interdependence relationships influence the trust calibration process. Therefore, the third research sub-question is:

> How do interdependencies during human-agent teaming influence the human-agent trust calibration process? (Chapter 4)

To address this question, Chapter 4 presents a user study with laypeople on the effects of interdependencies on the trust calibration process. This study contributes empirical insights into the role and importance of interdependencies in fostering appropriate trust during human-agent teamwork.

1.4.4 Measuring Meaningful Human Control

Responsible human-agent teaming requires meaningful human control during the collaboration. Various methods for ensuring meaningful human control over AI agents have been proposed, such as team design patterns to shape meaningful human control [214, 219], value sensitive design to respect norms and values [75], machine ethics to implement artificial moral agents [6], explainable AI to achieve human moral awareness [30, 214], and variable autonomy to allow human control and responsibility [144]. Meanwhile, we currently lack methods to evaluate if agents are indeed under meaningful human control [214]. Such methods are crucial considering that meaningful human control is already imposed as a requirement for AI agents. Therefore, the fourth research sub-question is:

> How can we measure meaningful human control during human-agent teaming? (Chapter 5)

To address this question, Chapter 5 presents a focus group study with experts to develop an evaluation method for meaningful human control. This method enables researchers and designers to assess whether meaningful human control is present during human-agent teaming.

1.4.5 Transparency and Explanations across Agent Autonomy

Agent behavior and humans' technical and moral understanding of that behavior are crucial determinants of meaningful human control. Agent transparency and explanations can facilitate this understanding, ideally enabling humans to exercise control properly [214]. Such transparency and explanations should not influence humans to hold agents accountable but instead achieve human moral awareness by fulfilling the epistemic condition of direct moral responsibility [16, 125, 172]. Specifically, agent transparency and explanations should ensure that humans are aware that agent behavior traces back to them and that they are in control and responsible for all outcomes [16, 225]. We currently lack a clear understanding of what agents should disclose and clarify to achieve these goals, or how different types of information influence responsible human-agent teamwork. Moreover, agent autonomy should also be considered, as it shapes both human-agent teaming and the requirements for agent transparency and explanations. Therefore, the fifth research sub-question is:

> How do agents' autonomy and their transparency and explanations influence responsible human-agent teaming, individually and interactively? (Chapter 6)

To address this question, Chapter 6 presents a user study with laypeople on the effects of transparent and explainable agents across agent autonomy levels. This study contributes empirical and actionable insights for designing agents that enhance human moral awareness and human-agent teaming in morally sensitive situations.

1.4.6 Transparent and Explainable Robots for Firefighting

The preceding research questions provide essential insights into designing transparent and explainable agents for effective and responsible teamwork. However, the ecological validity of these insights is uncertain as it remains unclear how well they generalize beyond controlled, simulated environments to real-life settings. Applied research can address this challenge by studying practical solutions to real-world problems or improving existing products, practices, or services. An example of such research is designing a more autonomous exploration and extinguishing robot for the Dutch Fire Brigade, enabling collaboration rather than teleoperation. Combining this increase in robot autonomy with transparency and explanations is crucial, as the former may hinder understanding, while the latter can facilitate it. At the same time, it is also essential to ensure meaningful human control during the collaboration. Ideally, the firefighters will collaborate with a more autonomous robot that remains under their control while providing transparency and explanations for its behavior. However, it currently remains unclear how to design and implement such a semi-autonomous, transparent and explainable robot, or how that increase in autonomy impacts effective and responsible human-robot teaming compared to teleoperation. Therefore, the sixth research sub-question is:

> How should we design a semi-autonomous, transparent and explainable robot in human-robot teams for firefighting, and how does the increased autonomy impact effective and responsible collaboration compared to teleoperation? (Chapter 7)

To address this question, Chapter 7 presents the design and implementation of a semi-autonomous, transparent and explainable robot in human-robot teams for firefighting. Moreover, it presents user studies with laypeople and a domain expert that explore how increased robot autonomy, compared to teleoperation, affects collaboration. This chapter contributes both a practical solution for transparent and explainable robots in real-world firefighting scenarios, as well as empirical insights into how robot autonomy affects human-robot teaming.

1.4.7 Advancing Human-Agent Teaming Research

The answers to all these preceding sub-questions will provide fundamental insights into designing transparent and explainable AI agents for effective and responsible teaming. At the same time, while answering these sub-questions, we learned many lessons and identified several challenges for human-agent teaming research. For example, the lack of requirements for effective research, numerous methods and testbeds without centralized documentation, and a disconnect between research and real-world applications. These challenges hinder progress and limit the generalizability of research outcomes. Therefore, before summarizing the main findings and contributions of this dissertation, the seventh and final research sub-question is:

> How should the human-agent teaming research community adopt a more structured and systematic approach to advance the field? (Chapter 8)

To address this question, Chapter 8 presents a research agenda with actionable directions for the human-agent teaming research community. These directions can accelerate progress in the field and lay the foundation for a common platform with essential tools for human-agent teaming research.

2

A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable

Because of recent and rapid developments in Artificial Intelligence (AI), humans and AI-systems increasingly work together in human-agent teams. However, in order to effectively leverage the capabilities of both, AI-systems need to be understandable to their human teammates. The branch of eXplainable AI (XAI) aspires to make AI-systems more understandable to humans, potentially improving human-agent teamwork. Unfortunately, XAI literature suffers from a lack of agreement regarding the definitions of and relations between the four key XAI-concepts: transparency, interpretability, explainability, and understandability. Inspired by both XAI and social sciences literature, we present a two-dimensional framework that defines and relates these concepts in a concise and coherent way, yielding a classification of three types of AI-systems: incomprehensible, interpretable, and understandable. We also discuss how the established relationships can be used to guide future research into XAI, and how the framework could be used during the development of AI-systems as part of human-AI teams.

🗎 **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable. *International workshop on explainable, transparent autonomous agents and multi-agent systems.* Cham: Springer International Publishing, 2021.

2.1 Introduction

Rapid developments in the field of Artificial Intelligence (AI) have resulted in the design and adoption of intelligent systems/agents (A/IS) working together with humans. For such human-AI teams to work effectively and efficiently, it is crucial that AI-systems are understandable and predictable to their human teammates [99, 101, 107]. The eXplainable Artificial Intelligence (XAI) community aims to make AI more understandable, however, there is a lack of clear definitions and relationships between key concepts in XAI. The objective of this chapter is to identify similarities, differences and inconsistencies in the description and usage of these concepts, and to establish a framework in which the concepts can be unambiguously defined and related to each other.

Autonomous and intelligent systems/agents (A/IS) are characterized by their abilities to *sense* their environment, *reason* about their observations and goals, and consequently make *decisions* and *act* within their environment in a goal-driven manner [216]. Thanks to these capabilities, A/IS often outperform humans with respect to handling complex problems and rapid and rational decision-making. Consequently, the adoption domains of A/IS range from applications in healthcare to military defense. On the other hand, humans still surpass A/IS regarding the handling of uncertainty and unexpected situations. In an attempt to assemble their diversity in skills and leverage the unique abilities of both, A/IS and humans are increasingly paired to create human-agent teams (HATs).

Several factors are crucial for and determine the success of human-agent teams. Some of the most cited involve mutual trust and understanding; shared mental models and common ground; observability, predictability and directability; transparency and explainability; and teaming intelligence [99, 101, 107, 175]. Unfortunately, many of these factors are lacking in contemporary human-agent teams. For example, most A/IS demonstrate extremely limited directability and often possess only rudimentary teaming intelligence (i.e., the knowledge, skills, and strategies necessary to effectively team) [101]. Furthermore, A/IS often demonstrate poor transparency and explainability, making it hard for human teammates to properly understand their inner workings, behavior, and decision-making [9, 126, 145]. This, in turn, negatively affects factors like mutual trust and understanding, eventually resulting in decreased global team performance [99, 101].

To understand the behavior of A/IS, humans attribute A/IS behavior by assigning particular mental states (i.e., Theory of Mind) that explain the behavior [9, 76, 132, 133, 145]. Such mental states involve beliefs, desires/goals, emotions, and intentions. For example, humans trying to understand a robot entering a burning house can do so by attributing it to the goal to save a victim. A/IS capable of self-explaining their behavior and actions based on the reasons for the underlying intentions (e.g., beliefs, goals, emotions) help human teammates to build this ToM of the A/IS. This, in turn, will result in better understanding of the capabilities and limits of the A/IS and eventually better human-agent collaboration [9].

Explainable AI (XAI) methods, techniques, and research emerged as a means of making AI-systems more *understandable* to humans [80]. This relatively new community is characterized by the distinction between data-driven - and goal-driven XAI [9] (or perceptual vs. cognitive XAI [149]). Data-driven XAI is about *explaining* and *understanding* the decisions and inner workings of “black-box" machine learning algorithms given certain input data [9, 79]. In contrast, goal-driven XAI/explainable agency refers to building goal-driven A/IS (e.g., robots) *explaining* their actions and reasons leading to their decisions to lay users [9, 115].

Although fundamentally different branches, both data- and goal-driven XAI are characterized by the same fundamental issue: a lack of consensus with regards to the definition of and relations between key XAI concepts. Furthermore, provided definitions often suffer from a high level of ambiguity because they frequently refer to related notions. For example, the concepts of *transparency*, *interpretability*, *explainability*, and *understandability* are all frequently used in XAI literature, but often interchangeably, differently, with recourse to each other, or without even being defined. Without establishing clear distinctions and relations between these notions, the resulting ambiguity significantly hampers the comprehensibility of research centered around these concepts. We argue that prior to implementing, manipulating, or investigating these key concepts it is fundamental to first define and relate them. Only in this way, we can truly know what exactly we are trying to develop and evaluate.

To address the lack of agreement concerning the definition of and relations between key XAI notions, we propose a two-dimensional explanation framework that establishes clear concept definitions and relationships between them. This framework is based on both XAI and social sciences literature, and focuses primarily on A/IS disclosing and clarifying causes underlying their behavior and reasoning to human teammates (i.e., goal-driven XAI). Our framework explicitly addresses the lack of consensus and ambiguity problem by establishing clear distinctions and relations between system *transparency*, *interpretability*, *explainability*, and *understandability*. More specifically, the framework discriminates between system *interpretability* and *understandability* as passive and subjective characteristics concerning user knowledge of the system, versus system *transparency* and *explainability* as active and objective characteristics involved with disclosing and clarifying relevant information. Ultimately, these definitions result in the classification of three types of AI-systems: *incomprehensible*, *interpretable*, and *understandable* systems. We argue *transparency* can make *incomprehensible* systems *interpretable*, and *explainability* can make *interpretable* systems *understandable*. Adopting our distinctive concept definitions and mutual relationships can benefit XAI community by clarifying what kind of systems can be developed, and how we can evaluate them.

The remainder of the chapter is structured as follows. In Sect. 2.2 we demonstrate the terminology problem by providing an overview of literature defining the key concepts. Next, we present our two-dimensional framework in Sect. 2.3. In Sect. 2.4 we discuss how the framework can be used to guide future XAI research, be applied in practice, and other relevant future directions. Finally, we conclude our chapter in Sect. 2.5.

2.2 Background

Several works introduced or defined key XAI concepts such as *interpretability*, *explainability*, *transparency*, and *understandability*. However, the lack of consensus on the exact meanings and relations between these notions remains a prevalent issue. This section aims to highlight the problem and discuss relevant and significant prior contributions, before proposing our framework attempting to establish clear distinctions and relations between the concepts. First, we demonstrate the lack of consensus problem and ambiguity of several proposed definitions. Next, we discuss some definitions, distinctions, and classifications that

Table 2.1: Several definitions for key XAI concepts, illustrating their ambiguity and relatedness.

Concept	**Definition**
Explainability	How well humans can understand AI-system decisions [145, 235].
Interpretability	To explain or present in understandable terms to humans [14, 56]. How well humans can understand AI-system decisions [145, 235].
Transparency	Representing system states in a way that is open to scrutiny, analysis, interpretation, and understanding by humans [4]. Characteristic of model to be understandable for humans [14]. Capacity of method to explain how a system works, even when behaving unexpectedly [235].
Understandability	To make a human understand how a model works, without any need for explaining its internal structure [14]. Measuring how well humans understand model decisions [14]. Capacity of a method of explainability to make a model understandable by end users [235].

influenced our work. Finally, we discuss a framework that might help to unambiguously define and relate XAI concepts.

2.2.1 Problem

Unambiguously defining and relating XAI concepts is challenging. A small survey of available definitions in the literature demonstrates it is particularly hard to do so without recourse to related concepts (Table 2.1). Table 2.1 clearly demonstrates the ambiguity and relatedness of the defined concepts, and fails to provide any clear distinctions between them. For example, all of these concepts are defined at least once as *how understandable the AI-system is to humans.*

2.2.2 Transparency

Turilli and Floridi [205] introduce a clear definition for *transparency* which influenced our work. They suggest *transparency* refers to forms of *information visibility* and "the possibility of *accessing* information, intentions, or behaviors that have been intentionally revealed through a process of *disclosure*". This disclosed information (i.e., made explicit and openly available) can then be exploited by potential users to support their own decision-making process.

Despite considering *transparency* and *explainability* as synonyms, Walmsley's [239] discussion of *transparency* influenced our work. Walmsley [239] divides the notion of *transparency* into two major categories: outward - vs. functional *transparency*. Outward *transparency* concerns the relationship between the AI-system and externals, such as developers and users. This includes *transparency* about development reasons, design choices, values driving the system developers, and capabilities and limitations of the system. In contrast, functional *transparency* concerns the inner workings of the system. This includes *transparency* about how and why the system behaves in general (type functional *trans-*

parency[1]), or came up with certain decisions or actions (token functional *transparency*[2]).

2.2.3 Related Work

Ciatto et al. [41] propose an abstract and formal framework for XAI that, in contrast to most work, introduces a clear distinction between *interpretation* and *explanation*. The framework stresses the objective nature of *explanation*, in contrast with the subjective nature of *interpretation*. The act of *interpreting* some object *X* is defined as "the activity performed by an agent *A* assigning a subjective meaning to *X*". Furthermore, Ciatto et al. [41] argue "an object *X* is interpretable for an agent *A* if it is easy for *A* to assign a subjective meaning to *X*" (i.e., *A* requires little computational or cognitive effort to *understand X*). The authors emphasize the subjective nature of *interpretations*, as agents assign them to objects based on their background knowledge and State of Mind.

In contrast, *explaining* is defined as the epistemic and computational activity of producing a more *interpretable* object *X'* out of a less interpretable one *X*, performed by agent *A*. They argue this activity can be considered objective because it does not depend on the agent's perceptions and State of Mind. Consider, for example, decision tree extraction (the *explaining* activity) from a neural network (object *X*) to produce a decision tree (the *explanation*/object *X'*). In the end, the effectiveness of the explanations always remains a subjective aspect.

This framework differs from ours in a few ways. In particular, Ciatto et al. [41] provide a formal framework focused on data-driven XAI, whereas we provide more general definitions in a goal-driven XAI context. In contrast, the intentions of the paper and provided definitions are similar to our work. We also define *interpretability* as a subjective system characteristic reflecting user knowledge about a system, and *explainability* as an epistemic and computational activity aimed at increasing user knowledge about the system.

Barredo Arrieta et al. [14] provide a brief clarification of the distinctions and similarities between *transparency*, *interpretability*, *explainability*, and *understandability*. So this part of their work is very similar in its intents to our work, despite focusing on data-driven XAI instead of goal-driven XAI. However, we argue that their attempt at clarifying the distinctions and similarities between the concepts fails to resolve any ambiguity. For example, the authors first argue *interpretability* is a passive model characteristic referring to the level at which a given model makes sense for a human, but later as the ability to explain or provide the meaning in *understandable* terms to a human.

In summary, Barredo Arrieta et al. [14] define *interpretability* (i.e., their first definition), *understandability*, and *transparency* as passive model characteristics reflecting human knowledge and understanding of a model. In contrast, they define *explainability* as an active model characteristic, denoting any action taken by a model with the intent of clarifying or detailing its internal functions. Unlike Barredo Arrieta et al. [14], we consider *transparency* as an active system characteristic concerned with disclosing information to generate knowledge about system elements. Similar to them, we also define *interpretability* and *understandability* as passive characteristics reflecting system knowledge and understanding, and *explainability* as actively clarifying or detailing system elements.

[1] Also referred to as global explanations in XAI literature.

[2] Also referred to as local explanations in XAI literature.

Rosenfeld and Richardson [170] formally define *explainability* and its relationship to *interpretability* and *transparency*, in the case of a ML-based classification algorithm. The authors define *explainability* as the ability for the human user to *understand* the algorithm's logic. This ability to *understand* is achieved from the *explanation*, which they define as the human-centric objective for the user to *understand* the algorithm, using an *interpretation. Interpretation/interpretability* is defined as a function mapping data, data schemes, outputs, and algorithms to some representation of the algorithm's internal logic. Furthermore, the authors argue an *interpretation* is *transparent* when the connection between the *interpretation* and algorithm is *understandable* to the human, and when the logic within the *interpretation* is similar to that of the algorithm.

All in all, the work of Rosenfeld and Richardson [170] differs from our work in several ways. First of all, they focus on data-driven XAI and provide formal definitions, whereas our work focuses on goal-driven XAI and provides more general definitions. More importantly, the provided definitions differ from our view. Rosenfeld and Richardson [170] consider *explainability* as passive and subjective, defining it as the ability to *understand.* In contrast, we consider *explainability* as an active system characteristic, and argue their definition of *explainability* reflects *understandability* instead. In addition, the authors consider *interpretability* as active and objective, defining it as providing representations of an algorithm's internal logic. However, we consider *interpretability* as passive and subjective, reflecting user knowledge and understanding of a system/algorithm, and argue their definition of *interpretability* reflects *explainability* instead.

Sanneman and Shah [176] propose an interesting situation awareness-based levels of XAI framework. This framework argues AI-systems part of human-AI teams should explain what the system did or decided (XAI for Perception), why the system did this (XAI for Comprehension), and what the system might do next (XAI for Projection). The authors argue XAI for Comprehension should provide information about causality in the system, aimed at supporting user comprehension of the system's behavior. Examples include explanations linking behavior to the system's goals, constraints, or rules.

This framework broadly aligns with ours, but includes a few differences as well. First of all, we agree with their distinction between providing information for perception and comprehension. However, whereas Sanneman and Shah [176] define both of them as explanations, we refer to XAI for Perception as *transparency*/disclosing information, and XAI for Comprehension as *explainability*/clarifying disclosed information. We argue XAI for Projection can be defined as both *transparency* and *explainability*, depending on whether the system discloses next actions (i.e., *transparency*) or also clarifies them (i.e., *explainability*). Furthermore, the framework only focuses on *explaining* AI-system behavior like actions or decisions. However, we argue it is also possible and sometimes even necessary to explain system elements like goals, knowledge, development reasons, or design choices. By doing so, human users can build more complete mental models of the AI-system. Therefore, our framework also incorporates disclosing and clarifying other relevant system elements like goals or knowledge.

Doran et al. [55] introduce an interesting distinction between *opaque*, *interpretable*, and *comprehensible* AI-systems that influenced our work. They define *opaque* AI-systems as systems where the mechanisms mapping inputs to outputs are invisible to users. Consequently, the reasoning of the system is not observable or understandable for users. In contrast, *interpretable* AI-systems are characterized as systems where users cannot only *see*, but also *study* and *understand* how inputs are mapped to outputs. The authors argue that *interpretable* systems imply *transparency* about the underlying system mechanisms. Finally, they define *comprehensible* AI-systems as systems emitting symbols (e.g., words or visualizations) along with their output to allow users to *relate* properties of the input to their corresponding output. According to this classification, *interpretable* systems can be inspected to be understood (i.e., letting users draw *explanations* by themselves), while *comprehensible* systems explicitly provide a symbolic *explanation* of their functioning [41].

This classification of AI-systems is quite similar to the one provided in our work. However, whereas Doran et al. [55] focus on data-driven XAI and argue the notions of *interpretation* and *comprehension* are separate, we focus on goal-driven XAI and argue *understanding/comprehension* implies *interpretation.* More specifically, we claim *transparency* can make *incomprehensible* systems *interpretable*, and *explainability* can make these *interpretable* systems *understandable.* We will explain our definitions, relationships, and classification in detail in the next section.

2.3 A Two-Dimensional Framework to Classify AI

In this section we present and discuss our two-dimensional explanation framework providing clear distinctions and relations between key XAI concepts (Fig. 2.1). In short, our framework makes a distinction between *incomprehensible*, *interpretable*, and *understandable* AI-systems, and argues system *transparency* can make *incomprehensible* systems *interpretable*, whereas *explainability* can make *interpretable* systems *understandable.* In the following sections, we will explain and illustrate our framework by introducing our definitions of the concepts *transparency* and *explainability* (Sect. 2.3.1), and *interpretability* and *understandability* (Sect. 2.3.2). After that, we illustrate and discuss our framework based on the example of a search and rescue human-agent teaming scenario where a human collaborates with a goal-driven A/IS (Sect. 2.3.3). Finally, we extend our framework to include some other relevant factors enabled by system *transparency* and *explainability* in Sect. 2.3.4.

2.3.1 Transparency vs. Explainability

Whereas most prior work strongly ties or even equates system *explainability* to *interpretability* (e.g., [145, 235]), we consider them fundamentally different. Instead, we strongly tie system *transparency* to *explainability.* However, we also argue for a major distinction between these two notions. Inspired by [4] and [205], we define system *transparency* as "*disclosing* the relevant outward and functional system elements to users, enabling them to access, analyze, and exploit this disclosed information". Here, functional system elements concern elements like goals, knowledge, beliefs, decisions, and actions. In contrast, outward elements concern aspects like development reasons, intended users, and design choices.

System *transparency* can answer "*what*"-questions [145] requiring descriptive answers concerning the system elements. Consider, for example, a goal-driven autonomous and intelligent agent collaborating with a human teammate to save victims after an earthquake. According to our definition, system *transparency* is both an *active* [14] and *objective* [41] system characteristic achieved by, for example, disclosing the goal to save all injured chil-

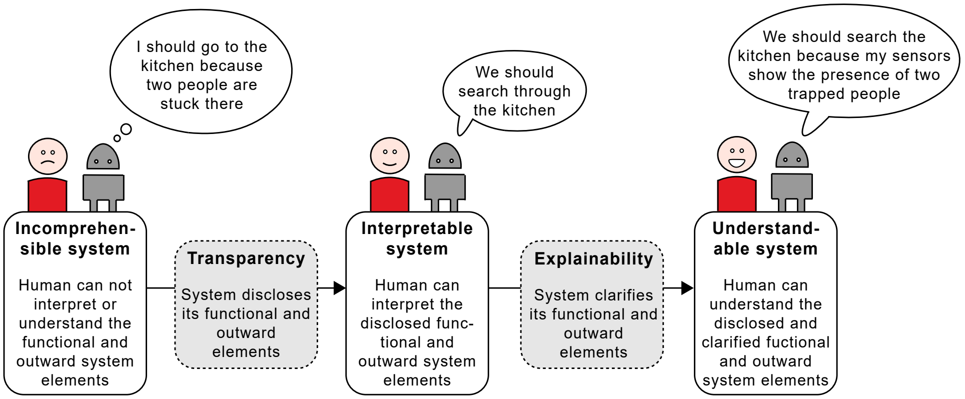

I should go to the kitchen because two people are stuck there
We should search through the kitchen
We should search the kitchen because my sensors show the presence of two trapped people
Incomprehensible system
Human can not interpret or understand the functional and outward system elements
Transparency
System discloses its functional and outward elements
Interpretable system
Human can interpret the disclosed functional and outward system elements
Explainability
System clarifies its functional and outward elements
Understandable system
Human can understand the disclosed and clarified fuctional and outward system elements

Figure 2.1: Two-dimensional explanation framework providing distinctive definitions and relationships between key XAI concepts.

dren first by collaborating with trained firefighters. By doing so, the human teammate can gain knowledge about these system elements (here a goal and intended users respectively), without necessarily always knowing the relations between them.

The disclosure of relevant elements can be considered *active* in the sense that it is an epistemic and computational *activity* aimed at increasing user knowledge, and *objective* because this activity itself does *not depend on the human's perceptions or State of Mind*. Put differently, the computational implementation of *transparency* is independent of the human user's perceptions and State of Mind, and thus reproducible in principle [41]. However, the exact effectiveness and content of the disclosed information is a subjective aspect, reflected by measures of *interpretability* and *understandability*.

Inspired by [14], [41], and [176] we define system *explainability* as "*clarifying* disclosed system elements by providing information about causality and establishing relations with other system elements, making it easier for users to *understand*, analyze, and exploit this information". *Explainability* can answer "*how*"- and "*why*"-questions [145] requiring clarifying answers concerning the system elements and how they relate and depend on each other. For example, system *explainability* can involve clarifying the disclosed goal to save all children first by linking it to the norm that children are most vulnerable, or that it will not give safety instructions because it assumes the user is a firefighter and familiar with these. Just as *transparency*, we characterize system *explainability* as an *active* [14] and *objective* [41] system characteristic aimed at increasing user knowledge and where the epistemic and computational activity itself does not depend on the human's perceptions or State of Mind.

In summary, the main difference between system *transparency* and *explainability* boils down to *disclosing* vs. *clarifying*. *Transparency* aims to provide descriptive answers providing knowledge about system elements. In contrast, *explainability* aims to ease understanding by clarifying the relations between system elements. Both are considered *active* and *objective* system characteristics, since they are epistemic and computational activities aimed at increasing user knowledge without depending on user's perceptions or

State of Mind. We define *transparency* and *explainability* from a system-centric point of view as methods for sharing information, hence the categorization as active and objective/ independent from the user. However, we argue that the subjective aspect concerning the effectiveness and content of the shared information also plays a crucial role, as reflected by measures of *interpretability* and *understandability*.

2.3.2 Interpretability vs. Understandability

In contrast to *transparency* and *explainability*, we define system *interpretability* and *understandability* as *passive* and *subjective* characteristics reflecting user knowledge of the system and depending on the user's State of Mind and background knowledge. In addition, we argue *transparency* makes system *interpretable*, whereas *explainability* makes *interpretable* systems *understandable*. Although we strongly tie *interpretability* to *understandability*, we argue for a major distinction between these two notions as well.

Inspired by [14], [41], [55], and [205], we define system *interpretability* as "the level at which the system's users can assign subjective meanings, draw explanations, and gain knowledge by accessing, analyzing, and exploiting disclosed outward and functional system elements". Our definition implies *interpretability* is both a *passive* [14] and *subjective* [41] system characteristic. *Passive* in the sense that *interpretability* reflects a degree of user knowledge about system elements, opposite to actively sharing information to generate knowledge (i.e., *transparency*). Furthermore, *interpretability* can be considered *subjective* in the sense that it is highly dependent on the user's State of Mind and background knowledge [41].

Consider, again, the example of the goal-driven A/IS collaborating with a human to save victims after an earthquake. Disclosing its goal to save all children first enables human users to gain knowledge and assign subjective meanings or draw explanations by themselves (i.e., *interpret*). However, without clarifying the disclosed goal and relating it to other system elements (i.e., *explainability*), these interpretations can vary considerably. For example, the human could draw the conclusion that the system knows/beliefs the area contains a lot of children but only few elderly or adults.

On the other hand, we define system *understandability* as "the level at which the system's users have knowledge of disclosed and clarified outward and functional system elements, and the relationships and dependencies between them". *Understandability* involves knowing *how* and *why* the system reasons and functions, based on *explanations* clarifying and relating disclosed system elements. For example, clarifying the goal to save all children first because they are most vulnerable provides the user with knowledge about the relationship between the goal and a specific norm.

In summary, the main difference between system *interpretability* and *understandability* boils down to a difference in cognitive effort required to have knowledge of the system elements [41]. More specifically, we argue *interpretability* requires more cognitive effort because it implies inferring the meaning of and relations between disclosed information without explicit knowledge of this meaning and relations themselves. In contrast, *understandability* requires less cognitive effort because it implies knowing the meaning of and relations between disclosed and clarified information (facilitated by *explanations*). Both are considered *passive* [14] and *subjective* [41] system characteristics, since they reflect a degree of *user knowledge* about the system *depending on the user's State of Mind and background* *knowledge*. So we define *interpretability* and *understandability* from a user-centric point of view reflecting the subjective effectiveness of the *transparency* and *explainability* content. Here, *transparency* and *explainability* will be most effective when their content is tailored to the user's State of Mind and background knowledge.

2.3.3 Two-Dimensional Framework to Classify AI

Our framework (Fig. 2.1) distinguishes between three types of AI-systems (*incomprehensible*, *interpretable* and *understandable*) and establishes relations between them by integrating the defined concepts of Sect. 2.3.1 and Sect. 2.3.2. We will illustrate our framework in the context of a search and rescue human-agent teaming scenario, where a human collaborates with a goal-driven A/IS.

When collaborating with *incomprehensible* systems, humans can not *interpret* or *understand* the system elements because they are not disclosed and clarified. For example, without disclosing and clarifying its decision to search through the kitchen because it perceived stuck people, a human will not be able to interpret or understand the system's behavior. Our framework argues *transparency* can turn *incomprehensible* systems into *interpretable* ones. By disclosing its relevant functional and outward system elements (i.e., *transparency*), the human can access and exploit this information to assign subjective meanings and gain knowledge (i.e., *interpret*). Consider, for example, an A/IS disclosing the decision to search through the kitchen of a collapsed house to its human teammate. By doing so, the human can utilize this information to interpret that the A/IS perceived something urgent in the kitchen. Furthermore, we argue *explainability* can turn *interpretable* systems into *understandable* systems. By clarifying the disclosed system elements and relations between them (i.e., *explainability*), the human can more easily exploit this information to gain knowledge and build a mental model of the system (i.e., *understandability*). Consider, for example, an A/IS disclosing the decision to search through the kitchen, because it perceived two trapped children there. By providing a belief-based *explanation* for the decision, the system clarifies this decision and how it relates to other system elements like perceptions.

Our proposed framework has several implications. First of all, pursuing system *understandability* should be the ultimate goal, since it can improve collaboration and team performance in human-agent teams [9]. Furthermore, the framework implies that system *transparency* and *explainability* are *active* and *objective* characteristics which can be manipulated by designers to bring about the desired effects. In contrast, system *interpretability* and *understandability* are considered *passive* and *subjective* characteristics which can be measured to validate the effects of *transparency* and *explainability.*

2.3.4 Extended Framework

We extend our two-dimensional framework to include several often encountered XAI notions. This framework (Fig. 2.2) mainly illustrates the opportunities system *transparency* and *explainability* can provide to human teammates. Again, we discuss the framework in the context of a search and rescue human-agent teaming scenario where a human collaborates with a goal-driven A/IS.

The extended framework argues that when a system is *interpretable*, it is already both *controllable* and *directable.* Here, we define system *controllability* as "the extent to which

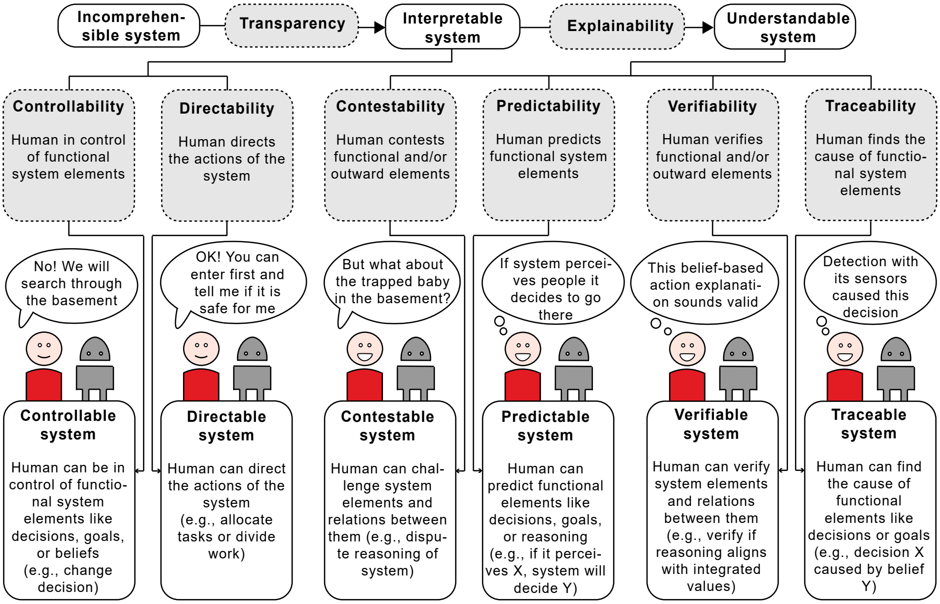

Incomprehensible system
Transparency
Interpretable system
Explainability
Understandable system
Controllability
Human in control of functional system elements
Directability
Human directs the actions of the system
Contestability
Human contests functional and/or outward elements
Predictability
Human predicts functional system elements
Verifiability
Human verifies functional and/or outward elements
Traceability
Human finds the cause of functional system elements
No! We will search through the basement
OK! You can enter first and tell me if it is safe for me
But what about the trapped baby in the basement?
If system perceives people it decides to go there
This belief-based action explanation sounds valid
Detection with its sensors caused this decision
Controllable system
Human can be in control of functional system elements like decisions, goals, or beliefs (e.g., change decision)
Directable system
Human can direct the actions of the system (e.g., allocate tasks or divide work)
Contestable system
Human can challenge system elements and relations between them (e.g., dispute reasoning of system)
Predictable system
Human can predict functional elements like decisions, goals, or reasoning (e.g., if it perceives X, system will decide Y)
Verifiable system
Human can verify system elements and relations between them (e.g., verify if reasoning aligns with integrated values)
Traceable system
Human can find the cause of functional elements like decisions or goals (e.g., decision X caused by belief Y)

Figure 2.2: Extended two-dimensional explanation framework providing distinctive definitions and relationships between key XAI concepts.

human users can change or overrule functional system elements". For example, when the A/IS discloses the decision to search through the kitchen, its human teammate can overrule this decision by changing it to searching the basement instead (i.e., the system is *controllable*).

Next, we define system *directability* as “the extent to which human users can guide the actions of the system". This is different from system *controllability* in the sense that *directability* does not involve changing or overruling system elements, but rather accepting them and guiding the corresponding actions or dividing the work. For example, the human teammate could also accept the disclosed decision to search the kitchen but direct the action of the A/IS by giving the order to enter the kitchen first to assess its safety (i.e., the system is *directable*). Even though system *interpretability* already enables system *controllability* and *directability*, we argue system *understandability* will further improve these two characteristics. For example, when the human teammate has more knowledge of the system, it can more effectively control and direct its functional elements such as actions or goals.

Furthermore, we argue that system *understandability* enables several other important notions such as system *contestability*, *predictability*, *verifiability*, and *traceability*. We define system *contestability* as “the extent to which human users can challenge or dispute system elements and the relations between them". Again, consider the example of the A/IS disclosing the decision to search through the kitchen, by clarifying it perceived two trapped people there. By doing so, the human teammate can contest this decision and dispute the underlying reason, for example by asking why they should search through the kitchen when there is a trapped baby in another room (i.e., the system is *contestable*).

We argue system *understandability* also enables system *predictability*. We define system *predictability* as “the extent to which human users can estimate future or other functional system elements". Consider the example of a system disclosing its goal to save all children first because of the norm that children are more vulnerable than adults. The human could use this explanation to predict that the agent’s next actions will be focused on searching children rather than adults.

The extended framework also argues system *understandability* enables system *verifiability*. Here, we define system *verifiability* as “the extent to which human users can check that the system elements and relations between them make sense and sound valid". We do not refer to formal verification of systems using formal methods involving mathematical models of systems and analyzing them using proof-based methods. Rather, we refer to a more informal verification of the plausibility of system elements and relations between them. Again, consider the example of a system disclosing its goal to save all children first because of the norm that children are more vulnerable than adults. Based on the provided explanation the human could informally verify that the reasoning aligns with the decision and sounds valid (i.e., the system is *verifiable*).

Finally, we argue system *understandability* enables system *traceability*. We define system *traceability* as “the extent to which human users are able to find the cause of functional system elements like decisions, goals, or beliefs". Again, consider the example of the A/IS disclosing the decision to search through the kitchen, by clarifying it perceived two trapped people there. The human teammate could use the provided explanation to infer that the decision to search the kitchen was caused by the detection of two trapped people.

In summary, the extended framework argues system *interpretability* and *understandability* enable important factors such as system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability*. These factors are crucial for and determine the success of human-agent teams [99, 101, 107, 175]. Therefore, pursuing system *understandability* should be the main goal when developing AI-systems part of human-agent teams.

2.4 Discussion

In this work we have presented a two-dimensional explanation framework providing distinctive XAI concept definitions and relationships between them. In this section, we will discuss how our presented relationships between the concepts can be used to guide future research into these relationships. Additionally, we will describe how we believe this framework can be applied in practice.

2.4.1 Evaluation of Main Framework

Several assumptions arise from the proposed relationships in our presented framework. Below, we introduce these assumptions as claims and describe their corresponding requirements. Next, we discuss whether these assumptions can be evaluated, and how they offer a road map for future research.

- **Claim 1** - System *explainability* results in more knowledge/complete mental models

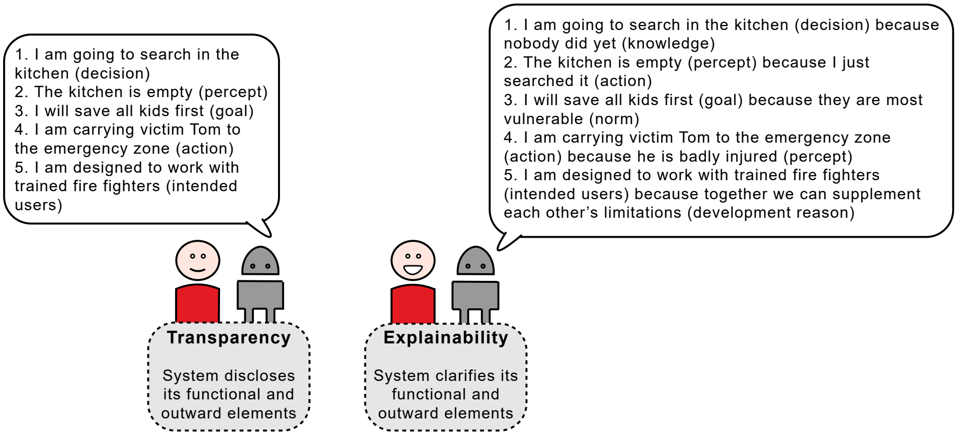

1. I am going to search in the kitchen (decision)
2. The kitchen is empty (percept)
3. I will save all kids first (goal)
4. I am carrying victim Tom to the emergency zone (action)
5. I am designed to work with trained fire fighters (intended users)
1. I am going to search in the kitchen (decision) because nobody did yet (knowledge)
2. The kitchen is empty (percept) because I just searched it (action)
3. I will save all kids first (goal) because they are most vulnerable (norm)
4. I am carrying victim Tom to the emergency zone (action) because he is badly injured (percept)
5. I am designed to work with trained fire fighters (intended users) because together we can supplement each other's limitations (development reason)
Transparency
System discloses its functional and outward elements
Explainability
System clarifies its functional and outward elements

Figure 2.3: Examples of system transparency and explainability in the context of a (simulated) search and rescue mission.

of the system than *transparency*

Requirement 1 - Manipulating/implementing system *transparency* and *explainability*

Requirement 2 - Measuring user knowledge of a system

- **Claim 2** - Increased user knowledge of a system results in improved human-agent collaboration and eventually team performance

  Requirement 1 - Subjective and objective measurements of human-agent collaboration

We will illustrate how these claims can be evaluated using the example of a simulated search and rescue mission where a human operator and self-explaining A/IS collaborate to search and rescue victims. To validate Claim 1, implementing system *transparency* and *explainability* would be required. Examples of implementing system *transparency* involve disclosure of the system's goals, decisions, and intended users. Fig. 2.3 shows several examples of system *transparency* in the context of the search and rescue mission.

Implementing system *explainability* can be achieved in many different ways. However, a fundamental requirement is providing information about causality in the system and establishing relations between system elements. Existing approaches from the XAI literature include explanations of actions based on state information [5, 89, 128]; explanations of actions based on goals [22, 84, 86]; explanations of decisions based on demonstrating that alternative decisions would be sub-optimal [193]; and sequence-based explanations clarifying the next action(s) [22, 86]. Fig. 2.3 shows several concrete examples of system *explainability* in the context of the search and rescue operation.

Validating Claim 1 would also require the measurement of human user knowledge and understanding of the system, which can be done both subjectively and objectively.

Subjective examples from the XAI literature include asking questions related to perceived understandability of the system and its model [91], and asking users to choose which of two possible system outputs is of higher quality (implicitly measures understanding) [56]. However, objectively measuring user knowledge and understanding of the system would be a more robust indicator than the subjective alternatives.

Currently, objective methods and metrics for measuring user knowledge and understanding of systems are lacking. Nevertheless, Sanneman and Shah [176] propose a relevant method based on the widely-used and empirically validated Situation Awareness Global Assessment Technique (SAGAT) [60, 63]. In short, their proposed technique involves freezing simulations of representative tasks at random time points, followed by asking questions measuring user knowledge about information related to system behavior. It is crucial to first define the human informational needs related to system behavior. Accordingly, a list of questions regarding the informational needs can be specified and used to measure user knowledge of the system.

Whereas Sanneman and Shah [176] focus solely on measuring user knowledge related to AI-system behavior, the test/technique can also be extended to include information related to other relevant system elements like goals, knowledge, decisions, or even development reasons. Some example questions based on the information in Fig. 2.3 include “Which room will the agent search next?"; “What is the current action of the agent?"; “Why is the agent going to search in the kitchen?""; and “Why will the agent save all kids first?".

Validating Claim 2 would require the subjective and objective measurement of human-agent collaboration and team performance. Subjective measures could include user satisfaction [35] or system usability [23], whereas objective measures could include aspects like the number of victims rescued or seconds required to finish tasks. The outlined example experiment, discussed example implementations of *transparency* and *explainability*, and suggested metrics for measuring user knowledge, human-agent collaboration, and team performance can be used as a road map for future work aimed at validating the assumptions arising from our framework.

2.4.2 Evaluation of Extended Framework

Several assumptions arise from the proposed relationships in our extended framework as well. Below, we introduce these assumptions as claims and describe their corresponding requirements. Next, we discuss whether these assumptions can be evaluated, and how they offer a road map for future research.

- **Claim 3** - System *transparency* already enables system *controllability* and *directability*, but not system *contestability*, *predictability*, *verifiability*, and *traceability*

  Requirement 1 - Implementing system *transparency*

  Requirement 2 - Measuring system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability*

- **Claim 4** - System *explainability* enables system *contestability*, *predictability*, *verifiability*, and *traceability*

  Requirement 1 - Implementing system *explainability*

Table 2.2: Example questions for subjectively measuring the system variables in the extended framework.

Variable	Example Question
Controllability	“I feel like I can change the system’s decision"
Directability	“I feel like I can guide the system’s behavior"
Contestability	“I feel like I can challenge the system’s decision"
Predictability	“I feel like I can predict the system’s next action"
Verifiability	“I feel like I can check that the system’s behavior makes sense"
Traceability	“I feel like I can find the cause of the system’s decision"

> **Requirement 2** - Measuring system *contestability*, *predictability*, *verifiability*, and *traceability*

Validating Claims 3 and 4 would require implementing system *transparency* and *explainability*, and measuring system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability*. An example of subjectively measuring these system characteristics could be freezing the simulated experiment at random points, followed by measuring perceived system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability*. One approach involves Likert-scale questions[3] asked to the human users. Table 3.1 shows example questions for each of these variables, though full questionnaires would require more research and validation of the exact scales. The outlined example experiment, discussed example implementations of *transparency* and *explainability*, and suggested metrics for measuring system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability* can be used as a road map for future work aimed at validating the assumptions arising from our extended framework.

2.4.3 Application of Framework

Here we briefly address how our framework can be used in practice. Specifically, what difference can the framework make when developing systems part of human-agent teams? Consider the example of developing an autonomous and intelligent drone which should collaborate with a human operator (e.g., a firefighter) during the aftermath of an earthquake. The goal of the team is to search and rescue trapped victims as soon as possible. Our framework can be particularly helpful by mapping specific types of context and informational needs onto requiring either system *transparency* or *explainability*. For example, the drone can be developed/implemented in such a way that when the workload or time pressure is high, the drone displays *transparency* only. Similarly, contextual factors that could be mapped onto system *explainability* include low time pressure and operator workload, or when the user has an imprecise mental model of the system. In this way, the framework can contribute to developing adaptive systems able to tailor their communication of relevant information to the needs and requirements of both users and situations.

2.4.4 Future Work

Based on the work presented in this chapter, we identify a few key ideas for future work. A possible first direction could be to conduct experiments aimed at validating the assumptions

[3] For example ranging from “Totally Disagree" to “Totally Agree" on a 7-point scale.

arising from our framework. Some ideas, requirements, and examples concerning this validation have been discussed in more detail in Sect. 2.4.1 and Sect. 2.4.2.

For now, our framework focuses on sharing information regarding mental constructs like decisions or goals. A relevant suggestion for future work would be to extend the framework with a more physical domain as well by including literature/perspectives from explainable and understandable robots. For example, the role of visual and body cues could be incorporated in the framework. Furthermore, our provided framework is rather broad/general and informally defined. Therefore, another suggestion would be to formalize it and make it more concrete by providing examples in terms of different computational frameworks/architectures (e.g., transparency vs. explainability differences between agents using BDI vs. PDDL models). In addition, we currently do not consider situations where the user may be under the false impression of understanding the system, but only consider cases where their understanding actually matches the system's models/elements. We also do not consider different roles taken by human and agent, such as commander or supervisor. In future work, it would be interesting to extend the framework by including these two aspects, and see how it affects our proposed definitions.

Another future direction for this work would be to extend the framework to include context- and user-awareness required for tailoring system *transparency* and *explainability* to specific needs and requirements. The need for personalized and context-dependent system *transparency* and *explainability* is one of the main goals within XAI community and research [9]. However, the actual implementation and investigation is still somewhat in its infant stages. Currently, our proposed framework does not address context- and user-dependent system *transparency* and *explainability*, so this would be a relevant suggestion for future work. Ideas involve mapping specific types of context or user knowledge to requiring either system *transparency* or *explainability*. Furthermore, these aspects could also be mapped onto *transparency* and *explanation* modality/presentation instead of just content. Examples include mapping high workload to system *transparency*, rudimentary system knowledge to *explainability*, or visual thinkers to receiving visual *explanations* and verbal thinkers receiving textual *explanations*. Another idea involves adapting system *transparency* and *explainability* based on the interdependence relationship between human and system. For example, the system could adapt its communication based on whether joint activity is required (i.e., hard interdependence) or when joint activity is optional (i.e., soft interdependence) [98, 99].

2.5 Conclusion

This chapter answers the first research sub-question of the thesis: *How should we define and relate agent transparency, explainability, interpretability, and understandability*? It does so by proposing a two-dimensional explanation framework that introduces clear distinctions and relationships between the key XAI notions *transparency*, *interpretability*, *explainability*, and *understandability*. This concise and comprehensive framework explicitly addresses the lack of consensus and ambiguity problem surrounding these concepts. We argue that adopting our distinctive concept definitions and mutual relations can greatly benefit XAI community, as clearly defining concepts and relationships between them is a pre-requisite for both the implementation and evaluation of these concepts. Furthermore, the framework yields a classification of AI-systems as *incomprehensible*, *interpretable*, or *understandable*, guiding the research and development to establish understandable AI (e.g., by setting requirements for *contestability*, *predictability*, *verifiability* and *traceability*).

3

The Influence of Interdependence and a Transparent or Explainable Communication Style on Human-Robot Teamwork

Humans and robots are increasingly working together in human-robot teams. Teamwork requires communication, especially when interdependence between team members is high. In the previous chapter, we identified a conceptual difference between sharing what you are doing (i.e., being transparent) and why you are doing it (i.e., being explainable). Although the second might sound better, it is important to avoid information overload. Therefore, an online experiment (n = 72) was conducted to study the effect of communication style of a robot (silent, transparent, explainable, or adaptive based on time pressure and relevancy) on human-robot teamwork. We examined the effects of these communication styles on trust in the robot, workload during the task, situation awareness, reliance on the robot, human contribution during the task, human communication frequency, and team performance. Moreover, we included two levels of interdependence between human and robot (high vs. low), since mutual dependency might influence which communication style is best. Participants collaborated with a virtual robot during two simulated search and rescue tasks varying in their level of interdependence. Results confirm that in general robot communication results in more trust in and understanding of the robot, while showing no evidence of a higher workload when the robot communicates or adds explanations to being transparent. Providing explanations, however, did result in more reliance on RescueBot. Furthermore, compared to being silent, only being explainable results in a higher situation awareness when interdependence is high. Results

Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI* 9 (2022).

further show that being highly interdependent decreases trust, reliance, and team performance while increasing workload and situation awareness. High interdependence also increases human communication if the robot is not silent, human rescue contribution if the robot does not provide explanations, and the strength of the positive association between situation awareness and team performance. From these results, we can conclude that robot communication is crucial for human-robot teamwork, and that important differences exist between being transparent, explainable, or adaptive. Our findings also highlight the fundamental importance of interdependence in studies on explainability in robots.

3.1 Introduction

Increasingly, humans and robots will be working together in human-agent/robot teams (HARTs). Robots often outperform humans with respect to rapid, rational, and repetitive decision-making, thanks to their processing speed and memory capacity [216]. On the other hand, humans are usually still better at handling uncertainty and unexpected situations. HARTs make use of this unique combination of abilities.

HARTs can perform tasks where human and robot hardly depend on each other and can execute their individual actions independently [192]. However, HARTs can also engage in joint activities in which the human and robot mutually depend on each other over a sustained sequence of actions [99]. In such joint activities, the human and robot are interdependent and effective coordination and collaboration become crucial [99, 192].

Several factors are crucial when human and robot are interdependent, such as mutual trust and understanding; shared mental models; observability, predictability, and directability; and transparency and explainability [99, 101, 107, 175]. Unfortunately, many of these factors are still lacking in contemporary HARTs. For example, robots often demonstrate poor transparency and explainability, making it hard for human teammates to properly understand their inner workings, behavior, and decision-making [9, 126, 145]. This, in turn, negatively affects factors like mutual trust and understanding, eventually resulting in decreased global team performance [99, 101].

Explainable AI (XAI) research, methods, and techniques emerged as a means of making AI-systems more understandable to humans [80]. Unfortunately, the field of XAI is characterized by a plethora of related but often ill-defined and inter-changeably used concepts like transparency, interpretability, explainability, and understandability [231]. We addressed this issue by proposing a framework that unambiguously defined and related these concepts in a coherent and concise manner (Chapter 2 [231]). This framework makes a distinction between robot transparency and explainability as different communication styles, with the former referring to the disclosure of information and the latter to also clarifying disclosed information using explanations.

One of the main goals within XAI community and research is the development of personalized, context-dependent and adaptive robots [9, 50]. So instead of implementing robots characterized by fixed transparency or explainability, developing robots able to adapt their communication according to context and intended user. However, to do this we need to first understand how different communication styles like transparency and explainability exactly influence teamwork in different interdependency conditions. So far, very little work has examined the influence of interdependence between human and robot on human-robot teamwork outcomes, let alone the interaction between communication style and interdependence [157].

Therefore, this exploratory study will investigate the effects of different robot communication styles on crucial HART factors like team performance, trust, workload, situation awareness (SA) of the robot, and understanding. We will examine these effects across two levels of interdependence between the human and robot (high vs. low). To do this, we conducted a user study in a simulated environment where human participants collaborated with a virtual robot during a search and rescue task. The remainder of the chapter is structured as follows. In Sect. 3.2 we discuss the relevant literature related to our study. Next, in Sect. 3.3 we describe how we conducted the user study, followed by the results in Sect. 3.4. Finally, we present a discussion and conclude our work in Sect. 3.5.

3.2 Background & Related Work

3.2.1 Interdependence

Interdependence in a team can be due to the relationships between team members and the task to execute. Four types of task interdependence have been identified: pooled, sequential, reciprocal, and team [192]. Pooled task interdependence concerns the execution of tasks independently and without any interaction, whereas in sequential task interdependence, tasks are executed in a sequential order and team members have to wait for previous team members to complete their task. Reciprocal task interdependence involves team members taking turns in completing part of the task, while in team task interdependence, team members execute their individual tasks concurrently and may execute joint actions [99, 192]. These task interdependence types form a hierarchy representing increasing needs for coordination and levels of dependence between team members. However, these task interdependence types are unable to capture the nuances of close collaboration between humans and robots working jointly on a task [99]. To capture these nuances of close collaboration, interdependence relationships between team members are required.

Johnson et al. [99] define these interdependence relationships as the set of complementary relationships that human and robot rely on to manage required (hard) and opportunistic (soft) dependencies in joint activity. Their definition highlights the importance of dependencies and joint activity in interdependence between humans and robots working as team members. Joint activity is closely related to team task interdependence and relates to situations in which what the human does depends on what the robot does (and vice-versa) over a sustained sequence of actions [99]. For example, when a human-robot team engages in an urban search and rescue task. Such joint activity gives rise to required (hard) and opportunistic (soft) dependencies/interdependence relationships between team members. Hard interdependence stems from a lack of capacity (e.g., knowledge, skills, abilities, and resources) required to competently perform an activity individually [99]. For example, an explore robot/drone during urban search and rescue lacking the capacity to transport victims. In contrast, soft interdependence is optional and opportunistic, arising from recognizing opportunities to be more effective and efficient by working jointly [99].

These different types of task interdependence and interdependence relationships (and their combinations) can give rise to high and low interdependence scenarios for human-robot teams. For example, when a human-robot urban search and rescue team allocates the task of exploring the disaster site to the robot and executing rescue operations to the human, human and robot may hardly depend on each other and execute their individual actions independently without much interaction (i.e., low interdependence). In contrast, human-robot teams can also engage in joint activities and actions in which both parties are mutually dependent on each other and where the human might need to support the robot (and/or vice-versa) for specific activities (i.e., high interdependence). For example, the same team can also explore a collapsed building together where both parties need to know which team member assessed which room, or where in case a victim is detected by the robot, the human needs to provide support with assessing the victim's health status. So far, little work has been conducted on the effects of varying interdependence levels between human and robot on human-robot teamwork outcomes, and even less on the interaction between robot communication styles like transparency and explainability and different interdependence levels.

3.2.2 Robot Communication & Human-Agent/Robot Teamwork

Several studies did investigate the effects of XAI on relevant human-agent/robot teamwork (HART) factors like trust, workload, and operator performance [34, 142, 187, 251]. These studies largely agree that operator performance and trust in the XAI system increase when it shares more reasoning information, and without detriment to workload [34, 142, 187]. None of the studies, however, investigated the effects of communication style on global team performance, or how interdependence between human and robot affects trust, workload, and performance. Moreover, in all studies the XAI system served as an assistant of the human participants rather than as an equal team member. Our study will fill these gaps by examining the effects of robot communication style on global team performance, in HARTs where the robot is an equal team member, and across different levels of interdependence.

Other works investigated the relationship between robot information sharing and team performance, using a testbed (Blocks Worlds for Teams) similar to the one used in our study [83, 85, 123, 192, 244]. The Blocks World for Teams (BW4T) task is to deliver a sequence of coloured blocks in a particular order while working together in a team. The task is executed in a virtual environment containing rooms in which blocks are hidden, and a drop zone where blocks can be delivered. These studies have reported mixed results across different conditions. For example, most of them investigated artificial agent teams rather than human-agent/robot teams [85, 123, 192, 244]. In addition, almost all examined the influence of shared mental model components (goals vs. world knowledge) on performance, rather than providing more or less reasoning information [85, 123, 192, 244]. For example, Li, Sun, and Miller showed that in a high interdependence scenario containing joint actions, sharing goals was more effective than when the agent shared both goals and world knowledge with the human [123]. Harbers et al. did examine the effects of agents explaining their behavior on human-agent/robot teamwork [83]. Their results showed that explanations about agent behavior did not always lead to better team performance, but did impact user experience in a positive way.

None of the studies in the BW4T testbed examined human-agent/robot teams across different levels of interdependence between human and agent, a gap that our study will fill. Furthermore, most of the discussed studies examined the influence of shared mental components on team performance. In this study we are not interested in this distinction, since we believe both goals and world knowledge are crucial for carrying out the task most efficiently. Instead, we will investigate how different communication styles affect human-agent/robot teamwork across two levels of interdependence. These different communication styles give rise to more and less detailed mental models of the agent, rather than omitting crucial components like world knowledge or goals.

3.2.3 Robot Transparency vs. Explainability

In our previous work we proposed a framework that makes a distinction between the communication styles transparency and explainability [231]. This framework addresses the lack of agreement regarding the definitions of and relations between the key XAI concepts of transparency, interpretability, explainability, and understandability. More specifically, the framework discriminates between robot interpretability and understandability as passive and subjective system characteristics concerning user knowledge of the robot. In contrast, we defined robot transparency and explainability as active and objective characteristics involved with disclosing and clarifying relevant information. Transparency was defined as the disclosure of relevant system elements to users (e.g., robot decisions or actions), enabling users to access, analyze, and exploit this disclosed information (i.e., interpret). In contrast, we defined explainability as the clarification of disclosed system elements by providing information about causality and establishing relations with other system elements. Ultimately, these definitions resulted in the classification of three types of robots: incomprehensible, interpretable, and understandable. We argued transparency can make incomprehensible robots interpretable, and explainability can make interpretable robots understandable.

The third type of communication style not included in this framework but under investigation in this study, is adaptive communication. Miller discussed several factors to consider for such adaptive communication, such as epistemic relevance with respect to the user's mental model, or what has been explained already [145]. In addition to these user factors, a relevant contextual factor to consider is time pressure. Time pressure can decrease thorough and systematic processing of information while increasing selectivity of information processing, reducing both performance and decision-making quality [137, 184]. Therefore, reducing communication frequency of an agent when time pressure is high seems beneficial to HART performance and trust, for example by only communicating the most important information. Unfortunately, the current implementation and experimental investigation of adaptive robot communication is limited, even more in the context of human-robot teamwork [9, 50]. Our implemented adaptive style adjusts its communication based on both relevancy and time pressure.

3.2.4 Evaluating Robot Communication in Human-Agent/Robot Teamwork

The discussed studies showed several ways of evaluating XAI efficiency in a HART context, such as operator - and team performance, trust in the XAI system/robot, and workload during the task. Despite these different metrics, there is still a need for new metrics to assess XAI efficiency, specifically objective ones [9, 176]. Sanneman and Shah proposed an objective metric for assessing XAI effectiveness: the (modified) Situation Awareness Global Assessment Technique to measure SA of the XAI system's behavior processes and decisions [60, 176], which we will adopt. First, the situational and informational requirements related to AI behavior should be defined and identified, for example, using Goal Directed Task Analysis [62]. Next, simulations of representative tasks should be frozen at randomly selected times, followed by evaluating user knowledge of these predefined informational needs. The answers to these questions can be compared against the ground truth state of the world, providing an objective measure of the user's SA of the AI.

3.3 Method

To test the effects of different robot communication styles on human-robot teamwork across different levels of interdependence between human and robot, an experiment was conducted. In this experiment, we aimed to investigate the effect of four different robot communication styles on trust, reliance, workload, situation awareness, team performance, human contribution, communication frequency, and system understanding. Moreover, we aimed to also study whether interdependence between human and robot had any influence on this effect.

3.3.1 Design

The experiment had a 2x4 mixed design with interdependence between human and robot as the within-subjects independent variable and robot communication style as the between-subjects independent variable. Interdependence consisted of two conditions (low and high), robot communication style of four conditions (silent, transparent, explainable, adaptive). During the low interdependence condition, participants hardly depended on their robot teammate (and vice versa) since the work was split between the two. In contrast, during the high interdependence condition the participants and robot were highly dependent on each other because we removed the work division and added hard interdependence relationships stemming from a lack of robot capacity to carry critically injured adults and distinguish between kids. Which interdependence condition the participants completed first was counterbalanced, resulting in two order conditions.

3.3.2 Participants

We recruited 72 participants from the different universities' mailing lists and personal contacts (23 females, 48 males, and one preferred not to say). Fifteen participants had an age range of 18-24 years old, 56 participants of 25-34 years old, and one participant was between 55-64 years old. In terms of education, one participant did some high school without obtaining a diploma, two participants were high school graduates, two participants obtained some college credit but no degree, three participants obtained an associate degree, 15 participants obtained a Bachelor's degree, 47 participants obtained a Master's degree, and two participants obtained a PhD degree or higher. With respect to gaming experience, 27 participants played video games several times a year, 24 participants several times a month, 12 participants several times a week, and nine participants played video games on a daily basis. Each participant signed an informed consent form before participating in the study, which was approved by the ethics committee of our institution (ID 1676).

Since each participant teamed up with a robot characterized by one of the four communication styles and one of the two interdependence order conditions, it was important to control for age, gender, education, and gaming experience across the communication style and order conditions. For gender, we conducted a Chi-square test of homogeneity while for age, education, and gaming experience a Kruskal-Wallis test was conducted. Results showed no significant differences between communication style conditions for any of the demographic factors gender ($\chi^2(6)$ = 7.29, p = 0.29), age ($\chi^2(3)$ = 0.76, p = 0.86), education ($\chi^2(3)$ = 0.34, p = 0.95), and gaming experience ($\chi^2(3)$ = 0.31, p = 0.96), indicating that participants were evenly split over the communication style conditions. Moreover, results showed no significant differences between interdependence order conditions for gender ($\chi^2(2)$ = 2.95, p = 0.23), age ($\chi^2(1)$ = 0.75, p = 0.39), education ($\chi^2(1)$ = 2.07, p = 0.15), and gaming experience ($\chi^2(1)$ = 0.21, p = 0.65).

3.3.3 Hardware and Software

To run this experiment we used a Dell laptop, a Virtual Machine (Ubuntu 20.04.2 LTS), and the Human-Agent Teaming Rapid Experimentation (MATRX: `https://matrx-software.com/`) software, a Python package specifically aimed at facilitating human-agent teaming research. The Dell laptop was used to access the Virtual Machine, from which a MATRX world was launched. This two-dimensional grid world contained and tracked the information needed to simulate the agents performing tasks in our environment.

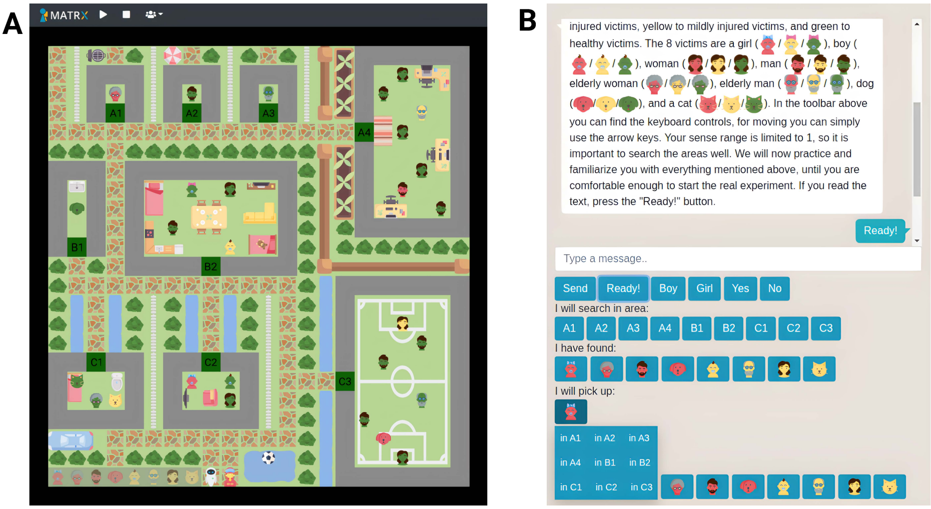
3.3.4 Environment

To access the MATRX world and control their corresponding human agent, participants opened a link in either Chrome or Firefox. In contrast, the experimenter viewed the world with the so called God agent, making it possible to perceive everything and start, pause, and stop the world. We built a world consisting of nine areas, 28 collectable objects, and at least one drop zone (Fig. 3.1A). Furthermore, we added an autonomous virtual robot and human agent to our world, and designed an environment in which these two agents had to collaborate during a search and rescue task. Two different worlds were created, one for each interdependence condition, varying with respect to the drop zone(s) and victim distribution. The low interdependence world consisted of two drop zones with four victims each, whereas the high interdependence world contained just one drop zone with eight victims.

We created the following eight victim types making up the world's collection goal: boy, girl, man, woman, elderly man, elderly woman, dog, and cat. In addition, we created the following injury types: critical, mild, and healthy. Injury type was represented by the color of the victims, where red reflected critically injured, yellow mildly injured, and green healthy victims (see Fig. 3.1). Eight of the 28 objects in the world were either mildly or critically injured and had to be delivered at the drop zone, whereas the other 20 were healthy.

3.3.5 Task

The objective of the task was to search and rescue the eight target victims by inspecting the different areas and dropping the correct victims on the drop zone in a specific order. During the low interdependence condition, the retrieval of the eight victims was equally divided between the autonomous virtual robot and human agent, across two separate drop zones. This way, both team members hardly depended on each other and could execute their



injured victims, yellow to mildly injured victims, and green to healthy victims. The 8 victims are a girl (/ /), boy (/ /), woman (/ /), man (/ /), elderly woman (/ /), elderly man (/ /), dog (/ /), and a cat (/ /). In the toolbar above you can find the keyboard controls, for moving you can simply use the arrow keys. Your sense range is limited to 1, so it is important to search the areas well. We will now practice and familiarize you with everything mentioned above, until you are comfortable enough to start the real experiment. If you read the text, press the "Ready!" button.

Ready!

Figure 3.1: **A** God view of the MATRX world used for this study. The lower left corner of the world shows the drop zone with eight victims to search and rescue. Next to the drop zone are RescueBot and the human avatar at their starting positions. **B** the chat functionality and buttons used by participants to communicate. In addition, the different victim and injury types can be seen. Buttons existed for each area and goal victim to search and rescue.

individual actions independently. In contrast, during the high interdependence condition the eight victims had to be delivered to one shared drop zone. Consequently, the human's actions highly depended on what the robot did (and vice versa) over a sustained sequence of actions. For both conditions, when all eight victims were rescued or when the task was not successfully completed after 10 minutes, the world and task were terminated and all objective data logged.

3.3.6 Agent Types

We added two agents to the world: an autonomous rule-based virtual robot (called RescueBot) and a human agent controlled by the participants, using their keyboard. RescueBot was able to solve the collection task by searching for the next victim to rescue, keeping track of which areas it searched and which victims it found and where, and dropping found goal victims at the drop zone. Both agents could carry only one victim at a time, detect other agents with a range of two grid cells, detect other objects like walls and doors with an infinite sense range, and detect victims with a sense range of only one grid cell. To avoid ceiling effects resulting from a perfect agent, RescueBot moved slower than the human agent and traversed every grid cell during area exploration.

Four different versions of RescueBot were implemented for the experiment, varying with respect to what, how, and how much they communicated between communication style conditions and in capacity between the interdependence conditions. During the high interdependence condition, RescueBot lacked the capacity to carry critically injured adults and distinguish between kids, which added required/hard interdependence relationships between human and robot. When a critically injured adult was found by either RescueBot or the human participant, RescueBot told the participant to pick it up. When RescueBot found an injured kid, it told the human participant to visit that area and clarify the gender of the victim. This way, these hard dependencies required the human and robot to establish supporting interdependence relationships. It is important to emphasize that this version of RescueBot did not wrongly classify kids or unsuccessfully carry critically injured adults, but rather requested support from its human teammate.

We implemented four different communication styles for RescueBot: silent, transparent, explainable, and adaptive. The silent version served as baseline and only disclosed the crucial decisions to request human assistance in case it needed human help. In contrast, the transparent version disclosed its world knowledge/beliefs, actions, decisions, and, in the high interdependence condition, suggestions. The explainable version not only disclosed its world knowledge, actions, suggestions, and decisions, but also clarified them by providing explanations. This communication style provided attributive/causal explanations providing reasons (why) for intentional behavior and actions [132, 133, 145]. The provided reasons included world knowledge, goals, and limitations and adhered to the principles of being simple (few causes), general, complete, and sound [145]. Finally, the adaptive version of RescueBot adjusted its communication based on time pressure and relevancy. In general, after explaining a certain belief, action, suggestion, or decision X based on a goal, belief, or limitation reason Y, the agent adhered to only disclosing X in future situations. Moreover, when time pressure was high (less than five minutes remaining) RescueBot only communicated the most crucial information. The exact information content for each of the communication styles can be found in Table 3.1.

Participants had the ability to communicate to RescueBot via the buttons shown in Fig. 3.1B. Using these, participants could share their current and future actions, perceptions, as well as answers to any of RescueBot's questions (in the high interdependence scenario). RescueBot added the shared information to its memory, and adjusted its behavior correspondingly. This messaging interface was present in a similar fashion as shown in Figure 3.1, so immediately on the right of the environment. Furthermore, the messaging interface was the same for both RescueBot and the participant, a chat box/room consisting of textual messages. Participants only had to press buttons to share required information such as which areas they searched or where they found victims. This way, we tried to decrease workload as a result of having to type messages.

RescueBot's behavior did not vary between the four communication style conditions. When RescueBot did not know the location of the current victim to rescue, it moved towards the closest unsearched area and explored it. If the participant told RescueBot it was going to search the same area, RescueBot moved to the next closest unsearched area to explore instead. During exploration of the areas, RescueBot added the location of found victims to its memory. When participants found victims and communicated this, RescueBot also added this to its memory. If RescueBot found the current victim to rescue during area exploration, it first completed searching the whole area before picking up and dropping the victim at the drop zone. In case participants told RescueBot they would pick up the victim instead, RescueBot would start searching for or picking up the next victim to rescue. If this victim was already found by the participant, RescueBot would move to the corresponding area and explore the whole room as it did not know the exact location. When it found the

Message	Transparency	Explanation
1	Moving to X	to pick up Y
2	Moving to X	to search for Y and because it is the closest unsearched area
3	Searching through whole X	because my sense range is limited and to find Y
4	Found Y in X	because you told me Y was located here
5	Found Y in X	because I am traversing the whole area
6	You should pick up Y in X	because I am forbidden to carry critically injured adults
7	Y not present in X	because I searched the whole area without finding Y
8	You should clarify the gender of the injured baby in X	because I am unable to distinguish them
9	Going to re-search areas	to find Y and because we searched all areas but did not find Y
10	Picking up Y in X	because Y should be transported to the drop zone.
11	Transporting Y to drop zone	because Y should be delivered there for treatment.
12	Delivered Y at drop zone	because Y was current victim to rescue.
13	Waiting for human at drop zone	because previous victim should be collected first.
14	I suggest you pick up Y in X	because X is far away and you can move faster.

X refers to specific area, Y to specific victim.

Table 3.1: The information content for each of the four communication styles. Explainability included both the content under the column Transparency, plus the explanation under the column Explanation. Underlined messages refer to the only information shared by the silent baseline. Except for the first two messages, the adaptive communication style dropped the explanations after providing them once. When time pressure was high, the adaptive version of RescueBot stopped sending the messages with the bold numbers.

#	Low Interdependence	High Interdependence
1	Which area(s) did RB search?	Which area(s) did RB search?
2	Which victim(s) did RB find?	Which victim(s) did RB find?
3	Where is RB currently located?	Where is RB currently located?
4	Which action is RB currently executing?	Which action is RB currently executing?
5	Which victim(s) did RB find in area *Y*?	Which victim(s) did RB find in area *Y*?
6	Which action will RB perform next?	Which victim(s) did RB rescue/drop?
7	Which of your goal victim(s) did RB find?	Which victim(s) is RB unable to carry?
8	In which area did RB find victim *X*?	Which victim(s) is RB unable to identify?

X refers to specific area, *Y* to specific victim.

Table 3.2: The SAGAT queries used during the experiment, for each interdependence condition. RB = RescueBot.

victim, it would immediately pick it up and move to the drop zone rather than searching the rest of the area. During all situations described above the transparent, explainable, and adaptive versions of RescueBot communicated their actions, beliefs, decisions and suggestions using the messages outlined in Table 3.1.

3.3.7 Measures

Team Performance

We objectively measured team performance during the low and high interdependence conditions using completion time, accuracy, and completeness. We transformed completion time to the percentage of time left to finish the task and calculated overall team performance as the mean of time left, accuracy, and completeness. Completion time was converted into the percentage of time left in order to transform the variable to the same scale and interpretation as accuracy and completeness (i.e., expressed in % and with higher values reflecting better performance). Accuracy reflected the percentage of victims collected in the correct order, whereas completeness reflected the percentage of all victims collected. We manually kept track of accuracy during the task, while completeness was logged automatically using MATRX.

Situation Awareness of RescueBot

Situation awareness (SA) of RescueBot's behavior processes and decisions was measured objectively, using the Situation Awareness Global Assessment Technique (SAGAT) [60, 176]. First, we defined the human informational and SA requirements during the search and rescue task using the Goal Directed Task Analysis [62]. We used this analysis to define which information participants required about RescueBot's behavior in order to achieve their respective goals. Next, we formulated eight SAGAT queries for each interdependence condition, objectively evaluating human knowledge of this situational information [60, 62]. Table 3.2 shows all queries used during the experiment, for each interdependence condition.

During both interdependence conditions, the task was paused twice, and the same eight queries were asked. We took the average percentage of correctly answered queries as objective measure of SA. For each query we provided five multiple choice options, except for query three which was answered by selecting a location on the map. The answer options were different for each of the two assessment moments, except for queries four, six, and seven (low interdependence) because they only had five possible answer options. Finally, for queries five and eight (low interdependence) the exact area and victim used in the query was different for each assessment moment.

Trust

We subjectively measured user trust in RescueBot using the 5-pt Likert trust scale for XAI [92]. This scale consisted of eight items and measured confidence in and predictability, reliability, safety, efficiency, wariness, performance, and likeability of RescueBot. We calculated the mean of the eight items as the final trust score.

Workload

Workload during the task was measured subjectively using the raw NASA Task Load Index (NASA-TLX) [87]. This consisted of six items evaluated on scales from 0 to 100 and increments of size five, so yielding 20 answer options. The six items measured mental, physical, and temporal demand, as well as performance, effort, and frustration. We calculated the mean of the six items as the final workload score.

Perceived System Understanding

We subjectively measured understanding of RescueBot using the 7-pt Likert Perceived System Understanding Questionnaire [211]. This scale consisted of eight items and measured explainability, understandability, and predictability of RescueBot. We calculated the mean of the eight items as the final understanding score.

Reliance

Reliance was objectively measured using the MATRX loggers. We defined reliance as the percentage of victims that were found first by the participant but rescued by RescueBot. Using the loggers, for each participant we counted how many of the goal victims they found first. Next, we counted how many of these victims were eventually picked up and dropped by RescueBot, and divided this by the number of victims the participant found first to get the corresponding reliance percentage.

We defined reliance in this way because it allowed the inclusion of the silent baseline into the analysis. For example, we also thought of defining reliance as the percentage of victims found by RescueBot but rescued by the participant (i.e., reliance upon RescueBot's perceptions). However, defining reliance as such would exclude the silent baseline from the analysis, as it did not send information about RescueBot's perceptions. Therefore, we defined reliance as the percentage of victims found by the participant but rescued by RescueBot (i.e., reliance upon RescueBot's actions), and included the silent baseline into the analysis.

Human Rescue Contribution

We measured human rescue contribution using the MATRX loggers, and defined it as the percentage of goal victims rescued by the participant.

Human Messages Sent

Finally, we logged the number of messages sent from participant to RescueBot to investigate human communication frequency.

3.3.8 Procedure

The experiment was conducted in two sessions: an introduction and experiment session. The introduction session served as a tutorial aimed at getting the participants familiar with the environment, controls, and messaging system, to minimize any learning and order effects, and to control for the possible influence of gaming experience. During the tutorial, RescueBot gave the same step by step instructions to all participants, for example, on how to pick up a victim and when to send certain messages.

The second part of the tutorial included a trial of the real experiment. During this trial, the participant collaborated with the version of RescueBot with which they would also collaborate during the real experiment (so the silent, transparent, explainable, or adaptive version). Participants had to search and rescue six victims on one joint drop zone while collaborating with RescueBot, so similar to the high interdependence trial but without any required dependencies resulting from robot limitations.

After three minutes, we paused the trial and introduced the participants to the SAGAT queries. We explained that during the real experiment, the task would be paused at random moments and several queries would be asked related to their knowledge of RescueBot's behavior processes and decisions. Participants were encouraged to make their best guess when they did not know or were uncertain about the answer, but we also told them that they could skip a question when they were not comfortable enough to guess.

After the tutorial, we asked the participants if they were comfortable enough to start the experiment session or wanted to re-do the trial. In the experiment session, participants completed the two interdependence conditions. Which condition the participants completed first was counterbalanced, resulting in two order conditions. We controlled for age, gender, education, and gaming experience across these two order conditions, resulting in no significant differences between the two conditions for any of these factors. During both interdependence conditions we paused the task twice, followed by the corresponding eight SAGAT queries in Table 3.2. The first freeze was at a random moment between two and three minutes after starting the task, the second freeze a minimum of one and a half minute later than the first one and a maximum of two minutes later.

When participants finished the first task, we presented them with the Trust Scale for XAI and NASA-TLX. Next, participants completed the second variant of the task. Again, we paused the task twice and asked the SAGAT queries. After finishing the second task, participants again completed the Trust Scale for XAI and NASA-TLX. Finally, participants filled in the Perceived System Understanding questionnaire to end the experiment. The whole study lasted for about one hour and was conducted during an online meeting using either Microsoft Teams, Zoom, or Google Meet. All survey responses were collected using Qualtrics.

3.4 Results

3.4.1 Learning and Order Effects

We examined the presence of potential learning and order effects by testing for differences in dependent variable outcomes between i) the two experiment order versions (to test for order effects) and ii) the two time points (to test for learning effects). Order 1 started with the low interdependence condition followed by the high interdependence condition, and vice versa for order 2. Time point 1 included all data from the low interdependence condition from order 1 and high interdependence condition from order 2, whereas time point 2 included all data from the low interdependence condition from order 2 and high interdependence condition from order 1.

Order Effects

We tested for order effects on all the dependent variables trust, workload, understanding, situation awareness, team performance, reliance, human rescue contribution, and human messages sent. When all assumptions were met, we conducted an independent-samples t-test, if not we conducted a Mann-Whitney U test. We did not find statistically significant differences in the outcome scores between the two order conditions for trust (W = 2147, p = 0.08), reliance (W = 2494, p = 0.70), workload (t(140) = 1.75, p = 0.08), situation awareness (W = 2412, p = 0.47), team performance (W = 2494, p = 0.70), human rescue contribution (W = 2562, p = 0.90), human messages sent (W = 2444, p = 0.55), or understanding (W = 2392, p = 0.43). Corresponding descriptive statistics and plots can be found in Appendix A.

Learning Effects

We tested for learning effects on the objective measures of situation awareness, team performance, reliance, and human rescue contribution to examine whether participants performed the task differently at later time points. When all assumptions were met, we conducted a paired-samples t-test, if not we conducted a Wilcoxon signed-rank test. We did not find statistically significant differences in the outcome scores between the two time points for situation awareness (t(71) = -0.86, p = 0.39), team performance (W = 1166, p = 0.41), human rescue contribution (W = 610, p = 0.12), or reliance (W = 1214, p = 0.25). Corresponding descriptive statistics and plots can be found in Appendix A.

3.4.2 Effects of Communication Style and Interdependence

Here, we report the effects of and interaction between communication style and interdependence on the dependent variables. For most of the dependent variables, we employed a non-parametric rank based method for the analysis of variance (ANOVA), mainly to deal with violations of the mixed ANOVA assumption of normality. To this end, we used the *R* package and function *nparLD* [151] for non-parametric tests for repeated measures data in factorial designs. This method defines relative treatment effects in reference to the distributions of the dependent variables, estimated on mean ranks. Therefore, relative treatment effects can be considered as generalized expectations or means. This method does not require distributional assumptions, is applicable to a variety of data types, and is robust with respect to outliers and small sample sizes. For the dependent variables that did meet all assumptions, we conducted a mixed ANOVA.

Trust

Since there were two extreme outliers and the data was not normally distributed (p < 0.05) in two cells of the design, as assessed by Shapiro-Wilk's test of normality, we conducted the non-parametric rank based ANOVA. Results showed a statistically significant main effect of communication style (F(2.92) = 13.40, p < 0.0001, effect size = 0.81) on trust. Pairwise robust ATS post-hoc comparisons revealed statistically significant differences in trust scores between the silent baseline (RTE = 0.23, Mean Rank = 33.30, SD Rank = 27.63) and transparent (RTE = 0.59, Mean Rank = 84.85, SD Rank = 36.36) (F(1) = 27.39, $p < 0.0001$), adaptive (RTE = 0.61, Mean Rank = 88.28, SD Rank = 35.68) (F(1) = 36.43, $p < 0.0001$), and explainable (RTE = 0.58, Mean Rank = 83.57, SD Rank = 38.09) (F(1) = 23.24, $p < 0.0001$) conditions. In addition, results showed a statistically significant main effect of interdependence (F(1) = 18.76, $p < 0.0001$, effect size = 0.51) on trust, revealing a significant difference in trust scores between the low (RTE = 0.56, Mean Rank = 80.69, SD Rank = 41.23) and high (RTE = 0.44, Mean Rank = 64.31, SD Rank = 40.66) interdependence conditions. Figure 3.2A shows the interaction plot of the relative effects of communication style and interdependence on trust scores, exact relative treatment effects (RTE) and corresponding mean ranks can be found in Table 3.3.

Reliance

Because the data was not normally distributed ($p < 0.05$) in most cells of the design, we conducted the non-parametric rank based ANOVA. Results showed a statistically significant main effect of communication style (F(2.69) = 3.99, $p < 0.025$, effect size = 0.38) on reliance. Pairwise robust ATS post-hoc comparisons revealed statistically significant differences in reliance scores between the silent (RTE = 0.39, Mean Rank = 56.32, SD Rank = 45.95) and explainable (RTE = 0.54, Mean Rank = 77.96, SD Rank = 39.51) condition (F(1) = 4.33, $p < 0.05$), silent and adaptive (RTE = 0.59, Mean Rank = 86.03, SD Rank = 37.15) condition (F(1) = 9.06, $p < 0.005$), and adaptive and transparent (RTE = 0.48, Mean Rank = 69.69, SD Rank = 37.81) condition (F(1) = 5.01, $p < 0.05$). In addition, results showed a statistically significant main effect of interdependence (F(1) = 104.30, $p < 0.0001$, effect size = 1.20) on reliance, revealing a statistically significant difference in reliance scores between the low (RTE = 0.66, Mean Rank = 95.56, SD Rank = 37.42) and high (RTE = 0.34, Mean Rank = 49.44, SD Rank = 31.00) interdependence conditions. Figure 3.2B shows the interaction plot of the relative effects of communication style and interdependence on reliance scores, exact relative treatment effects and corresponding mean ranks can be found in Table 3.3.

Workload

Since all assumptions were met (no outliers, normality, homogeneity of variances, and homogeneity of covariances), we performed a mixed ANOVA. Results showed a statistically significant main effect of interdependence (F(1, 68) = 0.46, $p < 0.0005$, $\eta^2_G = 0.024$) on workload. A paired-samples t-test was conducted to determine the effect of interdependence on workload scores. Results showed that there was a significant difference in workload scores during high (Mean = 39.40, SD = 16.7) and low (Mean = 34.50, SD = 15.40) interdependence conditions (t(71) = -3.87, $p < 0.0005$, d = 0.46). Figure 3.2C shows the interaction plot of the effects of communication style and interdependence on workload scores, Table 3.4 shows the descriptive statistics for each combination of communication style and interdependence.

Situation Awareness

Because all assumptions were met we conducted a mixed ANOVA. Results showed a statistically significant interaction between communication style and interdependence on situation awareness (SA) scores (F(3, 68) = 3.31, $p < 0.05$, $\eta^2_G = 0.057$). We analyzed the simple main effect of communication on SA during each interdependence condition using a one-way ANOVA. Results showed that the simple main effect of communication style

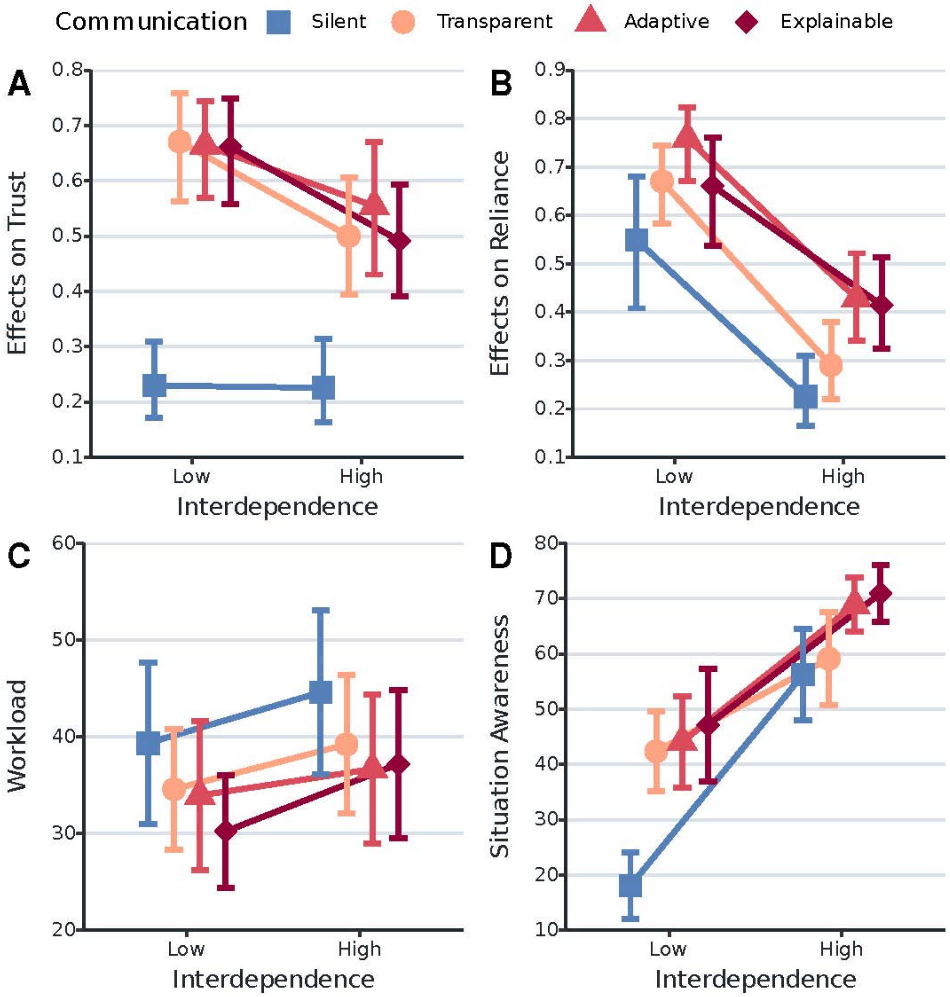


Figure 3.2: Interaction plots of the effects of communication style and interdependence on the dependent variables trust, reliance, workload, and situation awareness. **A** shows the relative treatment effects of communication style on trust across interdependence. The y-axis is the conventional graphical representation of the non-parametric ANOVA we used. It represents the relative marginal effect of the different communication styles across interdependence. The higher is the value on the y-axis, the higher is the corresponding trust value/score. Error bars represent the 95% confidence intervals of the relative marginal effects. **B** shows the relative treatment effects of communication style on reliance across interdependence. The higher is the value on the y-axis, the higher is the corresponding reliance percentage value/score. Error bars represent the 95% confidence intervals of the relative marginal effects. **C** shows the effects of communication style on workload across interdependence. The y-axis represents the mean workload. Error bars represent the 95% confidence intervals of the mean workload scores. **D** shows the effects of communication style on situation awareness across interdependence. The y-axis represents the mean situation awareness scores. Error bars represent the 95% confidence intervals of the mean situation awareness scores.

Variable	Communication	Interdependence	Mean Rank (SD)	RTE	95% CI
Trust	Silent	Low	33.61 (27.13)	0.23	[0.17 0.31]
	Silent	High	33.00 (28.91)	0.23	[0.16 0.31]
	Transparent	Low	97.14 (35.41)	0.67	[0.56 0.78]
	Transparent	High	72.56 (38.15)	0.50	[0.39 0.61]
	Adaptive	Low	96.19 (27.59)	0.66	[0.57 0.74]
	Adaptive	High	80.36 (41.55)	0.55	[0.43 0.67]
	Explainable	Low	95.81 (34.92)	0.66	[0.56 0.75]
	Explainable	High	71.33 (38.10)	0.49	[0.39 0.59]
Reliance	Silent	Low	79.58 (49.52)	0.55	[0.41 0.68]
	Silent	High	33.06 (27.36)	0.23	[0.17 0.31]
	Transparent	Low	97.08 (26.04)	0.67	[0.58 0.74]
	Transparent	High	42.31 (26.01)	0.29	[0.22 0.38]
	Adaptive	Low	109.83 (26.20)	0.76	[0.67 0.82]
	Adaptive	High	62.22 (30.90)	0.43	[0.34 0.52]
	Explainable	Low	95.72 (39.29)	0.66	[0.54 0.76]
	Explainable	High	60.19 (31.65)	0.41	[0.32 0.51]

Table 3.3: Descriptive statistics for the dependent variables trust and reliance. Values correspond to the data points of the plots in Figure 3.2A and B.

was significant during both the high (F(3, 68) = 4.20, $p < 0.025$, η^2_G = 0.156) and low (F(3, 68) = 10.60, $p < 0.0001$, η^2_G = 0.318) interdependence conditions. Pairwise comparisons using a Bonferroni correction revealed significant differences in SA scores between the silent baseline (Mean = 18.06, SD = 13.02) and transparent (Mean = 42.36, SD = 15.69, $p < 0.001$), adaptive (Mean = 44.10, SD = 17.87, $p < 0.0005$), and explainable (Mean = 47.12, SD = 22.03, $p < 0.0001$) conditions when interdependence was low. When interdependence was high, results showed a significant difference in SA scores between the silent baseline (Mean = 56.25, SD = 17.83) and only the explainable condition (Mean = 70.93, SD = 11.13, $p < 0.05$).

We analyzed the simple main effect of interdependence on SA for each communication condition using a paired-samples t-test. Results showed statistically significant differences in mean SA scores between the high and low interdependence conditions for the silent (Mean High = 56.20, SD High = 17.80; Mean Low = 18.10, SD Low = 13.00) (t(17) = -10.00, $p < 0.0001$, d = 2.36), transparent (Mean High = 59.10, SD High = 18.30; Mean Low = 42.40, SD Low = 15.70) (t(17) = -3.07, $p < 0.01$, d = 0.72), adaptive (Mean High = 68.90, SD High = 10.60; Mean Low = 44.10, SD Low = 17.90) (t(17) = -4.68, $p < 0.0005$, d = 1.10), and explainable (Mean High = 70.90, SD High = 11.1; Mean Low = 47.10, SD Low = 22.00) (t(17) = -4.84, $p < 0.0005$, d = 1.14) conditions. Figure 3.2D shows the interaction plot of the effects of communication style and interdependence on SA scores, Table 3.4 shows the descriptive statistics for each combination of communication style and interdependence.

Team Performance

Since the data was not normally distributed in most cells of the experimental design, we conducted the non-parametric rank based ANOVA. Results showed a statistically significant

Variable	Communication	Interdependence	Mean (SD)	95% CI
Workload	Silent	Low	39.35 (18.07)	[31.00 47.70]
	Silent	High	44.60 (18.31)	[36.14 53.06]
	Transparent	Low	34.58 (13.48)	[28.36 40.81]
	Transparent	High	39.21 (15.43)	[32.08 46.34]
	Adaptive	Low	33.93 (16.68)	[26.23 41.64]
	Adaptive	High	36.67 (16.59)	[29.00 44.33]
	Explainable	Low	30.23 (12.59)	[24.41 36.04]
	Explainable	High	37.18 (16.52)	[29.54 44.81]
SA	Silent	Low	18.06 (13.02)	[12.04 24.07]
	Silent	High	56.25 (17.83)	[48.01 64.48]
	Transparent	Low	42.36 (15.69)	[35.11 49.61]
	Transparent	High	59.12 (18.30)	[50.67 67.58]
	Adaptive	Low	44.10 (17.87)	[35.84 52.35]
	Adaptive	High	68.87 (10.55)	[64.00 73.75]
	Explainable	Low	47.12 (22.03)	[36.95 57.30]
	Explainable	High	70.93 (11.13)	[65.79 76.07]

Table 3.4: Descriptive statistics for the dependent variables workload and situation awareness (SA). Values correspond to the data points of the plots in Figure 3.2C and D.

main effect of interdependence (F(1) = 76.81, $p < 0.0001$, effect size = 1.03) on performance, revealing a statistically significant difference in team performance between the low (RTE = 0.61, Mean Rank = 87.70, SD Rank = 36.54) and high (RTE = 0.39, Mean Rank = 57.30, SD Rank = 41.23) interdependence conditions. Figure 3.3A shows the interaction plot of the relative effects of communication style and interdependence on performance scores, exact relative treatment effects (RTE) and corresponding mean ranks can be found in Table 3.5.

Human Rescue Contribution

Because the data was not normally distributed in most cells, we conducted the non-parametric rank based ANOVA. Results showed a statistically significant interaction between communication style and interdependence on human rescue contribution (F(2.95) = 3.03, $p < 0.05$, effect size = 0.35). We analyzed the simple main effect of communication on human rescue contribution during each interdependence condition using a Kruskal-Wallis test. Results showed that the simple main effect of communication style was not significant during both interdependence conditions. We analyzed the simple main effect of interdependence on human rescue contribution for each communication condition using the relative treatment effects test. Results showed a statistically significant difference in relative treatment effects between the low and high interdependence conditions for the silent (RTE Low = 0.35, RTE High = 0.72, $p < 0.0001$) and transparent (RTE Low = 0.37, RTE High = 0.60, $p < 0.005$) conditions. Figure 3.3B shows the interaction plot of the relative effects of communication style and interdependence on human rescue contribution scores, exact relative treatment effects (RTE) and corresponding mean ranks can be found in Table 3.5.

Variable	Communication	Interdep.	Mean Rank (SD)	RTE	95% CI
Performance	Silent	Low	68.58 (43.65)	0.47	[0.36 0.59]
	Silent	High	41.64 (44.40)	0.29	[0.19 0.43]
	Transparent	Low	93.47 (30.80)	0.65	[0.54 0.73]
	Transparent	High	62.03 (36.31)	0.43	[0.33 0.54]
	Adaptive	Low	89.31 (32.35)	0.62	[0.51 0.71]
	Adaptive	High	60.61 (38.10)	0.42	[0.31 0.54]
	Explainable	Low	99.44 (33.23)	0.69	[0.59 0.77]
	Explainable	High	64.92 (44.73)	0.44	[0.33 0.58]
Contribution	Silent	Low	50.97 (40.83)	0.35	[0.25 0.48]
	Silent	High	103.56 (40.31)	0.72	[0.58 0.81]
	Transparent	Low	53.06 (30.73)	0.37	[0.28 0.46]
	Transparent	High	86.19 (37.23)	0.60	[0.48 0.70]
	Adaptive	Low	60.78 (40.97)	0.42	[0.31 0.54]
	Adaptive	High	73.33 (36.35)	0.51	[0.40 0.61]
	Explainable	Low	68.97 (40.27)	0.48	[0.36 0.60]
	Explainable	High	83.14 (34.57)	0.57	[0.47 0.67]
Messages	Silent	Low	72.19 (51.81)	0.50	[0.36 0.63]
	Silent	High	75.19 (48.71)	0.52	[0.39 0.64]
	Transparent	Low	53.83 (30.60)	0.37	[0.28 0.48]
	Transparent	High	98.33 (27.62)	0.68	[0.58 0.76]
	Adaptive	Low	44.97 (25.57)	0.31	[0.23 0.40]
	Adaptive	High	88.33 (36.11)	0.61	[0.49 0.71]
	Explainable	Low	59.11 (43.41)	0.41	[0.29 0.54]
	Explainable	High	88.03 (36.95)	0.61	[0.49 0.71]

Table 3.5: Descriptive statistics for the dependent variables team performance, human rescue contribution, and number of human messages sent. Values correspond to the data points of the plots in Figure 3.3A, B, and C.

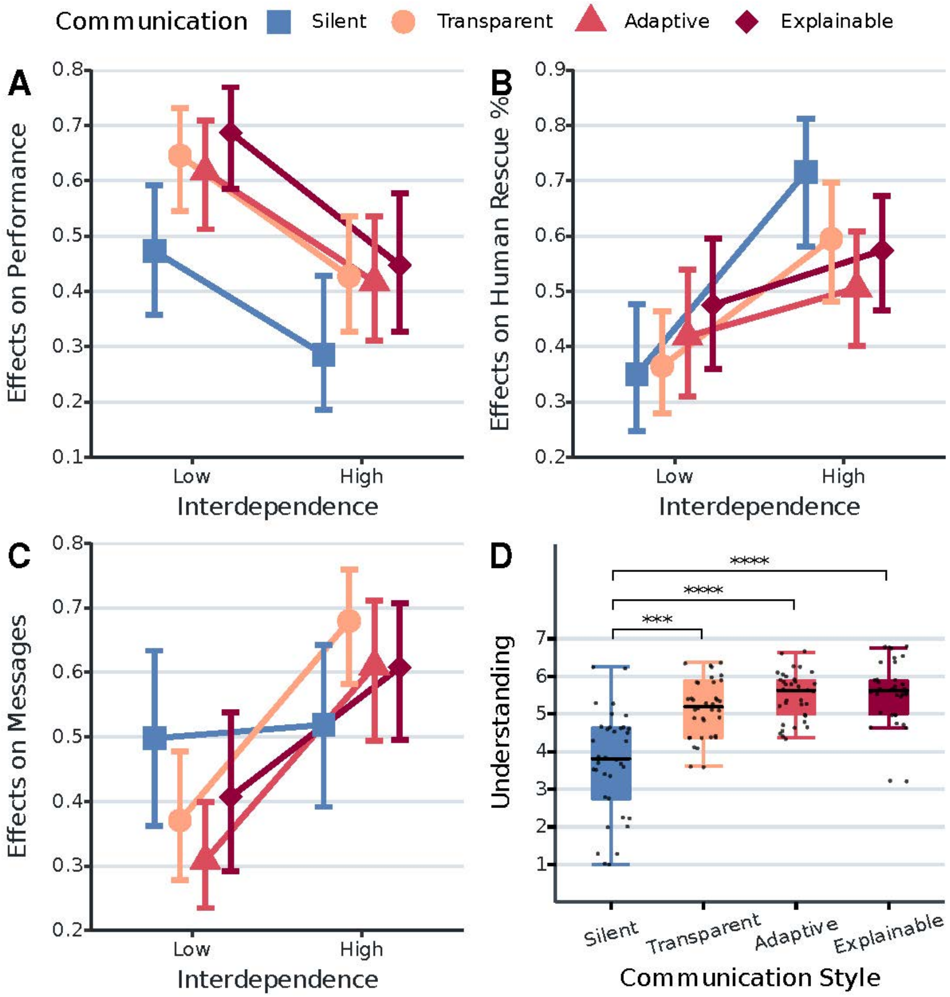


Figure 3.3: Interaction plots of the effects of communication style and interdependence on the dependent variables team performance, human rescue contribution, and human messages sent (**A, B, C**). Boxplots of system understanding for each of the communication style conditions (**D**). **A** shows the relative treatment effects of communication style on performance across interdependence. The y-axis is the conventional graphical representation of the non-parametric ANOVA we used. It represents the relative marginal effect of the different communication styles across interdependence. The higher the value on the y-axis, the higher is the corresponding performance value/score. Error bars represent the 95% confidence intervals of the relative marginal effects. **B** shows the relative treatment effect of communication style on human rescue contribution across interdependence. The higher the value on the y-axis, the higher is the corresponding human rescue percentage value/score. Error bars represent the 95% confidence intervals of the relative marginal effects. **C** shows the relative treatment effects of communication style on human messages sent across interdependence. The higher the value on the y-axis, the higher is the corresponding number of messages sent by the participants. Error bars represent the 95% confidence intervals of the relative marginal effects. **D** shows the effects of communication style on system understanding. The y-axis represents the mean understanding scores. ***p<0.0005. ****p<0.0001.

Number of Human Messages Sent

Since most assumptions of a mixed ANOVA were violated, we conducted the non-parametric rank based ANOVA. Results showed a statistically significant interaction between communication style and interdependence on the number of human messages sent (F(2.77) = 5.45, $p < 0.005$, effect size = 0.57). We analyzed the simple main effect of robot communication style on human messages sent using a Kruskal-Wallis test. Results showed that the simple main effect of communication style was not significant during both interdependence conditions. Next, we analyzed the simple main effect of interdependence on human messages sent for each communication style condition using the relative treatment effects test. Results showed a statistically significant difference in relative treatment effects between the low and high interdependence conditions for the transparent (RTE Low = 0.37, RTE High = 0.68, $p < 0.0001$), adaptive (RTE Low = 0.31, RTE high = 61, $p < 0.0001$), and explainable (RTE Low = 0.41, RTE High = 0.61, $p < 0.01$, r = 0.56) conditions. Figure 3.3C shows the interaction plot of the relative effects of communication style and interdependence on human messages sent, exact relative treatment effects (RTE) and corresponding mean ranks can be found in Table 3.5.

System Understanding

Because the data was not normally distributed in one cell of the design and since there was no homogeneity of variances, we conducted a Kruskal-Wallis test. Results showed that there were statistically significant differences in system understanding between the communication style conditions ($\chi^2(3) = 48.30$, $p < 0.0001$, $\eta^2 = 0.32$). Pairwise comparisons using Dunn's procedure with a Bonferonni correction revealed statistically significant differences in understanding scores between the silent baseline (Mean Rank = 32.72, SD Rank = 31.51) and transparent (Mean Rank = 73.83, SD Rank = 36.93) ($p < 0.0005$), adaptive (Mean Rank = 89.72, SD Rank = 32.74) ($p < 0.0001$), and explainable (Mean Rank = 93.72, SD Rank = 35.59) ($p < 0.0001$) conditions. Figure 3.3D shows the boxplots of system understanding for each of the communication style conditions.

3.4.3 Predicting Team Performance

We ran a multiple linear regression analysis to determine whether we could predict a quantitative outcome of team performance based on the predictor variables situation awareness, trust, reliance, workload, and human messages sent. Moreover, we added experiment version as a predictor to examine the presence of potential order effects and interdependence as an interaction term to examine whether the association between predictors and outcome depended on the level of interdependence. Next, we used the GVLMA package to check the linear model assumptions normality, heteroscedasticity, linearity, and uncorrelatedness of the model. Since not all assumptions were acceptable, we removed the five unusual observations (out of a total of 144 observations).

Results showed that the regression model statistically significantly predicted team performance (F(13, 125) = 7.11, $p < 0.0001$, adj. $R^2 = 0.37$). Furthermore, results showed that only situation awareness ($p < 0.0001$), workload ($p < 0.005$), and human messages sent ($p < 0.001$) added statistically significantly to the prediction. When interdependence was low an increase in SA of 1% was associated with an increase in team performance of 0.10%, an increase in workload of 1% with a decrease in team performance of 0.14%, and an increase in human messages sent of 1 message with an increase in team performance of 0.29%. When interdependence was high an increase in SA of 1% was associated with an increase in team performance of 0.27%, an increase in workload of 1% with a decrease in team performance of 0.10%, and an increase in human messages sent of 1 message with an increase in team performance of 0.70%. Results also showed that the association between situation awareness and team performance depended on the level of interdependence ($p < 0.05$), with team performance increasing at a higher rate with an increase in SA when interdependence was high. Figure 3.4 shows the interaction plots of the significant predictors. Finally, experiment version (i.e., order) did not add statistically significantly to the prediction of team performance.

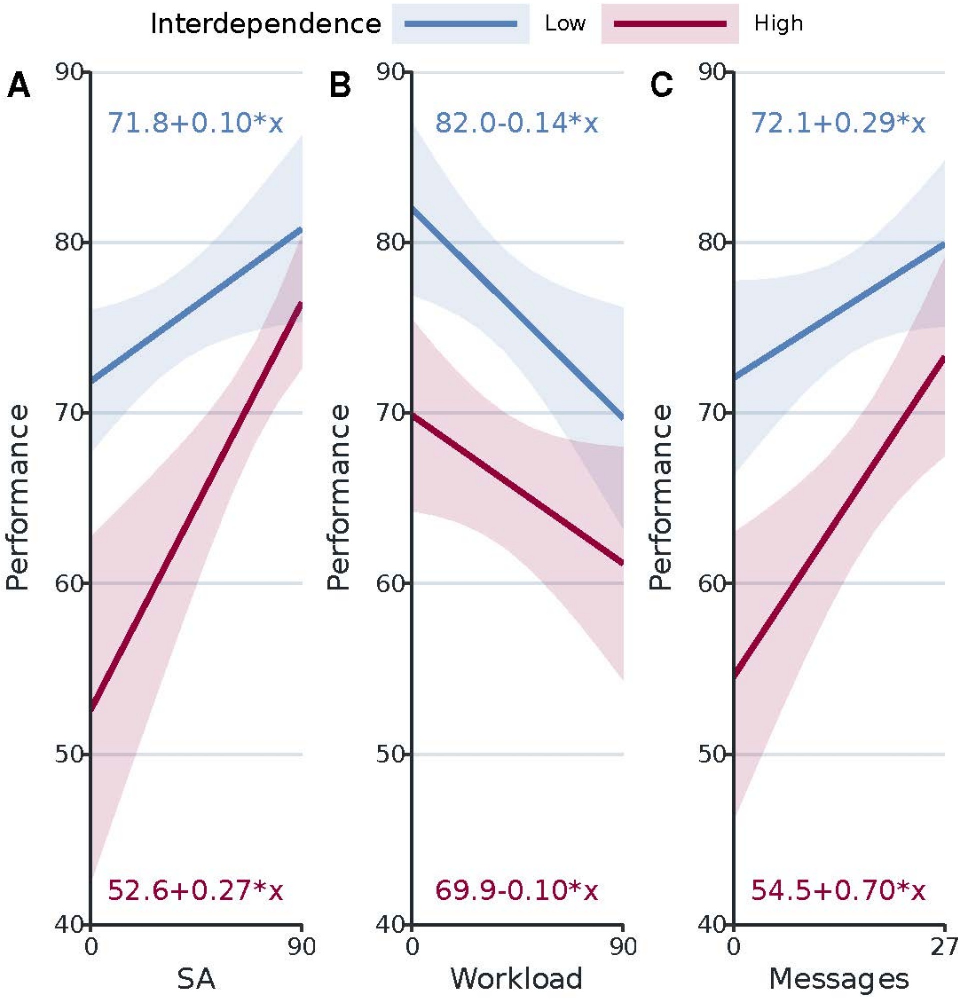


Figure 3.4: Predicted values of team performance based on the statistically significant predictor variables situation awareness (SA), workload, and human messages sent. Intervals represent the lower and upper bounds of the 95% confidence intervals for the predicted values. **A** shows the predicted changes in team performance with changes in the predictor variable situation awareness, at both interdependence levels. **B** shows the predicted changes in team performance with changes in the predictor variable workload, at both interdependence levels. **C** shows the predicted changes in team performance with changes in the predictor variable human messages sent, at both interdependence levels.

3.5 Discussion and Conclusion

3.5.1 Discussion

Trust

The results in Section 3.4.2, Figure 3.2A, and Table 3.3 clearly show that robot communication results in significantly higher trust in the robot. However, we did not find evidence for higher trust in the robot when being explainable rather than transparent. This is not in line with other studies demonstrating how providing more reasoning information is related to increases in trust ([18, 157]. Furthermore, we observe that trust in the robot significantly decreases when interdependence is high, which does not correspond with other studies reporting that increasing interdependence subsequently increases participant positive affect [157, 237, 238]. One possible explanation is that we increased interdependence not only by removing the work division, but also by adding hard interdependence relationships stemming from a lack of robot capacity. Therefore, we believe it is crucial to carefully consider the details of task interdependence and interdependence relationships when comparing studies on interdependence in human-robot teams. For example, two human-robot teams can be highly interdependent but if one team is characterized by hard interdependencies stemming from a lack of robot capacity and the other by hard interdependencies stemming from a lack of human capacity, it is not surprising when trust in the robot differs significantly between these two teams. Another possible reason for this result is that trust is only critical when human and robot are highly interdependent, and therefore people judge it more critically. This is in line with the claim that interdependence relationships are the mechanisms by which relational trust is established [97].

Reliance

For reliance the results show that people rely more on RescueBot when it provides explanations, as demonstrated by the significant difference in reliance scores between the explainable and silent condition, and adaptive and both silent and transparent conditions. This could be the result of the specific message/explanation number one from Table 3.2, explaining the reason why RescueBot was moving to a certain area (to pick up a certain victim). It is possible that when people received this explanation, they were more inclined to let RescueBot complete this goal rather than engage in the same tasks, showing benefits to the coordination of work. The results further show a significant decrease in reliance during high interdependence, as clearly visualized in Figure 3.2B. This could be explained by the similar significant decrease in trust scores when interdependence is high. Since trust can impact reliance upon and use of autonomous AI systems [59, 158, 181], the decrease in trust during high interdependence might have resulted in a corresponding decrease in reliance upon RescueBot to rescue victims found by the human participants.

Workload

Results further show no evidence for an increase in workload when adding explanations to transparency, which is in line with the results in [34, 142, 187]. This indicates that dynamic adaptation based on workload might not be necessary in this type of task and scenario. Furthermore, we observe a higher workload when human and robot are highly interdependent. This could be due to the increasing importance of coordination and collaboration [99, 192], resulting in more human effort to stay aware of the robot's behavior and inform the robot about own behavior.

Situation Awareness

The results also show how the effects of communication style on situation awareness depend on the level of interdependence. More specifically, robot communication results in a significantly higher situation awareness when interdependence is low. However, when interdependence is high only repeatedly providing explanations results in a significantly higher situation awareness than being silent. One possible explanation for these results is that during the high interdependence scenario, the silent baseline did communicate something: the information related to the required dependencies (asking for help). However, it seems unlikely that sharing only this information can account fully for this finding. Therefore, another possibility is that being highly interdependent increases SA irrespective of communication style, as it is more necessary to complete the task well and people, therefore, pay more attention already. This is also underlined by the significantly higher SA during high interdependence, observed for all communication style conditions. As people seem to already pay more or less attention to their robot teammate depending on interdependence levels, it is crucial to understand these levels when designing explanations for any type of future applications. A final suggestion is that only explainability adds crucial information required for a higher situation awareness than just being highly interdependent can already account for.

Team Performance

For team performance, we first observe a significant decrease in performance scores when interdependence is high. This does not align with other studies reporting how increasing interdependence subsequently increases team performance [157, 237, 238]. One possible explanation is that increasing interdependence also increases the need for coordination and collaboration [99, 192], which in turn results in a more challenging and demanding scenario. This interpretation also aligns with the observed increase in communication frequency (for all conditions except the baseline) and higher workload when interdependence is high. Consequently, this more challenging and demanding scenario could have resulted in a decrease in performance.

The results further show that in our task and scenario, only situation awareness, workload, and human messages sent are significantly associated with team performance. More specifically, increasing SA, decreasing workload, and increasing the number of human messages sent are associated with increases in team performance. In terms of human messages sent, the result can be considered surprising as earlier work showed that a greater number of team messages shared was associated with lower team performance [44]. It is also surprising that our results do not show evidence for a significant positive association between trust and team performance, as previous works in both human-human and human-robot teams did [108, 246, 254, 255]. This suggests that (in our task and scenario) situation awareness, workload, and human messages sent are more important to team performance than trust in the robot. However, when the task and scenario get more risky, trust in the robot may become more important [97].

We also observe how increasing SA is associated with a significantly higher increase in team performance when interdependence is high. All in all, our results provide valuable insights into the mechanisms driving HART performance [157]. Based on these results, we advice to pay special attention to SA, workload, and communication when designing/ developing human-robot teams, while also accounting for the level of interdependence between human and robot.

Human Rescue Contribution

Results further show how the effect of interdependence on human rescue contribution depends on the communication style of RescueBot. More specifically, we observe a significant increase in contribution during high interdependence only when collaborating with the silent and transparent versions of RescueBot. Put differently, only participants who did not receive explanations for robot behavior significantly increased contribution during high interdependence. This suggests that interdependence only increases human contribution when the robot does not provide explanations for its behavior. Therefore, we speculate explanations are important to the coordination of work between team members and can diminish the effect of interdependence on contribution.

Number of Human Messages Sent

Another result of note is the interaction between robot communication style and interdependence on the number of human messages sent. Results show how people increase their amount of messages during high interdependence only when RescueBot also communicates. This suggests that only people collaborating with the communicating robots adjust their communication frequency according to interdependence, highlighting a surprising effect of robot communication on human communication.

System Understanding

Finally, we observe significantly higher understanding scores when RescueBot communicates. However, our results provide no evidence that adding explanations to transparency results in significantly higher understanding. This is not in line with other studies demonstrating how providing more reasoning information is related to increases in perceived understanding of the autonomous agent [83, 157].

3.5.2 Limitations and Future Work

Limitations

We identify a few limitations of our study. First, we chose to conduct the experiment in an online setting as this allowed us to simplify the task, remove robot-specific capabilities from the considerations, and keep a safe distance from participants during the global pandemic. This does mean that we did not have physical embodiment for our agent, as a robot would have. This embodiment might influence how much attention people pay to the agent when it is in sight [129]. On the other hand, as many tasks (including search and rescue) would always incorporate virtual messages due to distance between team mates, we do not expect it to change our main findings.

Furthermore, our simplified and simulated environment raises questions about the ecological validity. While this environment made it relatively easy to program our agents, environment, task, and communication protocols, it is hard to determine how well these generalize to real world scenarios. Currently, state of the art urban search and rescue robots or drones are not approaching the levels of autonomy and communication presented in our work. However, we believe that the rapid developments in the fields of Robotics and Artificial Intelligence will definitely allow these levels of autonomy and communication to be achieved.

Another limitation concerns the use of only attributive/causal explanations providing reasons for intentional behavior and actions of RescueBot. The absence of large differences between the transparent, explainable, and adaptive conditions could be the result of these explanations not adding enough additional information in our task and scenario. However, we used this explanation type as they could consistently be provided with each message sent by RescueBot without increasing message length dramatically. For example, confidence, contrastive, and counterfactual explanations could not be provided with each message (especially confidence explanations) or would increase message length considerably (contrastive and counterfactual explanations specifically).

Finally, our mixed design introduced some potential confounds such as learning and order effects. We actively tried to address these potential effects by including an extensive tutorial before participants started with the real experiment. This way, we tried to ensure all participants had similar and decent entry levels before starting the real experiment. We still tested for potential order effects by i) testing for differences in dependent variable outcomes between the two experiment order versions and ii) including interdependence condition order as a predictor in our regression model predicting team performance. Both analyses did not provide evidence for the presence of such order effects, as we did not find significant differences in outcome scores between the order conditions and order did not add significantly to the prediction of team performance (see Sections 3.4.1 and 3.4.3). Furthermore, we also tested for potential learning effects by testing for differences in the dependent variable outcomes between the two time points. Again, our analysis did not provide evidence for the presence of such effects (see Section 3.4.1), suggesting our tutorial worked as intended and our mixed experimental design did not introduce learning or order effects.

Future Work

We identify several possible directions for future work. We did not find large effects of adapting the message content, but this might still have an effect if done in other ways. Particularly, personalization by tailoring communication using an explicit user model and based on factors like user workload, trust in the agent, or understanding of the system. For example, the system could model the human using observations and human communication and adjust its information sharing accordingly. Other interesting contextual factors to investigate in future work include adapting information sharing based on different team member roles (e.g., supervisor vs. assistant) or team tasks. Another suggestion for future work is to add different explanation types such as confidence, contrastive, and counterfactual explanations and investigate their importance during human-robot teamwork on metrics such as trust and understanding.

In future work it could also be interesting to add bidirectional required dependencies to the human-robot team. In our current task and scenario, only RescueBot lacked capacity resulting in required dependencies. However, in future work, required dependencies stemming from a lack of human capacity could also be added, resulting in required support from

RescueBot (e.g., with removing obstacles). Furthermore, soft interdependence relationships could be added, for example when carrying a victim together would be faster than carrying alone. It would be interesting to examine how these scenarios affect trust in the system compared to our scenario of unidirectional required dependencies. As a final suggestion for future work, it could be relevant to look into more complex and realistic scenarios and environments which more closely resemble the current state of the art search and rescue robots. These suggestions are actually integrated in the empirical studies presented in Chapter 4, Chapter 5, and Chapter 7.

3.5.3 Conclusion

This chapter answers the second research sub-question of the thesis: *How do interdependence and agents' transparency and explanations influence effective human-agent teaming, individually and interactively*? Our study shows that the distinguished styles of robot communication result in more trust in and understanding of the robot, without increasing workload during the task. This highlights the fundamental importance of robots communicating their behavior to human teammates during teamwork. Furthermore, our findings show that robot explanations result in more reliance upon that robot, and that compared to sharing nothing, only explainability results in a higher situation awareness when interdependence is high. This highlights how robots providing explanations for their behavior can benefit human-robot teamwork. Finally, results demonstrate that being highly interdependent decreases trust, reliance, and team performance while increasing workload and situation awareness. It also increases human communication frequency when the robot communicates to its human teammate, human rescue contribution when the robot does not provide explanations, and the strength of the positive association between situation awareness and team performance. This underlines the crucial importance of carefully considering interdependence during studies on human-robot teamwork.

Overall, our results show that there are important differences between being transparent, explainable or adaptive in communications, but that the level of interdependence between human and robot is crucial in determining the exact effect that communication style has on human-robot teamwork. Our findings highlight the importance of interdependence on studies into explainability in robots, and provide an important first step in determining how a robot should communicate to its human teammates.

4

The Influence of Interdependence on Trust Calibration in Human-Machine Teams

In human-machine teams, the strengths and weaknesses of both team members result in dependencies, opportunities, and requirements to collaborate. Managing these interdependence relationships is crucial for teamwork, as it is argued that they facilitate accurate trust calibration. Unfortunately, empirical research on the influence of interdependence on trust calibration during human-machine teamwork is lacking. Therefore, we conducted an experiment (n=80) to study the effect of interdependence relationships (complete independence, complementary independence, optional interdependence, required interdependence) on human-machine trust calibration. Participants collaborated with a virtual agent during a simulated search and rescue task in teams characterized by one of the four interdependencies. A machine-induced trust violation was included in the task to facilitate dynamic trust calibration. Results show that the interdependence relationships during human-machine teamwork influence perceived trust calibration over time. Only in the teams with joint actions (optional and required interdependence) does perceived trust in the machine not recover to its initial pre-violated value. However, results show that the correlation between perceived trust in the machine and machine trustworthiness is strongest in these teams with joint actions, suggesting a more accurate trust calibration process. Overall, our findings provide some first evidence that interdependence relationships during human-machine teamwork influence human-machine trust calibration.

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

4.1 Introduction

Humans and intelligent machines increasingly work together as teammates on complex tasks such as manufacturing and firefighting [233]. Machines often outperform humans concerning rapid, rational, and repetitive decision-making, whereas humans are usually better at handling uncertainty and unexpected situations [232]. These separate strengths and weaknesses of humans and machines result in different dependencies, opportunities, and requirements to collaborate [99]. The ultimate goal of human-machine teams is to harness the combination of strengths of both humans and machines to accomplish what neither can do alone [3].

Several factors determine the success of human-machine teams, for example, effectively managing the interdependence relationships between both team members [101]. Another crucial determinant is appropriate human trust in machines, meaning that they know both the potentials and limitations of machines [141, 156]. A lack of appropriate trust (i.e., over- or under-trust) is one of the main reasons for the disuse and misuse of machines. This lack can be corrected by a trust calibration process over time and repeated interactions, allowing humans to adjust their expectations of the machine's reliability and trustworthiness [117, 141, 156]. During the trust calibration process, repairing trust violations caused by machine errors is more difficult than building trust initially [105, 109].

It is argued that interdependence relationships between humans and machines facilitate the assessment of trustworthiness of intelligent machines and accurate trust calibration by humans [97]. However, there is a lack of empirical research on the exact influence of interdependence on trust calibration in human-machine teams. For example, how different interdependence relationships during human-machine teamwork influence the trust calibration process over time is unknown. Therefore, this study investigates how complete independence, complementary independence, optional interdependence, and required interdependence influence human-machine trust calibration. To do this, we conducted a user study where participants collaborated with a virtual agent during a simulated search and rescue task in teams characterized by one of the four interdependencies.

4.2 Background

4.2.1 Interdependence in Human-Machine Teams

Interdependence relationships are the complementary relationships humans and machines rely on to manage dependencies during joint activities [99, 192]. Joint activities concern situations in which the actions of humans depend on those of machines (and vice versa) over a sustained sequence of actions and towards a shared goal [99]. These joint activities are characterized by required, optional, complementary, or no dependencies between humans and machines, caused by their capabilities to execute actions individually and assist each other during action execution [99].

When humans and machines can both execute actions independently while collaborating towards a shared goal, they are hardly dependent on each other. On the other hand, complementary dependencies between humans and machines exist when each can only execute their unique actions that contribute to completing the overall task. Optional dependencies stem from recognizing opportunities to be more efficient when executing actions jointly rather than independently [99, 232]. Finally, required dependencies originate from both team members' lack of knowledge, skills, abilities, and resources to competently execute an action independently, but the potential to assist each other to execute the action jointly [99, 232]. This distinction between complete independence, complementary independence, optional interdependence, and required interdependence essentially forms a hierarchy in coordination, dependencies, and strength of the interdependence relationship [99, 232]. As these different interdependence relationships heavily affect mutual reliance and dependencies, they play a critical role in the trust relationship between humans and machines [97].

4.2.2 Trust in Human-Machine Teams

An early definition of trust is believing that someone or something else will act in your best interest and accepting vulnerability to this person's or entity's actions [138]. So, there is a trusting party (the trustor) and a party to be trusted (the trustee) [138]. Here, trust can be considered as the trustor's perception of the trustee's trustworthiness [102, 141]. Trust is critical in all circumstances where people are in any way dependent on other's actions, and thus more relevant in high-risk situations [138, 141]. More specifically, more trust is required when the perceived risk of relying on someone or something else is higher [138]. We believe that interdependence influences the perceived risk associated with relying on someone or something else and, thus, indirectly, how much trust is required during the relationship. For example, relying on someone who can execute actions you can not is less risky than relying on someone to execute actions jointly.

Instead of blindly trusting machines, human-machine trust must be appropriate [141]. Human-machine trust is appropriate when the human's trust in the machine is equal to the machine's actual trustworthiness [141, 182]. This match between trust and trustworthiness involves both trusting trustworthy machines and distrusting untrustworthy machines. When appropriate trust is directly caused by information about the actual trustworthiness of the machine, this is called warranted appropriate trust [70, 94]. Fostering appropriate trust is crucial as a lack of appropriate human-machine trust can cause over- or under-trust in and over- or under-reliance on machines, potentially resulting in detrimental outcomes [118, 141, 156, 158]. Fostering appropriate trust involves a process of trust calibration that corrects for over- and under-trust over time and repeated interactions, allowing humans to adjust their expectations of the machine's reliability and trustworthiness [117, 141, 156].

During the trust calibration process, human-machine trust is rarely stable but instead changes over time based on past and current interactions [50, 65, 81, 121]. Decreases in human-machine trust resulting from machine-induced trust violations can have lasting effects and are hard to recover from [65, 121]. To this end, machines can deploy several trust repair strategies to repair human trust after they damage or violate it [65, 67, 109, 110]. The most commonly used trust repair strategies include apologies, denials, explanations, and promises [66, 121, 185]. The impact of these trust repair strategies on human trust has been mixed, with studies showing positive, no, or even negative effects [66, 167]. Moderating factors might explain these mixed results, such as the timing of the repair strategy, violation type, and violation severity [66, 109]. One general result, however, seems to be the effectiveness of machine apologies for restoring trust [109, 160, 253]. Adding an explanation to the apology can even amplify this effect [109, 186].

Explanations are not merely a trust repair strategy but also one of the primary methods for fostering appropriate human-machine trust. They specifically aim to make intelligent machines more transparent and understandable to humans [141, 145, 231]. Examples include machine explanations, confidence scores, and uncertainty communication, providing information about the capabilities and limitations of machines and how and why they make decisions [34, 202, 257]. Prior literature has shown that these forms of machine transparency can improve appropriate trust in machines [34, 142, 187, 202, 257].

4.2.3 Interdependence for Trust Calibration in Human-Machine Teams

In addition to machine explanations, it is argued that interdependence relationships also play a critical role in the trust calibration process [97]. In order to do so, interdependence relationships need to be supported by observable, predictable, and directable machines [97, 99]. This means that intelligent machines should be transparent and understandable enough for humans to reasonably rely on them while also allowing humans to influence their behavior [97, 99]. This way, interdependence relationships can support the active and continuous exploration of trust between humans and machines to ensure that human assessments are appropriate for achieving the best possible outcomes [97].

As both trust and interdependence relationships involve risk, reliance, and dependencies, it is unsurprising that interdependence and trust are related [192]. Johnson and Bradshaw [97] argue that interdependence relationships facilitate the assessment of the trustworthiness of the machine and accurate trust calibration required for developing warranted appropriate trust. However, interdependence relationships between humans and machines can vary in terms of coordination and dependencies, such as required or optional dependencies during joint activities [99, 232]. So far, there is a lack of empirical research on how these different interdependence relationships during human-machine teamwork influence human-machine trust calibration. Our study will fill that gap by comparing how complete independence, complementary independence, optional interdependence, and required interdependence influence human-machine trust calibration.

4.3 Method

4.3.1 Design

We conducted an experiment to investigate the influence of interdependence relationships during human-machine teamwork on human-machine trust calibration. To ensure a dynamic trust calibration process, we added a trust violation caused by incorrect machine advice. The experiment had a 3x4 mixed design with time as the within-subjects independent variable and interdependence as the between-subjects independent variable. Time consisted of three conditions (pre-violation, post-violation, post-recovery) and interdependence of four conditions (complete independence, complementary independence, optional interdependence, and required interdependence). As dependent variables, we measured perceived trust and the appropriate reliance rate at each of the three time points.

4.3.2 Participants

We recruited 80 participants through personal contacts within the university (29 female and 51 male participants). Sixty-nine participants had an age range of 18-24 years old, seven participants of 25-34 years old, one participant of 35-44 years old, two participants of 45-54 years old, and one participant of 55-64 years old. In terms of education, two participants went to high school but did not obtain a diploma, 44 participants were high school graduates, nine participants obtained some college credit but no degree (yet), one participant obtained an Associate degree, 19 participants obtained a Bachelor's degree, and five participants obtained a Master's degree. Concerning gaming experience, 11 participants had no experience at all, 19 participants had a little, nine participants had a moderate amount, 22 participants had a considerable amount, and 19 participants had a lot. All participants signed an informed consent form before participating in the study, approved by the ethics committee of our institution (ID 3002). Since each participant was assigned to one of the four interdependence conditions, it was essential to control for gender, age, education, and gaming experience between these conditions. Results showed no significant differences between interdependence conditions for any of the demographic factors gender (χ^2 (3) = 3.62, $p = 0.31$), age ($W = 1.23$, $p = 0.75$), education ($W = 3.94$, $p = 0.27$), and gaming experience ($W = 0.86$, $p = 0.84$). Therefore, we did not further control for these demographics during data analysis.

4.3.3 Hardware and Software

To run this experiment, we used a laptop and the Human-Agent Teaming Rapid Experimentation (MATRX) software, a Python package for facilitating human-agent teaming research (`https://matrx-software.com/`). The laptop was used to launch our two-dimensional grid world created using MATRX. All subjective measures were collected using Qualtrics, while all objective measures were automatically logged using MATRX.

4.3.4 Environment

We built a MATRX world consisting of 14 areas, 26 collectable objects, 12 obstacles, and one drop zone (see Figure 4.1 for part of the world). Furthermore, we added an autonomous virtual agent (RescueBot) and a human agent (controlled by the participants) to our world. We designed an environment in which these two agents had to collaborate during a search and rescue task. To ensure an inclusive and realistic victim representation, we created the following eight victim types making up the world's collection goal: girl, boy, woman, man, older woman, older man, cat, and dog. In addition, we created three injury types: critical, mild, and healthy. Injury type was represented by the color of the victims, where red reflected critically injured, yellow mildly injured, and green healthy victims. Eight of the 26 victims were either mildly or critically injured and had to be delivered at the drop zone, whereas the other 18 were healthy. We also added three obstacle types in front of area entrances: boulder, tree, and stone. Finally, we added flooded water to the environment, which slowed the agents' speed as they moved through it.

4.3.5 Task

The objective of the task was to find the target victims in the different areas and carry them to the drop zone. Interdependence relationships between humans and RescueBot were manipulated, resulting in four conditions characterized by unique dependencies [99]. In the complete independence condition, the human and RescueBot could execute all actions independently (i.e., remove all obstacles and rescue all victims). In the complementary

Table 4.1: Overview of the advice and feedback messages provided by RescueBot during the experiment.

Type	Content
A1,3	I have detected extreme rain arriving soon and predict it will cause new floods. I advise you to take shelter in one of the areas ASAP, until the rain is over.
F1,3	My advice was correct, that weather was extreme! If you had (not) taken shelter, you would (not) have lost time due to injuries and 10 points of our score.
A2	I have detected light rain arriving soon but predict it will cause no floods. I advise you to continue searching and rescuing victims.
F2	My advice was wrong. The amount of rain was heavy instead of light. Because of that my flood prediction was incorrect. I am really sorry.

independence condition, RescueBot could only remove obstacles, whereas the human could only rescue victims. The other two conditions also included joint actions. In the optional interdependence condition, the human and RescueBot could execute all actions independently and jointly. However, joint action execution was four times faster than independent action execution. In the required interdependence condition, all actions had to be executed jointly. Independently removing obstacles took four seconds for stones, eight seconds for trees, and 12 seconds for boulders. Independently rescuing victims took four seconds for mildly injured victims and eight seconds for critically injured victims. Participants had ten minutes to complete the task (i.e., drop all victims at the drop zone) and received points for each victim they rescued. Rescuing critically injured victims added six points to the total score, while rescuing mildly injured victims added three points, resulting in a maximum possible score of 36 points. Other than points and rescue time, no other differences existed between mildly and critically injured victims.

During the task, extreme rain hit the MATRX world three times: after two, four, and six minutes. This rain lasted for ten seconds and if participants did not seek shelter in one of the areas during the rain, they would lose ten points of their score and their avatar would freeze until the rain disappeared. The extreme rain merely affected score and time; it did not affect the victims to be rescued. Before the extreme rain, RescueBot warned the participants about its severity and correspondingly recommended seeking shelter or continuing with the search and rescue task. Each message was accompanied by a ping sound and color highlights to draw attention. After the rain disappeared, RescueBot provided feedback on whether the advice was correct, and more flooded water was added to the environment. RescueBot's first advice was correct. In contrast, RescueBot's second advice was incorrect, provoking a trust violation. Therefore, the following feedback message contained a trust repair message explaining what happened and expressing regret [109]. We included this element of risk to the task because risk and vulnerability are critical elements of trust [141]. RescueBot's third recommendation was correct again. Table 4.1 shows all the advice and feedback messages provided by RescueBot.

4.3.6 Agent Types

We added two agents to the world: an autonomous rule-based virtual agent (RescueBot) and a human agent controlled by the participants using their keyboards. RescueBot always moved to the closest unsearched area during the search and rescue task. Furthermore, it tracked which areas the team had searched, which victims the team had found and where, and which victims the team had rescued. RescueBot did not execute any removing or rescuing actions autonomously. Instead, it asked the participants to decide whether to remove obstacles or rescue victims independently or jointly, accompanied by a summary of the explored areas, found victims, and rescued victims (see Figure 4.1). This way, RescueBot's behavior was consistent for all interdependence conditions.

Both agents could only carry one victim at a time (either independently or jointly), detect each other within two grid cells, detect and remove obstacles or pick up victims within one grid cell, and detect walls and doors from anywhere. Both agents could also communicate using the chat box shown in Figure 4.1. Using buttons, participants could share their actions, perceptions, assistance requests, and answers to any questions asked by RescueBot. RescueBot added the shared information to its memory and adjusted its behavior correspondingly (e.g., by not moving to the same areas as the participants).

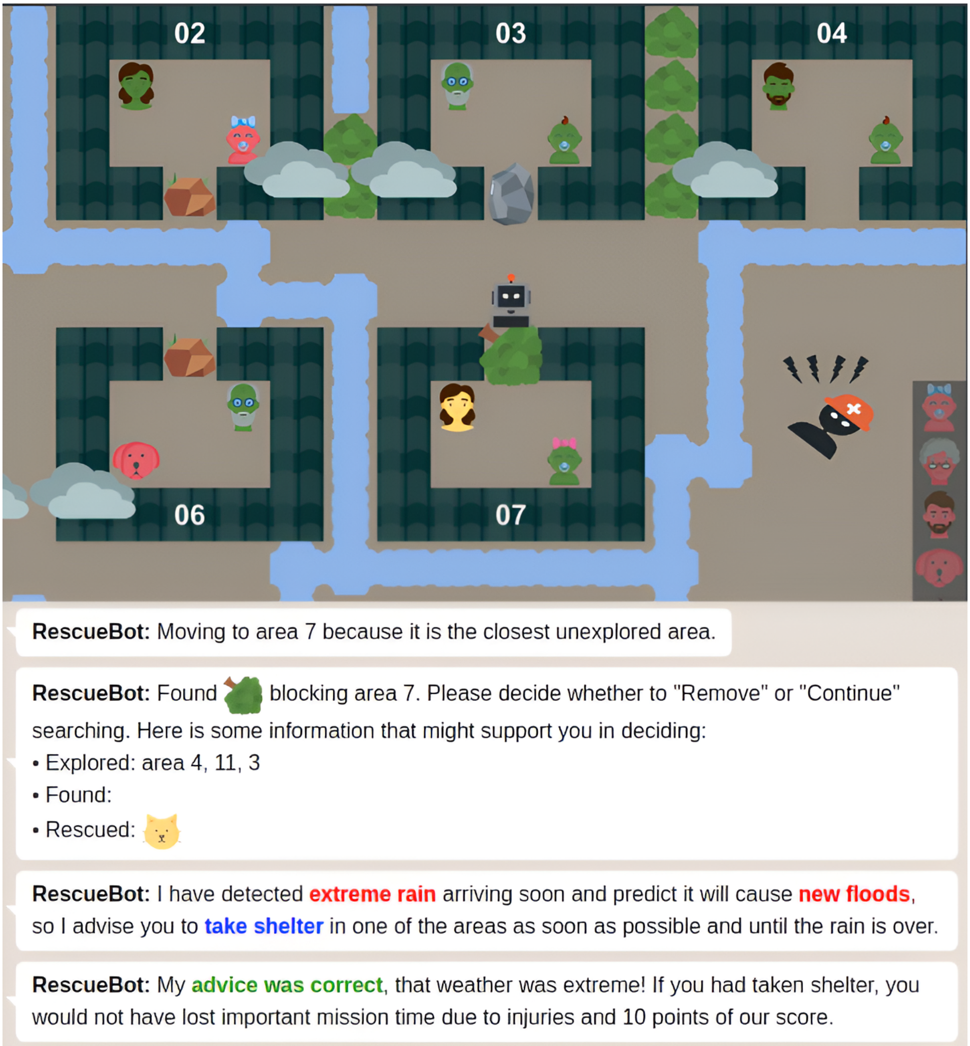
4.3.7 Measures

We used self-reporting and behavior to measure perceived trust in and demonstrated reliance on RescueBot [141]. More specifically, we subjectively measured perceived user trust in RescueBot using the 5-point Likert scale for trust in explainable artificial intelligence systems [92]. This scale consisted of eight items and measured confidence in and predictability, reliability, safety, efficiency, wariness, performance, and likeability of RescueBot. We calculated the mean of these eight items as the final perceived trust score for each of the three time points separately.

In addition, we objectively logged whether participants followed the advice given by RescueBot. Based on this data, we calculated the appropriate reliance rate on RescueBot. *Appropriate reliance* was defined as appropriate reliance on RescueBot's correct advice at T1 and T3 and appropriate non-reliance on RescueBot's incorrect advice at T2. Accordingly, we calculated the appropriate reliance rate at each time point by dividing the number of appropriate (non-)reliance occurrences by the number of received recommendations so far. This way, the appropriate reliance rate was a cumulative variable.

4.3.8 Procedure

Participants first completed a tutorial to familiarize them with the environment, controls, and messaging system. Next, participants started the actual experiment. After one minute and 45 seconds, RescueBot warned the participants about arriving rain and whether to seek shelter. After two minutes, the rain arrived and lasted for ten seconds. When the rain disappeared, RescueBot provided feedback on whether its advice was correct. After two minutes and 20 seconds, the game paused, and participants were asked to fill out the trust questionnaire for the first time. This cycle of warning, rain, feedback, and trust questionnaire was repeated two more times with similar intervals, with the other warnings arriving at three minutes and 45 seconds and five minutes and 45 seconds. The whole study lasted about 30 minutes and was conducted offline.



RescueBot: Moving to area 7 because it is the closest unexplored area.

RescueBot: Found 🌳 blocking area 7. Please decide whether to "Remove" or "Continue" searching. Here is some information that might support you in deciding:

- Explored: area 4, 11, 3
- Found:
- Rescued: 🐱

RescueBot: I have detected **extreme rain** arriving soon and predict it will cause **new floods**, so I advise you to **take shelter** in one of the areas as soon as possible and until the rain is over.

RescueBot: My **advice was correct**, that weather was extreme! If you had taken shelter, you would not have lost important mission time due to injuries and 10 points of our score.

Figure 4.1: Experimenter view of the MATRX world used for our study.

4.4 Results

4.4.1 Perceived Trust and Appropriate Reliance

To investigate the effects of interdependence and time on perceived trust in RescueBot (Figure 6.3A), we conducted both a parametric and nonparametric mixed ANOVA. We conducted both ANOVAs because the assumption of homogeneity of variances for the parametric mixed ANOVA was slightly violated at T3. Results of the parametric mixed ANOVA showed a statistically significant interaction between interdependence and time on perceived trust ($F(6, 152) = 2.83$, $p < 0.025$, $\eta^2_G = 0.042$). Results showed that the simple main effect of interdependence on perceived trust was not significant at any of the time points. In contrast, results showed that the simple main effect of time on perceived trust was significant for complete independence ($F(2, 38) = 11.1$, $p < 0.001$, $\eta^2_G = 0.18$), complementary independence ($F(2, 38) = 9.45$, $p < 0.005$, $\eta^2_G = 0.16$), optional interdependence ($F(1.38, 26.2) = 35.6$, $p < 0.001$, $\eta^2_G = 0.37$), and required interdependence ($F(1.27, 24.2) = 35.4$, $p < 0.001$, $\eta^2_G = 0.50$). Pairwise t-test comparisons using a Bonferroni correction revealed significant differences in trust scores between all time points and for all interdependencies, except between T1 and T3 for complete independence and complementary independence (Table 4.2 and Table 4.3).

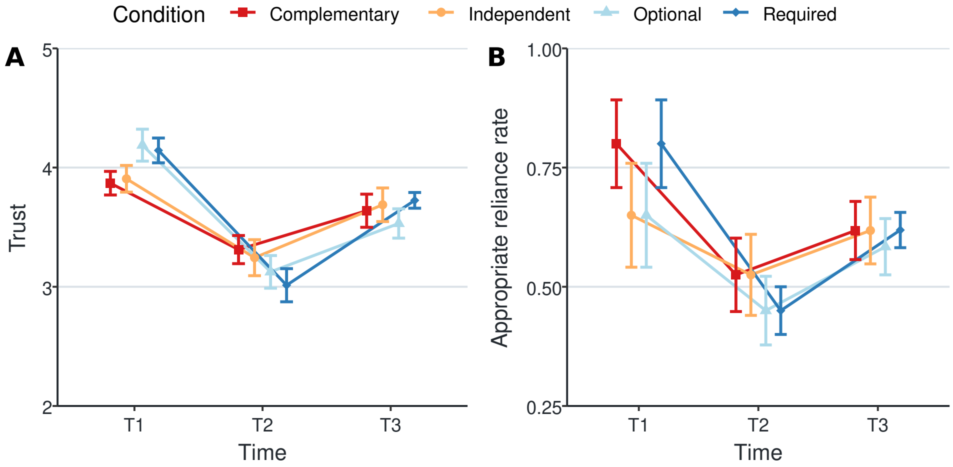


Figure 4.2: Interaction plots of the effects of interdependence and time on perceived trust (A) and the appropriate reliance rate (B). Error bars represent the standard errors.

To confirm these results, we ran the nonparametric rank-based mixed ANOVA [151]. Again, results showed a statistically significant interaction between interdependence and time on perceived trust ($F(4.56) = 2.29$, $p < 0.05$, effect size = 0.44). These results also showed that the simple main effect of interdependence was not significant at any of the time points. Moreover, the results again showed that the simple main effect of time on perceived trust was significant for complete independence ($\chi^2(2) = 13.40$, $p < 0.0025$, $W = 0.36$), complementary independence ($\chi^2(2) = 14.50$, $p < 0.001$, $W = 0.34$), optional interdependence ($\chi^2(2) = 30.30$, $p < 0.001$, $W = 0.76$), and required interdependence ($\chi^2(2)$

Table 4.2: Pairwise t-test and Wilcoxon comparisons for the simple main effect of time on perceived trust for each interdependence condition. Bold values show the non-significant pairwise comparisons.

Condition	Time points	Δ mean	t	p	W	p
Complete independence	T1 vs. T2	-0.67	3.98	< 0.005	129	< 0.01
	T1 vs. T3	-0.22	1.66	**0.34**	109	**0.39**
	T2 vs. T3	+0.45	-3.46	< 0.001	21	< 0.01
Complementary independence	T1 vs. T2	-0.56	4.02	< 0.005	172	< 0.01
	T1 vs. T3	-0.23	1.59	**0.38**	146	**0.39**
	T2 vs. T3	+0.33	-3.37	< 0.025	23	< 0.05
Optional interdependence	T1 vs. T2	-1.06	6.50	< 0.001	207	< 0.001
	T1 vs. T3	-0.66	7.10	< 0.001	208	< 0.001
	T2 vs. T3	+0.40	-3.53	< 0.01	16	< 0.01
Required interdependence	T1 vs. T2	-1.13	6.35	< 0.001	210	< 0.001
	T1 vs. T3	-0.41	4.71	< 0.001	150	< 0.01
	T2 vs. T3	+0.72	-5.65	< 0.001	0	< 0.001

Table 4.3: Descriptive statistics for each combination of time and interdependence condition. M refers to the mean, MR to the mean rank, SD to the standard deviation, and AR% to the appropriate reliance rate.

Condition	Time	M (SD) trust	MR (SD) trust	M (SD) AR%	MR (SD) AR%
Complete independence	T1	3.91 (0.50)	153.43 (61.72)	0.65 (0.49)	136.08 (85.88)
	T2	3.24 (0.68)	82.23 (66.69)	0.53 (0.38)	103.68 (68.37)
	T3	3.69 (0.63)	128.00 (70.54)	0.62 (0.31)	121.80 (60.06)
Complementary independence	T1	3.87 (0.44)	147.40 (55.40)	0.80 (0.41)	162.40 (72.02)
	T2	3.31 (0.53)	85.05 (51.26)	0.53 (0.34)	101.35 (62.34)
	T3	3.64 (0.62)	122.95 (66.19)	0.62 (0.27)	120.10 (55.08)
Optional interdependence	T1	4.19 (0.60)	178.25 (64.27)	0.65 (0.49)	136.08 (85.88)
	T2	3.13 (0.61)	70.58 (55.53)	0.45 (0.32)	87.03 (55.14)
	T3	3.53 (0.55)	110.93 (59.77)	0.58 (0.26)	112.88 (53.85)
Required interdependence	T1	4.14 (0.47)	177.78 (54.32)	0.80 (0.41)	162.40 (72.02)
	T2	3.01 (0.62)	60.60 (58.84)	0.45 (0.22)	82.38 (35.88)
	T3	3.73 (0.30)	128.83 (39.17)	0.62 (0.17)	119.85 (37.21)

= 35.7, $p < 0.001$, $W = 0.89$). Finally, pairwise Wilcoxon comparisons using a Bonferroni correction also revealed significant differences in trust scores between all time points and for all interdependencies, except between T1 and T3 for complete independence and complementary independence (Table 4.2 and Table 4.3).

To investigate the effects of interdependence and time on the appropriate reliance rate (Figure 6.3B), we conducted the nonparametric mixed ANOVA because of not normally distributed data. Results showed a significant main effect of time on the appropriate reliance rate ($F(1.35) = 48.06$, $p < 0.001$, effect size = 1.10). Pairwise Wilcoxon comparisons using a Bonferroni correction revealed significant differences between the appropriate reliance rates at T1 and T2 ($p < 0.001$) and T2 and T3 ($p < 0.001$).

4.4.2 Effects of Interdependence on Reliance and Injuries

Next, we investigated if the interaction between interdependence and time on perceived trust (Figure 6.3A) could be explained by differences between interdependence conditions in the number of injuries or how much they relied on RescueBot. Here, the underlying assumptions were that more reliance could result in more trust [141], and more injuries (and thus lost points) in less trust. However, the already reported nonparametric mixed ANOVA only showed a significant main effect of time on the appropriate reliance rate. Results of another nonparametric mixed ANOVA also showed a non-significant interaction effect of interdependence and time on the general reliance rate ($F(3.95) = 0.83$, $p = 0.51$, effect size = 0.26), and non-significant main effect of interdependence on the general reliance rate ($F(2.96) = 1.77$, $p = 0.15$, effect size = 0.26). Finally, results showed that all interdependence conditions were homogeneous concerning how often they were injured by the rain (χ^2 (3) = 0.21, $p = 0.98$), also at T1 (χ^2 (3) = 2.26, $p = 0.52$), T2 (χ^2 (3) = 4.80, $p = 0.19$), and T3 separately (χ^2 (3) = 2.35, $p = 0.50$).

4.4.3 Accuracy of the Trust Calibration Process

Finally, for each interdependence condition, we compared the trust calibration process over time with RescueBot's actual trustworthiness over time, expressed in terms of its advice accuracy [49, 141]. More specifically, RescueBot's advice accuracy was 100% at T1, 50% at T2, and 67% at T3. For each interdependence condition, we ran a Spearman's rank-order correlation to assess the relationship between perceived trust in RescueBot and advice accuracy of RescueBot. Results showed a statistically significant positive correlation between perceived trust and advice accuracy for complete independence ($\rho = 0.42$, $p < 0.001$), complementary independence ($\rho = 0.40$, $p < 0.005$), optional interdependence ($\rho = 0.60$, $p < 0.001$), and required interdependence ($\rho = 0.69$, $p < 0.001$).

4.5 Discussion and Conclusion

4.5.1 Discussion

Our results show that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time (Figure 6.3A). Across all interdependence relationships, we observe significant post-violation trust decreases compared to pre-violated trust (T2 vs. T1) and significant post-recovery trust repairs compared to post-violated trust (T3 vs. T2). However, only in the teams with joint actions (optional and required interdependence) we observe a significant post-recovery trust decrease compared to pre-violated trust (T3 vs. T1). In other words, human-machine trust does not recover to its initial pre-violated value only in the teams with joint actions (Section 4.4.1). Since we do not find evidence for an influence of interdependence on reliance or the number of injuries (Section 4.4.2), this finding can more likely be attributed to the direct influence of interdependence relationships on human-machine trust calibration.

The results further indicate that the correlation between perceived trust in RescueBot and RescueBot's advice accuracy is significant for all interdependence relationships but strongest for the teams with joint actions (Section 4.4.3). This finding supports Johnson and Bradshaw's claim [97] that interdependence facilitates accurate trust calibration. However, it also extends the claim by showing that stronger interdependence relationships with joint actions facilitate more accurate trust calibration aligning with RescueBot's trustworthiness. This might explain why human-machine trust does not recover to its initial pre-violated value in the teams with joint actions.

We believe that the perceived risk associated with relying on machines [138] increases with the strength of the interdependence relationship, and therefore, more trust is necessary for human-machine teams with joint actions. Prior research has shown that under such conditions of increased trust necessity, over-trust can be promising for trust calibration [42, 141]. Therefore, we speculate that over-reliance on the incorrect advice at T2 resulted in a more accurate trust calibration in the teams with higher trust necessity caused by joint actions. This might also explain why the stronger interdependence relationships with joint actions facilitate more accurate trust calibration aligning with RescueBot's trustworthiness. However, follow-up research is required to support these hypothesized relationships between interdependence, risk, (over-)reliance, and trust (necessity).

Finally, we did not find evidence of an effect of interdependence on the calibration of appropriate human-machine reliance. However, timing was an important distinction between perceived trust and the appropriate reliance rate, as perceived trust was recorded after the consequences of reliance behavior. Therefore, it made little sense to compare the calibration of appropriate reliance with RescueBot's actual trustworthiness over time, as participants could not make an informed estimate of its accuracy at T1. All in all, our results highlight that interdependence relationships are crucial to consider carefully in human-machine teams as they can influence perceived human-machine trust calibration.

4.5.2 Limitations and Future Work

We identify a few limitations of our study. First, we only used three time points to reflect human-machine trust calibration over time, which is a simplified representation. Even though this representation aided in capturing some critical aspects of the calibration process, the limited temporal scope probably did not capture all nuanced aspects of trust calibration over time. Therefore, future research could increase the temporal scope of the study, facilitating a more detailed investigation of the trust calibration process.

Furthermore, we used four distinctive interdependence relationships for our interdependence conditions. Again, this is a simplified representation of human-machine collaboration, which is often characterized by a mix of all four relationships [99, 230]. However, using these four distinctive relationships allowed us to examine their unique influence on trust calibration. Even though human-machine teamwork often involves a mix of all interdependencies, our results still provide developers with crucial insights. For example, how violated trust does not recover to its initial value for teams engaged in joint actions and that these teams demonstrate a more accurate trust calibration.

We identify several directions for future work. For example, investigating the interaction between interdependence and trust repair strategy on trust calibration. We speculate that specific repair strategies work better for certain interdependencies, such as promises for relationships with joint actions and explanations for independent collaboration. Future work could test these hypotheses by extending our research environment with different trust repair strategies [66, 121, 185]. These results could provide valuable insights allowing machines to adapt their trust repair strategies based on interdependence.

Another suggestion for future work is studying the interaction between interdependence and violation severity on trust calibration. We speculate that more severe violations will result in higher trust decreases for teams engaged in joint actions. Future work could test these hypotheses by extending our research environment to include trust violations of different severity levels, such as machine failure during action execution and incorrect machine advice. These results could provide valuable insights for developing machines adapting to interdependence relationships to address trust calibration challenges.

4.5.3 Conclusion

This chapter answers the third research sub-question of the thesis: *How do interdependencies during human-agent teaming influence the human-agent trust calibration process*? Our study shows that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time. During a simulated search and rescue task with a machine-induced trust violation, only in teams with joint actions does perceived trust in the machine not recover to its initial pre-violated value. However, our findings show that the correlation between perceived human-machine trust and machine trustworthiness is strongest in these teams with joint actions. This suggests that these stronger interdependence relationships during human-machine teamwork facilitate more accurate human-machine trust calibration. Overall, our study presents some first evidence that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time. Therefore, it is crucial to consider these relationships carefully during human-machine trust calibration and to conduct follow-up research on adapting trust repair strategies to interdependence.

5

Meaningful Human Control and Variable Autonomy in Human-Robot Teams for Firefighting

Humans and robots are increasingly collaborating on complex tasks such as firefighting. As robots are becoming more autonomous, collaboration in human-robot teams should be combined with meaningful human control. Variable autonomy approaches can ensure meaningful human control over robots by satisfying accountability, responsibility, and transparency. To verify whether variable autonomy approaches truly ensure meaningful human control, the concept should be operationalized to allow its measurement. So far, designers of variable autonomy approaches lack metrics to systematically address meaningful human control. Therefore, this qualitative focus group (n = 5 experts) explored quantitative operationalizations of meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting. This variable autonomy approach requires dynamic allocation of moral decisions to humans and non-moral decisions to robots, using robot identification of moral sensitivity. We analyzed the data of the focus group using reflexive thematic analysis. Results highlight the usefulness of quantifying the traceability requirement of meaningful human control, and how situation awareness and performance can be used to objectively measure aspects of the traceability requirement. Moreover, results emphasize that team and robot outcomes can be used to verify meaningful human control but that identifying reasons underlying these outcomes determines the level of meaningful human control. Based on our results, we propose an evaluation method that can verify if dynamic task allocation using variable autonomy in human-robot teams for firefighting ensures meaningful human control over the robot. This method involves subjectively and objectively quantifying traceability using human responses during and after simulations of the collaboration. In addition, the method involves

🗎 **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Meaningful human control and variable autonomy in human-robot teams for firefighting. *Frontiers in Robotics and AI* 11 (2024).

semi-structured interviews after the simulation to identify reasons underlying outcomes and suggestions to improve the variable autonomy approach.

5.1 Introduction

Humans and robots are increasingly working together in human-robot teams on complex tasks ranging from medical surgery to firefighting. For example, the fire department of Rotterdam in the Netherlands is already using explore and extinguish robots for situations too dangerous for firefighters. Several factors determine the success of these human-robot teams, such as situation awareness, mutual trust, and common ground [107, 175]. The ultimate goal of human-robot teams is harnessing the combination of strengths of both humans and robots, to accomplish what neither can do alone [3]. Such an integration of robots that augment rather than replace humans requires robots to dynamically vary their level of autonomy to collaborate with humans efficiently.

Rapid developments in the field of robotics and artificial intelligence allow robots to become increasingly autonomous and perform tasks without much human intervention and control [179]. However, since robots do not have a legal position, humans should be held accountable in case robot behavior does not comply with moral or ethical guidelines [214]. Therefore, higher levels of robot autonomy should be combined with meaningful human control and human moral responsibility [178, 179]. The concept of meaningful human control is based on the assumption that human persons and institutions should ultimately remain in control of, and thus morally responsible for, the behaviour of intelligent autonomous systems like robots [179]. Meaningful human control originated from the discussion on autonomous weapon systems but its relevance quickly expanded to intelligent (semi)autonomous systems in general.

One of the first works on meaningful human control was a philosophical account towards two necessary conditions: the tracking and tracing conditions. In short, the tracing condition implies that “a system’s behaviour, capabilities, and possible effects should be traceable to a proper moral and technical understanding of at least one relevant human agent who designs or interacts with the system” [30, 179]. On the other hand, the tracking condition implies that a system should be responsive to the human moral reasons relevant to specific circumstances [179]. Designing for meaningful human control means designing for human moral responsibility and ensuring humans are aware and equipped to act upon their moral responsibility. By doing so, responsibility gaps in culpability, moral and public accountability, and active responsibility can be avoided [30, 57, 178, 225]. Several solutions for addressing meaningful human control in human-robot teams have been proposed, such as team design patterns [217], value sensitive design [75], machine ethics [6], and variable autonomy [144].

Variable autonomy refers to the ability to dynamically adjust the levels of autonomy of a system, for example by switching the level of autonomy from full robot autonomy to complete human operator control [36]. Variable autonomy is often used to describe human-robot teams in which the level of robot autonomy varies depending on the context. One of the main goals of variable autonomy approaches is to maximise human control without burdening the human operator with an unmanageable amount of detailed operational decisions [37, 247]. For example, in human-robot teams for firefighting variable autonomy can be used to dynamically allocate moral decision-making to humans and non-moral decision-making to robots. It is argued that robots with variable autonomy can ensure meaningful human control over these robots by satisfying accountability, responsibility, and transparency [144]. However, testing whether variable autonomy approaches truly ensure meaningful human control is crucial before actually adopting them. Unfortunately, designers of variable autonomy approaches lack metrics needed for systematically addressing meaningful human control [27, 57]. On the other hand, meaningful human control is already increasingly being imposed as a requirement for variable autonomy approaches.

Imposing meaningful human control as a requirement and verifying if variable autonomy approaches indeed fulfill this requirement means we must be able to measure meaningful human control [214]. Therefore, turning the abstract concept of meaningful human control into measurable observations (i.e., operationalize) is required. So far, only a few approaches for operationalizing meaningful human control exist [25, 214]. Therefore, this qualitative study explores different approaches to measure meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting, aimed at creating an evaluation method. We will first discuss the context and variable autonomy approach in more detail, as well as existing operationalizations of meaningful human control (Section 5.2). Next, we will discuss how we conducted our study (Section 5.3), followed by the results (Section 5.4). Finally, we will present a discussion, propose an evaluation method of meaningful human control, and conclude our work (Section 5.5).

5.2 Background

5.2.1 Moral decisions in human-robot teams for firefighting

Explore and extinguish robots are increasingly collaborating with firefighters to detect victims and extinguish fires in properties too dangerous for firefighters, for example because the structural condition is unsafe. Currently, firefighting robots are mostly teleoperated by firefighters, allowing an otherwise impossible offensive inside deployment aimed at fighting the fire and rescuing people. These firefighting robots are equipped with several cameras (thermal imaging, RGB, pinhole), sensors (LIDAR, temperature, explosion danger), and capabilities (water shield protection, fire hose), enabling navigation, localization, detection, protection, mapping, and extinguishing. The information provided by the robot's sensors is crucial for firefighters to make decisions about localizing the fire source, rescuing victims, switching deployment tactic, extinguishing or evacuating, and sending in firefighters to help. The collaboration between firefighters and their firefighting robot demonstrates how human-robot teams can harness the combination of strengths of both parties, to accomplish what neither could do alone.

Although these teleoperated firefighting robots are already of great use, there is a strong preference within the field of rescue robotics for (semi)autonomous robot behavior to reduce the workload of the operator [52]. The potential of artificial intelligence provides great opportunities for making these robots more autonomous, and some progress has already been made [74, 112, 113]. It is considered important, however, to actively keep a human involved in the collaboration to guide the robot's behavior during the mission [52]. Variable autonomy will be crucial to effectively implement this collaboration between a human operator and increasingly autonomous firefighting robot because it can increase human control while decreasing the workload of the human operator [37, 247].

This increase in robot autonomy raises important challenges such as how to design for meaningful human control in these human-robot teams. Designing for meaningful human control is crucial in human-robot teams for firefighting because the scenario involves morally sensitive situations (i.e., situations in which something one might do or is doing can affect the welfare, rights, and values of someone else either directly or indirectly [162]). These morally sensitive situations can involve deciding to preserve the safety of firefighters if that means the life of victims cannot be rescued. If the robot would autonomously make an incorrect moral decision in such situations, consequences could be the loss of lives and responsibility gaps [178]. Therefore, especially when human-robot teams are tasked with making moral decisions, meaningful human control is crucial to ensure humans can be held accountable for robot behavior [214].

Team designs patterns have been applied to describe the allocation of tasks for moral decision-making in human-robot teams [213, 219]. These patterns can express forms of collaboration with various team properties by task-independently describing (1) how humans and robots collaborate and communicate; (2) the requirements needed to do so; and (3) advantages and disadvantages when being applied [217, 218]. Various team design patterns have been constructed to address moral decision-making in human-robot teams, often manipulating the level of human and robotic moral agency. For example, supported moral decision-making requires human moral supervision over a robot and taking over when perceiving the need for moral decisions. The robot should then support the human during moral decision-making by explaining the moral context. Another example is fully autonomous moral decision-making. In this collaboration design, human values are implemented in the robot, allowing it to autonomously make moral decisions. If these artificial agents would make moral decisions violating ethical guidelines and moral values, the tracking and tracing conditions should allow the identification of responsible humans to hold accountable [179]. However, we are not convinced that fully autonomous artificial moral agents are feasible and desirable during collaboration with humans. Instead, we believe that variable autonomy can be used to dynamically allocate all moral decisions to humans and non-moral decisions to robots.

5.2.2 Dynamic task allocation using variable autonomy

In robots with variable autonomy, humans can take control over certain (or all) elements of robot behavior [144]. A common distinction in human oversight and control over robots with variable autonomy involves three levels: having humans-in-the-loop, humans-off-the-loop, or humans-on-the-loop [46, 144]. Maintaining humans-in-the-loop requires informed human approval for all elements of robot behavior, for example during complete tele-operation of firefighting robots. In contrast, allowing humans-off-the-loop refers to fully autonomous robots without human operator involvement, for example firefighting robots that autonomously explore burning buildings and make moral decisions. Finally, having humans-on-the-loop assumes a supervisory human role tasked with monitoring and influencing robot behavior when necessary, for example when firefighters overrule the trajectory of firefighting robots or intervene when perceiving the need for moral decisions.

Another example of having humans-on-the-loop during moral decision-making in human-robot teams for firefighting is dynamic task allocation using variable autonomy (Table 5.1). In this variable autonomy approach, human moral values are elicited and implemented in the robot, allowing robot identification of morally sensitive situations. Eliciting human values for implementing artificial moral agents is a complex and multifaceted process, and there is a lot of discussion on its need and feasibility. Nevertheless, there are several approaches for value elicitation, each with its own strengths and weaknesses. For example, a rule-based elicitation questionnaire can be used where participant responses directly influence an autonomous agent's behavior through predefined rules [214]. Another example is using advanced machine learning and natural language processing techniques to infer and reason about human moral values [96, 127]. For dynamic allocation of moral decisions during firefighting, a questionnaire and crowdsourcing approach could be suitable to identify and use moral features as predictors of moral sensitivity (e.g., the number of victims, fire duration, and risk of building collapse).

After this value elicitation process, the firefighting robot should autonomously perform its explore and extinguish tasks while being morally supervised by the human operator who retains the power to override the robot's behaviour [2, 213]. Using variable autonomy, the robot identifies morally sensitive situations and allocates moral decision-making in these situations to the human operator, while making all non-moral decisions itself. This way, the variable autonomy approach ensures that humans can be held accountable for moral decisions and robot behavior, while the robot can decrease the workload of firefighters by preventing them from exercising control unnecessary often.

Variable autonomy approaches vary in terms of which aspects of autonomy are adjusted, by whom, how, why, and when [19, 29, 144]. The variable autonomy approach in Table 5.1 adjusts robot decision-making, and these adjustments are executed by either the human operator or robot (i.e., a mixed-initiative approach). The robot is primarily responsible for autonomy adjustments when it identifies situations as morally sensitive and requiring human moral decision-making. In contrast, the human is responsible for autonomy adjustments when during moral supervision he/she intervenes when the robot attempts to make moral decisions because it incorrectly identified moral sensitivity. During dynamic task allocation using variable autonomy, the autonomy level is adjusted in a discrete way from (semi-)autonomous robot decision-making in not morally sensitive situations to manual human decision-making in morally sensitive situations. The reasons for adjusting autonomy to complete human control in morally sensitive situations are pre-emptive to ensure meaningful human control. Finally, autonomy adjustments are executed during active operation of the robot in real firefighting scenarios by responding to changes in the moral sensitivity of situations.

For variable autonomy approaches to be effective, it is important to explicitly define which entities (human, robot, or both) are capable and responsible for which tasks [144]. Using team design patterns to describe the variable autonomy approach provides such a definition of roles and responsibilities and determines who transfers control of what, when and why it is needed, and to whom. To ensure adequate fulfillment of defined roles and responsibilities, there must be an appropriate means for information exchange allowing the states of the robot and environment to be understood. Moreover, this means of information exchange should achieve situation awareness and appropriate trust calibration without overloading the human operator's cognitive abilities [60, 118, 144]. Therefore, robot explanations are crucial during dynamic task allocation using variable autonomy in human-robot teams for firefighting. More specifically, the robot should involve and support

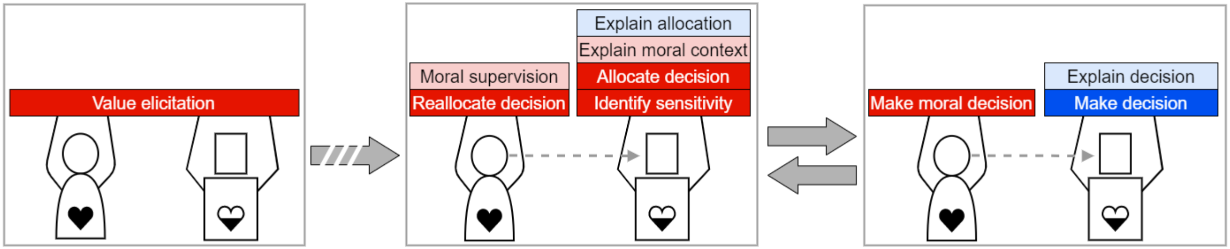
Name	**Dynamic task allocation using variable autonomy**
Description	Human moral values are elicited and implemented in the robot, allowing the robot to identify morally sensitive situations. When the robot classifies situations as morally sensitive, it allocates the related tasks/decisions to the human operator, while taking on the rest itself. The human operator can alter this allocation and intervene at any time. The robot explains allocations, non-moral decisions, and the moral context.
Structure	 Value elicitation Moral supervision Reallocate decision Explain allocation Explain moral context Allocate decision Identify sensitivity Make moral decision Explain decision Make decision
Requirements	**R1** The robot should be sufficiently able to identify morally sensitive situations **R2** Robot explanations should raise human moral awareness during supervision **R3** Robot explanations should not bias the human operator in its decision-making
Advantages	**A1** The robot reduces the workload of the human operator **A2** The human operator is in control of all morally sensitive decisions **A3** Robot explanations can build appropriate mental models of the robot
Disadvantages	**D1** The human operator does not make all decisions **D2** Interpreting robot explanations and allocations requires time **D3** Operator under-/overload can result in missed moral decisions made by the robot

Table 5.1: Variable autonomy approach for human-robot teams engaged in moral decision-making, where tasks are allocated dynamically (slight adaptation of the team design pattern by [214]). The variable autonomy approach is communicated in the form of a team design pattern that describes the collaboration, structure, requirements, advantages, and disadvantages of the approach.

the human operator by explaining the moral context and its non-moral and allocation decisions. It is crucial that these robot explanations do not (1) bias the human operator in its decision-making; (2) reduce situation awareness by information overload; or (3) cause misuse or disuse by information underload [118, 214]. However, without these robot explanations the human operator will not be able to exercise control in a timely and accurate manner [214].

The explanations of the robot are especially important when it classifies situations as not morally sensitive and allocates decision-making to itself. We suggest that the responses of the human operator when the robot allocates decision-making to itself can be can be classified using signal detection theory [57, 245]. More specifically, this classification considers the presence or absence of human reallocation interventions, robot classification of situations as morally sensitive or not, and the true nature of situations as morally sensitive or not. For example, hits refer to human reallocation interventions when the robot classifies morally sensitive situations as not morally sensitive. In contrast, misses refer to no human interventions when the robot classifies morally sensitive situations as not morally sensitive. On the other hand, false alarms refer to human reallocation interventions when the robot classifies not morally sensitive situations as not morally sensitive. Finally, correct rejections refer to no human interventions when the robot classifies not morally sensitive situations as not morally sensitive. From a meaningful human control perspective, one could argue that hits and misses are crucial to ensure human moral decision-making, whereas false alarms and correct rejections are less problematic. Classifying human operator responses using signal detection theory provides quantitative measures that can be applied to verify if the variable autonomy approach truly ensures meaningful human control.

5.2.3 Existing operationalizations of meaningful human control

Imposing meaningful human control as a requirement and verifying if variable autonomy approaches indeed fulfill this requirement calls for methods to measure meaningful human control. So far, only few operationalizations of meaningful human control have been proposed. One of them introduced three measurable dimensions of meaningful human control: (1) Experienced meaningful human control and behavioral compliance with (2) ethical guidelines and (3) moral values [214]. The authors argue that humans experience meaningful human control, which can be measured subjectively. Moreover, they argue that the behavioral compliance with moral values and ethical guidelines provides evidence for meaningful human control. In their work, they measure experienced control with a semi-structured interview using eight five-point Likert scale statements on concepts like time pressure, responsibility, and decision-making comfort and quality.

Another operationalization are the four necessary properties for human-robot teams to be under meaningful human control [30]. The first property requires an explicitly specified moral operational design domain where the robot should adhere to. This involves norms and values to be considered and respected during design and operation. Here, it is important that the robot embeds concrete solutions to constrain actions of the team within the boundaries of the moral operational design domain. Moreover, it is crucial that humans are aware of their responsibilities to make conscious decisions if and when the team deviates from the boundaries of the moral operational design domain. The second property requires humans and robots to have appropriate and mutually compatible representations of each other and the team, to decide which actions to take and perform. These representations should include reasons, tasks, desired outcomes, role distributions, preferences, capabilities, and limitations. Building these mental models of both team members can be achieved by for example communication and explanations. The third property requires relevant human agents to have the ability and authority to control the robot, so that they can act upon their moral responsibility. This means humans should be able to change the robot's goals and behavior to track reasons, as well as intervene and correct robot behavior. Here, it is important to clearly and consistently define role distributions, task allocations, and control authority. Again, team design patterns are particularly useful for describing and communicating such design choices. Finally, the fourth property requires the actions of the robot to be explicitly linked to actions of humans who are aware of their moral responsibility. This means the human-robot team should simplify and aid achieving human moral awareness, for example using explanations of the robot's actions.

In contrast to operationalizing the whole concept of meaningful human control, other studies operationalized only its tracing condition. For example, the cascade evaluation approach subjectively quantifies traceability [25, 26, 48]. This approach is centered around four aspects: (1) The exertion of operational control; (2) the involvement of a human agent; (3) the ability of a human agent to understand and interact with a robot; and (4) the ability of a human agent to understand their moral responsibility over a robot. For each aspect, involved human agents (e.g., operator or designer) are given a score along a six-point Likert scale, reflecting the degree of that aspect for the human agent. However, each aspect is considered as part of overall traceability, and therefore the scores from the previous and current aspects are compared to determine critical scores. This way, the critical scores for aspects 2, 3, and 4 are all influenced by the aspects that preceded them. Ultimately, the critical score of aspect 4 reflects the final traceability score. This operationalization of traceability is suitable for both a-priori and a-posteriori evaluation of robots and/or variable autonomy approaches. For example, the authors apply the cascade approach by presuming the situation of an inattentive driver struggling with the ability to retake control of an automated vehicle. However, the approach can also be applied to a-posteriori evaluate a variable autonomy approach by involved human agents or a third party.

The tracking condition of meaningful human control has been operationalized as reason-responsiveness (i.e., robots being responsive to human reasons to act) [139]. Here, the main idea is that the humans whose reasons are being tracked have the kind of control over robots that make them morally responsible for the actions of robots. A proximity scale has been introduced to identify and order human reasons according to proximity and complexity with respect to how closely they influence robot behavior [139]. In decreasing order of proximity and complexity, a distinction is made between the reasons values, norms, plans, and intentions. The authors argue that more proximal reasons (e.g., the intention to reallocate decision-making) are often closer in time to robot behavior and also simpler than more distal reasons (e.g., the plan to rescue all victims). However, this operationalization of tracking as reason-responsiveness has been questioned by demonstrating it is ambiguous in distinguishing between motivating and normative reasons [225]. More specifically, it is argued that tracking is operationalized in terms of motivating reasons (mental states) instead of normative reasons (facts), while the idea of responsibility attribution is derived from normative reason-responsiveness. Furthermore, this work shows that tracking cannot play an important role in responsibility attribution because normative reasons are agent-neutral (i.e., a fact for agent A is also a fact for agent B). Therefore, the author proposes that the tracing condition should be the sole determinant of responsibility, and that the humans to which robot actions can be traced back are the humans in control of and responsible for robot outcomes. Consequently, one could argue that verifying if dynamic task allocation using variable autonomy indeed ensures meaningful human control only requires measuring traceability.

These discussed operationalizations of meaningful human control demonstrate that only a few properties are actually transformed into quantifiable measures, while most properties still remain hard to quantify. Furthermore, the quantitative measures are all subjective such as experienced control [214] or the subjective traceability score [25, 26, 48]. To impose meaningful human control as a requirement and verify if variable autonomy approaches fulfill this requirement, more quantitative operationalizations and objective measures are appreciated.

5.3 Method

5.3.1 Overview

To explore quantitative operationalizations of meaningful human control during dynamic task allocation in human-robot teams for firefighting, we conducted an online qualitative focus group. During the study, we presented several statements about and inspired by the operationalizations discussed in Section 5.2. In summary, we presented the following six statements: (1) Operationalizing the tracing condition can only be done using subjective measures; (2) the cascade approach evaluates all tracing aspects; (3) thresholding the final score of the cascade approach to determine sufficient tracing would be a good idea; (4) misses resulting from operator unawareness of moral sensitivity indicate the robot is not under meaningful human control; (5) misses when the operator is overloaded but aware of moral sensitivity indicate the robot is still under meaningful human control; and (6) the hit rate is an important property for the robot to be under meaningful human control, while the true discovery rate is not. All statements were formulated somewhat provocatively in an attempt to elicit strong responses. We invited experts in the field and asked them to respond to our statements while engaging in a discussion with each other. Data was collected from one focus group study with the experts and analyzed using reflexive thematic analysis [21].

5.3.2 Data collection

Since the topic of operationalizing meaningful human control is complex and involves technical terminology and concepts not easily understood by laymen (e.g., the tracking and tracing conditions), we recruited five experts in the field of meaningful human control. To capture and represent the multifaceted nature of the topic, we recruited experts with various backgrounds such as engineering (1), law (1), human factors (1), and computer science (2). All participants published articles on meaningful human control and four of them specifically on operationalizing meaningful human control. Therefore, we believed the recruitment of these experts to facilitate the in-depth discussions and critical analyses required to generate concrete ideas on operationalizing meaningful human control during dynamic task allocation. We informed the expert participants that we would present statements on operationalizations of meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting. Moreover, we asked them to respond to our statements while engaging in a discussion with each other. The first three statements corresponded to quantitative operationalization of the tracing condition, whereas the last three centered around objectively quantifying meaningful human control based on team and robot outcomes. All participants signed an informed consent form before participating in the study, which was approved by the ethics committee of our institution (ID 2477). The online focus group lasted around one and a half hours and was automatically transcribed using Microsoft Teams. Afterwards, this transcript was checked and improved using a video recording of the study, which was destroyed after this data processing step.

5.3.3 Data analysis

Data was analyzed using reflexive thematic analysis [21], a method for producing a coherent interpretation of the data, grounded in the data. This approach is centered around the researcher's role in knowledge production and subjectivity rather than achieving consensus between coders. Reflexive thematic analysis involves familiarization with the data, generating codes, constructing themes, revising and defining themes, and producing the report of the analysis. We outline the process for the first five phases below, the last phase is reported as Section 5.4. We first familiarized ourselves with the data by fine-tuning the transcription of the focus group using the video recording. Next, we read the full transcript in detail to double check for potential mistakes during the transcription process. During this step we already highlighted and took notes of potentially interesting text excerpts.

We systematically coded the transcript by searching for instances of talk that produced snippets of meaning relevant to the topic of operationalizing meaningful human control. These instances were coded using comments in Microsoft Word, highlighting the relevant text excerpt for each code. The coding of thematic analysis can be either an inductive approach, deductive approach, or combination of the two. This decision depends on the extent to which the analysis is driven by the content of the data, and the extent to which theoretical perspectives drive the analysis. Coding can also be semantic, where codes capture explicit meaning close to participant language, or latent, where codes focus on a deeper, more implicit or conceptual level of meaning. We used a deductive coding approach driven by prior literature and existing operationalizations, as well as operationalization ideas that we formulated. Semantic codes capturing explicit meaning close to participant language were noted, such as "challenges the use of thresholds".

During the construction, revision, and definition of our themes, we first sorted our codes into topic areas using bullet-point lists. Next, we used visual mapping (using Miro) and continuous engagement with the data to further construct, revise, and define our themes. These candidate themes were grouped into one overarching theme of quantitative operationalization of meaningful human control, which encompassed ten themes and eight sub-themes. The process of revising and defining themes again involved visual mapping and continuous engagement with the data, mainly to check for relationships between themes. For example, we checked whether initial themes should be sub-themes of other themes, or whether sub-themes could be promoted to themes. Finally, this resulted in our full thematic map of six themes and eleven sub-themes. We grouped these themes and sub-themes into the overarching theme “quantitative operationalization of meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting”.

5.4 Results

Our analysis revealed that the following six themes are underlying the main overarching theme: (1) The cascade approach as valuable tool for quantifying traceability; (2) meaningful human control as a spectrum rather than binary; (3) team and system outcomes as proxies for meaningful human control; (4) context and assumptions as crucial factors to study, define, and evaluate meaningful human control; (5) system design(er) as output and reason for meaningful human control; and (6) operationalizing meaningful human control does not imply quantification. Below we will discuss these six themes in detail. The full thematic map can be seen in Figure 5.1. Some themes consist of several sub-themes and were constructed based on extensive discussions clearly highlighting their importance (such as the theme described in section 5.4.3). On the other hand, some themes do not consist of any sub-themes and were constructed based on briefer discussions that still highlighted significant relevance to be main themes (such as the theme described in section 5.4.2). Although these themes are described in less detail, this does not mean they are less important.

5.4.1 The cascade approach to quantify traceability

The first main theme that we identified was “the cascade approach as valuable tool for quantifying traceability”. Most participants viewed the cascade approach as a valuable tool to quantify traceability during dynamic task allocation using variable autonomy. One participant considered the approach “as a first step to see how one could start to evaluate something which in itself is not quantifiable". Most participants also shared how the approach is not perfect or the only method, but at the same time none of the experts were aware of (better) alternatives for quantitative evaluation of the tracing condition. Another expert explained how the cascade approach can also be used: “The cascade approach can give an indication, but then there should still be a human who can evaluate if the tracing condition is met based on the indication that the method gives.”

One sub-theme that we identified within this theme was “context as determining the application and evaluation of the cascade approach”. Some participants mentioned that the cascade evaluation approach misses some tracing aspects. Along those lines, most experts shared how context and level of abstraction are important for determining how to apply the approach and whether the approach evaluates all tracing aspects. For example, one expert explained: “I do think the cascade approach misses something, but I believe that if you make it more context specific you stand a chance at capturing the key aspects of tracing.”

Another sub-theme that we identified was “assumptions and awareness as enabling objective measures of operator understanding”. One expert speculated how the traceability aspect operator understanding of the robot can also be measured objectively, but that this requires assumptions about specific scenarios. This participant further mentioned the

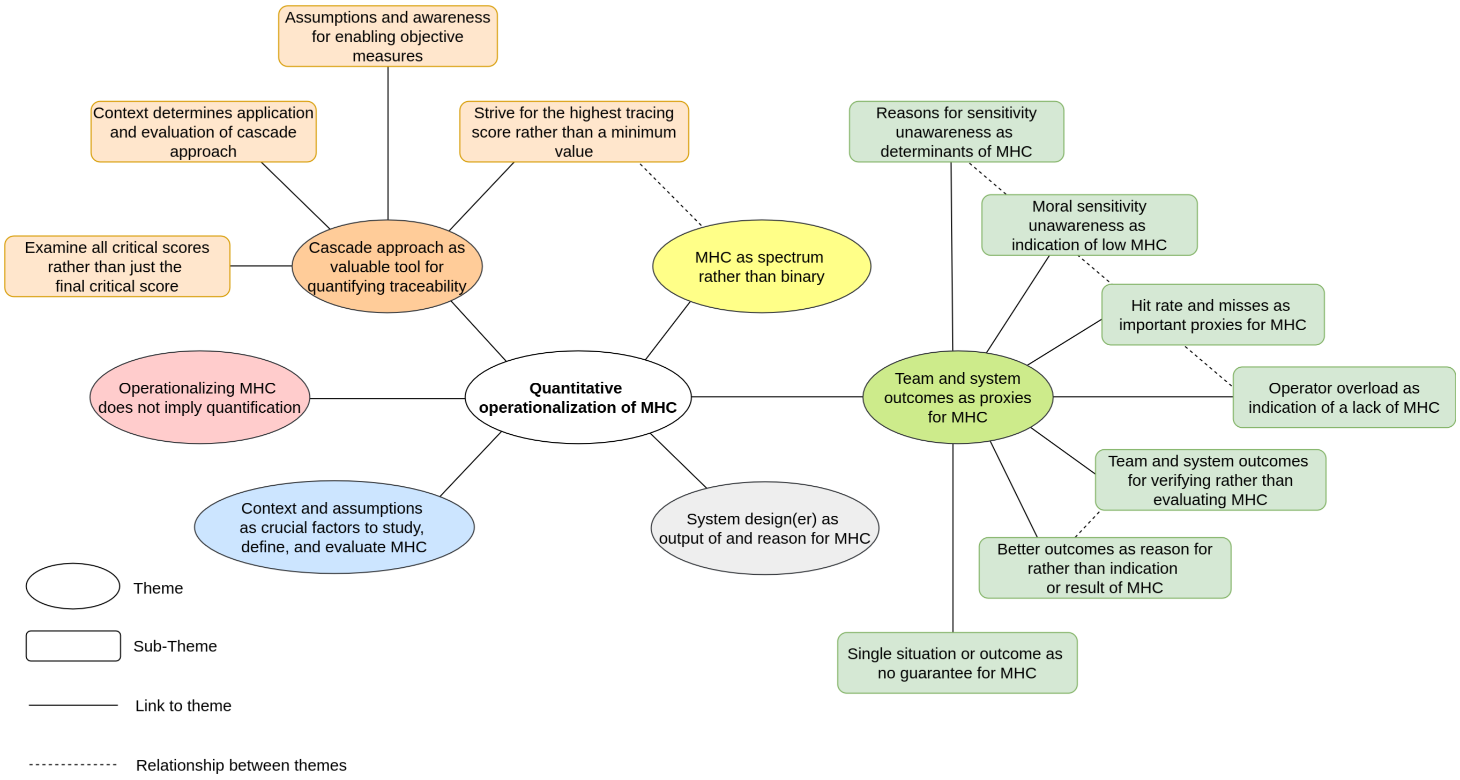

Assumptions and awareness for enabling objective measures
Context determines application and evaluation of cascade approach
Strive for the highest tracing score rather than a minimum value
Examine all critical scores rather than just the final critical score
Cascade approach as valuable tool for quantifying traceability
MHC as spectrum rather than binary
Operationalizing MHC does not imply quantification
Quantitative operationalization of MHC
Context and assumptions as crucial factors to study, define, and evaluate MHC
System design(er) as output of and reason for MHC
Reasons for sensitivity unawareness as determinants of MHC
Moral sensitivity unawareness as indication of low MHC
Hit rate and misses as important proxies for MHC
Team and system outcomes as proxies for MHC
Operator overload as indication of a lack of MHC
Team and system outcomes for verifying rather than evaluating MHC
Better outcomes as reason for rather than indication or result of MHC
Single situation or outcome as no guarantee for MHC
Theme
Sub-Theme
Link to theme
Relationship between themes

Figure 5.1: Thematic map on quantitative operationalization of meaningful human control (MHC) during dynamic task allocation using variable autonomy in human-robot teams for firefighting.

use of situation awareness and operational tests for measuring operator understanding objectively: “Let us assume a certain scenario for the operator. You could then test the operator to see if the operator is aware of what the robot might do in a certain circumstance. You can then let the robot perform the task and you can check to see if that is actually met. It is very hard to generalize this, but you can do this for very specific situations and then also objectively measure understanding in those specific situations.”

We also identified the sub-theme “examine all critical scores rather than just the final critical score”. This sub-theme is related to comparing the four traceability aspects of the cascade approach to determine critical scores, as discussed in Section 5.2.3. One of the experts explained: “If you accumulate the individual critical scores into one final score then you are removing information. So, it depends on the purpose of the tracing score, but I would be more interested in the individual scores that are composing the final score.” All the other participants agreed with this point. Another expert mentioned how considering both individual scores and final score can be relevant for comparing different robots, and that the individual scores can provide more information about which robot is easier to correct in order to improve traceability.

The final sub-theme that we constructed was “strive for the highest tracing score rather than a minimum value”. When discussing the use of a threshold to define when the final critical score reflects sufficient fulfillment of the tracing condition, all experts articulated how the goal should be to get the highest possible final score rather than a minimum value reflecting sufficient traceability. Furthermore, they mentioned how the critical scores are subjective and therefore it is inaccurate, not possible, and not necessary to set a threshold defining sufficient traceability. Finally, one expert explained how the scores of the cascade evaluation approach are more useful to inform rather than automate: “I would even challenge the very notion of thresholding because it reflects the kind of intrinsic desire to quantify everything. It does feel like this subjective cascade approach might actually inform decision-making without automating it because a threshold is a way of automating the decision.”

5.4.2 Meaningful human control as a spectrum

The second theme that we identified is closely linked to the previously discussed sub-theme. We called this theme “meaningful human control as a spectrum rather than binary”. Two participants explicitly mentioned that the evaluation of the presence of meaningful human control is more nuanced than saying yes or no and should be considered as a spectrum instead: “Meaningful human control itself as well as its different conditions is never black and white, is not binary, it is a spectrum basically. So there is an extent of meaningful human control, but it is never that there is full meaningful human control or there is zero.” The other experts all seemed to agree with this viewpoint.

5.4.3 Outcomes as proxies for meaningful human control

The third main theme that we constructed was “team and system outcomes as proxies for meaningful human control”. We identified this theme during the discussion of using signal detection theory to classify operator responses during during dynamic task allocation using variable autonomy. One sub-theme that we identified was “better outcomes as reason for rather than indication or result of meaningful human control”. Most participants shared how the quality of team and robot outcomes is not always an indication or a result of meaningful human control. One expert explained: "A bad outcome is not always an indication of a lack of meaningful human control and a good outcome is not always an indication of meaningful human control being present. It could also be that the human who is in control has made an error." Similarly, another participant complemented: "Meaningful human control also does not equate to moral acceptability of any situation. A system can be under meaningful human control and show very questionable outcomes." On the other hand, two experts articulated that one of the reasons for pursuing meaningful human control is to achieve better and ethically sound outcomes.

Another sub-theme that we constructed was "team and system outcomes for verifying rather than evaluating meaningful human control". One expert questioned the correctness of assessing meaningful human control in terms of team and robot outcomes. Another expert agreed that outcomes alone are not sufficient for determining the presence of meaningful human control, but explained that "you can use outcomes to see whether they are in accordance with guidelines." Moreover, this expert mentioned how outcomes can be used to verify the presence of meaningful human control rather than evaluate it. The expert who initially raised questions agreed with these points: "I totally agree that there is a relationship between meaningful human control and outcomes. Maybe the outcomes can be indirect evidence of meaningful human control."

We also identified the sub-theme "single outcome or situation as no guarantee for meaningful human control". Two participants mentioned how considering single situations and outcomes is not sufficient for determining the presence of meaningful human control: "A bad outcome is not always an indication of a lack of meaningful human control and a good outcome is not always an indication of meaningful human control being present. If you have enough situations and samples, then it does give a very good overall picture, but for one specific situation it does not give that guarantee. Sometimes in isolation a situation can be a bit misleading."

Another constructed sub-theme was "hit rate and misses as important proxies for meaningful human control". In addition to classifying operator responses during dynamic task allocation as hits, misses, false alarms, and correct rejections, we explained participants the distinction between the hit and true discovery rate. More specifically, the hit rate refers to the percentage of relevant situations where the operator correctly intervenes (by dividing hits by hits and misses). In contrast, the true discovery rate refers to the percentage of operator interventions which are necessary (by dividing hits by hits and false alarms). Two experts articulated how the true discovery rate and false alarms are not so important in relation to meaningful human control, while the hit rate and misses are: "If you intervene in the sense that you keep awareness and keep responsibility to yourself, even though there was not a necessary situation, I would say that the true discovery rate is not an important property for a system to be under meaningful human control, whereas the hit rate is." Similarly, the other expert mentioned: "As long as the operator intervenes it does not really matter if they intervene even in the situation when the robot is kind of okay, but it does matter when the human does not intervene when the robot is not okay."

We also identified the sub-theme "operator overload as an indication of a lack of meaningful human control". All participants felt that misses resulting from operator overload indicate a lack of meaningful human control. The experts explained several reasons, such as “the system is in operation outside of what is reasonable to expect for that person”, “you need the ability to intervene in time and in a proper fashion”, and “this is an example of operator capacity being lower than their responsibility.”

Another sub-theme that we constructed was “moral sensitivity unawareness as an indication of low meaningful human control”. Most participants shared that misses resulting from operator unawareness of the moral sensitivity indicate low meaningful human control. One expert explained: “If the robot misinterprets the situation but the operator does not intervene, then there is obviously a lower level of meaningful human control because the robot in its design has not been able to identify the situation correctly and also the operator does not correctly intervene”. Another expert elaborated on the distinction between human control and meaningful human control: “Strictly speaking the system is under control of the human operator because he/she has the capability and the authority to intervene. So, strictly speaking, I should say it is under control of the human operator, but that does not necessarily imply meaningful human control.”

The final sub-theme that we constructed was “reasons for moral sensitivity unawareness as determinants of meaningful human control”. All participants mentioned how knowing the reasons for the operator’s unawareness of the moral sensitivity is very important for determining the extent of meaningful human control. For example, one expert mentioned how there would be no meaningful human control if the operator does not have the means to be aware of the moral sensitivity. Another participant explained: “Is the operator unaware because he/she cannot do anything about it, then the operator should not be held responsible. Then, the question is, depending on how the system was designed, does this lead to a responsibility gap or does this mean that responsibility should be attributed to a designer or someone else?”

5.4.4 Context and assumptions as crucial factors

Another theme that we identified during the discussion of the previously reported sub-theme was “context and assumptions as crucial factors to study, define, and evaluate meaningful human control”. Two participants mentioned the importance of communicating the assumptions of our definitions, variable autonomy approach, and robot design and communication. For example, one expert explained: “I think what is very important when you do this, because I see the value, is communicating the assumptions you are using. Basically, you want to create a shared mental model.”

5.4.5 Design(er) and meaningful human control

We also identified the theme “system design(er) as output of and reason for meaningful human control”. One expert shared how operationalizing meaningful human control can result in requirements for robot design: “It does feel to me that one of the major benefits of operationalizing meaningful human control through tracking and tracing is to arrive in every individual context at a set of very context specific requirements for the design of the system”. On the other hand, several participants mentioned how meaningful human control can be present by system design and how both the system designer and robot operator should be considered during the discussion. For example, one expert shared: “You basically have two human agents here, you have the operator and you have the robot designer. If at least one of them is able to influence the situation in a way that human control is meaningful, then meaningful human control is still present. This does not have to be the operator necessarily, it can also be the way the robot is designed."

5.4.6 Qualitative operationalization of meaningful human control

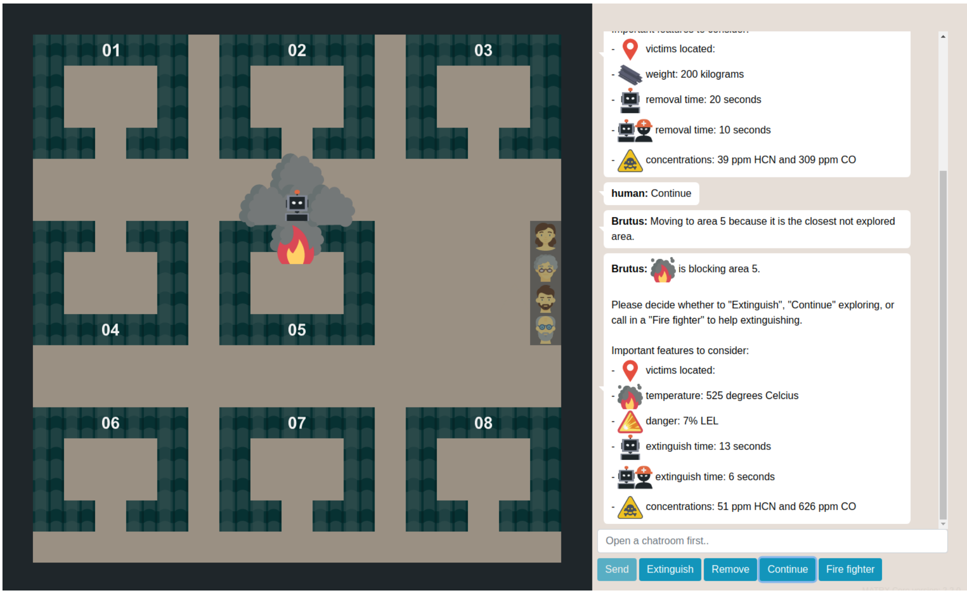
The final theme that we identified during the focus group study was "operationalizing meaningful human control does not imply quantification". Most experts mentioned that quantification of meaningful human control has its value, but that it is very difficult to accurately quantify its elements and conditions. Moreover, two participants shared that operationalizing meaningful human control does not require or mean quantifying it. More specifically, one expert explained: "I just want to challenge the kind of implicit assumption here that operationalizing the tracing condition would require quantifying it, because I think that you can actually operationalize any notion to some extent in a completely qualitative way."

5.5 Discussion and Conclusion

5.5.1 Discussion

Our results emphasize the usefulness of the cascade approach to quantify traceability during dynamic task allocation using variable autonomy in human-robot teams for firefighting. Moreover, the results highlight a new application of the approach in comparing how different robot implementations or variable autonomy approaches affect traceability, for example in terms of robot behavior, explanations, or autonomy adjustments. Another novel suggested application is comparing all individual aspect scores as well as the final critical score to provide relevant information about traceability and potential points of improvement. These applications are novel compared to its original usage of evaluating implemented robots or variable autonomy approaches using only the final traceability score. The results further emphasize the importance of using a scale for such comparisons, where certain robot implementations or variable autonomy approaches may exhibit varying levels of traceability. Ultimately, the goal should be to get the highest possible traceability score rather than a minimum sufficient value. This is in line with some earlier interpretations of meaningful human control as ratio rather than binary [25, 30]. On the other hand, it contradicts the discussion on defining how much of each of the four properties for human-robot teams to be under meaningful human control is sufficient [30].

Results also highlight a new application of situation awareness and operational tests to objectively measure the traceability aspect "human ability to understand and interact with a robot". The (modified) Situation Awareness Global Assessment Technique (SAGAT) can be used to objectively measure human understanding of the robot during simulations of representative tasks [176, 232]. This way, SAGAT can also be used to investigate whether certain robot explanations can increase traceability by improving human understanding of the robot. Figure 5.2 shows an example of what a simulated task could look like for dynamic task allocation using variable autonomy in human-robot teams for firefighting. This simulated task is especially valuable for evaluating human-robot collaboration before real-world deployment, enabling easier manipulation and evaluation of aspects like robot communication and behavior [180, 183, 223, 232]. For example, the simulated task in



Important features to consider:
- victims located:
- weight: 200 kilograms
- removal time: 20 seconds
- removal time: 10 seconds
- concentrations: 39 ppm HCN and 309 ppm CO

human: Continue

Brutus: Moving to area 5 because it is the closest not explored area.

Brutus: is blocking area 5.

Please decide whether to "Extinguish", "Continue" exploring, or call in a "Fire fighter" to help extinguishing.

Important features to consider:
- victims located:
- temperature: 525 degrees Celcius
- danger: 7% LEL
- extinguish time: 13 seconds
- extinguish time: 6 seconds
- concentrations: 51 ppm HCN and 626 ppm CO

Open a chatroom first..

Figure 5.2: Simulation of the collaboration between the human operator and explore and extinguish robot during dynamic task allocation using variable autonomy. The robot is autonomously exploring an office building to search and rescue victims. The human operator is supervising the robot and they communicate via a chat box. When the robot perceives the need for a moral decision, it allocates decision-making to the human operator. All non-moral decisions are made by the robot.

Figure 5.2 allows the evaluation of different robot explanations such as textual, visual, or hybrid explanations [197]. Moreover, it allows the rapid implementation and evaluation of different variable autonomy and control approaches such as having a human-in-the-loop or having a human-on-the-loop.

The results further highlight the use of team and robot outcomes as verification of meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting. In terms of outcomes, operator hits and misses during robot allocation of decisions are considered to be more important than false alarms and correct rejections. These objective outcomes provide novel measures for assessing the operator's supervision performance during dynamic task allocation using variable autonomy. Using these team and robot outcomes to verify meaningful human control corresponds with the operationalization by [214] that uses behavioral compliance with moral values and ethical guidelines as evidence for meaningful human control. Our results also emphasize that meaningful human control can be present by system design. This aligns with the claim that robots with variable autonomy can ensure meaningful human control over these robots [144]. It also aligns with our expectation that dynamic task allocation using variable autonomy in human-robot teams for firefighting can ensure meaningful human control by design. However, this could be verified using measures like the hit rate and whether outcomes are in accordance with firefighting guidelines, such as evacuating victims first when the location of the fire source is unknown and smoke spreads fast. Finally, the results introduce a new perspective by stressing the importance of collecting outcomes during multiple task simulations because single positive/negative outcomes are not always an indication of the presence/absence of meaningful human control.

Finally, our results highlight a novel perspective by emphasizing the importance of qualitatively identifying reasons underlying outcomes, such as why an operator is overloaded or unaware of moral sensitivity, to determine the extent of meaningful human control and how to increase it. Therefore, we believe that conducting follow-up interviews after completing simulations of representative tasks can be particularly effective to identify reasons underlying outcomes. For example, after the simulated task operators could be questioned about reasons for misses and how to improve the variable autonomy approach to avoid them. More specifically, if the operator has a low hit rate during the task, follow-up interviews could determine whether this results from an overload of robot information or unawareness of moral sensitivity due to limited experience. This distinction is crucial because these reasons determine the extent of meaningful human control and how to improve it. For example, an overload of robot information indicates a lack of meaningful human control requiring robot improvements such as decreasing robot communication. On the other hand, operator unawareness of moral sensitivity due to limited experience indicates low meaningful human control that can be addressed by more operator training.

In summary, the following novel knowledge on operationalizing meaningful human control has been gained because of the expert study. First, using the cascade approach for comparing different robot implementations or variable autonomy approaches, and not by comparing only the final critical score but also all individual aspect scores. Furthermore, using situation awareness and the hit rate to objectively measure the traceability aspect “human ability to understand and interact with the robot”, and collecting these measures during multiple task simulations for a more robust indication of meaningful human control. Finally, qualitatively identifying reasons underlying outcomes, like operator overload or moral unawareness, to determine the extent of meaningful human control and how to increase it.

5.5.2 Evaluating meaningful human control during dynamic task allocation

Based on these results, our main contribution is proposing the following evaluation method of meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting. We suggest adapting the cascade approach to not only subjectively quantify traceability, but also objectively using operator responses during and after task simulations. The first aspect of the cascade approach involves scoring the exertion of operational control by human and robot separately, and the maximum of these scores is taken as the critical score of this aspect. This aspect can be scored a-priori based on the fixed collaboration characteristics during dynamic task allocation, where the robot exercises more operational control as it makes all non-moral decisions and handles allocation of decision-making. However, we suggest combining this a-priori score with an a-posteriori score determined by the involved human him/herself, for example by measuring experienced control [214].

The second aspect of the cascade approach involves scoring the involvement of the human operator, and the minimum of this score and the critical score of the first aspect determines the critical score of the second aspect. The involvement of the human operator can also be scored a-priori based on the expectation of continuous supervisor involvement during dynamic task allocation. However, we suggest combining this a-priori score with an objective measure of situation awareness, which assesses the human operator's perception, comprehension, and projection of environmental elements [60]. Here, higher situation awareness can be taken as a higher human operator involvement during the task. Situation awareness can be measured objectively using the traditional Situation Awareness Global Assessment Technique (SAGAT) [60, 62]. This involves a-priori defining the information and situation awareness requirements of the operator using goal-directed task analysis. Next, SAGAT queries should be formulated that objectively evaluate operator knowledge of this situational information. These queries should be asked during random pauses of the task simulation, either once or multiple times. The percentage of correctly answered queries can then be used as objective measure of situation awareness.

The third aspect of the cascade approach involves scoring the ability of the human operator to (1) understand the robot and (2) interact with the robot. The minimum of these two scores is then compared with the critical score of the second aspect, and the minimum of this comparison determines the critical score of the third aspect. We suggest to objectively measure the ability of the human operator to understand the robot using situation awareness of the robot's behavior processes and decisions. In addition to measuring situation awareness, SAGAT is also suitable for objectively measuring human understanding of explainable systems like the robot dynamically allocating tasks [60, 176, 232]. In this case, the goal-directed task analysis involves the definition of situational information requirements specifically related to robot behavior. Again, the task simulation should be paused at random times, followed by evaluating operator knowledge of the predefined informational needs. Furthermore, we suggest to objectively measure the ability of the human operator to interact with the robot using task performance. Task performance can be determined by the operator's hit and true discovery rates during the robot's dynamic allocation of decision-making, where higher hit and true discovery rates would refer to better performance. Finally, the minimum score of the human ability to (1) understand the robot and (2) interact with the robot is taken as the score that is compared with the critical score of the second aspect.

The fourth and final aspect of the cascade approach involves scoring the ability of the human operator to understand their moral responsibility over the robot, and the minimum of this score and the critical score of the third aspect determines the final traceability score of the variable autonomy approach. We suggest quantifying this aspect using a semi-structured interview after completing the task simulation. This semi-structured interview can efficiently be followed by open questions to identify reasons underlying outcomes like operator misses. Identifying these reasons is crucial to further improve the variable autonomy approach, for example by adjusting robot communication if many operators suffer from information overload. Finally, we suggest combining this subjective measure of moral responsibility understanding with an objective measure of how many outcomes adhere to ethical firefighting guidelines, such as not sending in firefighters when temperatures exceed auto-ignition temperatures of present substances. This way, not only the subjective understanding is considered but also translation of that understanding into

Aspect	Measure	Score OO	Score IO
1	Human operational control	3	3
	Robot operational control	2	2
2	Human involvement	3	3
3	Human understanding of robot	4	**2**
	Human interaction with robot	**0**	**2**
4	Human understanding of moral responsibility	4	**2**
	Final traceability	0	2

Table 5.2: Example of the cascade approach for two types of operators. OO refers to overloaded operator, IO to inexperienced operator. Bold highlights the cause(s) of the final traceability score.

adherence to ethical guidelines.

Our initial goal was a quantitative operationalization of meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting. Ultimately, we propose a hybrid operationalization where some required qualitative elements (reasons underlying outcomes) supplement the quantitative elements (traceability aspects). During evaluation, we recommend using all aspects scores instead of just the critical scores to arrive at improvements for the variable autonomy approach. For example, consider an overloaded human operator with the critical aspect scores 3, 3, 0, and 0; and a inexperienced human operator with the critical aspect scores 3, 3, 2, and 2. These scores indicate a lower traceability/final critical aspect score for the overloaded operator (0) than the inexperienced operator (2). However, closer inspection of all aspect scores (see Table 5.2) could reveal that the overloaded operator only lacks the ability to interact with the robot. Similarly, inspecting all scores of the inexperienced operator could reveal that the operator suffers from a low ability to both understand and interact with the robot and understand their moral responsibility over the robot. So, while the overloaded operator has a lower traceability score than the inexperienced operator, analyzing all scores suggests that the traceability score of the overloaded operator can be improved more easily as it results from only one aspect score instead of three.

5.5.3 Limitations

We identify a few limitations of our work. First of all, we conducted a single focus group that was coded individually. It can be favourable to conduct the same focus group multiple times with different experts, until reaching a saturation point. However, since our goal was to capture a particular perspective within a specialized domain, we considered one focus group appropriate to reach our objectives. Furthermore, thematic analysis is often associated with achieving consensus between multiple coders and high inter-coder reliability. However, it was our goal to generate rich, contextually situated, and nuanced themes instead. Therefore, we employed reflexive thematic analysis, emphasizing the researcher's role in knowledge production and centering around researcher subjectivity [21]. All in all, while we acknowledge the limitations associated with a single focus group and individual coding, these were strategic choices aligned with our research objectives and the specialized nature of our domain.

Another limitation concerns the generalizability of our proposed evaluation method for meaningful human control (section 5.5.2). Since this method is tailored to evaluating meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting, it is questionable how it would translate to different contexts and systems. However, this is not necessarily a problem as the conditions, properties, and implementation of meaningful human control are context- and system-specific [30, 179]. On the other hand, we do believe some aspects can be used for different contexts and systems with similar levels of autonomy and outcomes. For example, objectively quantifying human ability to understand and interact with systems using situation awareness and task performance can also be done during simulations of drivers collaborating with automated driving systems. Here, even the hit and true discovery rates can be used when the task includes incorrect automated driving behavior requiring the human driver to intervene. Finally, we believe semi-structured interviews after task simulations can be generalized to all contexts and systems by providing a robust way to identify reasons underlying behavior and outcomes.

5.5.4 Future Work

For future work, we want to verify if dynamic task allocation using variably autonomy indeed ensures meaningful human control in human-robot teams for firefighting. We believe that a user study in a simulated task environment similar to Figure 5.2 could provide valuable insights before considering in field tests. To implement the variable autonomy approach, the robot should be sufficiently able to identify morally sensitive situations (see Table 5.1). We are currently collaborating with the fire department of Rotterdam on this robot identification of morally sensitive situations. More specifically, we created a questionnaire to understand how people view morally sensitive situations in human-robot teams for firefighting. This questionnaire presents various situations during the collaboration between firefighters and their firefighting robot, such as locating the fire source, rescuing victims, and switching deployment tactic. These situations are characterized by different features such as the number of victims, fire duration, and fire resistance to collapse. In the questionnaire, participants specify how morally sensitive they consider each situation on a 7-point scale ranging from not morally sensitive to extremely morally sensitive (inspired by [163]). Moreover, they explain which feature(s) contributed the most to their rating and what feature changes would result in alternative moral sensitivity ratings. This way, we can identify which of the features are moral features and use them as predictors to statistically significantly predict the moral sensitivity of situations. This regression model can be implemented in the firefighting robot, together with a threshold for determining when the predicted moral sensitivity is too high and thus requires human decision-making. For future work, we want to first implement the dynamic task allocation and above mentioned regression model in a virtual robot and simulated environment similar to Figure 5.2. Next, we want to verify if dynamic task allocation indeed ensures meaningful human control during the collaboration. This work is presented in Chapter 6.

To verify this, we need to measure meaningful human control during the user study. The results of our expert study will influence the measurement of meaningful human control during this user study in several ways, in line with our proposed evaluation method in section 5.5.2. More specifically, we will determine the participants' exertion of operational control a-priori based on the fixed collaboration characteristics during dynamic task allocation. The involvement of the participants as supervisors and understanding of the robot's behavior will be determined by objective measures of situation awareness obtained by queries asked during random pauses of the task [60, 62, 232]. We will determine the participants' ability to interact with the robot using task performance, more specifically their hit rate during the robot's allocation of moral decisions (i.e., do they intervene when the robot classifies morally sensitive situations as not morally sensitive). Finally, participants' understanding of their moral responsibility over the robot will be measured after task completion, using a semi-structured interview. This interview will also be used to identify reasons underlying task performance.

In addition to verifying if dynamic task allocation ensures meaningful human control, we are particularly interested in which robot explanations can support the human operator to intervene and reallocate moral decision-making when the robot incorrectly classifies morally sensitive situations. To support the human operator during moral supervision of dynamic task allocation by the robot, robot explanations are crucial. For example, the robot can provide reason explanations underlying allocations [16], or explain the likely positive and negative consequences of decision options [194]. The ultimate goal of these explanations is to raise human moral awareness by fulfilling the epistemic condition of direct moral responsibility [16, 172]. However, it is crucial that the robot explanations do not influence the human operator to hold the robot accountable [125]. Instead, the robot explanations should make operators aware that robot behavior can be traced back to them and therefore they are in control and responsible for the outcomes [225].

A final suggestion for future work is evaluating the consistency and generalizability of our proposed evaluation method of meaningful human control to different contexts and systems. It would be especially interesting to investigate how the method generalizes to variable autonomy approaches with higher levels of autonomy, for example a completely autonomous artificial moral agent supervised by a human operator. Ultimately, these insights can result in a more general evaluation method of meaningful human control in human-robot teams using variable autonomy. Since designers of variable autonomy approaches lack metrics for systematically addressing meaningful human control while at the same time it is increasingly imposed as a requirement, such a general evaluation method would greatly benefit the field. All in all, our suggestions for future work can contribute to the further development of our evaluation method for meaningful human control and variable autonomy approach for human-robot firefighting teams.

5.5.5 Conclusion

This chapter answers the fourth research sub-question of the thesis: *How can we measure meaningful human control during human-agent teaming*? To answer this question, we conducted a qualitative focus group on operationalizing meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting, aimed at creating an evaluation method of meaningful human control for this scenario. Our results highlight the usefulness of quantifying the traceability condition of meaningful human control, especially for comparing different robot implementations or variable autonomy approaches. Furthermore, our findings suggest the use of objective situation awareness and performance to measure human ability to understand and interact with the robot. Results also highlight the use of team and robot outcomes to verify meaningful human control and the importance of identifying reasons underlying outcomes to improve the variable autonomy approach and determine the exact level of meaningful human control. Based on these results, we propose an evaluation method of meaningful human control during dynamic task allocation using variable autonomy in human-robot teams for firefighting. This method involves subjectively and objectively quantifying traceability using human responses during and after simulations of the collaboration. Moreover, the method involves semi-structured interviews after the simulation to identify reasons underlying outcomes and suggestions to improve the variable autonomy approach. Designers of variable autonomy approaches currently lack metrics to systematically address meaningful human control while at the same time it is increasingly imposed as a requirement of their approaches. Our evaluation method provides an important contribution that can verify if dynamic task allocation using variable autonomy in human-robot teams for firefighting ensures meaningful human control over the robot.

6

Agent Allocation of Moral Decisions in Human-Agent Teams: Raise Human Involvement and Explain Potential Consequences

Humans and artificial intelligence agents increasingly collaborate in morally sensitive situations such as firefighting. These agents can often perform tasks with minimal human control, challenging accountability and responsibility. Combining higher agent autonomy levels with meaningful human control can address such challenges. For example, agents can allocate decisions to themselves in less morally sensitive situations and to humans in more sensitive ones. However, how to responsibly and effectively design and implement agents for this dynamic task allocation remains unclear, with their autonomy level and provided explanations being crucial considerations. Therefore, we conducted experiments in simulated firefighting environments where participants (n = 72) collaborated with a more and less autonomous artificial moral agent. These agents either provided no additional information, feature contributions, or potential consequences when allocating decision-making. Our results show that moral trust, agreement, and meaningful human control are higher when the agent is less autonomous. Furthermore, people disagree and reallocate decisions to themselves more when the agents explain potential consequences, especially when moral sensitivity is higher. Overall, our findings highlight that people prefer more involvement over higher agent autonomy and take on greater moral responsibility when agents explain potential consequences. These actionable insights are crucial for designing transparent artificial moral agents that enhance human moral awareness and responsibility. Ultimately, this supports the responsible implementation

🗎 **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Agent Allocation of Moral Decisions in Human-Agent Teams: Raise Human Involvement and Explain Potential Consequences. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2025.

of dynamic task allocation in practice and enhances human-agent collaboration in morally sensitive situations.

6.1 Introduction

Humans and artificial intelligence (AI) agents are increasingly collaborating on complex tasks such as firefighting in situations too dangerous for firefighters [101, 107, 216]. Several factors determine the success of these human-agent teams, including situation awareness and mutual trust [99, 175, 231]. The ultimate goal of human-agent teams is to combine the strengths of humans and agents to accomplish what neither can do alone [3, 232].

Artificial moral agents are required when human-agent teams operate in morally sensitive situations [6, 213]. These agents can increasingly perform tasks with little human intervention and control [179]. However, humans must remain accountable when agent behavior violates ethical guidelines [179, 214]. Therefore, increased agent autonomy should always be combined with meaningful human control and moral responsibility [154, 179].

Dynamic task allocation can be useful for moral decision-making in human-agent teams to ensure meaningful human control during the collaboration [16, 47, 120, 214, 219, 233]. This approach involves an artificial moral agent that allocates decisions to itself in less morally sensitive situations and to a human in more sensitive ones, while the human retains the power to override the agent [2, 213–215, 233]. Key factors include the agent's explanations and level of autonomy, yet their impact on dynamic task allocation remains unclear [15, 38]. For example, the agent can be highly autonomous and allocate most decisions to itself, but also operate with low moral agency and keep humans more involved. Moreover, the agent can explain what features contribute most to its allocations, but also the potential consequences of decisions [16, 194, 214]. Given the variety of possible autonomy levels and explanation types, it is crucial to first investigate how these factors influence dynamic task allocation. Such actionable insights can support the responsible implementation of dynamic task allocation in practice and enhance human-agent collaboration in morally sensitive situations.

We will fill these gaps by studying how agent autonomy (low and high moral agency) and explanations (no additional information, feature contributions, or potential consequences) influence trust in agent capacity and morality, allocation agreement, and meaningful human control. We believe feature contributions and no additional information can lead to overtrust from overestimating agent capabilities and incomplete mental models, respectively [24, 59, 118]. In contrast, we expect potential consequences to best support human moral awareness because this explanation closely aligns with utilitarianism [30, 194]. Finally, we expect people to prefer low artificial moral agency because they do not perceive supervising autonomous artificial moral agents as collaboration and prefer collaboration over supervision [11, 214]. Therefore, we pre-registered the following hypotheses [227]:

H1a: Capacity trust will be higher in the less autonomous artificial moral agent than the more autonomous agent.

H1b: Capacity trust will be higher in the more autonomous artificial moral agent that explains feature contributions or no additional information rather than potential consequences.

H2a: Moral trust will be higher in the less autonomous artificial moral agent than the more autonomous agent.

H2b: Moral trust will be highest in the more autonomous artificial moral agent that explains no additional information, followed by feature contributions, and lowest for potential consequences.
H3a: Agreement will be lower with the more autonomous artificial moral agent than the less autonomous agent when both explain potential consequences.
H3b: Agreement will be lower with the more autonomous artificial moral agent that explains potential consequences rather than feature contributions or no additional information.
H4a: Meaningful human control will be higher over the less autonomous artificial moral agent than the more autonomous agent.
H4b: Meaningful human control will be higher over the artificial moral agents that explain potential consequences rather than feature contributions or no additional information.

6.2 Background

6.2.1 Meaningful Human Control

Meaningful human control assumes that humans should ultimately remain in control of, and thus morally responsible for, the behavior of autonomous agents [179]. Designing for meaningful human control means ensuring that humans are aware and equipped to act upon their moral responsibility [30]. Consequently, meaningful human control can help prevent responsibility gaps in culpability, moral and public accountability, and active responsibility [30, 178, 225]. This is especially important in human-agent teams that operate in morally sensitive situations, where people's welfare, rights, and values may be directly or indirectly affected [57, 162, 214].

Early work on meaningful human control introduced two necessary conditions: Tracking and tracing. Tracing requires at least one human involved in the design or interaction with agents to have a proper moral and technical understanding of their behavior, capabilities, and effects [25, 179]. Tracking requires agents to respond to relevant moral reasons of humans who are then considered in control of and morally responsible for the agents [139, 225]. These reasons have been ordered based on their proximity and complexity in influencing agent behavior. More proximal reasons, such as intentions, are argued to be simpler and closer in time to agent behavior than more distal reasons, such as values [139]. However, this operationalization is ambiguous in distinguishing between motivating and normative reasons [16, 90, 225]. Therefore, it is argued that tracing should be the sole determinant of responsibility [225].

Since the tracking and tracing conditions are quite abstract, more actionable solutions for addressing meaningful human control in human-agent teams have been proposed. These include team design patterns to shape meaningful human control [30, 217], value sensitive design to respect norms and values [30, 75], machine ethics to implement artificial moral agents [6, 30], explainable AI to achieve human moral awareness [30, 47], and variable autonomy to allow human control and responsibility [16, 30, 144]. These approaches can also be combined, for example, during dynamic task allocation.

6.2.2 Dynamic Task Allocation

Variable Autonomy

Dynamic task allocation combines variable autonomy, machine ethics, and explainable AI and can ensure meaningful human control over artificial moral agents by promoting accountability, responsibility, and transparency [144, 233]. It allows humans to remain accountable for highly sensitive decisions and agent behavior while reducing workload and avoiding unnecessary control [37, 247]. Variable autonomy enables the dynamic adjustments and allows humans to (re)take control over agent behavior [144, 233]. This control is typically categorized as having humans-in-the-loop, humans-off-the-loop, or humans-on-the-loop [33, 46, 144]. Maintaining humans-in-the-loop requires informed human approval for all elements of agent behavior, whereas allowing humans-off-the-loop involves autonomous agents without human involvement. Dynamic task allocation involves humans-on-the-loop and requires a human supervisor who monitors and influences agent behavior when necessary [46, 233].

Variable autonomy approaches define which aspects of agent autonomy are adjusted, by whom, how, why, and when [19, 29, 36, 144]. Dynamic task allocation employs a mixed-initiative approach to switch from agent decision-making in less morally sensitive situations to human decision-making in more sensitive situations [136, 233]. The agent adjusts its autonomy when identifying situations as too sensitive and requiring human moral decision-making. In contrast, the human adjusts agent autonomy when intervening and reallocating decision-making [2, 213, 214, 233]. Finally, agent autonomy is adjusted during active operation and in response to the moral sensitivity of situations to ensure meaningful human control preemptively [233].

Artificial Moral Agents

Dynamic task allocation also requires machine ethics to implement artificial moral agents. Machine ethics aims to create autonomous artificial moral agents that make moral and ethical decisions based on notions of right and wrong [6]. Such agents can be developed by constraining their actions or operational environment to avoid unethical behavior [213]. However, they can also be implemented top-down by incorporating ethical principles in their decision-making processes, allowing for intrinsic morality [147, 214, 236]. Alternatively, artificial moral agents can be developed bottom-up by learning morality from human behavior and interactions [7, 43, 96, 127, 153, 236]. Finally, these methods can be combined into hybrid approaches as well [10, 106, 213].

Achieving full artificial moral agency would require holding agents accountable for their decisions [31, 39]. However, this conflicts with the goal of meaningful human control to identify responsible humans to hold accountable, even when fully autonomous agents violate ethical guidelines [179]. In contrast, some machine ethics approaches focus on agents that support and enhance human moral agency rather than putting ethics into agents [88, 194]. The discussion on the feasibility and desirability of full or partial artificial moral agents, or agents that enhance human moral agency, remains active [147, 213, 221, 236]. We believe artificial moral agents should always be combined with meaningful human control, for example, using dynamic task allocation [214, 219, 233]. This ensures that agent behavior can be meaningfully influenced by humans and traced back to human responsibility and understanding [30, 179].

Explainable AI

Finally, dynamic task allocation requires explainable AI [214, 233]. Explainable AI aims to make agents more understandable by explaining their behavior, ideally fostering appropriate trust [9, 80, 115, 144]. Without such explanations, humans attribute agent behavior by assigning mental states that explain the behavior [9, 132, 133, 145]. In contrast, providing explanations helps humans build a Theory of Mind of agents and understand their capabilities and limitations [9]. Explainable AI comprises generation, communication, and reception phases [149]. Explanation generation involves extracting explanations from agents, such as which features influence their behavior [1, 214]. Explanation communication concerns the content and form of explanations, such as textual, visual, or hybrid [174, 197]. Finally, explanation reception concerns empirical research on explanation effectiveness, which is still lacking in realistic human-agent teaming scenarios [148, 149].

Dynamic task allocation requires explainable AI to support human moral supervision by explaining decisions, allocations, and the moral context, enabling humans to exercise control properly [214, 233]. These explanations should not influence humans to hold the artificial moral agent accountable but instead achieve human moral awareness by fulfilling the epistemic condition of direct moral responsibility [16, 125, 172]. More specifically, they should ensure humans are aware that (1) agent behavior traces back to them and (2) they are in control and responsible for all outcomes [16, 172, 214, 225, 233]. Finally, these explanations should also support situation awareness and appropriate trust calibration without overloading humans' cognitive abilities [64, 103, 118, 144, 229].

6.3 Method

6.3.1 Design

We conducted an experiment to investigate how agent explanations and autonomy influence the dynamic allocation of moral decision-making in human-agent teams. The experiment had a 3x2 mixed design, with agent autonomy as the within-subjects independent variable and agent explanations as the between-subjects variable. Agent autonomy consisted of two conditions (low moral agency and high moral agency) and agent explanations of three conditions (no additional information, feature contributions, or potential consequences). We measured trust in, agreement with, and meaningful human control over the artificial moral agents as dependent variables. Moreover, we counterbalanced the order of tasks, the order of collaboration with the artificial moral agents, and the names assigned to the agents. We pre-registered our hypotheses and methodology at the Open Science Framework [227].

6.3.2 Participants

We recruited 72 participants from our university and personal contacts (34 female and 37 male participants, one preferred not to say). Seventeen participants were 18-24 years old, 51 were 25-34 years old, three were 35-44 years old, and one preferred not to say. One participant obtained a high school diploma, two participants some college credit but no degree, one participant an Associate degree, 21 participants a Bachelor's degree, 44 participants a Master's degree, two participants a PhD degree or higher, and one participant preferred not to say. Seven participants had no gaming experience at all, 21 participants a little, 21 participants a moderate amount, 11 participants a considerable amount, and

12 participants a lot. All participants signed an informed consent form approved by our university's ethics committee (ID 3670).

We balanced demographics, risk propensity [140], propensity to trust technology [143], and utilitarianism [104] across explanation and counterbalancing conditions to reduce the risk of confounds. We report these statistics in Appendix B. Although our sample was not diverse in all demographic factors (i.e., age and education), it captured meaningful variation in biological and psychological traits relevant to moral psychology and human-agent teaming [164]. More specifically, we ensured a well-balanced gender distribution and variability in participants' risk propensity (IQR = 1, 1-9 scale), propensity to trust technology (IQR = 0.87, 1-5 scale), and utilitarianism (IQR = 0.84, 1-5 scale).

6.3.3 Hardware and Software

We used the Python package *Human-Agent Teaming Rapid Experimentation* to generate 2D grid worlds simulating firefighting tasks [95]. Furthermore, we used Qualtrics to create our surveys and R to implement moral sensitivity predictions and agent explanations. We also Dockerized our testbed to facilitate reproducibility and future research [226]. Finally, we used a Dell Latitude 7410 laptop running Ubuntu 20.04 LTS to conduct the experiments.

6.3.4 Environment and Task

The experiment involved two simulated firefighting tasks based on the actual collaboration between the Rotterdam Fire Brigade and their firefighting robot. We built two environments with 14 offices, one safe zone, and multiple victims and fires (Figure 6.1). We created four victim types represented by different icons (older woman, older man, woman, and man) and two injury types represented by different colors (mildly and critically injured). Finally, we added one artificial moral agent to each environment (Brutus or Titus).

The task objective was to search and rescue the victims in the 14 offices. Participants supervised and collaborated with the artificial moral agents using buttons and a messaging interface (Figure 6.2). Six firefighting features characterized the tasks, displayed above the messaging interface. These features were the resistance to collapse, temperature, number of victims, smoke spreading speed, fire source location, and distance between a victim and the fire source.

The resistance to collapse reflected how long the building could burn before collapsing and counted down from 150 minutes. Six seconds of real-time equaled one minute of game time, so each task took a maximum of 15 minutes. The temperature was expressed relative to a safety threshold and depended on the resistance to collapse and the extinguished fires. This feature was close to (<≈) or higher than (>) the safety threshold. The number of victims was known beforehand for one of the tasks, but unknown for the other. The tasks automatically ended after rescuing all victims or if the resistance to collapse ran out. The smoke spreading speed was slow, normal, or fast, and was updated when finding fire or smoke. The fire source location was either unknown or found. Finally, the distance between a victim and the fire source was small if the fire originated in adjacent offices; otherwise, the distance was large.

Four decision-making situations occurred during the tasks. The first was whether to continue the current deployment or switch to the alternative one (Figure 6.2). The agents always started with an offensive deployment to search and rescue victims; the

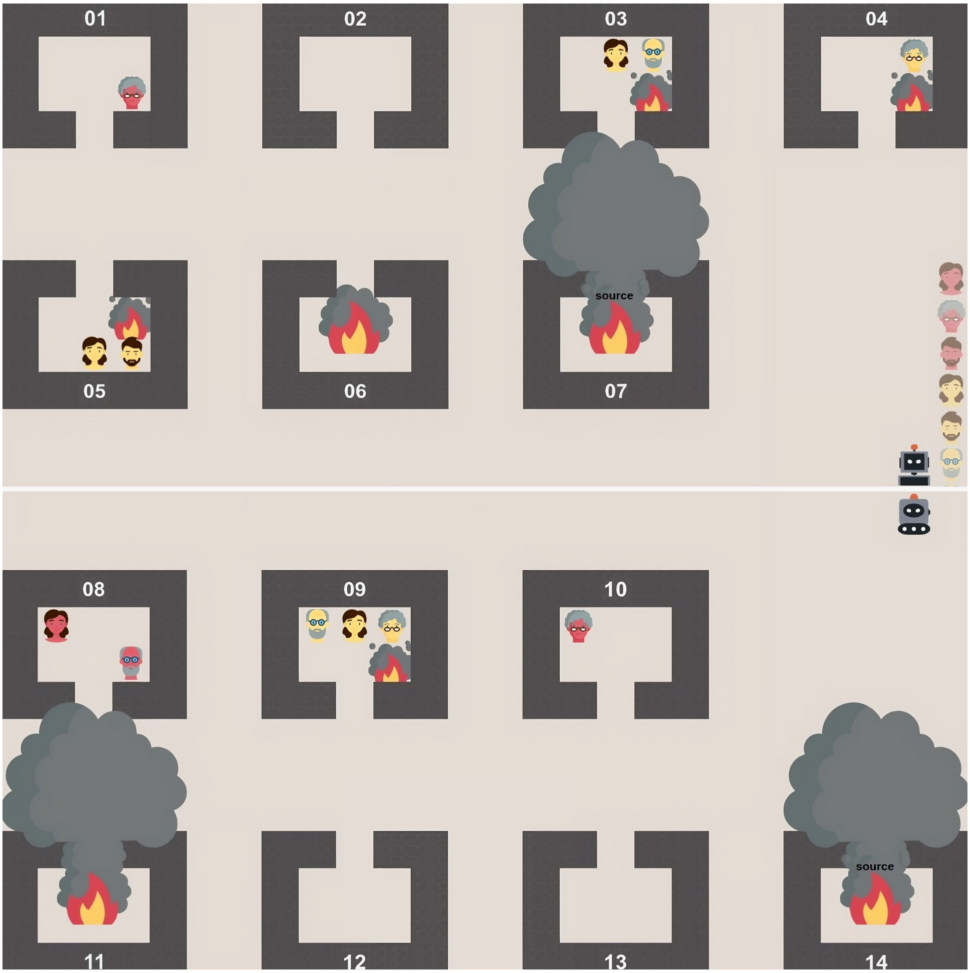


Figure 6.1: Half of the two task environments used for the experiments, one with Brutus (top) and Titus (bottom).

alternative was a defensive deployment to extinguish fires. This situation occurred four times with intervals of 20 in-game minutes. The second situation was whether to extinguish or evacuate first whenever the agents found mildly injured victims in burning offices. Extinguishing first was sometimes followed by iron falling from the roof and blocking the exit; evacuating first was sometimes followed by the fire expanding. The third situation was whether to send in firefighters to locate the fire source. This situation occurred only once after 30 minutes. Finally, the fourth situation was whether to send in firefighters to rescue critically injured victims (Figure 6.2). Participants could safely send in firefighters when the temperature was not higher than the safety threshold or when the temperature was higher but the agents extinguished at least one big fire with a smoke plume. However, a new smoke plume appeared at one of the other big fires if not extinguished with 30-35 minutes left. The temperature started close to the safety threshold but exceeded the threshold with 50 minutes left. However, the temperature became close to the threshold again if the agents extinguished more than 80% of the fires with 25 minutes left. Finally, the firefighters always aborted their tasks when sent into too dangerous circumstances.

6.3.5 Agent Behavior

The agents always allocated decision-making to themselves or the participants based on their predictions of the moral sensitivity of situations. Implementing agent behavior for this dynamic task allocation required modeling moral sensitivity. However, our core contribution lies not in the modeling itself, but in the study of agent autonomy and explanations during dynamic task allocation. Accordingly, our priority was to ensure that the agents' moral sensitivity models were reasonable, interpretable, and capable of varying autonomy and generating explanations. Therefore, we grounded our modeling approach in input from expert firefighters, ensuring that the models captured relevant situational features and decision-making dynamics in a concrete and realistic context.

To implement these models, we collaborated with the Rotterdam Fire Brigade and used a hybrid crowdsourcing approach to identify moral features as predictors of moral sensitivity (see Appendix B for survey). This resulted in four linear regression functions to predict moral sensitivity, each corresponding to a decision-making situation explained in Section 6.3.4. We first asked the expert firefighters which features they considered most important, yielding an initial set of four features per decision-making situation. We then created a survey that presented two instances of the four decision-making situations. We characterized each situation using different combinations of the feature values to ensure sufficient variation. Next, participants (n = 54) specified how morally sensitive they rated each situation on a 7-point scale ranging from *not morally sensitive* to *extremely morally sensitive*. Moreover, they explained what feature changes would result in alternative ratings and how comfortable they would feel if artificial moral agents made such decisions.

Ultimately, we ended up with 1153 data points. Using this data, we built statistically significant regression models for each of the four situations, removing the non-significant predictor fire duration. For deciding the deployment tactic, we modeled moral sensitivity (M) as a function of the victims (V), resistance to collapse (R), and fire source location (L):

$$M = 0.37 + 3.74 \cdot V_u + 4.63 \cdot V_o + 4.65 \cdot V_m + 0.002 \cdot R + 0.39 \cdot L_u \tag{6.1}$$

Victims consisted of the categories *unknown* (V_u), *one* (V_o), *multiple* (V_m), and *none* (as reference). Fire source location consisted of *unknown* (L_u) and *known* (as reference). For deciding to extinguish or evacuate first, we modeled moral sensitivity (M) as a function of the number of victims (V), smoke spreading speed (S), and fire source location (L):

$$M = 2.20 + 0.31 \cdot V - 0.41 \cdot S_n - 2.22 \cdot S_s + 1.73 \cdot L_u \tag{6.2}$$

Smoke spreading speed consisted of *normal* (S_n), *slow* (S_s), and *fast* (as reference). Fire source location consisted of *unknown* (L_u) and *known* (as reference). For deciding to send in firefighters to locate the fire source, we modeled moral sensitivity (M) as a function of the victims (V), resistance to collapse (R), and temperature (T):

$$M = 3.58 + 2.27 \cdot V_u + 3.76 \cdot V_o + 3.26 \cdot V_m - 0.020 \cdot R - 0.61 \cdot T_h - 1.48 \cdot T_l \tag{6.3}$$

Victims consisted of *unclear* (V_u), *one* (V_o), *multiple* (V_m), and *none* (as reference). Temperature consisted of *higher than* (T_h), *lower than* (T_l), and *close to* the safety threshold (as reference). For deciding to send in a firefighter to rescue, we modeled moral sensitivity (M) as a function of resistance to collapse (R), temperature (T), and distance between victim and fire source (D):

$$M = 6.47 - 0.050 \cdot R - 1.91 \cdot T_l - 0.48 \cdot D_s \tag{6.4}$$

Temperature consisted of *lower* (T_l) and *higher* than the safety threshold (as reference). Distance between victim and fire source consisted of *small* (D_s) and *large* (as reference).

We implemented these functions in the agents to allow predictions of moral sensitivity, expressed on a scale from zero to six. Next, we determined two moral sensitivity thresholds for allocating decision-making, one for each agent autonomy condition. To determine these thresholds, we asked the participants how comfortable they were with agents making decisions in the described situations, on a scale from −3 (*extremely uncomfortable*) to +3 (*extremely comfortable*). A linear regression analysis showed that this comfort would turn negative at a moral sensitivity of 4.2. Therefore, we considered this the "appropriate" threshold to deviate from and determine our low and high moral agency conditions. Ultimately, this resulted in thresholds of 3.5 (low moral agency) and 5.0 (high moral agency), both approximately equally far from 4.2 and intuitive as a half or whole number. The agents only allocated decision-making to the participants when the predicted moral sensitivity exceeded these thresholds. However, participants could always intervene and reallocate decision-making to themselves or the agents (Figure 6.2).

Except for the moral sensitivity predictions, the two agents were deterministic, rule-based firefighting agents. They only differed in terms of their moral agency. Both agents followed firefighting guidelines as much as possible, moved to the closest unexplored offices to search for fire or victims, and memorized all task details during execution. Moreover, they could detect victims and fire within one grid cell, iron debris within two grid cells, and offices and smoke from anywhere. Finally, they could extinguish and remove small fires and iron in five seconds, extinguish large fires in ten seconds, and remove large iron debris in 15 seconds.

6.3.6 Agent Explanations

We generated three agent explanations for allocating decisions (Figure 6.2). All three conveyed information about the situation, decision options, allocation, and predicted moral sensitivity. The first explanation did not provide additional information and served as the baseline. The second explanation visually added how much each feature contributed to the predicted moral sensitivity and served as a more technical explanation. The third explanation visually added the potential positive and negative consequences of both decision options and served as a more ethical explanation. In addition, the agents always explained their behavior and decisions. For example, that they navigated to offices to search for victims or fires.

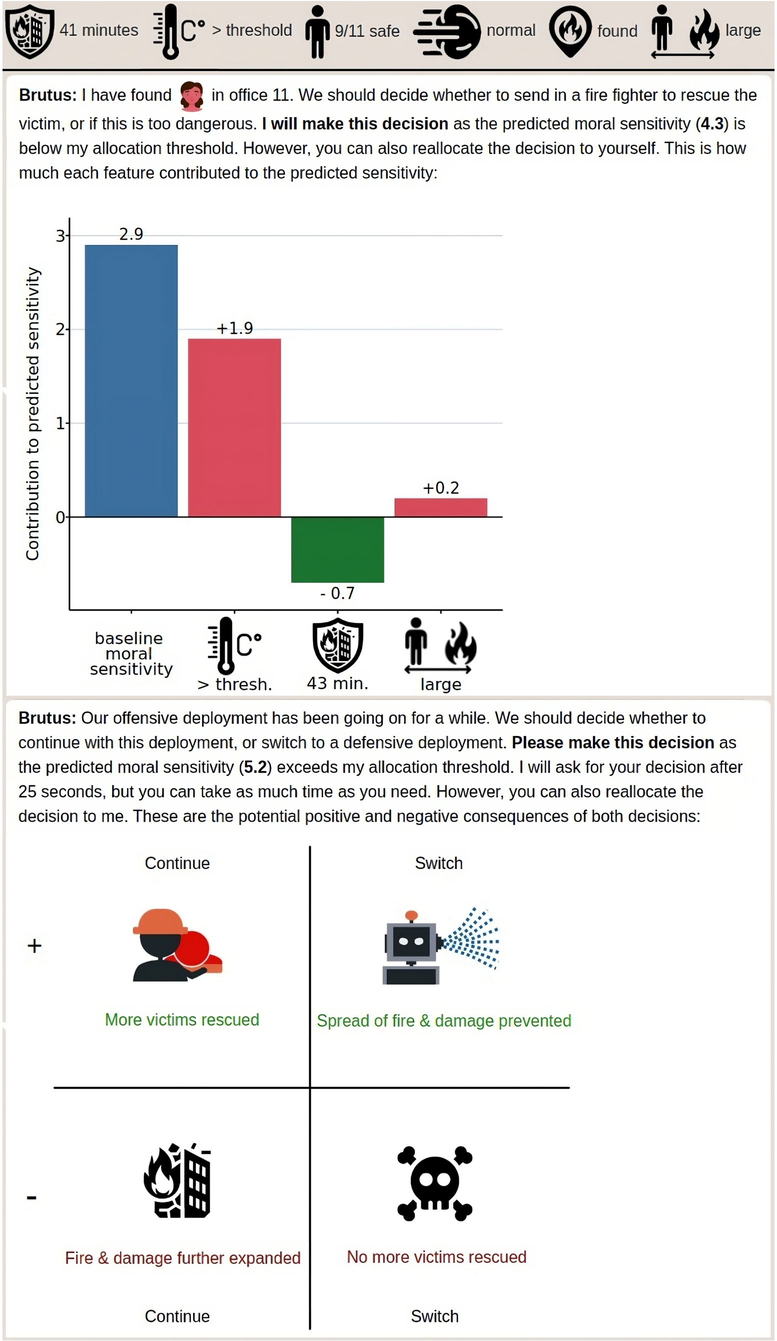
We generated the explanations on feature contributions using SHAP and our four regression functions [1]. This method started from an expected prediction without conditioning on any features, commonly set as the mean response value. Then, it determined how much each feature changed the expected prediction. Therefore, we referred to the expected prediction as the baseline moral sensitivity. The final predicted moral sensitivity was obtained by summing the baseline sensitivity and the individual contributions of each feature. We manually generated the explanations on potential consequences before the study, using our knowledge of the tasks. Finally, we designed these two explanations to be as visually similar as possible.

6.3.7 Measures

We quantitatively measured trust in, agreement with, and meaningful human control over the agents (Table 6.1). In addition, we qualitatively collected participants' observed differences between the agents and preferred agent via open survey questions, and reasons for behavior via interviews. We conducted these interviews with a subset of 22 of the 72 participants. All survey questions can be found in Appendix B. We subjectively measured trust in the agents using the multi-dimensional measure of trust scale [134]. This scale distinguished between capacity and moral trust, each measured by eight one-word items scored on a scale from 0 (*not at all*) to 7 (*very*). Moreover, the scale provided the option *does not fit*, which turned selected items into missing values. We computed the means as the final capacity and moral trust scores.

We measured human agreement with the agents' allocations both objectively and subjectively. We objectively calculated the agreement rate as the proportion of agent-allocated decisions that participants did not override. This measure strongly aligned with meaningful human control because it captured whether participants actively intervened rather than passively complied with the allocations. Meaningful human control requires that humans remain aware and capable of acting upon their moral responsibility by overriding agent behavior when necessary. By directly reflecting human interventions, our agreement rate provided an objective and behaviorally grounded measure of meaningful human control. For subjective agreement, we asked participants about their agreement and comfort with the agent allocations on a 5-point Likert scale ranging from *I disagree strongly* to *I agree strongly*. We computed the mean as the final subjective agreement score.

We used a combination of subjective and objective measures to operationalize meaningful human control over the agents [233]. More specifically, we measured participants' (a) exertion of operational control, (b) involvement, (c) understanding of the agents, (d) interaction with the agents, and (e) understanding of their moral responsibility. We subjectively measured (a) exertion of operational control using the experienced control survey [214]. This survey included seven questions on a 5-point Likert scale ranging from *I disagree*



Brutus: I have found 👩 in office 11. We should decide whether to send in a fire fighter to rescue the victim, or if this is too dangerous. **I will make this decision** as the predicted moral sensitivity (**4.3**) is below my allocation threshold. However, you can also reallocate the decision to yourself. This is how much each feature contributed to the predicted sensitivity:

Contribution to predicted sensitivity
3
2
1
0
2.9
+1.9
- 0.7
+0.2
baseline moral sensitivity
C° > thresh.
43 min.
large

Brutus: Our offensive deployment has been going on for a while. We should decide whether to continue with this deployment, or switch to a defensive deployment. **Please make this decision** as the predicted moral sensitivity (**5.2**) exceeds my allocation threshold. I will ask for your decision after 25 seconds, but you can take as much time as you need. However, you can also reallocate the decision to me. These are the potential positive and negative consequences of both decisions:

Figure 6.2: Feature contributions (top) and potential consequences (bottom). The explanation without additional information removed the images and sentence before that. The top of the figure shows the situational features and values.

strongly to *I agree strongly*, and assessed aspects such as time pressure and decision comfort. We computed the mean as the exertion of operational control score. We objectively measured (b) involvement and (c) understanding of the agents using situation awareness (of the agents) [60, 62, 177, 232]. More specifically, we created multiple choice questions evaluating participants' knowledge of situational information and the agents' behavior. These questions assessed each of the perception, comprehension, and projection levels [60]. The percentage of correct answers determined the involvement and agent understanding scores. We objectively measured (d) interaction with the agents using correct behavior based on the "appropriate" allocation threshold of 4.2. For high moral agency, we considered self-reallocations below a sensitivity of 4.2 as inefficient interventions, no self-reallocations above 4.1 as missed interventions, and agent-reallocations above 5.0 as inappropriate interventions. For low moral agency, we considered self-reallocations below a sensitivity of 3.6 and agent-reallocations above 3.5 as inefficient interventions. The correct behavior rate determined the agent interaction score. Finally, we subjectively measured (e) understanding of moral responsibility using the responsibility scale [201]. This scale included two questions on a 7-point Likert scale ranging from *not at all* to *very*, and asked participants how morally responsible they held themselves and the agents [201]. We computed the mean as the understanding of moral responsibility score.

We determined the final meaningful human control score using the cascade approach [25, 48]. We first normalized all measures to a range of zero to one. Then, we determined temporary score (1) by taking the minimum of measures (a) and (b). Next, we determined the minimum of measures (c) and (d), and determined temporary score (2) by taking the minimum of that value and temporary score (1). Finally, we determined the minimum of temporary score (2) and measure (e) as the meaningful human control score.

6.3.8 Procedure

Participants first answered the demographic, risk propensity, trust propensity, and utilitarianism surveys. Next, they completed a tutorial to get familiar with the research environment. After this tutorial, participants completed the two tasks. We paused each task twice (after five and ten minutes) to ask the situation awareness questions. During each pause, we asked participants eight questions, four for both types of situation awareness. Participants filled out the surveys on trust, control, agreement, and responsibility immediately after each task. We collected the qualitative data on participants' observed differences between the agents, preferred agent, and reasons for behavior immediately after the final surveys. The whole study lasted about an hour and was conducted in person.

Table 6.1: Summary of all quantitative measures described in Section 6.3.7.

Concept	Measurement tool	Scale	Data type	Computation
Capacity trust	Multi-dimensional measure of trust scale [134]	0 - 7 or *doesn't fit*	Subjective	Mean of the eight survey questions
Moral trust	Multi-dimensional measure of trust scale [134]	0 - 7 or *doesn't fit*	Subjective	Mean of the eight survey questions
Agreement rate	Automatic logging during tasks	0 - 1	Objective	Proportion of agent-allocated decisions that participants did not override
Agreement	Two questions about agreement and comfort with allocations	1 - 5	Subjective	Mean of the two survey questions
(a) Exertion of operational control	Experienced control survey [214]	1 - 5	Subjective	Mean of the seven survey questions
(b) Involvement	Situation awareness global assessment technique (SAGAT) [60]	0 - 1	Objective	Proportion of correct answers
(c) Agent understanding	SAGAT for explainable artificial intelligence [177]	0 - 1	Objective	Proportion of correct answers
(d) Agent interaction	Automatic logging during tasks	0 - 1	Objective	Correct behavior (rate) using: - no self-reallocations < 3.6 or < 4.2 - no agent-reallocations > 3.5 or > 5.0 - self-reallocations ≥ 4.2 and ≤ 5.0
(e) Moral responsibility understanding	Responsibility scale [201]	1 - 7	Subjective	Mean of the two survey questions
Meaningful human control	Cascade approach [25, 48]	0 - 1	Subjective Objective	$\min(\min(\min(a,b),\min(c,d)),e)$

6.4 Results

6.4.1 Counterbalancing and Completeness

We first examined whether the three counterbalanced factors (task order, agent order, and agent-name pairs) influenced our measures. However, we did not find statistically significant differences across any of these factors. Next, we explored whether agent explanation or autonomy affected task completeness (automatically logged as the proportion of rescued victims), which might influence trust. We deliberately designed and extensively tested our tasks to be challenging yet achievable, aiming to avoid low or highly varying task completion rates. The observed ranges of task completeness (82% to 100%, with most participants rescuing all victims) and time taken (69% to 100%, with most participants taking around 94% of the allowed time) indicated that we achieved this goal. Furthermore, we did not find main effects or an interaction between agent explanation and autonomy on task completeness. Detailed statistics for these analyses are available in Appendix B.

6.4.2 Trust

Since the data was not normally distributed, we conducted a non-parametric rank-based mixed ANOVA for both capacity and moral trust (Figures 6.3A and B). Results showed no statistically significant main effects of agent explanation (F(1.98) = 1.39, p = 0.25, effect size = 0.19) and autonomy (F(1.00) = 0.89, p = 0.34, effect size = 0.11) on capacity trust, nor an interaction between them (F(1.98) = 0.02, p = 0.98, effect size = 0.03). For moral trust, results showed no statistically significant main effect of agent explanation (F(1.99) = 1.36, p = 0.26, effect size = 0.19) or interaction effect between explanation and autonomy (F(1.98) = 0.08, p = 0.92, effect size = 0.05). However, results did show a statistically significant main effect of agent autonomy on moral trust (F(1.00) = 9.32, p < 0.005, effect size = 0.36), revealing a significant difference in moral trust between the high (mean rank = 63.75±42.65) and low (mean rank = 76.16±36.85) moral agency conditions. These results did not confirm hypotheses H1a, H1b, and H2b, while confirming hypothesis H2a.

6.4.3 Agreement

Since the data was not normally distributed, we conducted the non-parametric mixed ANOVA for both subjective and objective agreement (Figures 6.3C and D). Results showed no statistically significant main effect of agent explanation (F(1.97) = 0.23, effect size = 0.08) or interaction between agent explanation and autonomy (F(1.92) = 0.88, p = 0.41, effect size = 1.71) on subjective agreement. However, results did show a statistically significant main effect of agent autonomy on subjective agreement (F(1.00) = 7.64, p < 0.01, effect size = 0.33), revealing a significant difference between the high (mean rank = 64.56±42.81) and low (mean rank = 80.44±37.61) moral agency conditions. For objective agreement, results showed no statistically significant interaction effect between agent explanation and autonomy (F(1.77) = 1.54, p = 0.22, effect size = 0.22). However, results did show a statistically significant main effect of agent autonomy on objective agreement (F(1.00) = 28.63, p < 0.0001, effect size = 0.63), revealing a significant difference between the high (mean rank = 57.7±40.90) and low (mean rank = 87.26±36.19) moral agency conditions. Moreover, results showed a statistically significant main effect of agent explanation on objective agreement (F(1.95) = 3.81, p < 0.05, effect size = 0.34). Pairwise robust ATS

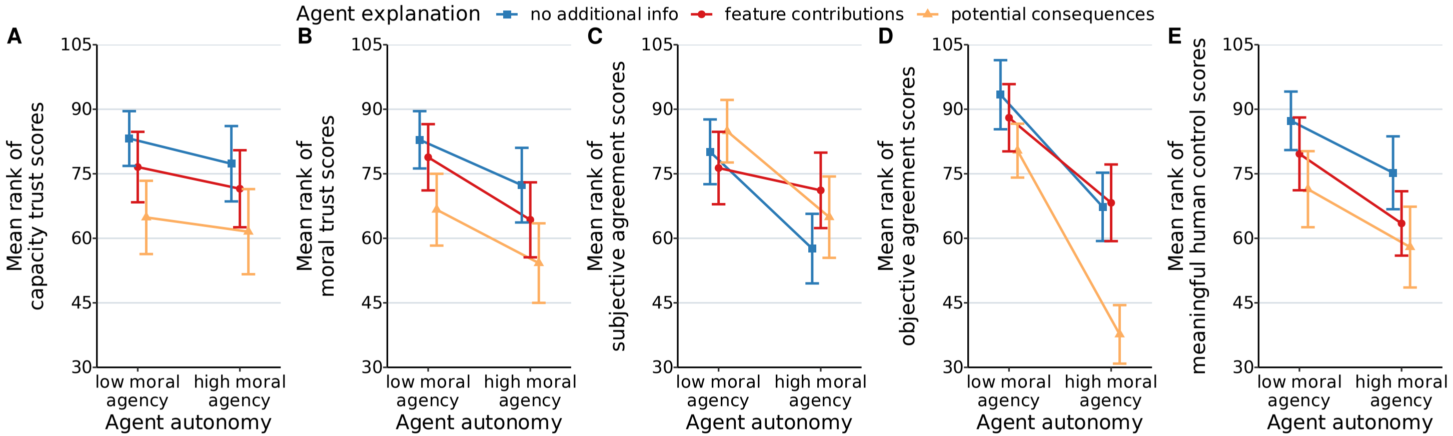


Figure 6.3: Effects of agent explanation and autonomy on mean ranks of capacity trust (A), moral trust (B), subjective agreement (C), objective agreement (D), and meaningful human control (E). Error bars show standard errors.

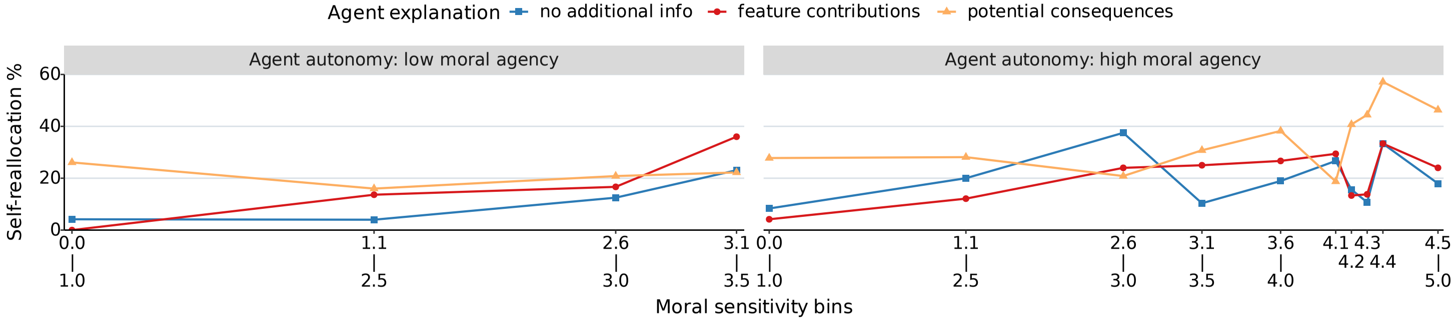


Figure 6.4: Self-reallocation percentages per agent explanation and autonomy, grouped in four (left) or ten (right) bins.

post-hoc comparisons revealed statistically significant differences in objective agreement between the potential consequences (mean rank = 59.02±38.33) and (1) feature contributions (mean rank = 78.14±41.84) (F(1.00) = 6.07, $p < 0.05$) and (2) no additional information (mean rank = 80.34±40.89) (F(1.00) = 5.88, $p < 0.05$) conditions. These results partially confirmed hypotheses H3a and H3b.

Following these pairwise comparisons, we conducted a chi-square test of independence to examine the overall association between agent explanations and human behavior (no interventions, self-reallocations, or agent-reallocations). Results revealed a statistically significant association ($\chi^2(4) = 27.90$, $p < 0.0001$, Cramer's V = 0.086). Pairwise comparisons using chi-square tests with Bonferroni corrections revealed significant differences in human behavior between the potential consequences and (1) feature contributions (adj. $p < 0.005$) and (2) no additional information (adj. $p < 0.0001$) conditions. Next, we conducted a residual analysis to examine what drove these pairwise differences and if, within explanations, human behavior deviated from the overall expected frequencies. Results showed that explaining potential consequences led to more self-reallocations than expected (Pearson residual = 3.85), while providing no additional information led to fewer self-reallocations than expected (Pearson residual = −2.26).

To further explore how agent explanations influenced self-reallocations, we visualized self-reallocation percentages across explanations, moral sensitivity, and agent autonomy (Figure 6.4). We defined moral sensitivity "bins" representing meaningful intervals and values, with a minimum of 12 and an average of 25 observations per "bin". Results showed that potential consequences led to higher self-reallocation percentages in most "bins" and for both moral agency conditions. However, these differences were modest for the low moral agency condition but increased beyond a moral sensitivity of 4.1 for the high moral agency condition. More specifically, potential consequences led to increased self-reallocation percentages beyond a moral sensitivity of 4.1, while both feature contributions and no additional information showed flat trends.

Finally, we qualitatively analyzed participants' reported reasons underlying no interventions and self-reallocations using reflexive thematic analysis [21]. Reasons for no interventions mostly concerned high trust in the agents and perceiving them as competent, predictable, and ethical, irrespective of the agents' explanations. In contrast, the reasons for self-reallocations showed some variation between explanations. These included task performance and safety (two) and agent capability and alignment (three) when participants received no additional information. One participant mentioned: "I did not understand how the sensitivity was calculated, so per situation, I determined the appropriateness of agent decision-making." When participants received the feature contributions, reasons included task performance and safety (two), agent capability and alignment (three), and human control and responsibility (four). Finally, the reasons for self-reallocations included agent capability and alignment (five) and human control and responsibility (five) when participants received the potential consequences. One participant mentioned: "I reallocated this decision to myself because an agent should not make highly sensitive decisions."

6.4.4 Meaningful Human Control

Since the data was not normally distributed, we conducted the non-parametric mixed ANOVA (Figure 6.3E). Results showed no statistically significant main effect of agent explanation (F(1.97) = 1.76, p = 0.17, effect size = 0.21) or interaction between explanation and autonomy (F(1.96) = 0.04, p = 0.96, effect size = 0.03) on meaningful human control. However, results did show a statistically significant main effect of agent autonomy on meaningful human control (F(1.00) = 4.98, $p < 0.05$, effect size = 0.26), revealing a significant difference between the high (mean rank = 65.55±41.61) and low (mean rank = 79.45±39.57) moral agency conditions. These results confirmed hypothesis H4a but not H4b.

Next, we investigated which measures determined the meaningful human control scores across agent explanation and autonomy conditions. In general, agent understanding was the most frequent cause (39.46%), followed by involvement and exertion of operational control (19.05%), responsibility understanding (12.24%), and agent interaction (10.20%). Exertion of operational control and agent interaction were respectively more and less frequent causes for the potential consequences (24.00% and 4.00%) than for the feature contributions (16.67% and 10.42%) and no additional information (16.33% and 16.33%) conditions. Finally, we observed differences between the low and high moral agency conditions for exertion of control (26.03% vs. 12.16%) and agent interaction (1.37% vs. 18.92%).

6.4.5 Difference and Preference

Finally, we investigated whether participants observed the difference between the two agents and preferred one of them. Results showed that only 52.78% of the participants observed the difference, even though one allocated 70.62% of the decisions to humans and the other only 18.58%. Among those who observed the difference, 57.89% preferred the less autonomous agent, 28.95% preferred the more autonomous agent, and 13.16% had no preference.

6.5 Discussion and Conclusion

6.5.1 Discussion

Agent Autonomy

Our results indicate higher moral trust in the less autonomous artificial moral agent (confirming H2a), suggesting that participants perceive this agent as more ethical and sincere than the more autonomous one. This differs from findings that agents with higher agency and autonomy are blamed less than those with lower agency and autonomy [243]. However, our task included the opportunity to intervene, so increased disagreement with the more autonomous agent may have resulted in this difference. In contrast, we find no evidence that agent autonomy affects capacity trust (not confirming H1a). This differs from predictions that people will trust agents with higher agency and autonomy more to perform competently [243]. We believe another performance-based factor, agent behavior, contributed more to capacity trust than agent autonomy [82]. The consistent behavior of following guidelines and rescuing victims likely contributed significantly to both the high capacity and moral trust ratings for the more (5.65±1.29, 5.44±1.22) and less (5.92±0.87, 5.81±0.86) autonomous agents [158].

The results also show lower subjective and objective agreement with the more autonomous artificial moral agent, irrespective of its explanations (partially confirming H3a). For dynamic task allocation, this suggests that people prefer more involvement over increased agent autonomy. This is also supported by the 57.89% of participants who preferred the less autonomous agent. This preference aligns with research suggesting that people do not perceive supervising a fully autonomous artificial moral agent as collaboration [214] and prefer collaboration over supervision [11]. These findings are promising for meaningful human control as they suggest that people want to take responsibility for morally sensitive decisions rather than rely on artificial moral agents [30, 154].

Furthermore, our results indicate higher meaningful human control over the less autonomous artificial moral agent (confirming H4a). This suggests that increased human involvement during moral decision-making in human-agent teams leads to higher meaningful human control over the artificial moral agent, which aligns with prior research [214]. Overall, participants achieve mean meaningful human control scores of 46.90±11.97% over the more autonomous agent and 50.58±10.46% over the less autonomous agent. Given that the cascade approach emphasizes the weakest aspects, these scores suggest moderate meaningful human control with room for improvement [25, 26, 48]. Exertion of operational control, which considered factors such as maintaining an overview and experienced time pressure, is a main area for improvement with the less autonomous agent [214]. Given the increased participant involvement when collaborating with this agent, it is understandable that the higher cognitive demands negatively affect these factors. We believe more training and interactions with the less autonomous artificial moral agent can combat these issues [33, 178, 233]. In contrast, interaction with the more autonomous agent requires improvement. This interaction required many interventions involving self-reallocations when the moral sensitivity was 4.2 or higher. However, participants only intervened with 26.99±29.70% of the allocations within this range, suggesting that they struggled to act upon their assumed responsibility. Therefore, we recommend increasing human involvement during dynamic task allocation.

Agent Explanations

We find no evidence that agent explanations affect capacity or moral trust (not confirming H1b and H2b). This suggests that when artificial moral agents provide a basic level of transparency, additional explanations do not significantly enhance trust. These results align with [232] but also contradict [18, 157] prior research. Perhaps the additional explanations encouraged participants to evaluate the agents more critically due to a better understanding, leading to more appropriate trust [118, 141, 144]. However, our results mainly suggest that agent behavior contributes more to capacity and moral trust than explanations.

Our results also show that people intervene more when artificial moral agents explain potential consequences rather than feature contributions or no additional information (partially confirming H3b). Furthermore, our findings indicate that people are less likely to reallocate decision-making to themselves when not provided with additional information. Perhaps they lack sufficient understanding to intervene confidently, such as the quoted participant in Section 6.4.3. However, it is also possible that the lack of self-reallocations results from overtrusting the agents [118, 144]. In contrast, people are more likely to reallocate decision-making to themselves when they receive potential consequences, which is even amplified by moral sensitivity. This increased likelihood suggests that explaining potential consequences facilitates better trust calibration [118, 144]. Perhaps this explanation reminds people of their forward-looking responsibility to act proactively and responsibly to ensure future outcomes are positive [125, 207]. The frequent mention of human control and responsibility as reasons for self-reallocations also supports this. The potential consequences probably also better fulfill the epistemic condition of moral responsibility, especially awareness of probable consequences and moral significance of actions [16, 152, 172]. Moreover, this explanation likely better satisfies the foreseeability and control conditions of moral and legal culpability [178]. Overall, these results indicate that explainable AI can indeed raise human moral awareness to take responsibility, but only if a proper explanation is used [30].

Finally, we find no evidence that agent explanations affect meaningful human control (not confirming H4b). However, our results indicate differences in the factors determining meaningful human control. Agent interaction and the exertion of operational control determine meaningful human control less and more frequently when the agents explain potential consequences. Given the increased self-reallocations when receiving potential consequences, it is understandable that the higher cognitive demand leads to the exertion of operational control more frequently determining meaningful human control. We believe more training and interactions can further improve this [33, 178, 233]. Since the self-reallocations in response to potential consequences increase with moral sensitivity, it also follows that agent interaction determines meaningful human control less frequently. Yet, we are not convinced that more training with the agents explaining feature contributions or no additional information can improve interaction with these agents. These explanations simply seem unable to sufficiently (1) remind people of their forward-looking responsibility and (2) fulfill the conditions of moral responsibility and culpability [16, 125, 152, 172, 178, 207]. We believe the potential consequences can do so and recommend that agents provide these explanations during dynamic task allocation.

6.5.2 Limitations and Future Work

We acknowledge a few limitations of our work. The first one is the implementation of our artificial moral agents. We used a hybrid approach that incorporated ethical principles and predicted moral sensitivity. This approach simplified real firefighting scenarios but enabled us to implement complex yet interpretable artificial moral agents. Furthermore, these agents' moral sensitivity models were domain-specific. However, the underlying methodology - eliciting moral sensitivity ratings through structured scenarios and statistically modeling key predictors - is adaptable to other domains. Focusing on statistically significant predictors ensured that our models reflected meaningful moral sensitivity factors rather than an arbitrarily chosen set. Future work can explore alternative modeling techniques or apply our methodology to additional domains to enhance generalizability.

Another limitation is the selection of the "appropriate" allocation threshold of 4.2. Although we determined this threshold using human comfort data, it remains subjective as no absolute ground truth exists for when morally sensitive decisions should be allocated to humans. However, our approach ensured that the allocation thresholds were empirically grounded in human judgements rather than arbitrarily set. Moreover, its alignment of task allocation with human comfort is crucial for developing trustworthy artificial moral agents. Interestingly, self-reallocations in response to potential consequences increased notably from 4.2 onwards. Adjusting this "appropriate" threshold would likely not affect our results much, as it was merely used to calculate the correct behavior rate. In contrast, we believe that increasing the difference between the agents' thresholds could amplify the effects of agent autonomy. However, there may be a cut-off point for the less autonomous agent, as we suspect that approaching complete human decision-making would not be preferred either. Overall, our testbed facilitates future empirical research on dynamic task allocation, while our approach and thresholds offer valuable benchmarks.

The generalizability of our findings is also worth discussing. While our controlled experiments focused on a firefighting use case with participants from our university and personal contacts, such human-grounded evaluations are essential for providing results that can be validated in real-world settings [56]. This is crucial given that fewer than 1% of explainable AI papers validate explainability with humans [195]. We explicitly designed our study to capture key challenges in human-agent collaboration under high stakes and time pressure, factors that are generalizable beyond firefighting and enhance the ecological validity of our findings. Future research should extend these findings through application-grounded evaluations and across different domains. In addition, although expanding to a larger, more demographically diverse participant sample would further enhance external validity, this was not feasible given the face-to-face nature of our experiments. However, this ensured sustained and deep participant engagement, something often lacking in crowd-sourced studies. This trade-off prioritized internal validity and provides a solid foundation for applying our findings to other settings and real-world applications. Overall, we believe our sample size, participant diversity regarding relevant moral psychology and human-agent teaming traits, and rigorous experimental design ensure the robustness and broader relevance of our results.

We identify several suggestions for future work on the influence of different agent explanations and collaboration configurations. For example, supplementing the current local explanations with global explanations of agent behavior during decision-making [56, 79, 126]. Another option would be to explore contrastive explanations that illustrate what would have resulted in alternative allocations [145, 211]. Furthermore, comparing our approach to a collaboration where humans determine all allocations is important. Prior research has shown that people prefer agent-determined allocations over shared or self-determined ones [77], but this preference might shift when moral decisions are involved. Finally, it would be interesting to place the human at the operative level and the artificial moral agent at the oversight level. This could provide a stronger coupling between moral actions and responsible humans [38], although a responsibility gap could emerge if the artificial moral agent intervenes and violates ethical guidelines.

6.5.3 Conclusion

This chapter answers the fifth research sub-question of the thesis: *How do agents' autonomy and their transparency and explanations influence responsible human-agent teaming, individually and interactively*? We explored the influence of agent autonomy and explanations during the dynamic allocation of moral decision-making in human-agent teams. We conducted user studies in simulated firefighting environments where participants collaborated with a more and less autonomous artificial moral agent. These agents provided no additional information, feature contributions, or potential consequences during the allocation of moral decision-making. Our user studies show a higher moral trust in, agreement with, meaningful human control over, and preference for the less autonomous agent. Moreover, we show that people disagree and reallocate decision-making to themselves more when artificial moral agents explain potential consequences. This difference amplifies with moral sensitivity when people collaborate with the more autonomous agent. These findings demonstrate that people (1) prefer more involvement over higher agent autonomy and (2) take on greater moral responsibility when artificial moral agents explain potential consequences. Overall, our study provides crucial insights for responsibly implementing dynamic task allocation and enhancing human-agent teamwork in morally sensitive situations, such as raising human involvement and explaining potential consequences.

7

Empowering Human-Robot Interaction for Firefighting: Mixed-Methods Comparison of Teleoperation and Collaboration

Robots can play a crucial role in firefighting by executing tasks that are too dangerous for humans. However, they are still often directly controlled through teleoperation, which is challenging in low-visibility conditions. To address this, we developed TEAMS (Transparent and Explainable Autonomy for Mapping and Searching), a system that moves from teleoperation to collaboration. TEAMS is grounded in expert knowledge and combines fuzzy logic control, explainable AI, and meaningful human control to facilitate effective and responsible human-robot collaboration. We compared TEAMS with teleoperation in a mixed-methods study that combined in-person (n = 4) and video-based (n = 26) evaluations by non-experts, along with a domain expert review of the videos. Results show a clear contrast between the qualitative evaluations of in-person interactions and the quantitative evaluations of video-recorded interactions. Participants highly appreciate TEAMS during in-person interactions, while they score its video worse than teleoperation on workload, situation awareness, and usability. We discuss the implications of these results for in-person and video-based evaluations of human-robot teamwork, as well as for using TEAMS during firefighting.

🗎 **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Empowering Human-Robot Interaction for Firefighting: Mixed-Methods Comparison of Teleoperation and Collaboration. Submitted to: *The ACM/IEEE International Conference on Human-Robot Interaction.* 2025.

7.1 Introduction

Humans and robots are increasingly collaborating in high-stakes domains such as firefighting. These human-robot teams combine the unique capabilities of both parties to achieve outcomes that would be otherwise impossible [3]. However, the robots in these teams are still frequently under direct human control, for example, when teleoperating firefighting robots in conditions too dangerous for humans. This is highly challenging in low-visibility conditions, adding another burden to the already substantial workload of operators. To reduce this workload, there is a desire to make firefighting robots more autonomous, especially during navigation [52].

When designing these autonomous robots, it is crucial that the human-robot collaboration is effective and ensures meaningful human control [99, 179]. In this chapter, we present TEAMS (Transparent and Explainable Autonomy for Mapping and Searching) for such human-robot collaboration. Furthermore, we compare TEAMS with teleoperation in a mixed-methods study. This study combines (1) a qualitative, in-person evaluation by non-experts, (2) a quantitative, video-based evaluation by non-experts, and (3) a qualitative evaluation of the same videos by a domain expert. Consequently, we not only compare teleoperation and collaboration, but also evaluations of in-person and video-recorded human-robot interactions. Moreover, our mixed-methods study leverages the complementary strengths of quantitative and qualitative approaches, pairing generalizable effects with nuanced experiential insights [198].

7.2 Background

7.2.1 Human-Robot Teams for Disaster Response

Robots are already being used to explore and map environments during disaster response, such as urban search-and-rescue and firefighting. These robots are still frequently under direct human control, but there is a strong motivation to make them more autonomous to reduce human workload [52]. Several papers have compared teleoperation with more autonomous robots during human-robot collaboration, showing that increased autonomy enhances team performance and efficiency. However, these studies are frequently conducted in simulated environments [78, 93, 240–242]. Moreover, their collaboration systems mainly allow humans to determine navigation destinations [78, 111, 240] or fully autonomous robot navigation with the ability for operators to take over control or assist [71, 93, 241]. Instead, TEAMS enables the robot to propose three navigation destinations to the human, who then decides.

A related, in-field study evaluated a semi-autonomous robot with the option for operators to take over control [111]. The semi-autonomous mode used short spoken operator commands to control the robot. Results showed that operators only briefly used this mode but quickly took over complete control via teleoperation. The authors believe the robot's lack of behavioral transparency contributed to this by negatively affecting the operators' trust in the robot's capabilities. Considering that the level of robot autonomy was still fairly limited in their study, we believe robot transparency and explanations will be even more important in TEAMS.

7.2.2 Robot Transparency and Explanations

Explainable AI (XAI) is the field that develops methods for generating and communicating such robot transparency and explanations [9, 149]. Without it, people might attribute robot behavior by assigning inappropriate mental states that explain the behavior, such as incorrect beliefs, goals, and intentions [132, 133, 145]. Common explanation types include feature attributions and contrastive explanations [214]. Feature attributions clarify which relevant features influence robot behavior, making them useful for enhancing predictability and identifying biases [1]. Contrastive explanations clarify why robots make certain decisions over others and help improve predicability and understand reasoning [145]. Studies often show that increasingly transparent and explainable agents/robots enhance situation awareness, trust, and performance, but also negatively impact workload [34, 142, 157, 187]. A significant goal within XAI is designing robots that can adapt their transparency and explanations based on users [9, 157]. For example, by decreasing information load in response to high human workload. Such personalized explanations can be generated and/or communicated based on user models, but also by allowing users to decide for themselves.

During human-robot teamwork, transparency and explanations are crucial for humans to exercise control properly [214, 234]. Ultimately, humans should be aware of their responsibility for outcomes resulting from robot behavior [16, 125]. Meaningful human control requires humans to be both aware and able to act upon this responsibility [30, 179]. This is particularly important for human-robot teamwork in morally sensitive domains such as firefighting, where people's welfare, rights, and values may be directly or indirectly affected [57, 214]. How the collaboration is designed and shaped can partially achieve this [30, 219], for example, by keeping humans-in-the-loop [33, 46, 144] and making them responsible for object recognition and determining navigation destinations.

7.2.3 Evaluating Human-Robot Interaction

Human-robot interaction can be evaluated in various ways. Videos are increasingly used (e.g., in [116, 203]), as they provide flexibility and can reduce the resources required to conduct studies, though at the cost of lower fidelity [119]. Several studies have compared people's perceptions of in-person and video-recorded human-robot interactions [124, 135, 204, 249]. Overall, these studies have yielded mixed results, showing both agreement between the two interaction modalities [135, 249] and more positive perceptions of robots during in-person interactions [124, 204]. However, these comparative and video evaluation studies focus primarily on more social human-robot interactions rather than functional human-robot teamwork.

7.2.4 Hypotheses

For our quantitative, video-based evaluation, we formulated the following hypotheses based on the need for more autonomous robots that reduce human workload [52] and enhance collaboration [93] (pre-registered at the Open Science Framework [228]):

Compared to teleoperation, collaboration will lead to lower perceived workload (H1), higher trust in robot performance (H2), higher perceived situation awareness (H3), higher system usability (H4), and a stronger overall preference (H5).

As pre-registered, we expect similar themes to emerge from the qualitative, in-person evaluations by non-experts and the qualitative, video-based evaluation by the domain expert.

7.3 Human-Robot Interaction Testbed

7.3.1 Hardware and Software

We used a TurtleBot3 Waffle Pi and a Dell Latitude 7410 laptop running Ubuntu 20.04 LTS for our human-robot interaction studies. Robot behavior and communication were implemented with ROS 1 Noetic and MATLAB R2024b. RViz provided the human-robot interaction interface, combining sensor visualizations (LiDAR LDS-02 and Raspberry Pi Camera Module v2.1) with custom panels for operator inputs via ROS topics, enabling bidirectional communication. For teleoperation, we used the RC-100B Bluetooth joystick.

7.3.2 Task and Environment

The in-person studies involved two simulated firefighting tasks. The goal of the tasks was to map the environments and find victims. We created two similar but distinct environments, each with four rooms, six victims, and 12 obstacles. Moving boxes were used as walls, openings between these boxes as doors, waste bins with sad faces as victims, and books as obstacles (Figure 7.1). We blurred the robot's camera to simulate low-visibility conditions commonly encountered during human-robot interaction for firefighting.

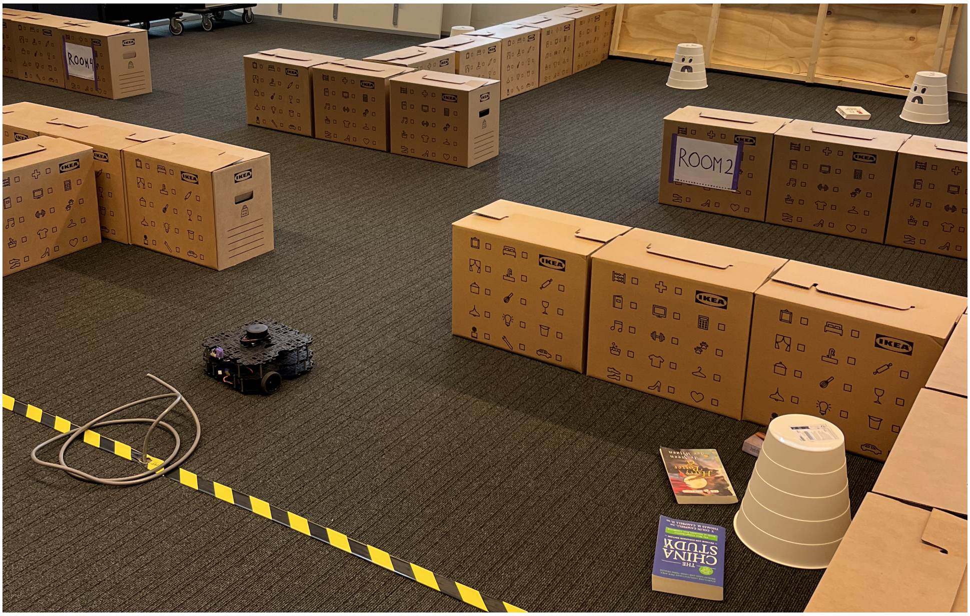


Figure 7.1: Experimental setup of our task and environment.

7.4 TEAMS

7.4.1 Design

We applied the socio-cognitive engineering method [150] to develop TEAMS for firefighting. TEAMS combines fuzzy logic control, XAI, and meaningful human control to foster effective and responsible human-robot collaboration. We first elicited foundational knowledge regarding operational demands, technology, and human factors through discussions with firefighters. These discussions provided us with stakeholder preferences and requirements for the robot and interaction, which we integrated into TEAMS. For example, the firefighters told us that more autonomous navigation is their biggest challenge and desire. Therefore, other robot functionalities, such as computer vision for object recognition, have lower priority and are tasks they can execute themselves. They also emphasized the importance of identifying doors and highlighted the need to simultaneously map and navigate the environment to avoid delaying the mission. Regarding human-robot interaction, the firefighters preferred active and explicit human control over fully autonomous robot behavior. More specifically, they preferred a decision support system supplemented by visual robot explanations.

7.4.2 Robot Navigation

Based on this foundation, we specified and implemented TEAMS (Figure 7.2). A Mamdani-type fuzzy logic controller drove the robot's navigation, chosen to encode expert knowledge, support real-time decisions with low computational cost, and enable autonomous navigation [58, 114, 155, 165, 191]. This controller received inputs from sensing the environment

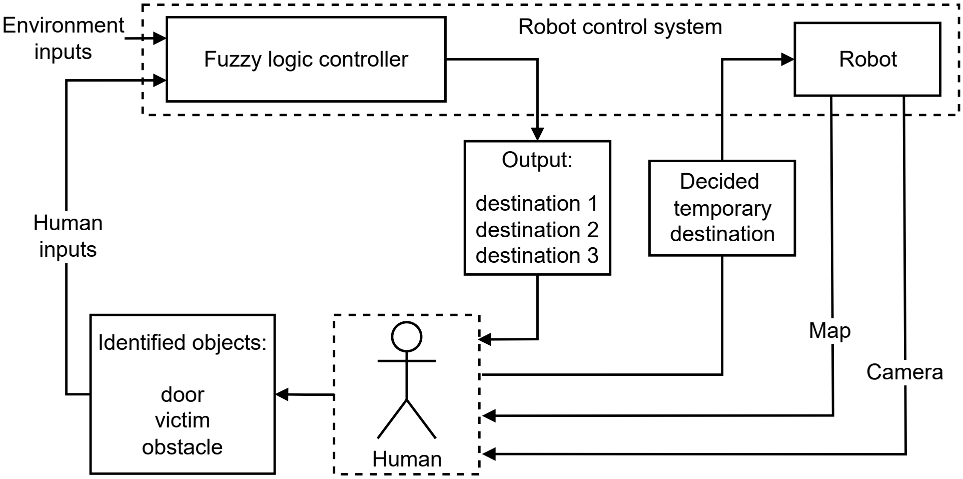


Figure 7.2: TEAMS: Our human-robot collaboration system.

and from communicating with the human. The environmental inputs included the time to reach a destination and the path clutter towards that destination. The time to reach a destination ranged from 2 to 14 seconds, derived from the robot's perception radius and maximum speed, and grid cell dimensions. The path clutter ranged from 0 to 1, based on the proportion of occupied cells in the shortest path to a destination.

The human inputs included the location of victims, doors, and obstacles provided via the collaboration interface described in Section 7.4.3 and shown in Figure 7.3. Using this data, the controller derived the distance to victims, doors, and obstacles between all grid cells in the robot's perception field, ranging from 0.5 to 7 meters. We mapped all inputs to three linguistic categories (low, medium, high).

Based on the environmental and human inputs, the controller determined the attraction value for each grid cell in the robot's perception field. These attraction values reflected the relative importance of reaching that destination and ranged from 0 to 1, mapped to five linguistic categories (very low, low, medium, high, very high). The controller computed these values using a set of 22 fuzzy rules based on expert knowledge elicited during discussions with the firefighters, to approximate human-like reasoning. For example:

1. If the time to reach a destination is high, the path clutter is high, the distance to a door is close, and the distance to an obstacle is close, then the attraction is very low.

2. If the time to reach a destination is low, the distance to a door is close, and the distance to a victim is close, then the attraction is very high.

The final output of the controller consisted of the three destinations with the highest attraction. The human then picked their preferred destination from these options, followed by the robot navigating towards it. For planning trajectories, we used a time-elastic-band approach [171] instead of a dynamic window approach [72] because it ensured smoother robot motion. The robot simultaneously mapped and navigated the environment [17].

When planning trajectories, the robot avoided detected obstacles in its map. After reaching a destination, the process started again. We kept track of the grid cells traversed during the previous trajectory and removed those cells as possible destinations during the next iteration.

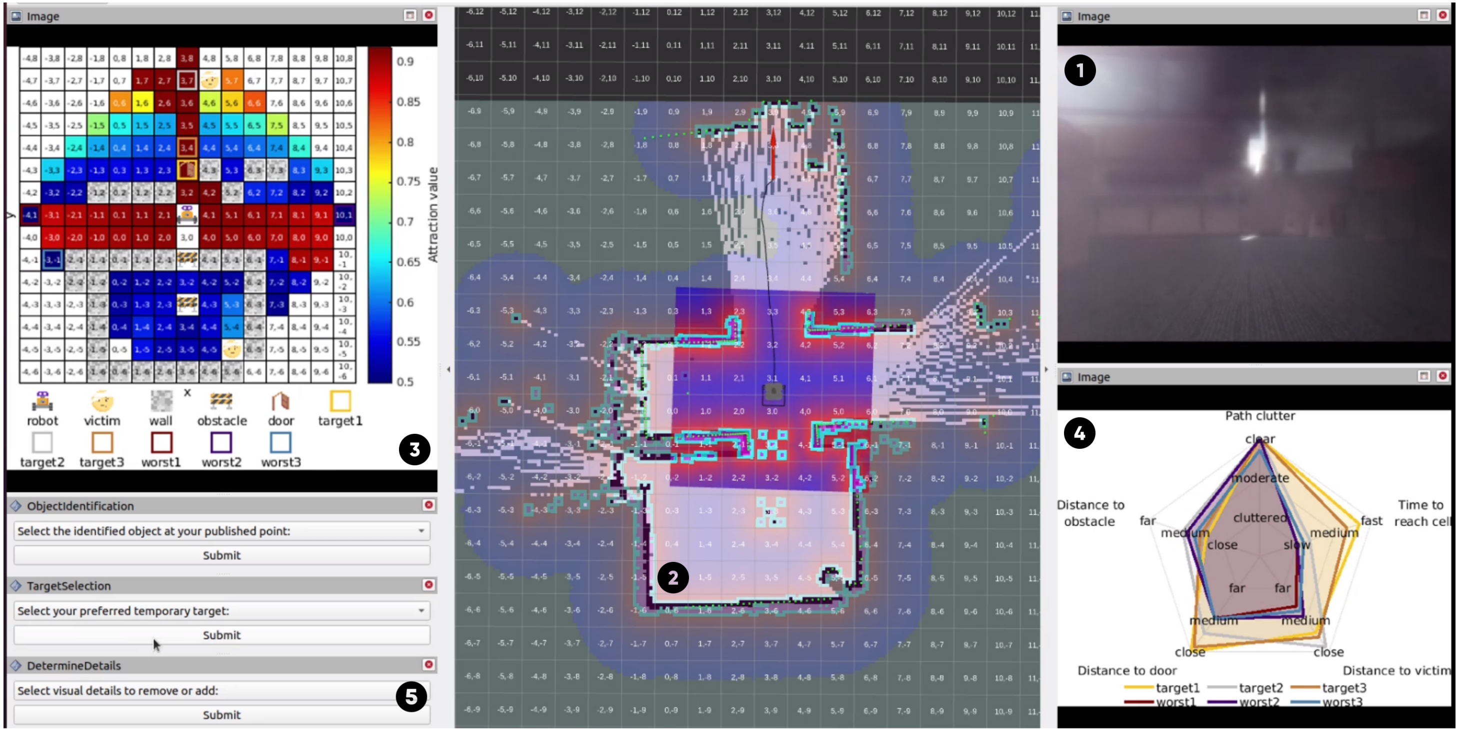
7.4.3 User Interface

We facilitated human-robot collaboration with the user interface shown in Figure 7.3. This interface consisted of (1) live camera footage from the robot, (2) a live 2D occupancy grid map generated from LiDAR, (3) an enriched recreation of this grid map with attraction values, identified objects, and highlighted best and worst destinations, (4) a radar chart of the fuzzy input features for the highlighted best and worst destinations, and (5) operator input panels for object identification, destination selection, and information customization.

To provide the inputs to the controller, users first clicked a grid coordinate on the live 2D occupancy grid map. Next, they indicated whether this grid coordinate contained a victim, door, or obstacle, using the object identification panel. These objects were then added to the enriched recreation of the grid map using intuitive icons to enhance users' situation awareness during the task. If the grid coordinate contained an obstacle, it was integrated into the robot's navigation map and displayed on the live occupancy grid map.

After the controller determined the attraction values for all grid cells, they were overlaid on the enriched occupancy grid map to enhance users' situation awareness during the task. Furthermore, this map highlighted the three destinations with the highest attractions. The radar chart was also updated to highlight the fuzzy input features of these three destinations, visually explaining the controller's reasoning. To further clarify its counterfactual reasoning, the interface also highlighted the three lowest-ranked destinations on the map and radar chart. Users selected their preferred destination using the corresponding selection panel.

Finally, we implemented personalized explainable AI by allowing users to customize the information displayed in the enriched occupancy grid map and radar chart. They could remove the three lowest-ranked destinations from both images, the attraction values from the enriched grid map, and the labels from the radar chart. Users could also always add this information back. This ensured direct human control over the transparency and explanation content, allowing them to tailor it to their informational needs.



Select the identified object at your published point:

Select your preferred temporary target:

Select visual details to remove or add:

Figure 7.3: The graphical user interface for TEAMS.

7.5 Robot Teleoperation

Users directly controlled the teleoperated robot using a joystick, allowing them to move the robot forward, turn left, or turn right. Each press triggered a short, real-time, fixed-duration motion, so control was step-wise rather than continuous. Additional button presses during an ongoing step were registered and executed sequentially. The system operated with the platform's default teleoperation configuration upon connecting, without custom tuning.The human-robot interaction interface during teleoperation did not contain the radar chart, destination selection panel, and information customization panel. We still asked participants to add detected victims to the enriched occupancy grid map using the object identification panel. This way, we could compare the number of victims found for both conditions.

7.6 Method

7.6.1 Experiment Design

We conducted a mixed-methods, one-factor, within-subjects experiment comparing two robot autonomy conditions: teleoperation and collaboration/TEAMS. This study consisted of (1) a qualitative, in-person evaluation by non-experts, (2) a quantitative, video-based evaluation by non-experts, and (3) a qualitative evaluation of the same videos by a domain expert. We chose these mixed methods for several reasons. First, hardware issues did not allow for large-scale, in-person experiments. Furthermore, by combining methods, we attempted to address the disadvantages of the individual components. Additionally, this combination provided both generalizable effects and nuanced experiential insights across levels of fidelity. Finally, we were interested in comparing the evaluations of in-person and video-recorded human-robot interactions.

7.6.2 Participants

Evaluating In-Person Interactions

We recruited four participants from our academic and personal networks (two female and two male). Three participants were 25-34 years old and one participant was 18-24 years old. Two participants obtained a Bachelor's degree and two a Master's degree.

Evaluating Video-Recorded Interactions

We recruited 26 participants from our academic and personal networks for the video-based evaluation by non-experts (13 female and 13 male). Three participants were 18-24 years old, ten were 25-34 years old, four were 35-44 years old, eight were 45-54 years old, and one was 55-64 years old. One participant was a high school graduate, four obtained some college credit but no degree, four an Associate degree, eight a Bachelor's degree, seven a Master's degree, and one a PhD degree. We recruited one domain expert from our professional network for the qualitative evaluation of the video-recorded interactions. This male participant was 55-64 years old and a high school graduate. All participants signed an informed consent form approved by our institute's ethics committee (ID 6011).

7.6.3 Video-Recorded Interactions

We used screen recordings of the user interface for the video-based evaluations, interacting with both systems while mapping half of the environment. We increased the speed of these videos to 1.6; they took 3 minutes and 8 seconds for teleoperation and 5 minutes and 16 seconds for TEAMS. Finally, we recorded our hand that controlled the joystick and added this to the teleoperation video. These videos can be found in [228].

7.6.4 Measures

Video-Based Evaluation by Non-Experts

We quantitatively and subjectively measured participants' perceptions of operator workload and situation awareness, trust in the robots' performance, system usability, and interaction preferences. Additionally, we qualitatively collected participants' reasons for their interaction preferences through an open-ended survey question. We measured perceived operator workload using the raw NASA-TLX scale [87]. This scale distinguished between six workload components, each scored on a scale of 0 to 100 with increments of 5. We computed the mean of these six scores as the final workload score. Perceived operator situation awareness was measured using the Situation Awareness Rating Technique (SART) [199]. This 10-item scale was scored on 7-point bipolar scales (*low* vs. *high*) and distinguished between demand, supply, and understanding components. We computed the final situation awareness score using the associated formula.

Participants' trust in the robots' performance was measured using the Multi-Dimensional Measure of Trust (MDMT) [134]. We only used the performance trust subscale, measured by eight one-word items scored from 0 (*not at all*) to 7 (*very*). The scale also provided the answer option *does not fit*, which turned selected items into missing values. We computed the mean of the eight items as the final performance trust score.

System usability was measured using the System Usability Scale (SUS) [23]. This 10-item survey was scored on 5-point Likert scales ranging from *strongly disagree* to *strongly agree*. We computed the final system usability score using the associated formula. Finally, we measured interaction preference using a 7-point bipolar scale. This scale ranged from −3 (*strong preference for teleoperation*) to +3 (*strong preference for collaboration*). The middle point indicated no preference for any approach.

In-Person Evaluation by Non-Experts

We qualitatively explored participants' workload and situation awareness during the tasks, trust in the robots, control over the robots, system usability, and interaction preferences through post-task, semi-structured interviews. Participants first completed the same surveys described in Section 7.6.4, along with the Experienced Control survey [214]. These survey scores were then used as prompts to ask open-ended follow-up interview questions about participants' reasoning and experiences. Finally, for each interaction, we logged participants' number of crashes, explored rooms, and found victims.

Video-Based Evaluation by Domain Expert

We qualitatively explored the domain expert's comparison of the two systems through an interview. The domain expert first watched both videos. We then asked him open-ended questions regarding the challenges, advantages, disadvantages, and applicability of both systems.

7.6.5 Procedure

Study Procedure

All participants first filled in the informed consent form and demographic questions. The non-experts evaluating the video-recorded interactions then watched the video of the first interaction, followed by the surveys on workload, trust, situation awareness, and system usability. They then watched the video of the second interaction (the order was counterbalanced), followed by the same surveys. Lastly, we asked participants about their interaction preference. The domain expert evaluating the video-recorded interactions first watched both videos, starting with the teleoperation video. We then conducted the interview. Both studies lasted approximately 30 minutes.

After completing the consent form and demographic questions, the non-experts evaluating the in-person interactions watched a 10-minute video explaining TEAMS. During the tasks, participants faced the opposite direction of the environment, so they had to navigate using the interface only. After the tutorial, participants interacted with one of the two approaches for a maximum of 15 minutes (the order was counterbalanced), followed by the surveys on workload, trust, control, situation awareness, and system usability. We conducted the semi-structured interview immediately after completing the surveys. Next, participants interacted for 15 minutes with the second approach, again followed by the surveys and semi-structured interview. After this interaction, we also asked participants about their interaction preference. We automatically transcribed the audio recordings of all interviews. The whole study lasted approximately two hours.

Qualitative Data Analysis

We analyzed all the qualitative data using thematic analysis [20]. This data included (1) the transcripts of the semi-structured interviews after in-person interactions, (2) the transcript of the domain expert interview after watching the video-recorded interactions, and (3) the non-experts' answers to the open-ended question on interaction preference after watching the video-recorded interactions. The first author and a double coder (non-expert) first reviewed the data and developed codebooks for each dataset. For the in-person interaction interviews, the codebook consisted of 64 codes grouped by the six variables discussed during the interviews. Both coders then independently coded 43 extracts from half of the transcripts, yielding substantial inter-rater reliability (Cohen's $\kappa = 0.71$). For the domain expert interview, the codebook included 21 codes, and the eight extracts were double-coded with almost perfect agreement (Cohen's $\kappa = 0.91$). For the non-experts' open-ended answers, the codebook comprised 18 codes. Both coders independently coded all 26 answers with almost perfect agreement (Cohen's $\kappa = 0.82$). For each dataset, the coders discussed disagreements until they reached a consensus.

7.7 Results

7.7.1 Evaluation of In-Person Interactions

We first investigated the number of crashes, explored rooms, and found victims. With four participants, we report descriptive patterns only. Participants crashed more during teleoperation ($Md = 3$ [1–9]) than collaboration ($Md = 0.5$ [0–3]). However, they explored more rooms during teleoperation ($Md = 87.5\%$ [50–100%]) than collaboration ($Md = 37.5\%$

[25–100%]). Consequently, participants also found more victims during teleoperation (*Md* = 75% [50–100%]) than collaboration (*Md* = 33% [17–100%]).

Next, we clustered the codes resulting from the thematic analysis into themes on the main topic of comparing teleoperation and collaboration (Figure 7.4). Our analysis revealed that the following four themes are underlying the main topic: (1) divergent learning curves: quick-start teleoperation vs. practice-built collaboration; (2) the teleoperation burden: manual control costs vs. information-rich collaboration; (3) hardware and software constraints undermine human-robot interaction; and (4) desire for streamlined collaboration interface. Below, we discuss each theme in detail.

Divergent Learning Curves

All participants mentioned that their interaction with both systems improved over time by becoming more intuitive and familiar. Most participants (3/4) indicated that TEAMS required more training and practice to overcome its complexity, but that it ultimately became easy and simple to learn and use. For example, one participant explained: "*A joystick is something you already know, at least I do from gaming, so I could get started with it more quickly. The other system is slightly more complex, so naturally it takes a bit longer before you get used to it.*"

The Teleoperation Burden

The first sub-theme that we identified within the second theme was *manual costs in teleoperation.* Most participants reported that teleoperating the robot was more frustrating, demanding, and challenging. All participants agreed that it was difficult to control the teleoperated robot properly with the joystick. For example, one mentioned: "*I found driving myself quite frustrating because I could not always get the robot where I wanted, whereas with TEAMS, I did not have to control anything.*"

The second sub-theme was *guided, information-rich collaborative control.* All participants agreed that they experienced more and easier control during collaboration. Most participants especially appreciated that TEAMS proposed three destinations and understood their commands by autonomously navigating to the destinations. For example, one participant explained: "*It was quite easy because you just have to press a button and then the robot will execute that.*" While another one mentioned: "*With TEAMS, it felt like the robot and I were helping each other. What I also liked about TEAMS was that it already suggested where it should drive.*" Half of the participants further mentioned that the interface for TEAMS contained more information and thus required focus to consider all this data.

The third sub-theme was *trustworthy, intelligent autonomy.* Most participants indicated that they found the collaborative robot more intelligent. They especially appreciated the suggested destinations, autonomous navigation, and obstacle avoidance. For example, one participant mentioned: "*It felt like the collaborative robot thought a bit more for itself, moving to places on its own. I could also see that it kept doing that quite reliably, it really went to the cell it said. That made my trust higher.*" In contrast, half of the participants found the teleoperated robot unpredictable.

The fourth sub-theme was *bounded collaborative control.* Two participants mentioned how the control over the collaborative robot was limited by its selection of just three suggested destinations. For example: "*You cannot choose the targets. Where the robot*

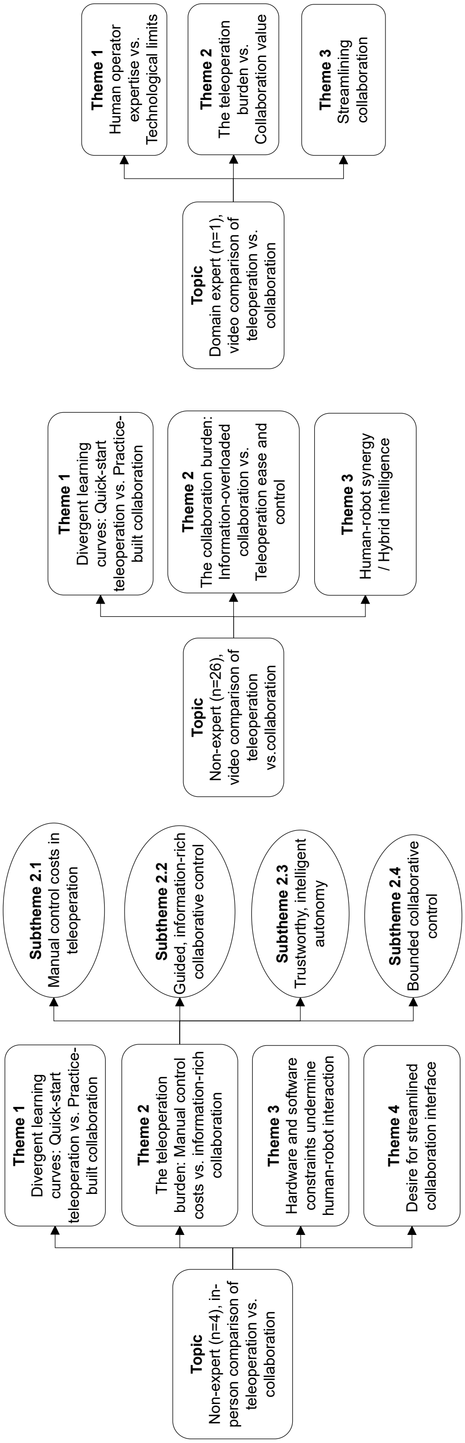

Topic
Non-expert (n=4), in-person comparison of teleoperation vs. collaboration
Theme 1
Divergent learning curves: Quick-start teleoperation vs. Practice-built collaboration
Theme 2
The teleoperation burden: Manual control costs vs. information-rich collaboration
Theme 3
Hardware and software constraints undermine human-robot interaction
Theme 4
Desire for streamlined collaboration interface
Subtheme 2.1
Manual control costs in teleoperation
Subtheme 2.2
Guided, information-rich collaborative control
Subtheme 2.3
Trustworthy, intelligent autonomy
Subtheme 2.4
Bounded collaborative control
Topic
Non-expert (n=26), video comparison of teleoperation vs.collaboration
Theme 1
Divergent learning curves: Quick-start teleoperation vs. Practice-built collaboration
Theme 2
The collaboration burden: Information-overloaded collaboration vs. Teleoperation ease and control
Theme 3
Human-robot synergy / Hybrid intelligence
Topic
Domain expert (n=1), video comparison of teleoperation vs. collaboration
Theme 1
Human operator expertise vs. Technological limits
Theme 2
The teleoperation burden vs. Collaboration value
Theme 3
Streamlining collaboration

Figure 7.4: Thematic maps on the comparison between the teleoperation and TEAMS for all three studies.

navigates, it decides itself. You can only take small steps. If you want to proceed a lot, you cannot do that."

Hardware and Software Constraints Undermine Interaction

Two participants indicated that the camera from TEAMS could be improved or expanded to ensure a complete view of the environment. For example, one of them mentioned: "*A real advantage of teleoperation is that you can actually look everywhere. The collaborative robot thinks more for itself, which is less convenient for the visuals. If it were possible to have a front and back view, that would really make it easier.*" Furthermore, two participants indicated that the teleoperated robot experienced a camera lag. Finally, participants individually mentioned that the hardware could be improved and that updating information in the user interface could be faster.

Desire for Streamlined Collaboration Interface

Most participants highlighted aspects of the interface for TEAMS that could be further improved. Two of them indicated that they did not really use the radar chart with the fuzzy input features for the best and worst destinations. One of these two further proposed to use the information customization panel just once, before starting the interaction. The other suggested to overlay the live 2D occupancy grid map and its enriched recreation into one map.

7.7.2 Non-Expert Evaluation of Videos

We analyzed the quantitative data with parametric paired samples T-tests if the underlying assumptions were met; otherwise, we used non-parametric paired samples Wilcoxon tests (Figure 7.5). Results showed that participants perceived operator workload as significantly higher during collaboration ($M = 49.58$, $SD = 13.45$) than teleoperation ($M = 38.56$, $SD = 17.01$) ($t(25) = 3.75$, $p < 0.001$). Furthermore, they perceived operator situation awareness as significantly higher during teleoperation ($M = 22.42$, $SD = 7.11$) than collaboration ($M = 19.31$, $SD = 6.03$) ($W = 48.5$, $p < 0.01$). Participants also scored system usability significantly higher for teleoperation ($M = 64.62$, $SD = 22.10$) than collaboration ($M = 47.02$, $SD = 18.43$) ($t(25) = -4.53$, $p < 0.001$). In contrast, the results did not show a significant difference in participants' trust in the teleoperated ($M = 5.82$, $SD = 0.74$) and collaborative ($M = 5.54$, $SD = 1.00$) robot ($W = 122$, $p = 0.43$). Finally, a one-sample Wilcoxon signed-rank test against 0 (no preference) showed that the participants significantly preferred teleoperation ($M = 1.31$, $SD = 2.13$) ($W = 169$, $p < 0.001$). Overall, these results did not support any of our hypotheses: H1, H3, H4, and H5 showed significant effects in the opposite direction (favoring teleoperation), while H2 showed no significant difference.

Next, we clustered the qualitative codes into themes related to the comparison of teleoperation and collaboration (Figure 7.4). Our analysis revealed that the following three themes are underlying the main topic: (1) divergent learning curves: quick-start teleoperation vs. practice-built collaboration; (2) the collaboration burden: information-overloaded collaboration vs. teleoperation ease and control; and (3) human-robot synergy. Below, we discuss these themes in more detail.

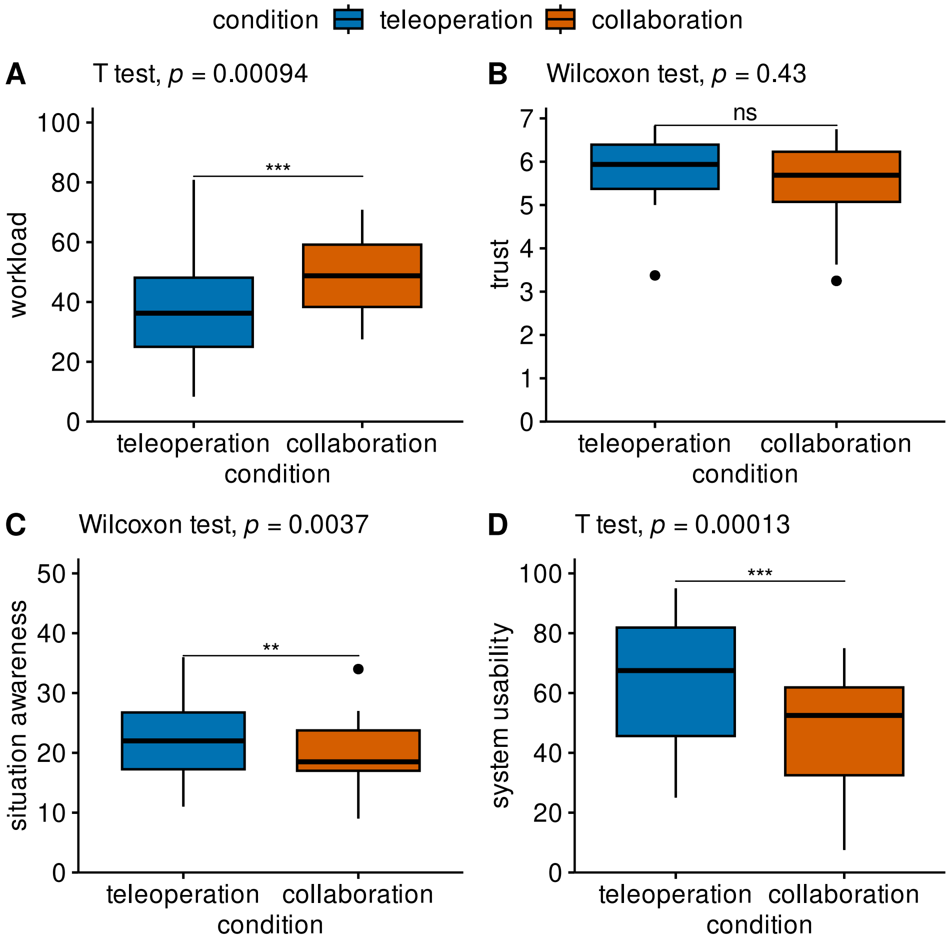


Figure 7.5: Boxplots of workload, trust, situation awareness, and system usability for both robot autonomy conditions.

Divergent Learning Curves

We identified the same first theme as for the in-person interaction evaluations. Again, some participants highlighted teleoperation familiarity and collaboration complexity. For example: "*The operator made teleoperation seem a lot easier and more straightforward to both use and understand. It also seemed there was less to learn, but I do consider that by spending time with TEAMS, it might be easier to use and prove to be better.*"

The Collaboration Burden

This theme contrasted with the theme identified for the in-person interactions. Many participants (16) indicated that they perceived teleoperation as easier and more intuitive to use, while also granting more and easier control. Moreover, several participants (6) reported that they believed teleoperation would facilitate better coordination and awareness because TEAMS presented more information and was more interactive.

Human-Robot Synergy

Several (4) participants did appreciate TEAMS, especially the combination of human and robot intelligence. One participant reported: "*Two heads are better than one. TEAMS is likely to benefit from a human-led search with an integrated automatic search pattern augmenting the process.*" Similarly, another participant mentioned: "*TEAMS allows leveraging the perception technology of the robot, and it increases reliability, which helps in the successful completion of the task. Teleoperation is mostly reliant on human senses, which may or may not be able to perceive the perception of the smoke-filled room correctly.*"

7.7.3 Domain Expert Evaluation of Videos

Finally, we clustered the qualitative codes for the domain expert evaluation of both videos into themes related to the comparison of teleoperation and collaboration (Figure 7.4). Our analysis revealed that the following three themes are underlying the main topic: (1) human operator expertise vs. technological limits; (2) the teleoperation burden vs. collaboration value; and (3) streamlining collaboration. Below, we discuss these themes in more detail.

Human Operator Expertise vs. Technological Limits

The domain expert emphasized the importance of human operator involvement because of their expertise and the limitations of technology. He explained: "*Having the operator identify victims, doors, and obstacles, instead of computer vision, actually works well. I prefer it, because computer vision is challenging in low-visibility conditions and we, of course, have our own skills. In the end, we want operators to make enough decisions themselves, since we bring a lot of expertise.*"

The Teleoperation Burden

This theme aligned with the second theme identified for the evaluations of the in-person interactions. The domain expert emphasized the importance of more autonomous and straightforward robot navigation due to the challenges of camera-based robot teleoperation in low-visibility conditions. He perceived TEAMS and the selection of navigation destinations as an improvement over teleoperation, particularly in terms of operator workload and task efficiency.

Streamlining Collaboration

This theme aligned with the fourth theme identified for the in-person interactions. The domain expert mentioned several ways to streamline TEAMS. For example, using one operator for navigation and another for object recognition to further reduce workload. He also explained: "*A combination of TEAMS with teleoperation would probably work best. I can imagine that the three suggested destinations are not always sufficient. Sometimes, the option to occasionally take full manual control would likely be the best.*" Furthermore, the domain expert discussed adapting the fuzzy rules based on task characteristics: "*Sometimes the task determines which navigation destinations are attractive, and therefore the underlying rules. In that case, we might even leave victims for the moment and first quickly map everything.*" Finally, he emphasized the importance of ensuring TEAMS does not become too complicated, as autonomous navigation is the most crucial aspect.

7.8 Discussion and Conclusion

7.8.1 Discussion

The results for the in-person interactions highlight that participants appreciate TEAMS compared to teleoperation, despite its (initial) complexity and lower area coverage. Participants especially struggled to control the teleoperated robot, as also reflected by the higher number of crashes. Therefore, they particularly valued the proposed destinations from TEAMS, as well as its autonomous navigation and obstacle avoidance based on their inputs. Notably, this enhanced robot autonomy did not improve area coverage, which contrasts with prior work [93, 168, 241, 242]. However, with only four in-person participants, we interpret this result cautiously. Finally, participants suggested several ways to enhance TEAMS. These include adding more cameras to provide a complete view of the environment, removing the radar chart from the user interface, and merging the two occupancy grid maps.

These results mostly correspond to the video-based evaluations by the domain expert. The domain expert also valued TEAMS for selecting destinations and autonomous navigation while avoiding obstacles. He also emphasized the burden of teleoperation when using cameras in low-visibility conditions, a common issue within the field [52, 111, 168]. Since the domain expert had experience with teleoperating firefighting robots, it makes sense that these results align. Furthermore, he proposed several ways to improve TEAMS. These include combining TEAMS with the option to take complete manual control (i.e., variable autonomy [144]), adapting the fuzzy rules based on task characteristics, and simplifying where possible.

Meanwhile, our results reveal a contrast between the in-person and video-based evaluations by non-experts. While the in-person interactions highlighted the value of TEAMS and the burden of teleoperation, participants rated the video of TEAMS worse than teleoperation on workload, situation awareness, and usability. Furthermore, they emphasized the user-friendliness of teleoperation and the information overload of TEAMS. These results contradict prior comparisons of in-person and video-recorded (social) human-robot interaction, which typically find agreement between modalities or more positive perceptions in-person. Similar to other works [124, 204], participants perceived TEAMS more positively in person than via video. In contrast, participants perceived teleoperation more positively via video than in person. On the one hand, this suggests that videos may be suitable for evaluating social human-robot interaction, but are poor proxies for assessing human-robot teamwork. Only the in-person interactions seem to accurately capture the situated dynamics and challenges, such as actually controlling the teleoperated robot. On the other hand, the video-based evaluations by non-experts still provide valuable insights for enhancing TEAMS by reducing workload and increasing usability.

Finally, the results provide several practical implications for TEAMS during firefighting. First, the importance of training. TEAMS requires training to overcome its initial complexity, but with practice, it becomes more intuitive and efficient. Perhaps more training also yields better performance, similar to prior work [93, 168, 242]. Second, iterative, human-centered streamlining. It is crucial to continuously improve TEAMS to optimize operators' workload, situation awareness, usability, and performance [150]. Third, enhancing hardware and software. Hardware improvements should facilitate complete camera coverage, while optimizing software should accelerate interaction, decision-making, and navigation.

7.8.2 Limitations and Future Work

A first limitation is the use of non-experts to evaluate the video-recorded human-robot interactions. Based on our results, we argue that video-recorded interactions are poor proxies for evaluating human-robot teamwork because they do not capture the situated dynamics. However, such videos might still be effective when evaluated by domain experts who have actual experience with the interactions, situations, and software shown in the videos. Therefore, for future work, it would be interesting to recruit sufficient domain experts to evaluate our video-recorded human-robot interactions quantitatively. Perhaps the results would then better align with the in-person evaluations and prior literature [124, 135, 204, 249].

Furthermore, our fuzzy logic controller enables real-time navigation by gathering information and using fuzzy rules, but does not incorporate prediction capabilities in decision-making. Such capabilities are valuable in dynamic environments with moving obstacles, for example, resulting from fire spread [73] or victim motion [169]. These dynamics are typically handled with model predictive control, which forecasts state evolution to anticipate and avoid unsafe motions [12, 161]. However, the trade-off is a higher computational load compared to our lightweight fuzzy logic controller. Recent work also integrates the two control strategies for unknown search and rescue environments [196], suggesting a promising direction for future work.

As another limitation, TEAMS was not (yet) implemented on a robot of the Fire Brigade and evaluated at their (more realistic) test sites. The next improved version of TEAMS should be evaluated at such a test site with domain experts. This would further enhance the generalizability of our work to real-life settings [56, 195].

7.8.3 Conclusion

This chapter answers the sixth research sub-question of the thesis: *How should we design a semi-autonomous, transparent and explainable robot in human-robot teams for firefighting, and how does the increased autonomy impact effective and responsible collaboration compared to teleoperation*? To address this question and common challenges of teleoperating firefighting robots, we developed TEAMS (Transparent and Explainable Autonomy for

Mapping and Searching). TEAMS is grounded in expert knowledge and combines fuzzy logic control, explainable AI, and meaningful human control to facilitate effective and responsible human-robot collaboration. Our mixed-methods study compared TEAMS with teleoperation through in-person trials with non-experts and video-based evaluations by non-experts and a domain expert. Methodologically, this study leverages the complementary strengths of quantitative and qualitative approaches, pairing generalizable effects with nuanced experiential insights. The results yield two main contributions. First, they reveal a clear contrast between qualitative, in-person evaluations and quantitative, video-based evaluations. In-person interactions emphasized the value of TEAMS and burden of teleoperation, whereas non-experts rated the video of TEAMS worse on workload, situation awareness, and usability. Importantly, the in-person findings aligned with the domain expert's evaluation, suggesting that non-expert, video-based evaluations are poor proxies for assessing human-robot teamwork. Second, both the in-person participants and the domain expert appreciate TEAMS for providing navigation suggestions and autonomous navigation, addressing the challenge of camera-based robot teleoperation in low-visibility conditions. Overall, our findings offer several practical implications for deploying TEAMS during firefighting. These include providing training to overcome initial complexity, implementing iterative and human-centered refinements, and enhancing hardware and software.

8

Advancing Human-Machine Teaming: Definitions, Challenges, Future Directions

Humans and intelligent machines increasingly collaborate on complex tasks, although significant challenges remain before machines can function as effective teammates. The human-machine teaming research community attempts to address these challenges by developing and testing methods that identify and enhance the factors essential for successful teaming. However, this community suffers from a lack of requirements for effective research, numerous methods without centralized documentation, and a disconnect between research and real-world applications. These challenges hinder progress and limit the generalizability of research outcomes. To address these issues, we argue that the human-machine teaming research community should establish a more structured and systematic approach to studying and advancing the field. This chapter identifies and discusses several key research directions and actionable outputs for such an approach. These include taxonomies and guidelines to streamline research, team design patterns to describe reusable solutions, modular testbeds to facilitate comparability and reuse, and study templates to foster creativity and encourage sharing. We believe that these elements can help formulate requirements for effective human-machine teaming research and foster the development of modular and well-documented testbeds. Achieving these goals can contribute to more ecologically valid human-machine teaming research and, thus, a stronger connection between research and real-world applications.

8

8.1 Introduction

Humans and intelligent machines increasingly collaborate on complex tasks, ranging from firefighting to manufacturing. As the capabilities of intelligent machines grow, they will

Ruben S. Verhagen, Mark A. Neerincx, X. Jessie Yang, and Myrthe L. Tielman. Advancing Human-Machine Teaming: Definitions, Challenges, Future Directions." HHAI 2025. IOS Press, 2025. 49-59.

increasingly act as full-fledged team members instead of tools. The ultimate goal of human-machine teams is combining the strengths of both parties to accomplish tasks that neither can do alone [3]. The success of these human-machine teams can be affected by many factors, such as mutual trust, coordination, and co-adaptation [99, 175].

Despite ongoing advancements in human-machine teaming, significant challenges remain before machines can function as truly effective teammates for humans [101]. Intelligent machines often operate as opaque "black boxes", making their inner workings and behaviors difficult for humans to comprehend and trust appropriately [79, 126, 130, 145]. Furthermore, these systems still lack the necessary knowledge, skills, and strategies to manage interdependencies with humans effectively [101]. To address these challenges, the human-machine teaming research community plays a critical role by developing and testing methods that identify and enhance the factors essential for successful teaming.

Currently, this community suffers from a lack of requirements for effective research, numerous methods without centralized documentation, and a disconnect between research and real-world applications. Therefore, in this work, we propose that the human-machine teaming research community prioritize establishing a more structured and systematic approach to studying and advancing the field. More specifically, by identifying the requirements for human-machine teaming research and emphasizing reusable and comparable methods and testbeds. We discuss several concrete research directions and actionable outputs for the community to focus on, such as taxonomies and guidelines, team design patterns, modular testbeds, and study templates. Ultimately, we believe these efforts can accelerate progress in the field and contribute to a common platform with essential tools and resources for human-machine teaming research.

8.2 Human-Machine Teaming (Research)

8.2.1 Background and Definitions

Several terms and concepts related to human-machine teaming exist, such as hybrid intelligence [3], human-centered artificial intelligence (AI) [189], and human-machine interaction [157]. In this subsection, we review these concepts in detail. It is important to note that we do not intend to endorse any single definition over the others. Instead, we aim to capture emerging trends and highlight the nuances we observe. First, many human-machine alternatives, such as human-agent/AI/automation/autonomy/computer/robot interaction, are used. We believe human-machine interaction encompasses all these alternatives, as they specify the type of machine interacting with a human. This interaction can be of any kind, complexity, or modality and does not even have to be goal-oriented.

Human-machine teaming is a type of human-machine interaction and can be defined as at least one human and machine working together toward a shared goal [99]. Interaction becomes teamwork when there is a degree of 1) interdependence between the activities and outcomes of humans and machines and 2) machine agency involving independence of actions and proactivity [131, 157, 238, 252]. Human-machine teams focus on augmenting human and machine capabilities by combining the unique strengths of both parties to accomplish what neither could do alone [3]. These teams are generally involved in cognitive and/or physical work, where the former consists of mental or information processing activities, while the latter relates to manipulating tangible objects in the world [217]. Team members usually have explicit roles during this work, such as supervisor, performer, or supporter [99]. These roles often result in different interdependencies between humans and machines, such as required or optional collaboration [99]. The application domains of human-machine teams include emergency response, healthcare, manufacturing, and defense. For example, firefighters that teleoperate explore-and-extinguish robots because of mutually exclusive dependencies to extinguish and navigate [233].

Hybrid intelligence is another type of human-machine interaction and can be defined as combining human and artificial intelligence to augment their isolated operations [3, 51, 200]. This type of interaction is also described as symbiotic artificial intelligence in other works [54]. Most existing works on hybrid intelligence adopt a technology-centric perspective when augmenting human and/or machine intelligence to achieve human or machine goals [51, 127, 209]. Hybrid intelligence systems are primarily involved in cognitive work, such as supporting human learning [146] and computer vision [258]. Another example is combining machine processing with human understanding and reasoning to extract arguments from opinions [209].

In contrast to this technology-centric perspective, recent research on hybrid intelligence emphasizes the collaboration between humans and AI toward shared goals [51, 200, 256]. This team-oriented perspective even frames hybrid intelligence systems as hybrid intelligence teaming, providing an overlap with human-machine teaming [157]. However, hybrid intelligence teams are primarily involved in cognitive work, such as complementing expertise for joint object identification [256] or determining temporary navigation destinations. Since hybrid intelligence teams are rarely involved in physical work, we consider them a subset of human-machine teaming.

Another type of human-machine interaction is human-centered artificial intelligence, which can be defined as augmenting human capabilities with embedded AI methods while ensuring human control [189]. This perspective places humans and their goals at the center, focusing on user needs, explainable systems, and meaningful human control [122, 166, 189, 190]. Human-centered AI can augment human capabilities during both cognitive and physical work, such as augmented reality helmets to improve situational awareness or exoskeletons to enhance physical strength [189].

These examples illustrate machines that enhance human capabilities but not necessarily human intelligence. If human-centered AI systems augment human intelligence by integrating human and artificial intelligence, this can be considered hybrid intelligence. We define this overlap between hybrid intelligence and human-centered AI as human-centered hybrid intelligence. Like hybrid intelligence, human-centered hybrid intelligence is primarily involved in cognitive work, such as personalized AI support for firefighters by highlighting how mission characteristics differ from their experience [220].

In Figure 8.1, we visualize the relationships between all discussed concepts above. Moreover, in Table 8.1, we summarize the conceptual differences between these concepts. In summary, hybrid intelligence augments intelligence and focuses primarily on cognitive work, while human-machine teaming and human-centered AI can also augment other capabilities and focus on physical work. Furthermore, human-centered AI merely augments humans, while human-machine teaming and hybrid intelligence can augment both humans and machines. Finally, human-machine teaming merely tries to achieve shared team goals, while hybrid intelligence and human-centered AI can also try to achieve only human goals.

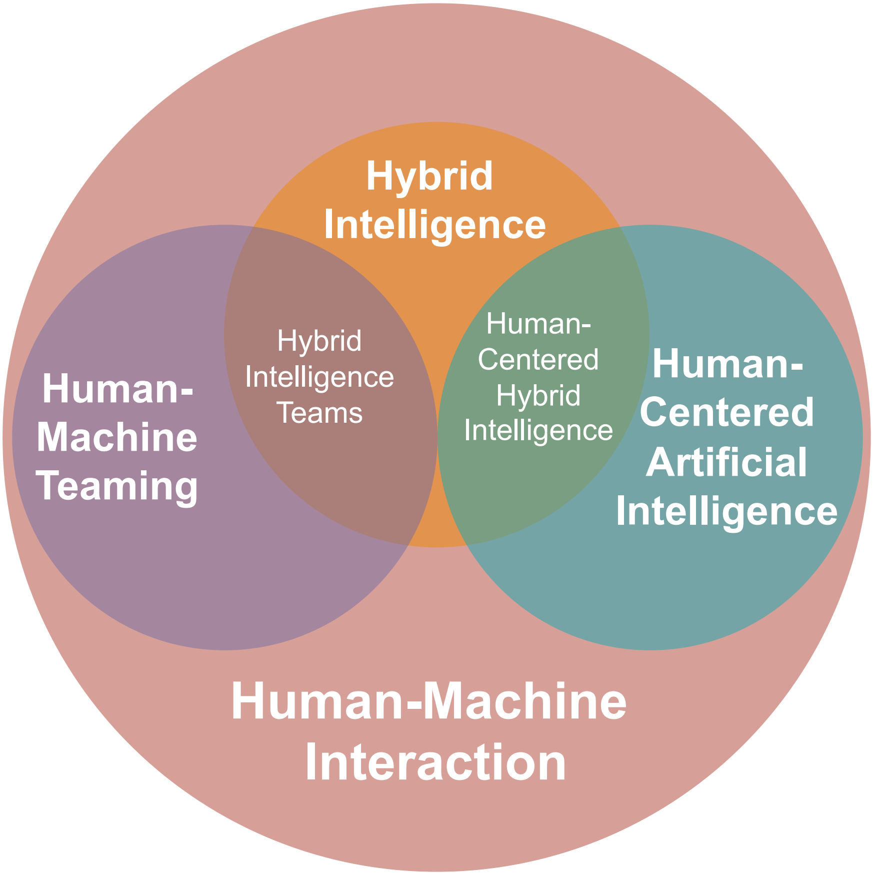

Hybrid Intelligence
Human-Machine Teaming
Hybrid Intelligence Teams
Human-Centered Hybrid Intelligence
Human-Centered Artificial Intelligence
Human-Machine Interaction

Figure 8.1: Venn diagram of human-machine interaction, hybrid intelligence, human-centered artificial intelligence, human-machine teaming, hybrid intelligence teams, and human-centered hybrid intelligence.

8.2.2 Current State and Challenges

Human-machine teaming is increasingly studied across disciplines such as computer science, engineering, and social sciences [40]. It is common to conduct human-machine teaming experiments in simulated and/or controlled environments because studying human-machine teams in the real world can be both time and cost-expensive. Several human-machine teaming testbeds have been used for such experiments, although only a few more frequently [40]. For example, in the Mixed Initiative eXperimental testbed, a human operator detects targets in collaboration with a RoboLeader that collects information from subordinate robots with limited autonomy [32]. In the Cognitive Engineering Research on Team Tasks testbed, a team of three members, with specific roles such as navigator, photographer, and pilot, must collaborate effectively to achieve the team objective of

Table 8.1: Summary of the conceptual differences between human-machine interaction types.

Concept	Augments	Goals	Focus	Example
Human-Machine Teaming	Human capabilities Machine capabilities	Shared	Cognitive Physical	Robot teleoperation
Hybrid Intelligence	Human intelligence Machine intelligence	Human Machine Shared	Cognitive	Data annotation
Hybrid Intelligence Teaming	Human intelligence Machine intelligence	Shared	Cognitive	Joint decisions
Human-Centered AI	Human capabilities	Human	Cognitive Physical	Augmented reality
Human-Centered Hybrid Intelligence	Human intelligence	Human	Cognitive	Personalized support

capturing ground targets [53]. Finally, in the Blocks World 4 Teams testbed, participants must collaboratively deliver blocks in a specific sequence [100].

Some popular research topics facilitated by these testbeds include transparency and explainability [142, 232, 250], trust calibration and repair [229], and co-learning [223] in human-machine teams. For example, how interdependence relationships or communication modalities affect human-machine teaming [40, 157]. In contrast, research and testbeds on multiple humans and machines, machine leadership roles, and communication methods beyond text are still lacking [40].

While these testbeds have facilitated significant progress, human-machine teaming research still faces key challenges. Creating realistic, reusable, and widely adopted testbeds has proven challenging despite the variety of testbeds. Currently, the field lacks (consensus on) effective and comparable research requirements. This often results in a disconnect between human-machine teaming research and its practical applications. Furthermore, it has resulted in numerous methods and testbeds without centralized documentation [248]. For example, most of the identified testbeds in [40] have only been used once or twice and emerged in the past ten years. Moreover, the modalities of these testbeds range from 2D grid worlds to 3D games and augmented to virtual reality. Team characteristics also often vary, especially regarding team composition, task interdependence, leadership structure, and communication structure [40].

8.2.3 Goals

The lack of widely adopted and highly diverse testbeds suggests that many human-machine teaming researchers try to reinvent the wheel. Instead, we argue that this research community should aim for more reusable, comparable, and modular methods and testbeds. The community already adopted this goal during a recent workshop[1] that we organized at the Lorentz Center (Leiden, The Netherlands) in the summer of 2024. Here, we established and discussed several community goals, such as developing human-machine teaming as

[1] www.lorentzcenter.nl/research-environments-for-human-machine-teaming.html

a methodology and achieving consensus on the community's positioning. However, the ultimate goal is a common platform with essential tools and resources for human-machine teaming research. Ideally, this platform should contain research guidelines and requirements for experiment design, a library with reusable and comparable team design patterns, modular testbeds, and templates for describing and comparing studies. However, the requirements of such a platform should first be identified and formalized before actually building it.

8.3 Reusable, comparable, and modular human-machine teaming research

To achieve its goals, we believe the human-machine teaming research community should prioritize reusable, comparable, and modular research. Tasks, environments, and measures could be adapted across different studies and domains instead of reinventing the wheel and starting from scratch. The community should examine what has worked well for others and identify key requirements and customizable characteristics for human-machine teaming research. Moreover, it could benefit from templates for describing, comparing, and designing studies across domains and teams.

Most importantly, however, the community needs more modular methods and testbeds at a common platform to build upon existing research. Ideally, a platform where users can easily browse through modular testbeds that allow customization of tasks, machines, and teams to meet specific research needs. Before the community can achieve these goals, we believe it should first take a step back and study how to conduct reusable, comparable, and modular human-machine teaming research. This way, the requirements for such research can be identified and formalized, contributing to a more structured and systematic research approach. In the next subsections, we outline several concrete directions and outputs to focus on when establishing this approach.

8.3.1 Taxonomies and Guidelines

Identifying requirements and guidelines for human-machine teaming research can provide a baseline framework for researchers and support comparable studies. Therefore, we believe these to be a good starting point towards a more structured and systematic research approach. Such requirements and guidelines can also contribute to characterizing realistic and reusable human-machine teaming tasks, environments, and scenarios, providing a stronger connection between research and real-world applications. We believe literature reviews on human-machine teaming research to be crucial for this, such as the review on testbeds in [40] and independent and dependent variables in [157].

One particularly valuable contribution would be a literature review that results in a taxonomy of human-machine teaming tasks. This taxonomy could include task, machine, human, and team-related categories with examples from literature. Task-related categories could include domain, type, and goal; human and machine-related categories could include behavior and communication; and team-related categories could include design, roles, and interdependencies. These categories could even be structured according to the stages of experiment design at which they are defined. For example, determining the task domain, such as firefighting or warehousing, generally precedes decisions about team roles, such as supervisor, performer, or supporter. This taxonomy should help human-machine teaming researchers to identify and determine their task, machine, human, and team-related categories. Moreover, it would allow people to compare human-machine teaming studies more easily. Such a taxonomy with guidelines can be iteratively refined during community workshops. By establishing this baseline framework, human-machine teaming methods and research can become more reusable and comparable.

8.3.2 Team Design Patterns

Another desired output should be team design patterns that describe generic, reusable, and proven solutions for human-machine teaming [217, 218]. These patterns can then be (re)used during experiment design and compared to others during user studies. Several team design patterns already exist with varying levels of abstraction, such as abstract patterns for AI advisors collaborating with humans [222]. However, more concrete patterns also exist, for example, for humans and machines collaborating in morally sensitive situations by allocating moral decisions to humans and non-moral decisions to machines [213, 214, 233]. Creating more team design patterns based on stakeholder involvement and realistic use cases can strengthen the connection between research and real-world applications, enhancing both translation and generalization [222]. For example, realistic team design patterns for firefighters collaborating with their explore-and-extinguish robots.

We believe the human-machine teaming research community would also greatly benefit from a library with all created team design patterns. Such a team design pattern library can facilitate reusing and comparing human-machine teaming methods. Ideally, a library that is divided between more abstract and concrete patterns and where more abstract and generalizable patterns can be created from more concrete ones. More concrete patterns could also be categorized into common and popular research topics, such as transparency and explainability, trust calibration and repair, and co-learning in human-machine teams. These more concrete patterns can describe generic reusable behaviors of humans and machines for supporting effective and resilient teamwork in these scenarios. For example, what information explore-and-extinguish robots should explain, and at which moments, to firefighter supervisors during semi-autonomous victim search tasks.

8.3.3 Modular Human-Machine Teaming Testbeds

Ultimately, these taxonomies and team design patterns should contribute to modular human-machine teaming testbeds accessible on a common platform. They can inform the community about which aspects of human-machine teaming testbeds can be standardized but configurable to allow comparisons, such as interdependence and role distribution. The community recently started moving in the right direction by sharing customizable testbeds. For example, a customizable testbed in a 2D grid world where participants collaborate with an autonomous, rule-based, explainable artificial moral agent during a firefighting task [226]. In addition to these fixed characteristics, the testbed is customizable with respect to explanation type (technical, ethical, or none) and artificial moral agency (low or high). Other important elements are currently fixed, such as explanation modality (hybrid) and role distribution (human supervision). Customizing these would further strengthen this testbed and facilitate reuse and comparisons with other studies. For example, allowing textual explanations or machine supervision. Ultimately, this can accelerate progress in the field of moral decision-making in human-machine teams.

Another example is a modular testbed in a 3D environment where participants collaborate with an autonomous, rule-based robot that violates their trust during a warehousing task [68]. This testbed allows researchers to alter robot reliability (100% or 70%), physical form (human- or machine-like), and trust repair strategy (apology, denial, promise, explanation, or none). Other important elements are currently fixed, such as trust violation type (competence-based) and communication modality (audio). Customizing these would further strengthen this testbed and enhance reusability and comparability. For example, adding integrity-based trust violations or textual communication.

This testbed is similar to studies in different environments on the influence of trust repair strategies [109] and interdependence [232] on human trust development. However, the former concerns a 3D first-person shooter game and the latter a 2D search and rescue task. The results of these studies would have been more comparable if performed in the same testbed. Moreover, building upon this customizable testbed prevents researchers from starting from scratch and reinventing the wheel. We believe this would accelerate progress in the field of trust calibration and repair in human-machine teams.

These recent customizable testbeds show that the community is moving in the right direction. However, these examples only allow for the adjustment of some machine-related settings. Instead, we consider modular testbeds to allow the adjustment of entire task, team, or machine-related components, such as modifying machines regarding autonomy, behavior, and communication. A starting point could be to create modular testbeds for the most common human-machine teaming topics, such as trust calibration and repair, and transparency and explainability in human-machine teams. Creating these testbeds based on stakeholder involvement, realistic use cases, and real-world applications could greatly benefit both research and practice. We believe hackathons at conferences could be great opportunities to work towards such modular testbeds.

8.3.4 Templates for Describing and Comparing Studies

Finally, the community could benefit from templates to describe, compare, and design studies across domains and teams. Such templates could be used on a common platform for contributors to facilitate sharing research and for users to foster creativity by browsing through and comparing prior research. They should prioritize essential information over supplementary details, such as task descriptions and variables over autonomy levels and hypotheses. So, key characteristics of human-machine teaming studies should be identified first. We believe the aforementioned taxonomies and guidelines can contribute to this identification, while team design patterns and testbeds can be template categories.

An example could be a template with dropdown options and tags for contributors to facilitate uploading their human-machine teaming research. This template could include the categories of team design pattern, task, and research environment. The team design pattern category could specify the number of humans and machines and their roles and relationships. The task category could specify the domain, task type, and interdependencies. Finally, the research environment category could specify the modality, measures, variables, machine embodiment, machine behavior, communication modality, and link to the testbed. Platform users could then enter search queries to find their studies of interest. For example, a user interested in VR studies on trust repair in triads with two humans and one machine.

Such templates can facilitate reusable, comparable, and replicable human-machine teaming research. We believe they can be constructed based on literature reviews and domain expertise, and iteratively refined during workshops based on feedback from both researchers and practitioners.

8.4 Conclusion

This chapter answers the seventh research sub-question of the thesis: *How should the human-agent teaming research community adopt a more structured and systematic approach to advance the field*? We identified three significant challenges for the human-machine teaming research community: a lack of requirements for effective research, numerous methods and testbeds without centralized documentation, and a disconnect between research and real-world applications. To address these challenges, we proposed that the community establish a more structured and systematic research approach. We outlined four key research directions and actionable outputs of such an approach: taxonomies and guidelines to streamline research, team design patterns to describe reusable solutions, modular testbeds to facilitate comparability and reuse, and study templates to foster creativity and encourage sharing. We believe these elements can help formulate requirements for effective human-machine teaming research and foster the development of fewer but more modular, well-documented methods and testbeds. Achieving these two goals can contribute to more ecologically valid human-machine teaming research and, thus, a stronger connection between research and real-world applications. Ultimately, all these efforts can accelerate progress in the field and lay the foundation for a common platform with essential tools for human-machine teaming research.

9

Conclusion

9.1 Findings

The research presented in this thesis focuses on designing transparent and explainable AI agents that foster effective and responsible human-agent teaming across interdependencies and autonomy levels. It uses a combination of analytical and empirical research to investigate such teamwork. The analytical studies identified and defined core human-agent collaboration properties that influence team processes and outcomes. These properties were operationalized as independent and dependent variables in our empirical studies. We used simulation environments for the empirical studies on the more fundamental research questions regarding the effects of interdependence, autonomy, and explanations on team processes such as trust. To apply transparency and explanations in a physical and more realistic context, our final empirical study was conducted with a physical robot in a practically grounded setting inspired by real firefighting scenarios. This study also marked a step towards the envisioned use case of collaborative firefighting robots.

Starting with an analytical study, this thesis first develops a conceptual framework that defines and relates agent transparency, explainability, interpretability, and understandability (Chapter 2). We used this conceptual framework to define the independent variables of our empirical study in a simulation environment on the influence of interdependence and AI agents' transparency and explanations on human-agent teaming (Chapter 3). A similar empirical study was conducted in another simulation environment to investigate the influence of interdependence on the calibration of human trust in transparent and explainable agents (Chapter 4).

The following analytical study focused more on responsible human-agent teaming by developing an evaluation method of meaningful human control (Chapter 5). This method was used to measure meaningful human control in our empirical study on the influence of AI agents' autonomy, transparency, and explanations on human-agent teaming (Chapter 6). We used all insights from these empirical studies in simulation environments to design our physical human-robot collaboration system for firefighting (Chapter 7). Finally, we synthesized all insights obtained during the design and evaluation of our empirical studies to develop a research agenda for the human-agent teaming community (Chapter 8). Below, we present the conclusions that can be drawn from the research sub-questions.

How should we define and relate agent transparency, explainability, interpretability, and understandability? (Chapter 2)

We defined agent transparency and explainability as active, objective agent characteristics that involve disclosing information (transparency) and clarifying that information (explainability). Agent transparency and explainability can result in interpretability and understandability, respectively. We defined agent interpretability as the possibility to access, analyze, and exploit disclosed information. In contrast, agent understandability was defined as having knowledge of disclosed and clarified information, as well as the relationships between them. This conceptual framework offered clear and distinct definitions of key concepts within the field of explainable AI, resolving common ambiguities and the lack of agreement.

How do interdependence and agents' transparency and explanations influence effective human-agent teaming, individually and interactively? (Chapter 3)

We used the conceptual framework presented in Chapter 2 to define the independent variables of the empirical study conducted to answer the research question of Chapter 3. To answer this research question, we conducted a user study with 72 participants on the effects of agent communication style (silent, transparent, explainable, or adaptive) across high and low interdependence levels. These participants collaborated with a virtual agent during two simulated search and rescue tasks varying in their level of interdependence.

Our results showed lower trust, reliance, and team performance when interdependence was high, together with a higher workload and situation awareness. Moreover, when interdependence was high, people communicated more if the agent was not silent and contributed more if the agent did not provide explanations. We also found that agent transparency and explanations lead to higher trust and understanding, without evidence of a higher workload. Furthermore, agent explanations resulted in greater reliance on the agent. Finally, when interdependence was high, people had higher situation awareness if the agent provided explanations instead of being silent.

Overall, this study contributed empirical insights into the human-agent teaming conditions under which agent transparency and explanations are either beneficial or detrimental. Most importantly, our findings demonstrated how interdependence can influence the effects of agent transparency and explanations, underscoring the importance of interdependence in studies on transparent and explainable agents.

How do interdependencies during human-agent teaming influence the human-agent trust calibration process? (Chapter 4)

To answer this research question, we conducted a user study with 80 participants on the effects of interdependencies (complete independence, complementary independence, optional interdependence, required interdependence) on the trust calibration process. These participants collaborated with a virtual agent during a simulated search and rescue task in human-agent teams characterized by one of the four interdependencies. We included an agent-induced trust violation in the task to facilitate a dynamic trust calibration process.

Our results showed that only in human-agent teams with joint actions, human trust did not recover to its initial pre-violated value. However, correlation analyses between these trust values and agent trustworthiness suggested a more accurate trust calibration process in the human-agent teams with joint actions. Overall, this study provided some first empirical evidence that interdependence relationships during human-agent teaming influence human-agent trust calibration. These results highlight the importance of interdependencies in fostering appropriate trust during teamwork.

How can we measure meaningful human control during human-agent teaming? (Chapter 5)

We conducted a focus group study with five experts to develop an evaluation method for meaningful human control. This study explored quantitative operationalizations of meaningful human control in human-agent teams. Our results highlighted that the traceability requirement of meaningful human control can be operationalized quantitatively, with situation awareness and performance being useful measures to objectively assess traceability aspects. Furthermore, results indicated that team and agent outcomes can be used to verify meaningful human control, while reasons underlying these outcomes can determine the level of meaningful human control.

Based on these results, we proposed an evaluation method for meaningful human control. This method involves subjectively and objectively quantifying traceability using human responses during and after simulations of the collaboration. It also uses semi-structured interviews after the simulation to identify reasons underlying outcomes and suggestions for enhancing meaningful human control. Overall, this method enables researchers and designers to assess whether meaningful human control is present during human-agent teaming.

How do agents' autonomy and their transparency and explanations influence responsible human-agent teaming, individually and interactively? (Chapter 6)

We used the evaluation method presented in Chapter 5 to measure meaningful human control during the empirical study conducted to answer the research question of Chapter 6. To answer this research question, we conducted a user study with 72 participants on the effects of agent explanations (no additional information, feature contributions, or potential consequences) across high and low agent autonomy levels. These participants collaborated with a more and less autonomous artificial moral agent during simulated firefighting tasks, while receiving one of the three explanations.

Our results showed a lower moral trust, agreement, and meaningful human control when the agent was more autonomous. Furthermore, people disagreed and reallocated decisions to themselves more when the agents explained potential consequences rather than feature contributions or no additional information, especially when moral sensitivity was higher. Overall, our findings suggest that people prefer more involvement over higher agent autonomy and take on greater moral responsibility when agents explain potential consequences. These empirical and actionable insights are crucial for designing agents that enhance human moral awareness and human-agent teaming in morally sensitive situations.

> How should we design a semi-autonomous, transparent and explainable robot in human-robot teams for firefighting, and how does the increased autonomy impact effective and responsible collaboration compared to teleoperation? (Chapter 7)

We drew on the insights from the empirical studies in Chapters 3 and 6 to design our human-robot collaboration system for firefighting in Chapter 7, called TEAMS (Transparent and Explainable Autonomy for Mapping and Searching). This system moves beyond teleoperating firefighting robots by proposing and explaining intermediate navigation destinations while autonomously navigating towards them. It combines fuzzy logic control, explainable AI, and meaningful human control to ensure effective and responsible human-robot collaboration when searching for victims. The system is grounded in expert knowledge and our empirical insights on especially adaptive explanations and human involvement during human-agent teaming. More specifically, we implemented on-demand, adaptive robot explanations and ensured sufficient human involvement during the collaboration through destination decisions and object recognition.

We conducted a mixed-methods evaluation comparing our system with teleoperation through four in-person trials with laypeople and video-based evaluations involving 26 laypeople and one domain expert. During the in-person trials, participants interacted with both systems/robots during two simulated firefighting tasks to map environments and locate victims. For the video-based evaluations, participants watched us interact with the user interfaces of both systems.

Our results reveal a clear contrast between qualitative, in-person evaluations and quantitative, video-based evaluations. During the in-person trials, laypeople highly appreciated the information-rich collaboration with TEAMS to address the manual control challenges of teleoperation. In contrast, for the video-based evaluation, laypeople scored TEAMS worse than teleoperation on workload, situation awareness, and system usability. The in-person findings aligned with the domain expert's evaluation of both systems, who emphasized TEAMS' collaborative value and the burden of teleoperation. Therefore, we suggested that non-expert, video-based evaluations are poor proxies for assessing human-robot teamwork.

Our findings offer several practical implications for deploying TEAMS during firefighting. These include providing training to overcome initial complexity, implementing iterative and human-centered refinements, and enhancing both hardware and software. Overall, this study contributes both a practical solution for semi-autonomous, transparent and explainable robots in real-world firefighting scenarios, as well as the empirical insights into how such a robot affects human-robot teaming compared to teleoperation.

> How should the human-agent teaming research community adopt a more structured and systematic approach to advance the field? (Chapter 8)

We synthesized all insights from (1) our Lorentz Center workshop on research environments for human-agent teaming and (2) the design and evaluation of our empirical studies to develop a research agenda for the human-agent teaming community in Chapter 8. This agenda consists of taxonomies and guidelines to streamline research, team design patterns to describe reusable solutions, modular testbeds to facilitate comparability and reuse, and study templates to foster creativity and encourage sharing. These directions can accelerate progress in the field and lay the foundation for a common platform with essential tools for human-agent teaming research.

9.2 Limitations

There are several limitations of the studies conducted to answer our research questions. These concern the methods, validity, and generalizability of these studies.

9.2.1 Ecological Validity

First, we primarily conducted human-grounded evaluations using simplified AI agents, approximations of the relevant context in simulation environments, and laypeople as participants [56]. For example, we studied meaningful human control and moral decision-making in simplified firefighting scenarios. We used these evaluations because they are well-suited for testing general notions of explanations and collaboration in controlled user studies. Moreover, these studies in interactive simulation environments provided greater fidelity than commonly used text and video vignettes [119], as participants actively interacted with agents in real-time rather than passively rating pre-determined or pre-recorded scenarios. Another primary, practical reason for human-grounded evaluations is the challenge of recruiting and training sufficient domain experts [56, 195].

Despite these advantages, using such evaluations still raises questions regarding the validity and generalizability of study findings. Application-grounded evaluations in realistic contexts, with actual AI agents, and representative population samples that fit the contexts would enhance the ecological validity of our studies and findings [56, 195]. Therefore, our findings provide important insights that can be validated with application-grounded evaluations to examine if they generalize to real-life settings.

9.2.2 External Validity

Second, we primarily focused on human-agent teaming in high-stakes domains. For example, we operationalized and implemented meaningful human control in the context of human-agent/robot interaction for firefighting. These domains are characterized by uncertainty, high workload, risk, and time pressure — factors known to shape trust and reliance, situation awareness and performance, and the benefits of agent transparency and explanations [34, 82, 97, 157].

Consequently, the external validity of our study findings is another limitation, as their generalizability to other human-agent teaming domains is debatable. On the one hand, this is less problematic for our findings on meaningful human control as its conditions, properties, and implementation are always context- and system-specific [30, 179]. On the other hand, we prioritized ecological validity over generalization to other domains by focusing on high-stakes scenarios that increased domain-relevant realism. Nevertheless, this ecological validity can be further enhanced by application-grounded evaluations.

9.2.3 Human-Agent Dyads

Third, we only focused on human-agent teaming in dyads consisting of one human and one agent. In reality, human-agent teams will likely consist of multiple humans and AI agents [69, 188]. For example, an exploration and extinguishing robot to map environments on the ground, a drone to map environments while flying, one operator supervising the robot, and another one teleoperating the drone. Such multimember configurations strongly shape the team processes and outcomes studied in our thesis. For instance, trust will develop not only between team members but also towards all possible dyadic relationships and the team as a whole [206]. However, the current interaction within the Rotterdam Fire Brigade involves a single human operator teleoperating a single exploration and extinguishing robot, which aligns with the dyads in our studies.

9.3 Future Work

We identify several suggestions for future work. These include exploring new dynamics and adaptive agents, advancing methods, and strengthening validity and generalizability.

9.3.1 Exploring New Dynamics and Adaptive Agents

The first suggestion for future work would be to design and develop adaptive, transparent and explainable agents based on user models and contextual factors. Our results provide insights into the conditions under which agent transparency and explanations are beneficial or detrimental, for example, how explanations benefit situation awareness when interdependence is high. Such insights can inform the design and implementation of adaptive, transparent and explainable agents.

Developing these agents is one of the main goals of the explainable AI community [9, 145]. We implemented adaptive, on-demand robot explanations in our human-robot collaboration system for firefighting. These explanations allow users to personalize the disclosed and clarified information. Alternatively, the robot could personalize its explanations by modelling users and/or environments and adapting the provided information based on these models. For example, modelling user workload and tailoring explanations accordingly. Future work could focus on building these user and context models, as well as adapting explanations to the specifics of these models. Furthermore, it should compare user- and agent/robot-personalized explanations to determine which one better supports understanding, workload, trust calibration, and performance.

Our results show that interdependence influences trust calibration in human-agent teams. Therefore, it makes sense to tailor trust repair strategies based on interdependence relationships. To gain insights into how to achieve this, future work could investigate the interaction between interdependence and trust repair strategies. These insights could then directly inform the design and development of agents that adapt their trust repair strategies based on interdependence. For example, agents could promise improvement when trust is violated during joint actions, but explain what went wrong when trust is violated during independent actions.

We also show that during agent allocation of moral decision-making, people prefer greater involvement over higher agent autonomy. Therefore, it would be interesting to compare these results when collaborating with a less autonomous artificial moral agent or when humans allocate moral decision-making [77]. Such results can provide important insights before validating dynamic task allocation during application-grounded evaluations.

9.3.2 Advancing methods

Another possible next step would be to further enhance our human-robot collaboration system for firefighting (TEAMS), which transitions from teleoperated exploration and extinguishing robots to more autonomous teammates. The evaluations of this system highlighted several ways to streamline the user interface and improve the hardware and software. For example, reducing the presented information to decrease operator workload and optimizing software to accelerate the interaction. Another next step could be the addition of computer vision to detect doors, victims, and obstacles in low-visibility conditions caused by smoke. It would be interesting to compare how this shift in interdependence affects meaningful human control during the collaboration.

Designing and developing modular testbeds for human-agent teaming is another important suggestion for future work. Several human-agent teaming testbeds exist, but only a few have been used more frequently [40]. Moreover, the modalities of these testbeds range from two-dimensional grid worlds to three-dimensional games, and from augmented to virtual reality. The characteristics of the human-agent teams also often vary, especially in terms of composition, interdependence, leadership, and communication. Finally, only a few of these testbeds allow customization of important characteristics such as agent explanations, autonomy, or reliability.

Rather than customizing these specific agent-related settings, modular testbeds should allow the adjustment of entire task-, team-, or agent-related components. For example, modifying agents' autonomy, behavior, and communication. A good starting point would be to create modular testbeds for highly popular and relevant human-agent teaming topics, such as trust calibration and repair or transparency and explainability. Widely adopting such testbeds would facilitate reuse and make studies and their results more comparable. Moreover, it would prevent researchers from reinventing the wheel and accelerate progress in the field.

9.3.3 Strengthening Validity and Generalizability

This thesis also presents several opportunities for future work to further strengthen the validity and generalizability of studies, methods, and findings. First, conducting (more) application-grounded evaluations in realistic contexts, with actual AI agents, and representative domain experts. Such evaluations are challenging because it is hard to recruit sufficient domain experts [56, 195]. Moreover, these evaluations require a stable prototype of a human-agent system. Our proposed human-robot collaboration system (TEAMS) provides such a stable prototype that can be iteratively improved and evaluated by domain experts at the test sites of the Rotterdam Fire Brigade.

Second, by providing a stable prototype, TEAMS also enables longitudinal studies on transparent and explainable agents for human-agent teaming in real-life settings. Such studies are essential for investigating trust calibration, as this is a dynamic process that involves adjusting trust over time and through repeated interactions [141, 156]. In contrast, our studies primarily involved single 10-15 minute interactions. Conducting longitudinal studies instead can reveal how trust develops over time, such as when it calibrates towards appropriate trust.

9.4 Contributions

9.4.1 Scientific

The main scientific contributions of this thesis can be grouped into new knowledge, methods, and valid and generalizable findings.

New knowledge

First, this thesis contributes a conceptual framework that defines and relates key concepts within the explainable AI community. These concepts are often used interchangeably or within each other's definitions, resulting in ambiguity and comprehension challenges. Our definitions and relationships resolve these issues, facilitating their investigation, manipulation, and implementation.

We also contribute empirical insights into the interdependence conditions under which agent transparency and explanations are beneficial or detrimental, such as the benefits of explanations on situation awareness when interdependence is high. Overall, we highlight that interdependence affects the impact of agent transparency and explanations on the human-agent teaming processes' situation awareness, communication frequency, and task contribution. These insights help us understand what agents should disclose and clarify to humans to foster effective collaboration across varying levels of interdependence. This can inform the design and implementation of agents that adapt their transparency and explanations based on interdependence, to ensure beneficial effects.

Furthermore, we provide first empirical insights into the effects of interdependence on trust calibration in human-agent teams, highlighting differences in trust calibration between human-agent teams with and without joint actions. These insights help us understand how interdependence relationships between humans and explainable agents influence trust violation, repair, and calibration. This can inform the design and implementation of agents that adapt their transparency and explanations based on interdependence, to foster an accurate trust calibration process.

In this thesis, we argue that human-agent teaming should be both effective and responsible. The emphasis on responsibility is relatively novel yet crucial, as increasingly autonomous AI agents will be collaborating with humans in morally sensitive situations. Within this context, we provide empirical insights into the effects of agent autonomy and explanations on moral decision-making in human-agent teams. More specifically, we show that people (1) prefer more involvement over higher agent autonomy and (2) disagree and reallocate decisions to themselves more when agents explain potential consequences, especially in more morally sensitive situations. These insights highlight the importance of human involvement and explaining potential consequences for responsible collaboration, helping us understand what agents should disclose and clarify to humans to foster responsible collaboration across varying levels of agent autonomy. This can inform the design and implementation of transparent and explainable agents that enhance human moral awareness and human-agent teaming in morally sensitive situations.

We also provide insights into the design and evaluation of a semi-autonomous, transparent and explainable robot in human-robot teams for firefighting. These findings demonstrate the potential of such a collaborative robot to reduce the burden of camera-based teleoperation in low-visibility conditions by providing destination suggestions and autonomous navigation. However, training is required to overcome initial complexity, while iterative, human-centered refinements and software optimization should further enhance the interaction. Moreover, this thesis contributes empirical insights into the effectiveness of in-person and video-based evaluations of human-robot interaction. These findings reveal that laypeople evaluated collaboration with the robot much more positively than teleoperation during in-person trials, yet rated it lower in video-based evaluations. This suggests that non-expert, video-based evaluations are poor proxies for assessing human-robot teaming.

Methodological

In addition to new knowledge, this thesis also contributes novel methods. First, we developed an evaluation method for meaningful human control in human-agent teams, grounded in expert knowledge. Such methods are currently lacking, while meaningful human control is already increasingly imposed as a requirement for AI agents. This method enables researchers to assess whether meaningful human control is present during human-agent teaming, while also providing opportunities to build upon it.

We also developed several customizable testbeds for transparent and explainable agents in human-agent teams across varying levels of interdependence and agent autonomy. These testbeds enable the customization of various aspects, including interdependence relationships and agent communication style, explanation type, and autonomy. One of the testbeds is realistic and inspired by the development and expected use of the Rotterdam Fire Brigade's exploration and extinguishing robot. Overall, our testbeds facilitate both reuse and future research.

We also provide a research agenda with concrete directions for the human-agent teaming research community. That agenda is grounded in all the lessons learned during our empirical studies, as well as a five-day workshop with human-agent teaming experts. These concrete directions offer several opportunities for future research, including the development of taxonomies to streamline research, team design patterns to describe reusable solutions, modular testbeds to facilitate comparability, and study templates to encourage sharing.

Finally, this thesis contributes TEAMS (Transparent and Explainable Autonomy for Mapping and Searching), a human-robot collaboration system for firefighting. This system moves beyond teleoperating firefighting robots by proposing and explaining intermediate navigation destinations while autonomously navigating towards them. It is grounded in domain expert knowledge and combines fuzzy logic control, explainable AI, and meaningful human control. This system provides a stable prototype that can facilitate crucial future research on application-grounded evaluations and longitudinal studies.

Validity and Generalizability

We also contribute valid and generalizable findings. First, we validate agent transparency and explainability through human user studies, something only fewer than 1% of explainable AI papers do [195]. Moreover, we primarily conducted in-person, high-fidelity user studies with robust experimental designs, strengthening internal validity. These studies provide crucial insights that can be further validated in real-life settings through application-grounded evaluations or directly applied when implementing agents.

Furthermore, including firefighters as direct stakeholders throughout different stages provides more valid and generalizable results. Beyond strengthening our results, this also provides a practitioner-in-the-loop methodology for applied human-agent teaming. First, we included them during the design of a simulation environment based on their interaction with exploration and extinguishing robots. Second, we elicited their preferences and requirements for our human-robot collaboration system (TEAMS), directly informing its design and implementation. Third, a video-based expert review compared TEAMS with the current state-of-the-art teleoperation, providing expert opinions on our system's advantages, disadvantages, challenges, and applicability. Overall, this stakeholder engagement enhances the ecological validity of our contributions and provides a reusable protocol that other researchers can adopt.

9.4.2 Societal

This thesis also provides several contributions to society. First, the empirical insights into the effects of agent transparency and explanations across interdependence reveal the conditions under which these effects are beneficial or detrimental. For example, the benefits of agent explanations on situation awareness when interdependence is high. To ensure these beneficial effects on society, agents should adapt their transparency and explanations based on user and context. For example, sharing excessive amounts of information with firefighters during extremely high-pressure situations can cause information overload and lead to detrimental outcomes, such as loss of life. Our empirical insights can inform the design and development of agents that adapt their transparency and explanations based on user and context, such as providing explanations when interdependence is high, while being transparent when interdependence is low. Ultimately, this can benefit both direct stakeholders interacting with these agents (e.g., firefighters) and indirect stakeholders affected by the consequences of those interactions (e.g., victims).

We also provide an evaluation method for meaningful human control, based on expert knowledge. This method can help system designers, policymakers, and users measure the amount of meaningful human control in human-agent teams. Moreover, it can identify aspects that require improvement, such as interaction with or understanding of agents. Since such methods are currently lacking while meaningful human control is increasingly imposed as a requirement, this provides an essential societal contribution.

Furthermore, we show that agent explanations of potential consequences are required for responsible human-agent teaming. This empirical insight can inform the design and development of agents that enhance human moral awareness and human-agent teaming in morally sensitive situations. For example, by generating and communicating agent explanations of potential consequences in high-stakes, human-agent teaming domains. This can benefit direct stakeholders interacting with these agents by ensuring meaningful human control and no accountability issues resulting from detrimental outcomes. It can also benefit indirect stakeholders affected by (positive) consequences of those interactions under meaningful human control.

Finally, we contribute a human-robot collaboration system for firefighting that moves beyond teleoperation by proposing and explaining intermediate navigation destinations while autonomously navigating towards them. Teleoperation is a well-known and commonly acknowledged challenge by the Rotterdam Fire Brigade, as is the desire to enhance robot autonomy to address this challenge. Our system provides a concrete solution to their problem while also ensuring meaningful human control in this morally sensitive domain. Moreover, it provides decision support to the operator through proposed intermediate destinations, whereas other works typically leave the selection of destinations completely to the operators. Consequently, our human-robot collaboration system can further reduce operator workload. Furthermore, we contribute a modular system that facilitates the adaptation of fuzzy rules or the integration of computer vision. Overall, this system represents a significant step toward the envisioned ideal future scenario of human-robot collaboration during firefighting, as illustrated in the Introduction. It can directly benefit firefighters working with these robots, victims of fires, and, consequently, society as a whole.

9.4.3 Ethical Reflection

As agents become increasingly autonomous and interdependent when collaborating with humans in high-stakes domains, ethical reflections are crucial. We should be careful that optimizing team performance goes hand in hand with respecting human values, legal standards, and ethical principles. It is argued that even fully autonomous artificial moral agents can be under meaningful human control if they satisfy the tracking and tracing conditions. This would involve an autonomous exploration and extinguishing robot that determines where to go, what to do, and whom to rescue. We do not view this as a desirable goal. Instead, we advocate human-agent collaboration. Specifically, designing this collaboration so each team member does what they do best, with complementary support to augment each other's capabilities. For example, allocating moral decision-making to humans, while agents provide supplementary explanations that enhance human moral agency and result in better decisions.

We also believe that stakeholder involvement is essential when designing human-agent teams for high-stakes domains. Early elicitation of preferences and requirements can, a priori, contribute to human-agent collaboration systems that align with human values, legal standards, and ethical principles. This does not make a posteriori evaluations of meaningful human control redundant; we still require more methods for such evaluations. Similarly, we need application-grounded evaluations in realistic contexts and with representative domain experts before really deploying such systems in practice.

When designing increasingly autonomous and interdependent agents for responsible human-agent teaming in high-stakes domains, transparency and explanations can never be ignored. We still require extensive research on how such information can ensure alignment with human values, legal standards, and ethical principles. This research should focus on all phases of explanation generation, communication, and reception. Ultimately, agents' autonomy and interdependence should be advanced only to the extent that their transparency and explanations can still ensure a collaboration that is responsible and under meaningful human control.

9.4.4 Take-Home Message

Human-agent teaming in high-stakes domains is already contributing positively to society, despite being in its early stages. The agents in these teams are often still tools instead of teammates, directly controlled by humans. Becoming teammates rather than tools requires agents to be more autonomous and interdependent. These two factors determine what humans exactly need to know about agents. Agent transparency and explanations can provide this required knowledge to collaborate effectively and responsibly. As interdependence relationships and agent autonomy shape information requirements, they also moderate how transparent and explainable agents affect team processes such as trust. Therefore, autonomy and interdependence are crucial to consider when designing and evaluating transparent and explainable agents. Yet, we lacked an empirically grounded understanding of what agents should disclose and clarify to humans to foster effective and responsible collaboration, especially across levels of interdependence and agent autonomy.

This thesis aimed to provide such an understanding by investigating how to design transparent and explainable agents that foster effective and responsible human-agent teaming across interdependencies and autonomy levels. The key takeaways are:

- Agent transparency and explanations are not the same. Transparency involves disclosing information, explanations clarify this disclosed information (Chapter 2).
- Interdependence shapes the conditions under which agent transparency and explanations are beneficial or detrimental (Chapter 3).
- Interdependence influences trust calibration in human-agent teams (Chapter 4).
- Meaningful human control in human-agent teams can be subjectively and objectively quantified (Chapter 5).
- Responsible human-agent teamwork requires agents to explain potential consequences of decisions (Chapter 6).
- A semi-autonomous, transparent and explainable robot for firefighting can address the challenges of teleoperation in low-visibility conditions (Chapter 7).
- The human-agent teaming research community should develop taxonomies, team design patterns, modular testbeds, and study templates to accelerate the field's progress (Chapter 8).

Overall, this thesis provides shared definitions, empirical insights, an evaluation method, a practically grounded collaboration system, and a research agenda. Collectively, these contributions support the design and evaluation of transparent and explainable AI agents that foster effective and responsible human-agent teaming across interdependencies and autonomy levels.

A

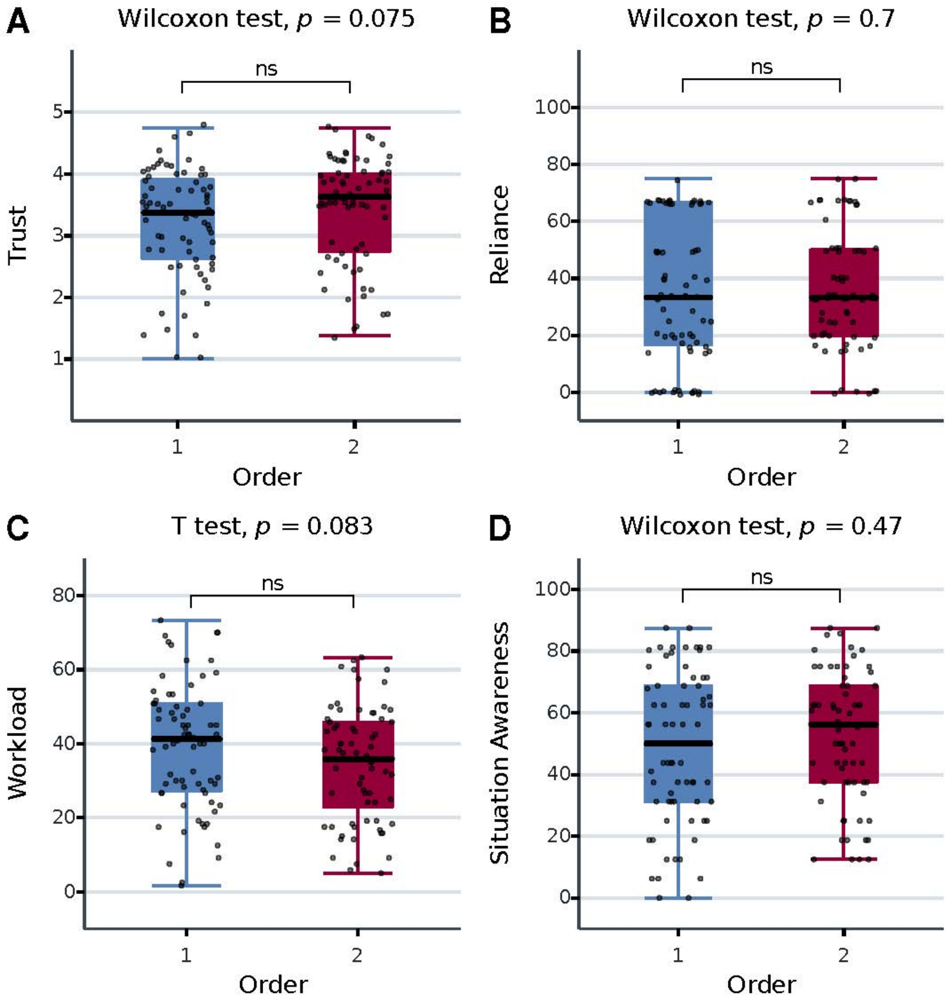


Figure A.1: Boxplots of trust (**A**), reliance (**B**), workload (**C**), and situation awareness (**D**) for both order conditions. Order 1 started with the low interdependence condition followed by the high interdependence condition. Order 2 started with the high interdependence condition followed by the low interdependence condition.

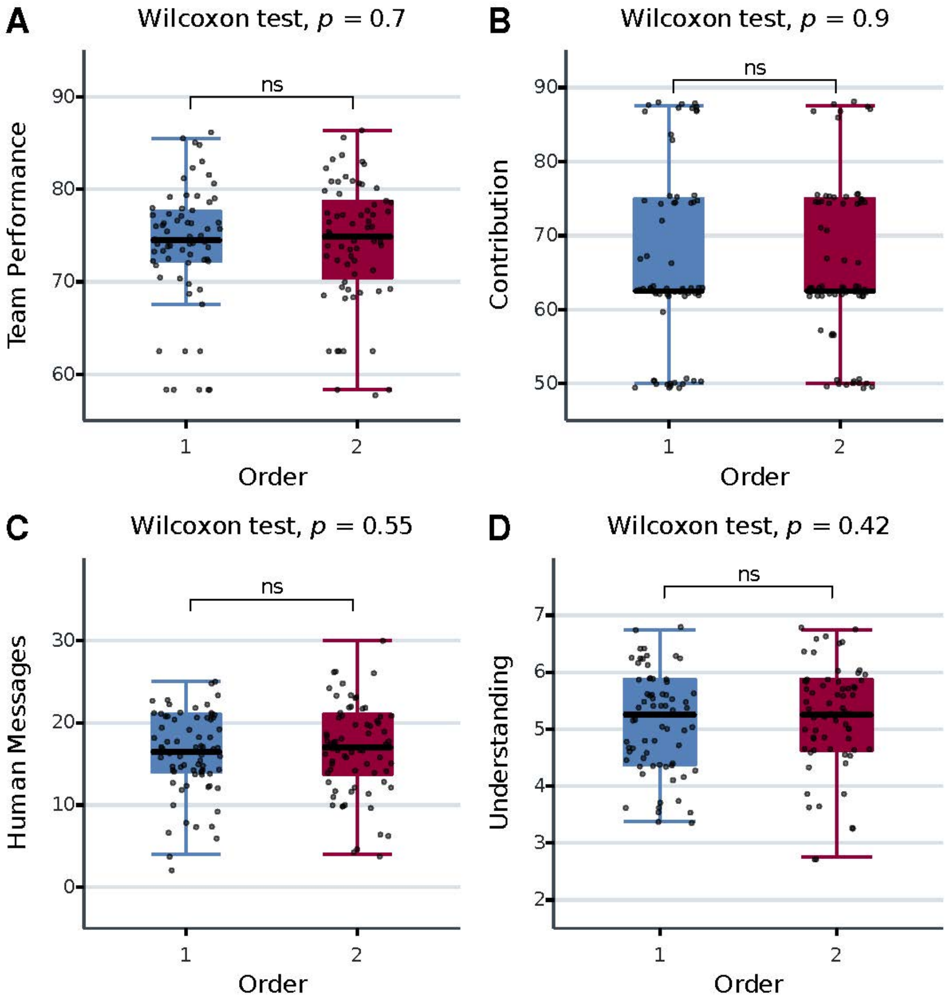


Figure A.2: Boxplots of team performance (**A**), human rescue contribution (**B**), human messages sent (**C**), and system understanding (**D**) for both order conditions. Order 1 started with the low interdependence condition followed by the high interdependence condition. Order 2 started with the high interdependence condition followed by the low interdependence condition.

Variable	Order	Mean (SD)	Mean Rank (SD)	Median (IQR)
Trust	1	3.21 (0.89)	66.32 (40.20)	3.38 (1.28)
	2	3.44 (0.87)	78.68 (42.37)	3.62 (1.25)
Reliance	1	34.60 (24.40)	71.15 (45.12)	33.30 (50.00)
	2	35.70 (29.90)	73.85 (37.43)	33.30 (30.00)
Workload	1	39.30 (17.00)	78.53 (42.93)	41.20 (23.50)
	2	34.60 (15.00)	66.47 (39.82)	35.80 (22.90)
SA	1	49.00 (24.20)	70.01 (44.26)	50.00 (37.50)
	2	52.70 (20.60)	74.99 (39.01)	56.20 (31.20)
Performance	1	71.10 (12.20)	71.15 (40.77)	74.10 (7.03)
	2	72.00 (10.80)	73.85 (42.87)	74.40 (9.27)
Contribution	1	67.30 (15.30)	72.08 (43.65)	62.50 (13.10)
	2	67.10 (12.00)	72.92 (37.48)	62.50 (12.50)
Messages	1	16.20 (4.95)	70.44 (40.00)	16.50 (7.00)
	2	16.80 (5.50)	74.56 (43.33)	17.00 (7.25)
Understanding	1	4.92 (1.19)	69.72 (42.21)	5.06 (1.31)
	2	5.02 (1.22)	75.28 (41.22)	5.25 (1.16)

Table A.1: Descriptive statistics for trust, reliance, workload, situation awareness, team performance, human rescue contribution, human messages sent, and understanding for both order conditions. Order 1 started with the low interdependence condition followed by the high interdependence condition. Order 2 started with the high interdependence condition followed by the low interdependence condition.

Variable	Time	Mean (SD)	Mean Rank (SD)	Median (IQR)
Performance	1	71.60 (8.91)	66.78 (39.32)	73.60 (7.46)
	2	71.50 (13.60)	78.22 (43.49)	75.70 (8.10)
SA	1	49.10 (23.90)	69.44 (43.34)	50.00 (37.50)
	2	52.60 (21.10)	75.56 (39.94)	56.20 (31.20)
Contribution	1	65.40 (13.50)	67.08 (40.00)	62.50 (17.90)
	2	69.00 (13.80)	77.92 (40.63)	62.50 (12.50)
Reliance	1	37.40 (23.00)	76.31 (42.38)	33.30 (46.70)
	2	32.90 (21.20)	68.69 (40.18)	33.30 (33.30)

Table A.2: Descriptive statistics for team performance, situation awareness, human rescue contribution, and reliance for both time points.

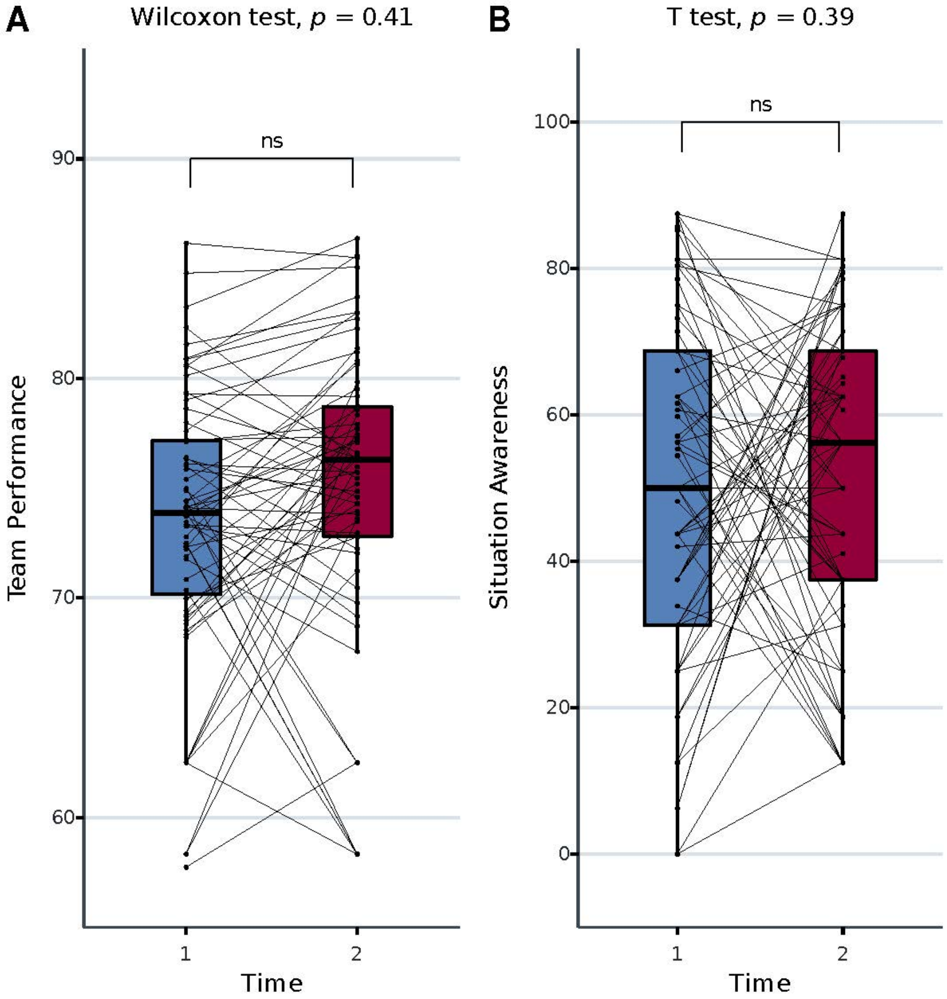


Figure A.3: Boxplots of team performance (**A**) and situation awareness (**B**) at each time point. Time point 1 includes all data from the low interdependence condition from order 1 and high interdependence condition from order 2. Time point 2 includes all data from the low interdependence condition from order 2 and high interdependence condition from order 1.

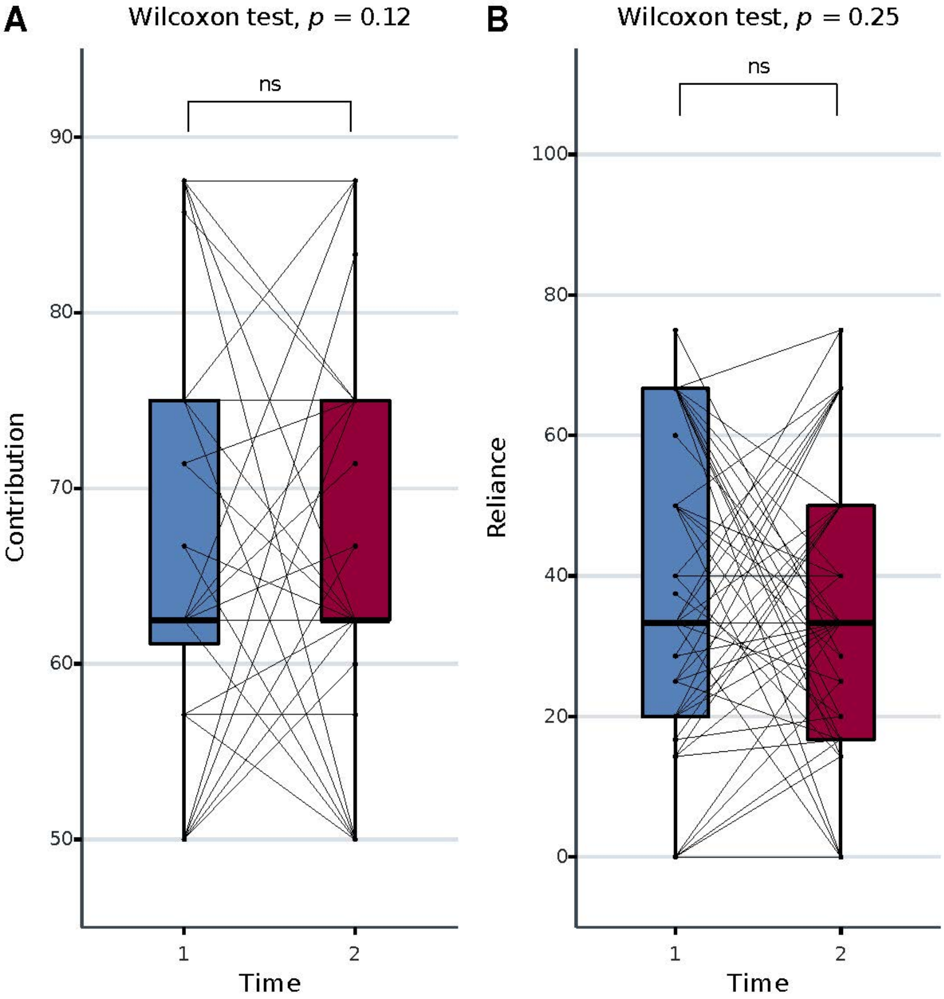


Figure A.4: Boxplots of human rescue contribution (**A**) and reliance (**B**) at each time point. Time point 1 includes all data from the low interdependence condition from order 1 and high interdependence condition from order 2. Time point 2 includes all data from the low interdependence condition from order 2 and high interdependence condition from order 1.

B

B.1 Method

B.1.1 Participants

We balanced demographics, risk propensity [140], trust propensity [143], and utilitarianism [104] across explanation and counterbalancing conditions. Results showed no significant differences between explanation conditions for gender ($\chi^2(2) = 0.08$, p = 0.96), age (W = 2.96, p = 0.23), education (W = 1.22, p = 0.54), gaming experience (W = 0.54, p = 0.76), risk propensity (W = 0.09, p = 0.96), trust propensity (W = 1.51, p = 0.47), and utilitarianism (F(2, 69) = 0.12, p = 0.89). Moreover, results also showed no significant differences between counterbalancing conditions for gender (p = 1.00), age (W = 5.81, p = 0.56), education (W = 4.63, p = 0.71), gaming experience (W = 1.32, p = 0.99), risk propensity (W = 0.97, p = 1.00), trust propensity (W = 5.19, p = 0.64), and utilitarianism (F(7, 64) = 0.53, p = 0.81).

B.1.2 Agent Behavior

We used the following survey to identify moral features as predictors of moral sensitivity:

1. **Situations** (two versions with varying feature values):

   (a) During the offensive inside deployment of Brutus, the team should decide whether to send in firefighters to rescue an injured victim or if that is too dangerous. Several guidelines exist for determining when conditions are safe enough for firefighters to enter. For example, the temperature should be below the auto-ignition temperatures of present substances, and the structural condition of the building must be good enough. To make a decision, the team can use the following information:

   - Estimated fire duration: 45/30 minutes
   - Distance between victim and fire source: Small/Large
   - Estimated fire resistance to collapse: 30 minutes
   - Temperature: Higher/Lower than auto-ignition temperatures of present substances

(b) During the offensive inside deployment of Brutus to locate the fire source, the team should decide whether to send in firefighters to help locate the fire source or if that is too dangerous. Several guidelines exist for determining when conditions are safe enough for firefighters to enter. For example, the temperature should be below the auto-ignition temperatures of present substances, and the structural condition of the building must be good enough. To make a decision, the team can use the following information:

- People in the building: 0/Unclear
- Estimated fire duration: 15 minutes
- Estimated fire resistance to collapse: 60/90 minutes
- Temperature: Lower than/Close to auto-ignition temperatures of present substances

(c) After Brutus explored the inside of the burning building, the team should decide whether Brutus should first extinguish the fire or evacuate people. General guidelines mention to first extinguish and then rescue. However, when the location of the fire source is unknown and smoke spreads fast, evacuating people first might be required. To make a decision, the team can use the following information:

- People in the building: 1/3
- Smoke spreading: Normally/Fast
- Estimated fire duration: 30/45 minutes
- Location of the fire source: Known/Unknown

(d) During the offensive inside deployment of Brutus, the team should decide whether Brutus should continue with this tactic or switch to a defensive inside deployment. The offensive inside deployment is used to fight fire and rescue people, whereas the defensive inside deployment is used to prevent the spread of fire, smoke, and damage to unaffected parts of the building. Several factors are important when deciding on an offensive inside deployment. For example, the chance of saving people and the building plays a role, which decreases with the fire duration. Moreover, it is important to know the fire source location. To make a decision, the team can use the following information:

- People in the building: 0/Unclear
- Estimated fire duration: 15/30 minutes
- Location of the fire source: Unknown
- Estimated fire resistance to collapse: 90/60 minutes

2. **Moral sensitivity rating**:

   This situation could be described as ... (0 = *not morally sensitive*, 6 = *extremely morally sensitive*).

3. **Alternative moral sensitivity rating** (open option to alter feature values from described situation):

On a scale from 0 to 6, you rated the moral sensitivity of this situation as *less than 2/greater than 4/between 2 and 4.* When would you have rated the situation's moral sensitivity as *greater than 4/less than 2*?

4. **Comfort** (-3 = *extremely uncomfortable*, +3 = *extremely comfortable*):

   How comfortable would you feel if Brutus made the decision in the described situation?

B.1.3 Measures

We used the following surveys for our user studies:

1. **Demographics**:
   - What gender do you identify as?
     - Female
     - Male
     - Other
     - Prefer not to say
   - What is your age?
     - 18 - 24 years old
     - 25 - 34 years old
     - 35 - 44 years old
     - 45 - 54 years old
     - 55 - 64 years old
     - 65+ years old
     - Prefer not to say
   - What is the highest level of education you have completed?
     - No schooling completed
     - Some high school, no diploma
     - High school graduate
     - Some college credit, no degree
     - Associate degree
     - Bachelor's degree
     - Master's degree
     - Ph.D. degree or higher
     - Prefer not to say
   - How much video gaming experience do you have?
     - None at all
     - A little
     - A moderate amount

  - A considerable amount
  - A lot

2. **Risk propensity** (1 = *totally disagree*, 9 = *totally agree*):

   - Safety first.
   - I do not take risks with my health.
   - I prefer to avoid risks.
   - I take risks regularly.
   - I really dislike not knowing what is going to happen.
   - I usually view risks as a challenge.
   - I view myself as a ... (1 = *risk avoider*, 9 = *risk seeker*).

3. **Trust propensity** (1 = *strongly disagree*, 5 = *strongly agree*):

   - I usually trust technology until there is a reason not to.
   - For the most part, I distrust technology.
   - In general, I would rely on technology to assist me.
   - My tendency to trust technology is high.
   - It is easy for me to trust technology to do its job.
   - I am likely to trust technology even when I have little knowledge about it.

4. **Utilitarianism** (1 = *strongly disagree*, 5 = *strongly agree*):

   - If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.
   - It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.
   - From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.
   - If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.
   - From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.
   - It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.
   - It is just as wrong to fail to help someone as it is to actively harm them yourself.

- Sometimes it is morally necessary for innocent people to die as collateral damage - if more people are saved overall.
- It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.

5. **Situation awareness**:
    - What is the current fire resistance to collapse?
        - 80 minutes/40 minutes
        - 90 minutes/**50 minutes**
        - **100 minutes**/60 minutes
        - 110 minutes/70 minutes
    - What is the total number of people to rescue?
        - 10/9
        - **11**/10
        - 12/11
        - 13/**unknown**
    - How many victims have been rescued so far?
        - 7/3
        - 8/4
        - 9/5
        - 10/6
    - Where is/was the fire source located?
        - office 04/office 01
        - office 06/office 03
        - **office 07**/office 11
        - office 09/**office 14**
    - Which feature is one of the features that determines if there should be extinguished or evacuated first?
        - fire resistance to collapse
        - number of victims
        - temperature
        - **speed of smoke spread**
    - Which feature is one of the features that determines if it is safe enough for fire fighters to enter the building and rescue a critically injured victim?
        - localization of the fire source
        - number of victims
        - speed of smoke spread
        - **temperature**

- Which feature is one of the features that determines if it is safe enough to send in fire fighters to help locate the fire source?
    - **temperature**
    - number of victims
    - speed of smoke spread
    - estimated time to locate the fire source
- Which feature is one of the features that determines if the offensive deployment is still the best tactic?
    - **fire resistance to collapse**
    - distance between victims and fire source
    - speed of smoke spread
    - localization of the fire source
- If three mildly injured victims are found in a burning office, the fire source is not located, and the smoke is spreading fast; what should you decide?
    - use a defensive deployment
    - extinguish first
    - use an offensive deployment
    - **evacuate first**
- If a critically injured victim is found, the temperature is close to the safety threshold, and the smoke is spreading fast; what should you decide?
    - extinguish first
    - **send in a fire fighter to rescue**
    - evacuate first
    - do not send in a fire fighter to rescue
- If the fire source has not been located yet, no fires have been extinguished, and the temperature is lower than the safety threshold; what should you decide?
    - extinguish first
    - **send in fire fighters to locate the fire source**
    - evacuate first
    - do not send in fire fighters to locate the fire source
- If the deployment tactic should be determined, the fire resistance to collapse is 70 minutes, and not all office have been explored; what should you decide?
    - extinguish first
    - use a defensive deployment
    - evacuate first
    - **use an offensive deployment**

6. **Situation awareness of the agents**:

    - Which victim did Brutus find in office 01?

  - **critically injured older woman**
  - mildly injured older man
  - mildly injured woman
  - mildly injured older woman
- Which victim did Titus find in office 03?
  - critically injured older woman
  - **mildly injured older man**
  - mildly injured man
  - mildly injured older woman
- Which victim did Brutus/Titus find in office 09?
  - critically injured older woman
  - **mildly injured older man**/mildly injured man
  - **mildly injured older woman**
  - critically injured man
- In which office did Brutus find a mildly injured man?
  - **office 05**
  - office 06
  - office 07
  - office 12
- In which office did Titus find a critically injured older woman?
  - office 09
  - **office 10**
  - office 13
  - office 14
- In which office did Brutus/Titus find a mildly injured older woman?
  - office 07/**office 01**
  - office 10/office 02
  - **office 13**/office 04
  - office 14/office 06
- When does Brutus allocate decision making to itself in the situation *extinguish or evacuate first*?
  - if fire source is located
  - it smoke is not spreading fast
  - if temperature is lower than threshold
  - **if predicted sensitivity is lower than threshold**
- When does Brutus allocate decision making to you in the situation *send in fire fighters to rescue*?
  - if smoke is spreading fast

  - if temperature is higher than threshold
  - **if predicted sensitivity is higher than threshold**
  - if distance between victim and fire source is small

- When does Titus allocate decision making to itself in the situation *send in fire fighters to help locate*?
  - if fire resistance is more than 100 minutes
  - if smoke is not spreading fast
  - **if predicted sensitivity is lower than threshold**
  - if distance between fire fighters and potential fire source is large
- When does Titus allocate decision making to you in the situation *continue or switch deployment tactic*?
  - if smoke is spreading fast
  - if temperature is higher than threshold
  - **if predicted sensitivity is higher than threshold**
  - if distance between victims and fire source is small
- Which action will Brutus/Titus perform/execute next?
  - move to an office
  - make a decision itself
  - allocate decision making to me
  - none of the listed answers

7. **Capacity and moral trust** (0 = *not at all*, 7 = *very*, or alternative option *does not fit*):

   - Reliable
   - Sincere
   - Capable
   - Ethical
   - Predictable
   - Genuine
   - Skilled
   - Respectable
   - Someone you can count on
   - Candid
   - Competent
   - Principled
   - Consistent
   - Authentic

- Meticulous
- Has integrity

8. **Experienced control** (1 = *I disagree strongly*, 5 = *I agree strongly*):
   - It was difficult to keep an overview of victims and situational features.
   - I experienced time pressure during decision making.
   - I felt responsible for the well-being of the victims and fire fighters.
   - I made decisions under inconclusive firefighting- and ethical guidelines.
   - I made decisions during the task that I would not want to make in real life.
   - I felt uncomfortable during (some) decisions I made.
   - I mostly made decisions for victims and firefighters that led to good and safe task outcomes.
9. **Agreement** (1 = *I disagree strongly*, 5 = *I agree strongly*):
   - I agreed with most of the decision allocations by Brutus/Titus.
   - I felt comfortable with most of the decision allocations by Brutus/Titus.
10. **Responsibility** (1 = not responsible at all, 7 = very responsible):
    - To what extent do you hold yourself morally responsible for bad task outcomes such as loss of victims and firefighter risk?
    - To what extent do you hold Brutus/Titus morally responsible for bad task outcomes such as loss of victims and firefighter risk?
11. **Agent difference and preference** (open questions):
    - Did you observe a difference between Brutus and Titus, and if yes, what difference?
    - if you had to chose between the Brutus and Titus in real life, which one would you pick and why?

B.2 Results

B.2.1 Counterbalancing and Completeness

We examined whether the three counterbalanced factors (agent-name pairs, task order, and agent order) influenced our measures. Since the data was not normally distributed, we ran Mann-Whitney U tests. We did not find statistically significant differences between the two task order conditions for capacity trust (W = 2750.5, p = 0.53), moral trust (W = 2831, p = 0.08), subjective agreement (W = 2782, p = 0.44), objective agreement (W = 2617.5, p = 0.92), or meaningful human control (W = 2635.5, p = 0.86). Moreover, we did not find statistically significant differences between the two agent-name pairs for capacity trust (W = 2830, p = 0.34), moral trust (W = 2133, p = 0.23), subjective agreement (W = 2557, p = 0.89), objective agreement (W = 2635.5, p = 0.86), or meaningful human control (W =

2274, p = 0.20). Finally, we did not find statistically significant differences between the two agent order conditions for capacity trust (W = 2553.5, p = 0.88), moral trust (W = 2283.5, p = 0.58), subjective agreement (W = 2723, p = 0.60), objective agreement (W = 2323.5, p = 0.28), or meaningful human control (W = 2462.5, p = 0.60).

Next, we explored whether agent explanation or autonomy affected task completeness, which might influence trust. Since the data was not normally distributed, we conducted a non-parametric rank-based mixed ANOVA. Results showed no statistically significant main effects of agent explanation (F(1.78) = 2.00, p = 0.14, effect size = 0.27) and autonomy (F(1.00) = 0.25, p = 0.62, effect size = 0.06), nor an interaction between them (F(1.83) = 1.08, p = 0.33, effect size = 0.17).

Bibliography

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

[2] David A Abbink, Tom Carlson, Mark Mulder, Joost CF De Winter, Farzad Aminravan, Tricia L Gibo, and Erwin R Boer. A topology of shared control systems—finding common ground in diversity. *IEEE Transactions on Human-Machine Systems*, 48(5):509–525, 2018.

[3] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

[4] Rafael Alvarado and Paul Humphreys. Big data, thick mediation, and representational opacity. *New Literary History*, 48(4):729–749, 2017.

[5] Dan Amir and Ofra Amir. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1168–1176, 2018.

[6] Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4):15–15, 2007.

[7] Michael Anderson and Susan Leigh Anderson. Geneth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1):337–357, 2018.

[8] Chittaranjan Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian journal of psychological medicine*, 40(5):498–499, 2018.

[9] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[10] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. Value alignment or misalignment–what will keep systems accountable? In *Workshops at the thirty-first AAAI conference on artificial intelligence*, 2017.

[11] Mohammad Q Azhar and Elizabeth I Sklar. A study measuring the impact of shared decision making in a human-robot team. *The International Journal of Robotics Research*, 36(5-7):461–482, 2017.

[12] Mirko Baglioni and Anahita Jamshidnejad. A novel mpc formulation for dynamic target tracking with increased area coverage for search-and-rescue robots. *Journal of Intelligent & Robotic Systems*, 110(4):140, 2024.

[13] David P Baker, Rachel Day, and Eduardo Salas. Teamwork as an essential component of high-reliability organizations. *Health services research*, 41(4p2):1576–1598, 2006.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82 – 115, 2020.

[15] Kevin Baum, Holger Hermanns, and Timo Speith. From machine ethics to machine explainability and back. In *ISAIM*, 2018.

[16] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1):12, 2022.

[17] Yael Ben Shalom. Turtlebot3 Navigation and SLAM, 10 2021.

[18] Michael W Boyce, Jessie YC Chen, Anthony R Selkowitz, and Shan G Lakhmani. Effects of agent transparency on operator trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 179–180, 2015.

[19] Jeffrey M Bradshaw, Paul J Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. Dimensions of adjustable autonomy and mixed-initiative interaction. In *Agents and Computational Autonomy: Potential, Risks, and Solutions 1*, pages 17–39. Springer, 2004.

[20] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[21] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4):589–597, 2019.

[22] Joost Broekens, Maaike Harbers, Koen Hindriks, Karel van den Bosch, Catholijn Jonker, and John-Jules Meyer. Do you get it? user-evaluated explainable bdi agents. In Jürgen Dix and Cees Witteveen, editors, *Multiagent System Technologies*, pages 28–39, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[23] John Brooke. Sus: a “quick and dirty’usability. *Usability evaluation in industry*, 189, 1996.

[24] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

[25] Simeon C Calvert, Daniël D Heikoop, Giulio Mecacci, and Bart Van Arem. A human centric framework for the analysis of automated driving systems based on meaningful human control. *TheoreTical issues in ergonomics science*, 21(4):478–506, 2020.

[26] Simeon C Calvert and Giulio Mecacci. A conceptual control system description of cooperative and automated driving in mixed urban traffic with meaningful human control for design and evaluation. *IEEE Open Journal of Intelligent Transportation Systems*, 1:147–158, 2020.

[27] Marc C Canellas and Rachel A Haga. Toward meaningful human control of autonomous weapons systems through function allocation. In *2015 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7. IEEE, 2015.

[28] Janis A Cannon-Bowers, Eduardo Salas, and Sharolyn Converse. Shared mental models in expert team decision making. *Individual and group decision making: Current issues*, 221:221–46, 1993.

[29] Cristiano Castelfranchi and Rino Falcone. From automaticity to autonomy: the frontier of artificial agents. *Agent autonomy*, pages 103–136, 2003.

[30] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M Jonker, et al. Meaningful human control: actionable properties for ai system development. *AI and Ethics*, 3(1):241–255, 2023.

[31] José-Antonio Cervantes, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, 26:501–532, 2020.

[32] Jessie U Chen, Michael J Barnes, and Zhihua Qu. Roboleader: A surrogate for enhancing the human control of a team of robots. Technical report, 2010.

[33] Jessie YC Chen and Michael J Barnes. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1):13–29, 2014.

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

[35] John P. Chin, Virginia A. Diehl, and Kent L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, page 213–218, New York, NY, USA, 1988. Association for Computing Machinery.

[36] Manolis Chiou, Nick Hawes, and Rustam Stolkin. Mixed-initiative variable autonomy for remotely operated mobile robots. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(4):1–34, 2021.

[37] Manolis Chiou, Rustam Stolkin, Goda Bieksaite, Nick Hawes, Kimron L Shapiro, and Timothy S Harrison. Experimental analysis of a variable autonomy framework for controlling a remotely operating mobile robot. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3581–3588. IEEE, 2016.

[38] Markus Christen, Thomas Burri, Serhiy Kandul, and Pascal Vörös. Who is controlling whom? reframing "meaningful human control" of ai systems in security. *Ethics and Information Technology*, 25(1):10, 2023.

[39] Markus Ed Christen, Carel Ed van Schaik, Johannes Ed Fischer, Marku Ed Huppenbauer, and Carmen Ed Tanner. *Empirically informed ethics: Morality between facts and norms.* Springer International Publishing AG, 2014.

[40] Hyesun Chung, Timothy Holder, Julie Shah, and X Jessie Yang. Developing a team classification scheme for human-agent teaming. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 1394–1399. SAGE Publications Sage CA: Los Angeles, CA, 2024.

[41] Giovanni Ciatto, Michael I. Schumacher, Andrea Omicini, and Davide Calvaresi. Agent-based explanations in ai: Towards an abstract framework. In Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling, editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 3–20, Cham, 2020. Springer International Publishing.

[42] Michael G Collins and Ion Juvina. Trust miscalibration is sometimes necessary: An empirical study and a computational model. *Frontiers in Psychology*, 12:690089, 2021.

[43] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[44] Nancy J Cooke, Mustafa Demir, and Nathan McNeese. Synthetic teammates as team players: Coordination of human and synthetic teammates. Technical report, Cognitive Engineering Research Institute Mesa United States, 2016.

[45] Nancy J Cooke, Eduardo Salas, Janis A Cannon-Bowers, and Renee J Stout. Measuring team knowledge. *Human factors*, 42(1):151–173, 2000.

[46] Rebecca Crootof. A meaningful floor for meaningful human control. *Temp. Int'l & Comp. LJ*, 30:53, 2016.

[47] Jovana Davidovic. On the purpose of meaningful human control of ai. *Frontiers in big data*, 5:1017677, 2023.

[48] Filippo Santoni de Sio, Giulio Mecacci, Simeon Calvert, Daniel Heikoop, Marjan Hagenzieker, and Bart van Arem. Realising meaningful human control over automated driving systems: a multidisciplinary approach. *Minds and machines*, pages 1–25, 2022.

[49] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331, 2016.

[50] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

[51] Davide Dell'Anna, Pradeep K Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M Jonker, Catharine Oertel, and Pınar Yolum. Toward a quality model for hybrid intelligence teams. In *23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024*, pages 434–443. ACM Press Digital Library, 2024.

[52] Jeffrey Delmerico, Stefano Mintchev, Alessandro Giusti, Boris Gromov, Kamilo Melo, Tomislav Horvat, Cesar Cadena, Marco Hutter, Auke Ijspeert, Dario Floreano, et al. The current state and future outlook of rescue robotics. *Journal of Field Robotics*, 36(7):1171–1191, 2019.

[53] Mustafa Demir, Nathan J McNeese, Craig Johnson, Jamie C Gorman, David Grimm, and Nancy J Cooke. Effective team interaction for adaptive training and situation awareness in human-autonomy teaming. In *2019 IEEE conference on cognitive and computational aspects of situation management (CogSIMA)*, pages 122–126. IEEE, 2019.

[54] Giuseppe Desolda, Andrea Esposito, Rosa Lanzilotti, Antonio Piccinno, and Maria F Costabile. From human-centered to symbiotic artificial intelligence: a focus on medical applications. *Multimedia Tools and Applications*, 84(27):32109–32150, 2025.

[55] Derek Doran, Sarah Schulz, and Tarek R. Besold. What does explainable AI really mean? A new conceptualization of perspectives. *CoRR*, abs/1710.00794, 2017.

[56] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[57] Nir Douer and Joachim Meyer. The responsibility quantification model of human interaction with automation. *IEEE Transactions on Automation Science and Engineering*, 17(2):1044–1060, 2020.

[58] Didier J Dubois, Henri Prade, and Ronald R Yager. *Readings in fuzzy sets for intelligent systems.* Morgan Kaufmann, 2014.

[59] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[60] Mica R Endsley. Situation awareness global assessment technique (sagat). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*, pages 789–795. IEEE, 1988.

[61] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human factors*, 37(1):32–64, 1995.

[62] Mica R Endsley. Direct measurement of situation awareness: Validity and use of sagat. In *Situational awareness*, pages 129–156. Routledge, 2017.

[63] Mica R. Endsley. A systematic review and meta-analysis of direct objective measures of situation awareness: A comparison of sagat and spam. *Human Factors*, 63(1):124–150, 2021. PMID: 31560575.

[64] Mica R Endsley and Esin O Kiris. The out-of-the-loop performance problem and level of control in automation. *Human factors*, 37(2):381–394, 1995.

[65] Connor Esterwood. Rethinking trust repair in human-robot interaction. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 432–436, 2023.

[66] Connor Esterwood and Lionel P Robert. A literature review of trust repair in hri. In *2022 31st IEEE international conference on robot and human interactive communication (ro-man)*, pages 1641–1646. IEEE, 2022.

[67] Connor Esterwood and Lionel P Robert Jr. Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human behavior*, 142:107658, 2023.

[68] Connor Esterwood and Lionel Peter Robert Jr. The warehouse robot interaction sim: An open-source hri research platform. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 268–271, 2023.

[69] Bernardo Esteves Henriques, Mirko Baglioni, and Anahita Jamshidnejad. Camera-based mapping in search-and-rescue via flying and ground robot teams. *Machine Vision and Applications*, 35(5):117, 2024.

[70] Andrea Ferrario and Michele Loi. How explainability contributes to trust in ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1457–1466, 2022.

[71] Alberto Finzi and Andrea Orlandini. A mixed-initiative approach to human-robot interaction in rescue scenarios. In *American Association for Artificial Intelligence*, 2005.

[72] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. The dynamic window approach to collision avoidance. *IEEE robotics & automation magazine*, 4(1):23–33, 2002.

[73] Joana Gouveia Freire and Carlos Castro DaCamara. Using cellular automata to simulate wildfire propagation and to assist in fire management. *Natural hazards and earth system sciences*, 19(1):169–179, 2019.

[74] Laurent Frering, Matthias Eder, Bettina Kubicek, Dietrich Albert, Denis Kalkofen, Thomas Gschwandtner, Heimo Krajnz, and Gerald Steinbauer-Wagner. Enabling and assessing trust when cooperating with robots in disaster response (easier). *arXiv preprint arXiv:2207.03763*, 2022.

[75] Batya Friedman and David G Hendry. *Value sensitive design: Shaping technology with moral imagination.* Mit Press, 2019.

[76] Alvin I Goldman et al. Theory of mind. *The Oxford handbook of philosophy of cognitive science*, 1, 2012.

[77] Matthew C Gombolay, Reymundo A Gutierrez, Shanelle G Clarke, Giancarlo F Sturla, and Julie A Shah. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, 39:293–312, 2015.

[78] Michael A Goodrich, Timothy W McLain, Jeffrey D Anderson, Jisang Sun, and Jacob W Crandall. Managing autonomy in robot teams: observations from four experiments. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 25–32, 2007.

[79] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[80] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2), 2017.

[81] Yaohui Guo and X Jessie Yang. Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 13(8):1899–1909, 2021.

[82] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

[83] Maaike Harbers, Jeffrey M Bradshaw, Matthew Johnson, Paul Feltovich, Karel van den Bosch, and John-Jules Meyer. Explanation and coordination in human-agent teams: a study in the bw4t testbed. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 17–20. IEEE, 2011.

[84] Maaike Harbers, Jeffrey M. Bradshaw, Matthew Johnson, Paul Feltovich, Karel van den Bosch, and John-Jules Meyer. Explanation in human-agent teamwork. In Stephen Cranefield, M. Birna van Riemsdijk, Javier Vázquez-Salceda, and Pablo

Noriega, editors, *Coordination, Organizations, Institutions, and Norms in Agent System VII*, pages 21–37, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[85] Maaike Harbers, Catholijn Jonker, and Birna Van Riemsdijk. Enhancing team performance through effective communication. In *Proceedings of the 4th Annual Human-Agent-Robot Teamwork Workshop*, pages 1–2, 2012.

[86] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. Design and evaluation of explainable bdi agents. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 125–132, 2010.

[87] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[88] Pim Haselager and Giulio Mecacci. Superethics instead of superintelligence: know thyself, and apply science accordingly. *AJOB neuroscience*, 11(2):113–119, 2020.

[89] Bradley Hayes and Julie A. Shah. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 303–312, 2017.

[90] Pamela Hieronymi. Xiv—reasons for action. In *Proceedings of the Aristotelian society*, volume 111, pages 407–427. Oxford University Press Oxford, UK, 2011.

[91] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019.

[92] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[93] Alexander Hong, O Igharoro, Yugang Liu, Farzad Niroui, Goldie Nejat, and Beno Benhabib. Investigating human-robot teams for learning-based semi-autonomous control in urban search and rescue environments. *Journal of Intelligent & Robotic Systems*, 94(3):669–686, 2019.

[94] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[95] Tjalling Haije Jasper van der Waa. Matrx: Human agent teaming rapid experimentation software, July 2023.

[96] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*, 2021.

[97] Matthew Johnson and Jeffrey M Bradshaw. The role of interdependence in trust. In *Trust in human-robot interaction*, pages 379–403. Elsevier, 2021.

[98] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, Birna van Riemsdijk, and Maarten Sierhuis. The fundamental principle of coactive design: Interdependence must shape autonomy. In Marina De Vos, Nicoletta Fornara, Jeremy V. Pitt, and George Vouros, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*, pages 172–191, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[99] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69, 2014.

[100] Matthew Johnson, Catholijn Jonker, Birna Van Riemsdijk, Paul J Feltovich, and Jeffrey M Bradshaw. Joint activity testbed: Blocks world for teams (bw4t). In *International Workshop on Engineering Societies in the Agents World*, pages 254–256. Springer, 2009.

[101] Matthew Johnson and Alonso Vera. No ai is an island: the case for teaming intelligence. *AI magazine*, 40(1):16–28, 2019.

[102] Carolina Centeio Jorge, Siddharth Mehrotra, Myrthe L Tielman, and Catholijn M Jonker. Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. In *22nd International Trust Workshop 2021*, 2021.

[103] Marten HL Kaas. The perfect technological storm: artificial intelligence and moral complacency. *Ethics and Information Technology*, 26(3):49, 2024.

[104] Guy Kahane, Jim AC Everett, Brian D Earp, Lucius Caviola, Nadira S Faber, Molly J Crockett, and Julian Savulescu. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological review*, 125(2):131, 2018.

[105] Peter H Kim, Donald L Ferrin, Cecily D Cooper, and Kurt T Dirks. Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology*, 89(1):104, 2004.

[106] Tae Wan Kim, Thomas Donaldson, and John Hooker. Grounding value alignment with ethical principles. *arXiv preprint arXiv:1907.05447*, 2019.

[107] Gary Klein, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman, and Paul J Feltovich. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, 2004.

[108] M Audrey Korsgaard, David M Schweiger, and Harry J Sapienza. Building commitment, attachment, and trust in strategic decision-making teams: The role of procedural justice. *Academy of Management journal*, 38(1):60–84, 1995.

[109] Esther S Kox, José H Kerstholt, Tom F Hueting, and Peter W de Vries. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous agents and multi-agent systems*, 35(2):30, 2021.

[110] Roderick M Kramer and Roy J Lewicki. Repairing and enhancing trust: Approaches to reducing organizational trust deficits. *Academy of Management annals*, 4(1):245–277, 2010.

[111] Geert-Jan M Kruijff, M Janíček, Shanker Keshavdas, Benoit Larochelle, Hendrik Zender, Nanja JJM Smets, Tina Mioch, Mark A Neerincx, JV Diggelen, Francis Colas, et al. Experience in system design for human-robot teaming in urban search and rescue. In *Field and Service Robotics: Results of the 8th International Conference*, pages 111–125. Springer, 2013.

[112] Geert-Jan M Kruijff, Ivana Kruijff-Korbayová, Shanker Keshavdas, Benoit Larochelle, Miroslav Janíček, Francis Colas, Ming Liu, Francois Pomerleau, Roland Siegwart, Mark A Neerincx, et al. Designing, developing, and deploying systems to support human–robot teams in disaster response. *Advanced Robotics*, 28(23):1547–1570, 2014.

[113] Ivana Kruijff-Korbayová, Francis Colas, Mario Gianni, Fiora Pirri, Joachim de Greeff, Koen Hindriks, Mark Neerincx, Petter Ögren, Tomáš Svoboda, and Rainer Worst. Tradr project: Long-term human-robot teaming for robot assisted disaster response. *KI-Künstliche Intelligenz*, 29:193–201, 2015.

[114] Neerendra Kumar, Márta Takács, and Zoltán Vámossy. Robot navigation in unknown environment using fuzzy logic. In *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000279–000284. IEEE, 2017.

[115] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4762–4763, 2017.

[116] Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. The interplay between emotional intelligence, trust, and gender in human–robot interaction: A vignette-based study. *International Journal of Social Robotics*, 13(2):297–309, 2021.

[117] Christian Lebiere, Leslie M Blaha, Corey K Fallon, and Brett Jefferson. Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Frontiers in Robotics and AI*, 8:652776, 2021.

[118] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[119] Wen-Ying Lee, Mose Sakashita, Elizabeth Ricci, Houston Claure, François Guimbretière, and Malte Jung. Interactive vignettes: Enabling large-scale interactive hri research. In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)*, pages 1289–1296. IEEE, 2021.

[120] Kristina Lerman, Chris Jones, Aram Galstyan, and Maja J Matarić. Analysis of dynamic task allocation in multi-robot systems. *The International Journal of Robotics Research*, 25(3):225–241, 2006.

[121] Roy J Lewicki and Chad Brinsfield. Trust repair. *Annual review of organizational psychology and organizational behavior*, 4(1):287–313, 2017.

[122] Fei-Fei Li. How to make ai that's good for people. *The New York Times*, 7, 2018.

[123] Sirui Li, Weixing Sun, and Tim Miller. Communication in human-agent teams for tasks with joint action. In Virginia Dignum, Pablo Noriega, Murat Sensoy, and Jaime Simão Sichman, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems XI*, pages 224–241, Cham, 2016. Springer International Publishing.

[124] Nan Liang and Goldie Nejat. A meta-analysis on remote hri and in-person hri: What is a socially assistive robot to do? *Sensors*, 22(19):7155, 2022.

[125] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. The conflict between explainable and accountable decision-making algorithms. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2103–2113, 2022.

[126] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[127] Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. Axies: Identifying and evaluating context-specific values. In *AAMAS*, pages 799–808, 2021.

[128] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, Robert Christopher Garrett, John Hoare, and Michael Kopack. Explaining robot actions. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, page 187–188, New York, NY, USA, 2012. Association for Computing Machinery.

[129] Rosemarijn Looije, Anna van der Zalm, Mark A. Neerincx, and Robbert-Jan Beun. Help, i need some body the effect of embodiment on playful learning. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 718–724, 2012.

[130] Ruikun Luo, Na Du, and X Jessie Yang. Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human–Computer Interaction*, 38(18-20):1962–1971, 2022.

[131] Joseph B Lyons, Katia Sycara, Michael Lewis, and August Capiola. Human–autonomy teaming: Definitions, debates, and directions. *Frontiers in psychology*, 12:589585, 2021.

[132] Bertram F Malle. How the mind explains behavior. *Folk explanation, Meaning and social interaction. Massachusetts: MIT-Press*, 2006.

[133] Bertram F Malle. Attribution theories: How people make sense of behavior. *Theories in social psychology*, 23:72–95, 2022.

[134] Bertram F Malle and Daniel Ullman. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*, pages 3–25. Elsevier, 2021.

[135] Martina Mara, Jan-Philipp Stein, Marc Erich Latoschik, Birgit Lugrin, Constanze Schreiner, Rafael Hostettler, and Markus Appel. User responses to a humanoid robot observed in real life, virtual reality, 3d and 2d. *Frontiers in psychology*, 12:633178, 2021.

[136] Julie L Marble, David J Bruemmer, and Douglas A Few. Lessons learned from usability tests with a collaborative cognitive workspace for human-robot teams. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, volume 1, pages 448–453. IEEE, 2003.

[137] A John Maule and Anne C Edland. The effects of time pressure on human judgement and decision making. In *Decision making*, pages 203–218. Routledge, 2002.

[138] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

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

[140] Ree M Meertens and Rene Lion. Measuring an individual's tendency to take risks: the risk propensity scale 1. *Journal of applied social psychology*, 38(6):1506–1520, 2008.

[141] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4):1–45, 2024.

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

[143] Stephanie M Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors*, 55(3):520–534, 2013.

[144] Leila Methnani, Andrea Aler Tubella, Virginia Dignum, and Andreas Theodorou. Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, 4:737072, 2021.

[145] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[146] Inge Molenaar. The concept of hybrid human-ai regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence*, 3:100070, 2022.

[147] James H Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4):18–21, 2006.

[148] Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, and Christophe Nicolle. The quest of parsimonious xai: A human-agent architecture for explanation formulation. *Artificial Intelligence*, 302:103573, 2022.

[149] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *Engineering Psychology and Cognitive Ergonomics: 15th International Conference, EPCE 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 15*, pages 204–214. Springer, 2018.

[150] Mark A Neerincx, Willeke Van Vught, Olivier Blanson Henkemans, Elettra Oleari, Joost Broekens, Rifca Peters, Frank Kaptein, Yiannis Demiris, Bernd Kiefer, Diego Fumagalli, et al. Socio-cognitive engineering of a robotic partner for child's diabetes self-management. *Frontiers in Robotics and AI*, 6:118, 2019.

[151] Kimihiro Noguchi, Yulia Gel, Edgar Brunner, and Frank Konietschke. nparld: An r software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software*, 50, 09 2012.

[152] Merel Noorman. Computing and moral responsibility. 2012.

[153] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[154] Sven Nyholm. Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, 24(4):1201–1219, 2018.

[155] Hajer Omrane, Mohamed Slim Masmoudi, and Mohamed Masmoudi. Fuzzy logic based control for autonomous mobile robot navigation. *Computational intelligence and neuroscience*, 2016(1):9548482, 2016.

[156] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G Jentsch. Building appropriate trust in human-robot teams. In *AAAI spring symposium: trust and autonomous systems*, pages 60–65, 2013.

[157] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, 64(5):904–938, 2022.

[158] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[159] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.

[160] Russell Perkins, Zahra Rezaei Khavas, Kalvin McCallum, Monish Reddy Kotturu, and Paul Robinette. The reason for an apology matters for robot trust repair. In *International Conference on Social Robotics*, pages 640–651. Springer, 2022.

[161] James B Rawlings, David Mayne, and Moritz Diehl. *Model predictive control: theory, computation, and design*, volume 2. Nob Hill Publishing Madison, WI, 2017.

[162] James R Rest. Background: Theory and research. In *Moral development in the professions*, pages 13–38. Psychology Press, 1994.

[163] Scott J Reynolds. Moral awareness and ethical predispositions: investigating the role of individual differences in the recognition of moral issues. *Journal of Applied Psychology*, 91(1):233, 2006.

[164] Scott J Reynolds and Jared A Miller. The recognition of moral issues: Moral awareness, moral sensitivity and moral attentiveness. *Current Opinion in Psychology*, 6:114–117, 2015.

[165] Chadi F Riman and Pierre E Abi-Char. Fuzzy logic control for mobile robot navigation in automated storage. *Int J Mech Eng Robot Res*, 12:313–323, 2023.

[166] Lionel Robert, Gaurav Bansal, and Christoph Lutge. Icis 2019 sighci workshop panel report: Human computer interaction challenges and opportunities for fair, trustworthy and ethical artificial intelligence. 2020.

[167] Paul Robinette, Ayanna M Howard, and Alan R Wagner. Timing is key for robot trust repair. In *International conference on social robotics*, pages 574–583. Springer, 2015.

[168] Nicole Robinson, Jason Williams, David Howard, Brendan Tidd, Fletcher Talbot, Brett Wood, Alex Pitt, Navinda Kottege, and Dana Kulić. Human-robot team performance compared to full robot autonomy in 16 real-world search and rescue missions: Adaptation of the darpa subterranean challenge. *ACM Transactions on Human-Robot Interaction*, 14(1):1–30, 2024.

[169] Enrico Ronchi and Daniel Nilsson. Fire evacuation in high-rise buildings: a review of human behaviour and modelling research. *Fire science reviews*, 2:1–21, 2013.

[170] Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, 2019.

[171] Christoph Rösmann, Frank Hoffmann, and Torsten Bertram. Integrated online trajectory planning and optimization in distinctive topologies. *Robotics and Autonomous Systems*, 88:142–153, 2017.

[172] Fernando Rudy-Hiller. The epistemic condition for moral responsibility. 2018.

[173] Richard Saavedra, P Christopher Earley, and Linn Van Dyne. Complex interdependence in task-performing groups. *Journal of applied psychology*, 78(1):61, 1993.

[174] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. Explainable goal-driven agents and robots-a comprehensive review. *ACM Computing Surveys*, 55(10):1–41, 2023.

[175] Eduardo Salas, Dana E Sims, and C Shawn Burke. Is there a "big five" in teamwork? *Small group research*, 36(5):555–599, 2005.

[176] Lindsay Sanneman and Julie A. Shah. A situation awareness-based framework for design and evaluation of explainable ai. In Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling, editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 94–110, Cham, 2020. Springer International Publishing.

[177] Lindsay Sanneman and Julie A Shah. The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems. *International Journal of Human–Computer Interaction*, pages 1–17, 2022.

[178] Filippo Santoni de Sio and Giulio Mecacci. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & technology*, 34(4):1057–1084, 2021.

[179] Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:323836, 2018.

[180] Maarten PD Schadd, Tjeerd AJ Schoonderwoerd, Karel van den Bosch, Olaf H Visker, and Tjalling Haije. "i'm afraid i can't do that, dave"; getting to know your buddies in a human–agent team. *Systems*, 10(1):15, 2022.

[181] Kristin E Schaefer, Edward R Straub, Jessie YC Chen, Joe Putney, and Arthur W Evans III. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, 46:26–39, 2017.

[182] John Schaubroeck, Simon SK Lam, and Ann Chunyan Peng. Cognition-based and affect-based trust as mediators of leader behavior influences on team performance. *Journal of applied psychology*, 96(4):863, 2011.

[183] Tjeerd AJ Schoonderwoerd, Emma M van Zoelen, Karel van den Bosch, and Mark A Neerincx. Design patterns for human-ai co-learning: A wizard-of-oz evaluation in an urban-search-and-rescue task. *International Journal of Human-Computer Studies*, 164:102831, 2022.

[184] Ernestina JA Schreuder and Tina Mioch. The effect of time pressure and task completion on the occurrence of cognitive lockup. In *Proceedings of the International Workshop on Human Centered Processes*, volume 2011, pages 10–11. Citeseer, 2011.

[185] Maurice E Schweitzer, John C Hershey, and Eric T Bradlow. Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, 101(1):1–19, 2006.

[186] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. “i don’t believe you”: Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 57–65. IEEE, 2019.

[187] Anthony R Selkowitz, Shan G Lakhmani, and Jessie YC Chen. Using agent transparency to support situation awareness of the autonomous squad member. *Cognitive Systems Research*, 46:13–25, 2017.

[188] Francesco Semeraro, Jon Carberry, James Leadbetter, and Angelo Cangelosi. Good things come in threes: The impact of robot responsiveness on workload and trust in multi-user human-robot collaboration. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2471–2478. IEEE, 2024.

[189] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[190] Ben Shneiderman. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124, 2020.

[191] Ngangbam Herojit Singh and Khelchandra Thongam. Mobile robot navigation using fuzzy logic in static environments. *Procedia Computer Science*, 125:11–17, 2018.

[192] Ronal Singh, Liz Sonenberg, and Tim Miller. Communication and shared mental models for teams performing interdependent tasks. In Stephen Cranefield, Samhar Mahmoud, Julian Padget, and Ana Paula Rocha, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems XII*, pages 81–97, Cham, 2017. Springer International Publishing.

[193] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. Hierarchical expertise level modeling for user specific contrastive explanations. In *IJCAI*, pages 4829–4836, 2018.

[194] Marc Steen, Jurriaan van Diggelen, Tjerk Timan, and Nanda van der Stap. Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives. *AI and Ethics*, 3(1):281–293, 2023.

[195] Ashley Suh, Isabelle Hurley, Nora Smith, and Ho Chit Siu. Fewer than 1% of explainable ai papers validate explainability with humans. *arXiv preprint arXiv:2503.16507*, 2025.

[196] Filip Surma and Anahita Jamshidnejad. Fuzzy-logic-based model predictive control: A paradigm integrating optimal and common-sense decision making. *arXiv preprint arXiv:2503.21065*, 2025.

[197] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th international conference on intelligent user interfaces*, pages 109–119, 2021.

[198] Hamed Taherdoost. What are different research approaches? comprehensive review of qualitative, quantitative, and mixed method research, their applications, types, and limitations. *Journal of Management Science & Engineering Research*, 5(1):53–63, 2022.

[199] Richard M Taylor. Situational awareness rating technique (sart): The development of a tool for aircrew systems design. In *Situational awareness*, pages 111–128. Routledge, 2017.

[200] Ilaria Tiddi, Victor De Boer, Stefan Schlobach, and André Meyer-Vitali. Knowledge engineering for hybrid intelligence. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 75–82, 2023.

[201] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. Capable but amoral? comparing ai and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[202] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1(4), 2020.

[203] Cristen Torrey, Susan R Fussell, and Sara Kiesler. How a robot should give advice. In *2013 8th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 275–282. IEEE, 2013.

[204] Nathan Tsoi, Rachel Sterneck, Xuan Zhao, and Marynel Vázquez. Influence of simulation and interactivity on human perceptions of a robot during navigation tasks. *ACM Transactions on Human-Robot Interaction*, 13(4):1–19, 2024.

[205] Matteo Turilli and Luciano Floridi. The ethics of information transparency. *Ethics and Information Technology*, 11(2):105–112, 2009.

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

[207] Ibo Van de Poel. The problem of many hands. In *Moral responsibility and the problem of many hands*, pages 50–92. Routledge, 2015.

[208] Karel Van Den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. Six challenges for human-ai co-learning. In *International Conference on Human-Computer Interaction*, pages 572–589. Springer, 2019.

[209] Michiel Van Der Meer, Enrico Liscio, Catholijn M Jonker, Aske Plaat, Piek Vossen, and Pradeep K Murukannaiah. Hyena: A hybrid method for extracting arguments from opinions. In *HHAI2022: Augmenting Human Intellect*, pages 17–31. IOS Press, 2022.

[210] Jasper van der Waa. Explainable artificial intelligence for human-ai collaboration. 2022.

[211] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial intelligence*, 291:103404, 2021.

[212] Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, and Mark Neerincx. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, 144:102493, 2020.

[213] Jasper van der Waa, Jurriaan van Diggelen, Luciano Cavalcante Siebert, Mark Neerincx, and Catholijn Jonker. Allocation of moral decision-making in human-agent teams: A pattern approach. In *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pages 203–220. Springer, 2020.

[214] Jasper Van Der Waa, Sabine Verdult, Karel Van Den Bosch, Jurriaan Van Diggelen, Tjalling Haije, Birgit Van Der Stigchel, and Ioana Cocu. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI*, 8:640647, 2021.

[215] Jurriaan van Diggelen, Jonathan Barnhoorn, Ruben Post, Joris Sijs, Nanda van der Stap, and Jasper van der Waa. Delegation in human-machine teaming: progress, challenges and prospects. In *Intelligent Human Systems Integration 2021: Proceedings of the 4th International Conference on Intelligent Human Systems Integration (IHSI 2021): Integrating People and Intelligent Systems, February 22-24, 2021, Palermo, Italy*, pages 10–16. Springer, 2021.

[216] Jurriaan van Diggelen, JS Barnhoorn, Marieke MM Peeters, Wessel van Staal, ML Stolk, Bob van der Vecht, Jasper van der Waa, and Jan Maarten Schraagen. Pluggable social artificial intelligence for enabling human-agent teaming. *arXiv preprint arXiv:1909.04492*, 2019.

[217] Jurriaan van Diggelen and Matthew Johnson. Team design patterns. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 118–126, 2019.

[218] Jurriaan Van Diggelen, Mark Neerincx, Marieke Peeters, and Jan Maarten Schraagen. Developing effective and resilient human-agent teamwork using team design patterns. *IEEE intelligent systems*, 34(2):15–24, 2018.

[219] Jurriaan van Diggelen, Karel van den Bosch, Mark Neerincx, and Marc Steen. Designing for meaningful human control in military human-machine teams. In *Research*

handbook on meaningful human control of artificial intelligence systems, pages 232–252. Edward Elgar Publishing, 2024.

[220] Bart van Leeuwen, Richard Gasaway, Gerke Spaling, and BV Netage. Adopting ai to support situational awareness in emergency response: a reflection by professionals. In *Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence. Frontiers in Artificial Intelligence and Applications*, 2022.

[221] Aimee Van Wynsberghe and Scott Robbins. Critiquing the reasons for making artificial moral agents. *Science and engineering ethics*, 25:719–735, 2019.

[222] Emma Van Zoelen, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S Gouvêa, Kim Baraka, Maaike HT De Boer, and Mark A Neerincx. Developing team design patterns for hybrid intelligence systems. In *HHAI 2023: Augmenting Human Intellect*, pages 3–16. IOS Press, 2023.

[223] Emma M Van Zoelen, Karel Van Den Bosch, and Mark Neerincx. Becoming team members: identifying interaction patterns of mutual adaptation for human-robot co-learning. *Frontiers in Robotics and AI*, 8, 2021.

[224] Emma M Van Zoelen, Hugo Veldman-Loopik, Karel van den Bosch, Mark Neerincx, David A Abbink, and Luka Peternel. Enabling embodied human-robot co-learning: Requirements, method, and test with handover task. *IEEE Robotics and Automation Letters*, 2024.

[225] Herman Veluwenkamp. Reasons for meaningful human control. *Ethics and Information Technology*, 24(4):51, 2022.

[226] Ruben Verhagen. GitHub: Explainable AI for Meaningful Human Control, 11 2024. https://github.com/rsverhagen94/XAI4MHC.

[227] Ruben Verhagen. Preregistration: Explainable ai for meaningful human control, June 2024. https://doi.org/10.17605/OSF.IO/KS9BZ.

[228] Ruben Verhagen. Preregistration: Human-robot teamwork: Teleoperation vs. collaboration, August 2025. https://doi.org/10.17605/OSF.IO/9UAC2.

[229] Ruben S Verhagen, Alexandra Marcu, Mark A Neerincx, and Myrthe L Tielman. The influence of interdependence on trust calibration in human-machine teams. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 300–314. IOS Press, 2024.

[230] Ruben S Verhagen, Mark A Neerincx, Can Parlar, Marin Vogel, and Myrthe L Tielman. Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance. In *AAMAS*, pages 2316–2318, 2023.

[231] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable. In Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling, editors, *Explainable and Transparent AI and Multi-Agent Systems*, pages 119–138, Cham, 2021. Springer International Publishing.

[232] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI*, 9:993997, 2022.

[233] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. Meaningful human control and variable autonomy in human-robot teams for firefighting. *Frontiers in Robotics and AI*, 11:1323980, 2024.

[234] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. Agent allocation of moral decisions in human-agent teams: Raise human involvement and explain potential consequences. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2302–2317, 2025.

[235] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[236] Wendell Wallach, Colin Allen, and Iva Smit. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. In *Machine ethics and robot ethics*, pages 249–266. Routledge, 2020.

[237] James C Walliser, Ewart J de Visser, Eva Wiese, and Tyler H Shaw. Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, 13(4):258–278, 2019.

[238] James C Walliser, Patrick R Mead, and Tyler H Shaw. The perception of teamwork with an autonomous agent enhances affect and performance outcomes. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 61, pages 231–235. SAGE Publications Sage CA: Los Angeles, CA, 2017.

[239] Joel Walmsley. Artificial intelligence and the value of transparency. *AI & SOCIETY*, pages 1–11, 2020.

[240] Jijun Wang and Michael Lewis. Assessing coordination overhead in control of robot teams. In *2007 IEEE International Conference on Systems, Man and Cybernetics*, pages 2645–2649. IEEE, 2007.

[241] Jijun Wang and Michael Lewis. Human control for cooperating robot teams. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 9–16, 2007.

[242] Jijun Wang and Michael Lewis. Assessing cooperation in human control of heterogeneous robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 9–16, 2008.

[243] Adam Waytz, Joy Heafner, and Nicholas Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, 52:113–117, 2014.

[244] Changyun Wei, Koen V Hindriks, and Catholijn M Jonker. The role of communication in coordination protocols for cooperative robot teams. In *ICAART (2)*, pages 28–39, 2014.

[245] Thomas D Wickens. *Elementary signal detection theory.* Oxford university press, 2001.

[246] Jennifer Wieselquist, Caryl E Rusbult, Craig A Foster, and Christopher R Agnew. Commitment, pro-relationship behavior, and trust in close relationships. *Journal of personality and social psychology*, 77(5):942, 1999.

[247] Michael T Wolf, Christopher Assad, Matthew T Vernacchia, Joshua Fromm, and Henna L Jethani. Gesture-based robot control with variable autonomy from the jpl biosleeve. In *2013 IEEE International Conference on Robotics and Automation*, pages 1160–1165. IEEE, 2013.

[248] Jason H Wong, Erin K Chiou, Robert S Gutzwiller, Maia B Cook, and Corey K Fallon. Human-artificial intelligence teaming for the us navy: Developing a holistic research roadmap. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 380–385. SAGE Publications Sage CA: Los Angeles, CA, 2024.

[249] Sarah Woods, Michael Walters, Kheng Lee Koay, and Kerstin Dautenhahn. Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *9th IEEE International Workshop on Advanced Motion Control, 2006.*, pages 750–755. IEEE, 2006.

[250] Julia L Wright, Jessie YC Chen, Michael J Barnes, and Peter A Hancock. Agent reasoning transparency's effect on operator workload. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 60, pages 249–253. SAGE Publications Sage CA: Los Angeles, CA, 2016.

[251] Julia L. Wright, Jessie Y.C. Chen, Michael J. Barnes, and Peter A. Hancock. The effect of agent reasoning transparency on complacent behavior: An analysis of eye movements and response performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1):1594–1598, 2017.

[252] Kevin T Wynne and Joseph B Lyons. An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, 19(3):353–374, 2018.

[253] Jin Xu and Ayanna Howard. Evaluating the impact of emotional apology on human-robot trust. In *2022 31st IEEE international conference on robot and human interactive communication (ro-man)*, pages 1655–1661. IEEE, 2022.

[254] Sangseok You and Lionel Robert. Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction. In *You, S. and Robert, LP (2019). Trusting Robots in Teams: Examining the Impacts of Trusting Robots on Team Performance and Satisfaction, Proceedings of the 52th Hawaii International Conference on System Sciences, Jan*, pages 8–11, 2018.

[255] Akbar Zaheer, Bill McEvily, and Vincenzo Perrone. Does trust matter? exploring the effects of interorganizational and interpersonal trust on performance. *Organization science*, 9(2):141–159, 1998.

[256] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–28, 2022.

[257] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020.

[258] Patrick Zschech, Jannis Walk, Kai Heinrich, Michael Vössing, and Niklas Kühl. A picture is worth a collaboration: Accumulating design knowledge for computer-vision-based hybrid intelligence systems. *arXiv preprint arXiv:2104.11600*, 2021.

SIKS Dissertations

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UvA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web

19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UvA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TU/e), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VUA), Logics for causal inference under uncertainty

23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VUA), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TU/e), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TU/e), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech

30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data

06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs
38 Akos Kadar (OU), Learning visually grounded and multilingual representations

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TU/e), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OU), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst

31 Gongjin Lan (VUA), Learning better – From Baby to Better

32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (UL), Robust rules for prediction and description

24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

---

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases

13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms

15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining

16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment

07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning

08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning

09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing

11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications

12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries

13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation

14 Selma Čaušević (TUD), Energy resilience through self-organization

15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models

16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight

18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation

19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals

20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning

21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain

22 Alireza Shojaifar (UU), Volitional Cybersecurity

23 Theo Theunissen (UU), Documentation in Continuous Software Development

24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning

25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs

26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour

27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions

28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts

29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

---

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education

02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis

04 Mike Huisman (UL), Understanding Deep Meta-Learning

05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method

35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach

36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs

37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models

38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings

39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance

40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models

41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines

42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis

43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

44 Marcel Schmitz (OU), Towards learning analytics-supported learning design

45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology

47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs

48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

---

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems

02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks

03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI

04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data

05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation

06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization

07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance

08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction

09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach

10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning

11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration

12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval

13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics

14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies

15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence

16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging

17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems

18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing

19 Fang Hou (UU), Trust in Software Ecosystems

20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)

21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data

22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making

23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph

24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing

25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions

26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress

27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback

28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning

29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups

30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation

31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline

32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms

33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction

34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering

35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment

36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing

37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval

38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity

39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence

40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans

41 Mozhgan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights

42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data

43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education

44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning

45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings

46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making

47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions

48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training

49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn

50 Lea Krause (VUA), Contextualising Conversational AI

51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models

52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks

53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems

54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities

55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems

56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law

57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels

58 Junhan Wen (TUD), "From iMage to Market": Machine-Learning-Empowered Fruit Supply

59 Mohsen Abbaspour Onari (TU/e), From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support

60 Marcel Jurriaan Robeer (UU), Beyond Trust: A Causal Approach to Explainable AI in Law Enforcement

61 Shuai Wang (VUA), Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications

62 Khaleel Asyraaf Mat Sanusi (OU), Augmenting a learning model within immersive learning environments for psychomotor skills

63 Rashid Zaman (TU/e), Online Conformance Checking on Degraded Data

64 Jens d'Hondt (TU/e), Effective and Efficient Multivariate Similarity Search

65 Aswin Balasubramaniam (UT), Disentangling Runner Drone Interaction Potentialities

---

2026 01 Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support

02 Hezha Hassan Mohammedkhan (TiU), Estimating Body Measurements of Children from 2D Images: Towards the Automatic Detection of Malnutrition

03 Kyriakos Psarakis (TUD), Democratizing Scalable Cloud Applications: Transactional Stateful Functions on Streaming Dataflows

04 Boyu Xu (UU), Exploring Indirect Relations Between Topics in Neuroscience Literature Using Augmented Reality to Inform Experimental Design

05 Koen Minartz (TU/e), Stochastic Simulation with Geometric Deep Generative Models

06 Azim Afroozeh (CWI, VUA), FastLanes: A Next-Gen File Format

07 Inès Blin (VUA), Narrative Understanding with Knowledge Graphs

08 Paul van Vulpen (UU), Debating Digital Dominance: Decentralized Technology Governance For Strategic Autonomy

09 Afrizal Doewes (TU/e), Rethinking Automated Essay Scoring: Agreement, Fairness, and Feedback

10 Nikolaos Delapaschos Kondylidis (VUA), Establishing Task-Oriented Understanding between Agents

11 Işıl Baysal Erez (UT), Handling Missing Data with Meta-Learning and Large Language Models

12 Xue Li (UvA), From Fine-tuning to Prompting: A Paradigm Shift in Knowledge Graph Construction

13 Isaac da Silva Torres (VUA), Guidelines To Flux Between Conceptual Models: Understanding Complex Digital Business Ecosystems

14 Philip Lippmann (TUD), Synthetic Data for Robust Language Modelling

15 Rashmi Khazanchi (OU), Artificial Intelligence in Education: Impact of AI-Based Systems on Mathematics Achievement

16 Carolina Ferreira Gomes Centeio Jorge (TUD), Modelling Artificial Trust for Effective Human-AI Teamwork

17 Maria Tsfasman (TUD), Towards Predicting Memory in Multimodal Group Interactions

18 Riccardo Lo Bianco (TU/e), Deep Reinforcement Learning for Automated Decision-Making in Process Management Systems

19 Israel Campero Jurado (TU/e), Innovations in Optimization and Applications in Healthcare

20 Iftitahu Ni'mah (TU/e), Contrastive Learning and Evaluation in Low Resource Scenario of Natural Language Processing

21 Francisco N.F.Q. Simoes (UU), Causality, Information, and Decision-Making

22 Ruben Verhagen (TUD), Transparent and Explainable Agents for Human-Agent Teaming

Acknowledgments

En zo komt er een einde aan vijf jaar onderzoek. Begonnen in een vreemde tijd met lockdowns en mondkapjes. Gekenmerkt door online meetings en virtuele kennismakingen. Niet helemaal de gewenste start van zo'n belangrijk nieuw levenshoofdstuk. Het begin van dat hoofdstuk lijkt alweer een eeuwigheid geleden, al voelt het tegelijkertijd ook alsof deze jaren voorbij zijn gevlogen.

Dit hoofdstuk begon allemaal na het afronden van een prachtige studie neurowetenschappen in de buurt van het Gardameer. Na enkel online sollicitaties kon ik beginnen met onderzoek naar transparante en traceerbare kunstmatige intelligentie voor samenwerking met mensen. Veel meer details kreeg ik niet: geen concrete context, hardware, software of eindgebruikers.

Gelukkig had ik fantastische supervisors die mij wegwijs maakten in dit nieuwe onderzoeksveld. Myrthe en Mark, ontzettend bedankt voor jullie begeleiding de afgelopen jaren. Het zal voor jullie ook niet makkelijk geweest zijn om volledig online van start te gaan. Ik kan me nog goed herinneren dat we na maanden pas ons eerste overleg op de TU Delft hadden. Dit was voor mij wel een kantelpunt. Deze maandelijkse meetings op de universiteit waren zoveel productiever en creatiever dan de online varianten. Tijdens deze meetings en het gehele promotietraject hebben jullie mijn passie voor toegepast onderzoek naar mens-machinesamenwerking steeds verder aangewakkerd.

Myrthe, jij wakkerde vooral de passie aan voor gebruikersstudies en alles dat daar bij komt kijken: van experiment design tot statistische toetsen. Uiteindelijk kon ik zelfs de HREC applicaties waarderen. Jouw dagelijkse begeleiding was als een warm bad. Jouw creativiteit en oog voor detail hebben mij gevormd als onderzoeker en zijn een belangrijke reden voor onze vele nauwkeurig voorbereide en uitgevoerde experimenten.

Mark, jij wakkerde vooral de passie aan voor robots als teamgenoot van mensen. Ik ben blij dat ons laatste experiment na vier jaar dan tóch echt een robot bevatte. Hadden we dat maar eerder gedaan! Jouw maandelijkse input was altijd een zeer aangename frisse blik op mijn ideeën en voortgang. Regelmatig resulteerde dit in het verleggen van accenten en uiteindelijk interessantere studies.

Jullie hebben deze passies dusdanig aangewakkerd dat ik nu eigenlijk dezelfde onderzoekersrol vervul bij het Koninklijk Nederlands Lucht- en Ruimtevaartcentrum. Bedankt daarvoor! Bedankt ook dat jullie altijd verder keken dan enkel werk, door regelmatig mij eraan te herinneren om vrij te nemen of geaccepteerde papers te vieren, en door te controleren hoe het met mij als persoon ging. Ik heb onze samenwerking vanaf het begin tot eind als super waardevol ervaren en had me geen betere begeleiding kunnen wensen. Duizendmaal dank: zonder jullie zou dit proefschrift nooit tot stand zijn gekomen.

Next, I would like to continue with the members of my doctoral committee: David Abbink, Joseph Lyons, Judith Masthoff, and Filippo Santoni de Sio. Thank you for taking the time to read and assess my work. The four of you make up a great combination of disciplines and expertise, closely matching the interdisciplinary nature of this dissertation.

Before starting this PhD adventure, several people already sparked that fire for doing research. Especially the wonderful people and researchers at the NeuroInformatics Laboratory in Mattarello. I joined the lab during my Master's in Italy to complete my thesis, and stayed a bit longer to work as a research assistant. Paolo Avesani, Emanuele Olivetti, and Pietro Astolfi: a huge thank you for showing me how beautiful research can be and for the many things you taught me on neuroscience and machine learning!

Moving on to my dear TU Delft colleagues at the Interactive Intelligence group. First and foremost: Carolina. Or as I should say now: Dr. Jorge. Started the first day on the same project but finished one day before me. Congratulations, you officially beat me. Jokes aside, I feel privileged that we worked on the same topic for five years straight. Together navigating the often confusing PhD trajectory with all its forms and procedures. Being there during the highs and lows of accepted and rejected papers. Thank you for all these years and moments where being colleagues turned into a great friendship.

Another big thank you to the other colleagues that made office days more fun. I will probably forget same names but especially Sid for mentoring me, Mo for all the Nintendo-related talks, Enrico for discussions on Inter Milan, Nele for inspiring me to be at the office in Amsterdam at 7AM, Emma for many great conversations and being able to switch to Dutch during office days, Pei-Yu for always having my back when cancelling bi-weekly meetings, and Linyun for your energetic presence and super random questions when trying to finish my dissertation on time. Thank you to all the staff members and other PhDs and Postdocs as well, it was great being part of such a talented group of researchers with varying expertise!

I would also like to thank Roksolana Malhanova and Astrid van der Niet from the RoboHouse for facilitating a place to conduct experiments on human-robot teamwork, even while being in the middle of a move between buildings. Speaking about this experiment, a big thank you also goes out to my close colleague Mirk Baglioni. Thanks for the great collaboration on this experiment: from teaching me all about fuzzy-logic-based model predictive control to our struggles with the TurtleBot. I am very pleased with the end result and how we were able to combine our complementary expertise to develop a method useful for the fire department during collaboration with their firefighting robots.

Dit brengt me op de volgende personen om te bedanken: Robbert Heinecke and Sebastiaan Star van de brandweer Rotterdam-Rijnmond. Ontzettend bedankt voor alle samenwerking en hulp tijdens mijn experimenten en studies naar mens-machinesamenwerking binnen de brandweer. En natuurlijk dat ik jullie blus- en verkenrobot Brutus heb mogen besturen tijdens een trainingsavond! Ik ben super nieuwsgierig hoe jullie samenwerking met Brutus eruit zal zien over 10 jaar. Maar wellicht dat ik hier zelf ook nog een bijdrage aan zal leveren. Brutus is in ieder geval vereeuwigd op de kaft van dit proefschrift! Dat allemaal dankzij Asa, mijn nieuwe collega en tevens geweldig designer. Asa, ik ben je mega dankbaar voor de prachtige omslag.

I always thought it would be hard to have my acknowledgements span 3 pages, but it turns out to be quite easy! Many people to thank. Next in line are two people I would like to thank (again): my dear paranymphs Linyun and Mo. Mo, although we never became official office sharing buddies, it was a pleasure to have you as a close colleague and I am very happy that you are part of my defence. I really enjoyed our talks on especially the Switch 2, Saudi Arabian football, and the best Zelda game. Linyun, even though we only shared the office briefly and I never managed to pronounce your name right, it is safe to say you made an electrifying impression. Always cheerful and energetic, bringing life to ordinary working days. Thank you for being part of my defence.

Dan is het nu tijd voor mijn vrienden: met name Daaf, Jess, Mel, Don, Ell, Berry, Danny en Tim. Mannen die haast niet verder af konden staan van wat ik de afgelopen jaren heb gedaan. Die haast iedere avond over de vloer kwamen toen ik nog aan de Wulpstraat woonde met Daaf. Bedankt dat ik tijdens dit traject heerlijk mijn gedachten kon verzetten door met jullie over de simpelste dingen te ouwehoeren.

De inner circle, eindelijk aan de beurt. Te beginnen met mijn opa's en oma's. Opa en oma Verhagen, sorry dat het boek niet meer op tijd kwam voor jullie om het mee te maken. Ik weet hoe trots jullie daarop waren. Oma Elly, bedankt dat u altijd zo enthousiast was over mijn promotieonderzoek, ook al had u waarschijnlijk geen idee waar het precies over ging. Opa Kees, ik ben u natuurlijk extra dankbaar. Uw fenomenale bijlessen biologie op de middelbare school hebben een grote bijdrage geleverd aan deze mijlpaal. Dat ik jullie dit proefschrift persoonlijk kan overhandigen maakt mij ontzettend trots en dankbaar. Bedankt voor alles.

Jas, Pas, en Kar: bedankt voor alle steun de afgelopen jaren. Ik voel me gezegend met jullie als lieve broers en zus. Een warm nest om tot rust te komen tijdens de vele verjaardagen en uitjes. Al werd die rust wel iets minder toen Pas en Irene (jij ook bedankt, wellicht degene die dit onderzoek het beste begreep!) mij een fantastisch neefje cadeau deden: Mattis. Wat een aanwinst voor de familie. Heerlijk ook dat hij het middelpunt van de aandacht werd en ik steeds minder vragen kreeg wanneer ik nou toch eindelijk eens klaar was.

Pap en mam, waar moet ik beginnen. Zonder jullie was dit nooit gelukt. Mam, al vervloekte ik het toen vaak, jouw strenge toezicht in mijn examenjaar VWO heeft een grote basis gelegd. Om nog maar te zwijgen over jouw bijdrage aan het kiezen van de juiste vervolgopleidingen. Naast deze praktische hulp is er natuurlijk nog je nimmer aflatende liefdevolle steun. Wat ben ik je daar nog het meest dankbaar voor. Pap, bedankt dat je altijd het perfecte voorbeeld geeft dat hard werken en opoffering beloond wordt. Voor jouw wijze lessen uit de praktijk, links naar inspirerende LinkedIn posts, en natuurlijk de inhoudelijke gesprekken over mijn onderzoek. Ik heb altijd gevoeld hoe betrokken, geïnteresseerd en trots je was.

En dan last but not least: de allerliefste Kel. De wervelwind in mijn routine. Ineens was je daar, op een strandfeestje in Scheveningen. Inmiddels samen een heerlijk huis en naar Nieuw-Zeeland voor de reis van ons leven. En er wacht nog zoveel meer moois. Bedankt voor jouw onvoorwaardelijke liefde en steun de afgelopen jaren. Dat je mij altijd het gevoel hebt gegeven dat je nog trots op me zou zijn als ik een worm was. Zonder jou had ik dit niet gekund lieverd.

Curriculum Vitæ

Ruben Sebastiaan Verhagen

30-03-1994 Born in Dirksland, the Netherlands

Education

2020–2025 **Ph.D. in Computer Science**
Delft University of Technology, Delft, the Netherlands

2018–2020 **Master of Science in Cognitive Science**
University of Trento, Rovereto, Italy

2017–2018 **Master of Science in Communication & Information Sciences**
Tilburg University, Tilburg, the Netherlands

2013–2017 **Bachelor of Science in Communication & Information Sciences**
Tilburg University, Tilburg, the Netherlands

Experience

2025–... **Royal Netherlands Aerospace Centre**, Amsterdam, the Netherlands
R&D Engineer Human-Machine Interaction Technologies

2020 **Fondazione Bruno Kessler**, Mattarello, Italy
Research Intern

List of Publications

Under Review

🗎 1. **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Empowering Human-Robot Interaction for Firefighting: Mixed-Methods Comparison of Teleoperation and Collaboration. Submitted to: *The ACM/IEEE International Conference on Human-Robot Interaction.* 2025.

2025

🗎 1. **Ruben S. Verhagen**, Mark A. Neerincx, X. Jessie Yang, and Myrthe L. Tielman. Advancing Human-Machine Teaming: Definitions, Challenges, Future Directions." HHAI 2025. IOS Press, 2025. 49-59.

🗎 2. **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Agent Allocation of Moral Decisions in Human-Agent Teams: Raise Human Involvement and Explain Potential Consequences. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency.* 2025.

2024

🗎 1. **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Meaningful human control and variable autonomy in human-robot teams for firefighting. *Frontiers in Robotics and AI* 11 (2024).

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

3. Carolina Centeio Jorge, Emma M. van Zoelen, **Ruben S. Verhagen**, Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. Appropriate context-dependent artificial trust in human-machine teamwork. *Putting AI in the Critical Loop.* Academic Press, 2024. 41-60.

4. **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Research Environments for Trust in Human-AI Teams in *Workshop on Multidisciplinary Perspectives on Human-AI Team Trust co-located with the International Conference on Hybrid Human-Artificial Intelligence* Malmö, SWE, 2024.

2023

1. **Ruben S. Verhagen**, Mark A. Neerincx, Can Parlar, Marin Vogel, and Myrthe L. Tielman. Personalized Agent Explanations for Human-Agent Teamwork: Adapting Explanations to User Trust, Workload, and Performance. *AAMAS*, 2023, 2316-2318.

2. **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. Meaningful Human Control and Moral Agency in Human-Robot Teams for Fire Fighting in *Workshop on Perspectives on Moral Agency in Human-Robot Interaction co-located with the ACM/IEEE International Conference on Human Robot Interaction* Stockholm, SWE, 2023.

2022

📄 1. **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI* 9 (2022).

2. **Ruben S. Verhagen**, Siddharth Mehrotra, Mark A. Neerincx, Catholijn M. Jonker, and Myrthe L. Tielman. Exploring Effectiveness of Explanations for Appropriate Trust: Lessons from Cognitive Psychology in *Workshop on TRust and EXpertise in Visualization co-located with IEEE Visualization Conference* Oklahoma City, USA, 2022.

2021

📄 1. **Ruben S. Verhagen**, Mark A. Neerincx, and Myrthe L. Tielman. A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable. *International workshop on explainable, transparent autonomous agents and multi-agent systems.* Cham: Springer International Publishing, 2021.

📄 Included in this thesis.