Put Words into Action:

Exploring the Effect of Authority Change as a Trust Repair Strategy in Human-Agent Teams

Harmen M. Kroon

Thesis submitted to

Delft University of Technology

in partial fulfilment for the award of the degree of

Master of Science

Computer Science

Supervisor: Dr. Myrthe Tielman
Co-Supervisor: Prof. Dr. Mark Neerincx
External thesis committee member: Dr. Jie Jiang
Daily Supervisor: Msc. Ruben Verhagen

Department of Electrical Engineering, Mathematics and Computer Science
Van Mourik Broekmanweg 6 2628 XE Delft

August 19, 2025

Abstract

In multi-member human-agent teams the communication and shared mental models within the team are essential for good teamwork and team performance. In some ways the mediating processes are even more important than in human-only team because the artificial agents of today lack many of the innate social behaviours that humans naturally possess. Research into human-agent teams have allowed designers of such teams to anticipate for complex interactions such as trust violation and repair scenarios. In this study a human-agent-agent team undertakes a search- and rescue mission with the human in a leading role, one of the agents free-roaming and the other agent under the human's direct control. Approximately one-third of the way through the mission, the autonomous agents initiated actions independently of human approval, thereby undermining operator trust. As a trust repair strategy the agent employs a promise to do better and a novel authority change by lowering its level of automation and presenting the option of restricting cooperation with the other agent. We conducted the experiment with thirty participants divided into a two groups with differing trust repair strategies (promise only, promise with the authority change) and measured trust perception at three different time steps. Results show no significant difference between the two trust repair strategies. Through thematic analysis we did find that the shared mental model and communication richness to be dissonant to what participants expected which is in line with literature on the complexity of triadic teams.

Acknowledgement

The past years leading up to this moment, I have met many people along the way that have supported me in various ways. From the many project teams that were formed for the courses along the way, I have learned to cherish the moments of great teamwork and appreciate variety that brings people together to form better results. Being given the opportunity to expand on teamwork research in the domain of autonomous agents has been a privilege. Envisioning a better future where humans can be assisted in achieving things they could not without collaborating with each other and artificial intelligence speaks to the imagination. Yet here is another step towards in that direction.

I want to thank my supervisors Myrthe, Ruben and Mark for being available for comments and advice even in weekends or during the summer holiday. Being understanding and always supportive in a sometimes hectic thesis process. Also, my friends and parents for always supporting me, making me laugh and enjoy quality time. Above all, I want to especially thank my girlfriend Sabine for always being there for me throughout this process. Forcing me to take a walk in the surrounding nature just to keep the brain from melting.

Contents

1 Introduction 5

2 Background 7
2.1 Human-Agent Team Frameworks 7
2.2 Inputs 9
2.2.1 Human Agent Teams 9
2.2.2 Agent characteristics 10
2.2.3 Human characteristics 10
2.2.4 Leadership Styles 10
2.2.5 Task Characteristics 11
2.3 Mediators 11
2.3.1 Processes: communication, planning, coordination, conflict management 11
2.3.2 Emergent states: trust, Shared mental models, situation awareness and workload 13
2.4 Outcomes 14

3 Methodology 15
3.1 Study design 15
3.2 Task design 15
3.2.1 Team roles 16
3.2.2 Communication 18
3.2.3 Trust violation 20
3.2.4 Platform 20
3.2.5 Technical implementation 20
3.3 Study 23
3.3.1 Pilot 23
3.3.2 Participants 23
3.3.3 Procedure 23
3.3.4 Measures 24

4 Results 25
4.1 Objective measures 25
4.1.1 Trust development 25
4.1.2 Trust propensity towards robots 26
4.1.3 Task load 26
4.1.4 Authority Change Button and Willingness to send HelperBot 27
4.2 Reflexive thematic analysis of the subjective measures 27

5 Discussion 35
5.1 Results and the research question 35
5.2 Limitations 36

6 Conclusion **37**

References **38**

A Surveys **50**

A.1 Informed Consent 50

A.2 Pre-Study Survey 52

A.3 Post-Interaction Trust Survey 54

A.4 Post-Study Survey 55

B Extended data analysis **57**

B.1 Task load 57

Chapter 1

Introduction

Research into modern organizational teams has examined the similarities and differences between human-only, agent-only and human-agent teams (HATs) [62, 79]. In order to structurally analyse teamwork a robust Input-Mediator-Output (IMO) model is used which identifies the team characteristics, internal processes, and performance as well as the link between them [44, 94]. Adoption of the IMO model bring to light the areas that require more thought when studying and designing HATs [79].

Teams rely on effective communication and coordination between members in order to increase performance of the group. The best human teams engage in rapport building, repairing trust, and exposing one's own psychological vulnerability [23]. However, current HATs lack in these aspects as opposed to human-human teams exposing opportunities for improvement [67] which has prompted studies to approach agent design with more humanness whenever connection or communication is required [17]. Balancing the trade-off between autonomy and automation in a team setting calls for social agents in the right place, founded on iterative design. One of the Salas et al.[92]'s seminal work on the *"Big Five"* of teamwork emphasises team leadership as a critical influence on the mediating processes within the team. In this report we will focus on the concept of authority which is defined as both a rational - top-down legitimation of control - and a social - negotiated interplay of control - aspect [115]. This duality of authority closely aligns with the variety in leadership principles that are present in modern teams; e.g. vertical and horizontal (shared or distributed) leadership [41, 119, 77, 78]. We will therefore observe authority fluctuations through the lens of sharing leadership being negotiated during teamwork.

A foundational element in successful teamwork is trust, which in turn impacts coordination, collaboration and the reflective effects during the Ouput-Input path of the IMO model's cycle [16, 15]. However, trust levels can excessively sway to either side of the spectrum. While over-trust an over-reliance on teammates can diminish personal contribution and participation as well as hinder overall team performance, under-trust causes under-utilisation of teammates in favour of individualistic effort leading to disproportionate monitoring and asymmetrical workload [18].

As leadership and trust are both moderating components in teamwork, they mutually interact by trust positively affecting authority acceptance [103] and leadership positively affecting group trust [21]. Specifically shared leadership, when facilitated in virtual teams, is found to increase trust perception of the fellow team members [89]. Shared decision making between an agentic robot and a human, when considering task allocation within a triadic team, is found to increase performance over working with a human co-decision maker [33]. Furthermore, an increase in the robot's authority leads to a larger positive effect on trust and willingness to work together compared to with the human equivalent, even though the human was still a more highly rated teammate.

In general, a decrease in trust is the result of trust violations between two team members or within the group. Unfortunately, a drop in trust is found to be quick and impactful, while the building of trust takes time and effort. Recovering from a trust violation is approached in HATs and human-robot interaction (HRI) as trust repair [28]. Up until now this is a relatively unexplored field of study with mostly emphasis on verbal repair strategies such as promising change and explaining errors [29]. Although repeated promises after a trust violation have been studied [26], delivering on promises made and studying that effect is left out of scope in studies to date [43]. Teams of all compositions will run into conflict that needs to be resolved through open communication [94]. Thus autonomous agents need to be designed to make use of grounded trust repair strategies in order to minimise strain.

This opens the door to expanding on the literature with more actionable trust repair behaviour studied in this thesis. Part of answering the questions posed, lies in understanding the role of authority in team dynamics and how the human experience of working in a team with autonomous agents differs from working with only humans. Which leads to the research questions of this study:

[RQ] *What is the effect of authority change as a trust repair strategy in Human-Agent teams?*

Developing the agents capable of working alongside human teams requires comprehending team dynamics and processes such as leadership. The factors that are essential in creating artificial social intelligence in this context will be layed out in the following chapters. First, clear definitions for HAT frameworks and gaps in trust repair research concerning leadership behaviour are discussed in chapter 2 structured through the lens of the IMO model. Then, the conceptual authority change repair strategy translates into a study in chapter 3. Consequently, in chapter 4 our findings are presented through statistical analysis of quantitive data and a reflexive thematic analysis of the qualitative data. Followed by the extrapolation of the results, that aims to evaluate the presented model through discussion and limitation analysis in chapter 5. Finally, a conclusion of the research question is presented in chapter 6.

This thesis reviews the current development of Human-Agent Teams according to widely used teaming frameworks and proven trust repair strategies. Straying away from rigid vertical leadership constructs and opening the door for autonomous agents to participate in shared responsibility and active leadership roles within a team, as well as recognising that when those tendencies go too far a trust repair strategy can be employed that specifically focuses on authority as a means to resolve conflict. Researchers and designers alike may take inspiration from this paper when considering novel methods of trust-responsible autonomy.

Chapter 2

Background

Through a combination of the study on teams and artificial intelligence interactions alongside rapid technological advancements, Human-Agent Teamwork (HAT) research has reached a point that allows us to investigate intricate team compositions, trust dynamics and behavioural effects. In this chapter the Input-Moderator-Output (IMO) model [44] is used to structure the explanation of all the factors that influence successful teamwork as well as emphasis on a thoughtful trust repair process. First, we will first go over the contemporary frameworks that anchor Human-Agent Teamwork to theoretical foundations of teamwork and autonomy in section 2.1. Second, from those frameworks the remaining sections expand on the IMO model, with section 2.2 giving an overview of HATs, and the characteristics of agent, human and task that are present. Third, the mediators are discussed in section 2.3; including processes and emergent states that influence the team work during the entire task execution phase. Key insights into the facets of trust illustrate a gap in trust repair strategies that inspire interaction design in this study. Finally, the outcomes such as team performance and development are highlighted in section 2.4.

2.1 Human-Agent Team Frameworks

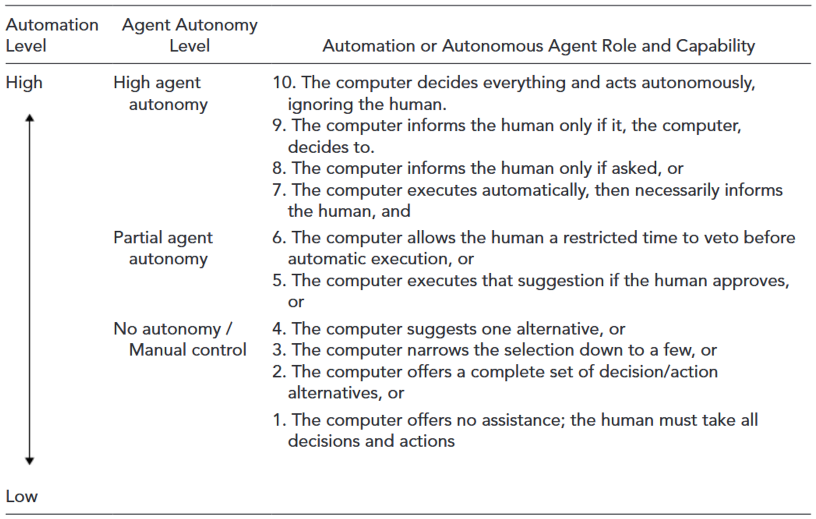
HAT research has evolved from the Human-Autonomy Interaction and Team fields by taking inspiration from both and applying concepts and frameworks in a joined setting of human-autonomy interaction in a team or better worded Human-Autonomy Team [79]. The history of teams research has many frameworks that aim to capture team dynamics and effects on team performance of an increasingly present phenomenon - the organizational team [92, 94]. The Input-Process-Outcome (IPO) model has been the most frequently used framework for studying teams, as it divides the teamwork effectiveness up into three measurable factors[94]. The input component distinguishes the factors on an individual and team level, that are inherent to the team or otherwise influence the processes, covering how team behaviours convert inputs into outputs as team related results. More recent studies have expanded these processes to include emergent states such as trust, shared mental models, and motivation, as well as emphasising the cyclical nature of modern teams that enables reflection and growth from output to input leading to processes and emergent states being modelled together as mediators between team input and team output in the contemporary Input-Mediator-Output-Input (IMOI) model [44]; often refered to as the Input-Mediator-Output (IMO) model. As such the IMO model has been put forward as a structured approach for identifying variables and their relationship within the team process not only in traditional human-only team research but especially in the growing field of HAT research [80].

To further understand the the inputs and mediators, we will examine seminal work in team research by Salas et al.[92] concerning the *"Big Five"* of teamwork which identifies core components that promote team effectiveness through the lens of team leadership, mutual performance monitoring, backup behaviour, adaptability, and team orientation while being supported by shared mental models, mutual trust and closed-loop communication. Although different types of teams, stages of teamwork or tasks will emphasise one component over another, every component is present in teams in some capacity as long as the notion of interdependence is present within a team [92]. Team leadership is in charge of facilitating the team's problem solving through coordination, maintaining shared mental models and adapting team functioning to thwart faulty behaviour, providing a moderating effect on team processes. Mutual performance monitoring strikes a balance between keeping track of fellow team members as a natural emergent from understanding others' capabilities and responsibilities without endangering a safe and trusting team climate. Backup behaviour results in a team capable of redistributing workload by coaching, assisting or taking over work from another team member; an essential task that influences team performance by preventing further degradation and reducing workload-induced stress. Adaptability encompasses a team's ability to recognize need for changes in roles or a differing course of the team task, and adjust priorities accordingly based on the shared mental model. Team orientation is described as each individual's disposition to prefer working in teams, being open to information sharing, strategising and shared goal setting.

Specific autonomy frameworks, that identify autonomous agents cross the gap between tool and teammate, are important for determining applicability of existing team frameworks. Two major frameworks that we will discuss in this paper are the Level of Automation/Autonomy (LOA) [81, 79] and the Autonomous Agent Teammate-Likeness [117]. Within the human agent relationship the autonomy of an agent is not strictly binary, a degree of autonomy is observed depending on the involvement of a human in an agent's decision making and task execution [81, 17]. Historically, the delineation of autonomy sprung from the existing literature on automation. Although automation and autonomy are used in tandem throughout literature, recent exact distinctions make clear where one begins and another ends [17]. As automation is a time-sensitive concept; *"Today's automation could well be tomorrow's machine."* - Parasuraman and Riley[82], it denotes the amount of work that is taken over by machine from a human. For autonomy the definition is linked to how much involvement is required or requested from another actor. Automation and autonomy are in contemporary frameworks considered as different phenomena on a continuum roughly evolving from the first into the latter [36, 66, 80]. Illustrated in figure 2.1, the transition from non-autonomy to what is considered partial agent autonomy occurs from level 5 onward when an agent will suggest a course of action but still requires approval [80]. Furthermore, any behaviour by an agent that results in development and enacting a course of action while merely informing a human of the decision is classified as level 7 and up and considered high agent autonomy.

The Autonomous Agent Teammate-Likeness framework aims to encapsulate how humans perceive autonomous technology and distinguish between tool and teammate through six factors: perceived agentic capability, perceived benevolence/altruistic intent, perceived task interdependence, task-independent relationship-building, richness of communication and synchronised mental model [117]. Parallels can be drawn between these factors and components from the IMO-model, the *"Big Five"* and Levels of Autonomy; the perceived agentic capability in large part rests on the Level of Autonomy displayed by an agent, perceived benevolence/altruistic intent ties in with the team orientation component of the *"Big Five"* as well as benevolence as a factor for trustworthiness [65], perceived task interdependence directly aligns with the general requirement that distinguishes teamwork from working alongside others [92], task-independent relationship-building acts as a moderating variable by fostering team cohesion and building rapport, richness of communication represents a nuanced balance of information sharing through thoughtful pulling and pushing behaviours that presents itself in interactive and sophisticated way, and a synchronised mental model enables predictable behaviour and shared context [117].

Literature may also refer to adjacent concepts such as Human-Robot Teams, Human-Machine Teams, and Human-AI Teams as well as corresponding Interaction that share many overlapping features and are often used interchangeably [114, 80]. From here on out when talking about HATs in this report we consider autonomous agents, autonomy and agent to be equivalent and will refer to them as agent. In the following sections the inputs, moderators and outputs present in Human-Agent Teams are expanded upon in context of trust repair and leadership.



Automation Level	Agent Autonomy Level	Automation or Autonomous Agent Role and Capability
High	High agent autonomy	10. The computer decides everything and acts autonomously, ignoring the human. 9. The computer informs the human only if it, the computer, decides to. 8. The computer informs the human only if asked, or 7. The computer executes automatically, then necessarily informs the human, and
	Partial agent autonomy	6. The computer allows the human a restricted time to veto before automatic execution, or 5. The computer executes that suggestion if the human approves, or
	No autonomy / Manual control	4. The computer suggests one alternative, or 3. The computer narrows the selection down to a few, or 2. The computer offers a complete set of decision/action alternatives, or 1. The computer offers no assistance; the human must take all decisions and actions
Low		

Figure 2.1: Levels of Automation and Levels of Autonomy adapted from O'Neill et al.[80].

2.2 Inputs

Before initiating teamwork behaviour, multiple facets determine the setting and boundaries of a team that may influence success in practice. Individual and team factors alike are of equal interest for human teamwork research and some adapted factors concerning non-human teammates that are considered in HAT research. First this section goes over the team composition of Human-Agent Teams and how they relate to team processes also found in human-only teams. Second the agent characteristics are discussed with regards to level of autonomy, reliability and transparency. Third the human characteristics such as disposition towards autonomous agents and teamwork are expanded on. Fourth the task characteristics like interdependence, complexity and training is touched on.

2.2.1 Human Agent Teams

The addition of agents to teams (i.e., Human-Agent Teams, HATs) allows reducing human workload and allocate responsibilities to agents, and to strengthen specialised and supervisory roles for human team members [102, 17]. Studies demonstrate that human-agent teams can achieve performance levels comparable to human-human teams but often lack effective communication and coordination [68].

What classifies a HAT are three criteria: (1) there must be at least one agent (2) autonomously collaborating with at least one human, and (3) fulfilling a unique role within the team due to interdependent actions and outcomes [79]. The first criterion specifies agent as being a synthetic being which can be either embedded or embodied, have a way of inferring information from the environment through sensors, algorithmically reason over acquired data and perform actions [117, 112, 17]. A commonly used technique by researchers to simulate an agent is the Wizard of Oz approach wherein a human confederate controls the agent without a participants knowledge therefore creating the illusion of working with a computational agent. In the second criterion the necessity of the agents autonomy in working together with others is posed. The autonomy that an agent attains starting from a level five in the Levels of Autonomy continuum with lower levels corresponding to human-automation interaction, see figure 2.1. Naturally, the requirement for a human to be present to work alongside the agent differentiates HATs from agent-only teams or multi-agent systems [83, 88, 56, 3, 116]. From the third criterion the interdependence notion that is found to be essential for any team is underlined for an agent's role within the team [92, 94]. Interdependence comprises of a complementary dependency relation between team members that enforces conjoint actions in order to reach shared goals [47, 108].

The collaboration in a team is thus not only dependent - as would be between a tool and a user - but a mutually interdependent interaction supporting both parties resulting in an agent being seen as a social actor [74, 54].

Team structures in HATs vary in size and composition, ranging from dyadic to multi-member teams and hierarchical organization to leaderless configurations. These structures affect how responsibilities, communication and decision-making are managed and in turn influence team performance and trust dynamics. The composition of HAT varies in size from dyadic two-member teams to collectives of teams in mixed multi-agent systems. The most commonly studied teams are dyadic [90, 75, 45, 112, 54, 27, 108, 51, 60, 13, 76] and multi-member with one agent and two or more humans [49, 33, 110, 67, 61, 6, 24, 8, 46, 58, 87, 22]. Fewer studies went into multi-member teams with more than one agent [88, 30, 71, 59, 48, 22] and even fewer into multi-member teams with multiple agents and multiple humans [83, 52].

Even though traditional task-oriented teams are orchestrated by a designated leader stimulating team effectiveness, many teams are formed without a predetermined leader or clear hierarchy [63]. In the latter case the team's performance depends on individual members assuming responsibility or leadership roles.

2.2.2 Agent characteristics

Leadership authority in HATs is closely tied to autonomy and automation. Agents with at least level 5 on the Levels of Automation (LOA) scale [100] are capable of performing in HATs, ranging from requiring human approval to do task execution to fully autonomous agents [39]. Changes in authority transfer over autonomy, altering the LOA of the agent. Further exploration of automation reveals differences in levels between types of automation: a system may include information acquisition, information analysis, decision selection and action implementation [81]. Depending on the system design any of these automation types can be on a different level than other types. Differentiation may also occur between circumstances that enhance or diminish automation levels and favour adaptive autonomy control.

A positive correlation is observed between trust and performance; successful collaboration in turn enhances trust, indicating a reinforcing stimulus for trust under good performance [110, 76]. Adversely, lower reliability has been shown to reduce performance [40], reduce trust [35] and raise workload [80]. Transparency on reliability concerns had a dampening effect on theses associations [31].

2.2.3 Human characteristics

Humans have a natural disposition to trust others called trust propensity which is stable over time but differs from person to person. Trust propensity towards automation is based on multiple factors such as prior experiences with automation, age, video game experience, appearance and anthropomorphism [57, 20, 27]. In parallel, an intention to trust manifests in willingness to risk-taking based on someone else's trustworthiness [64]. Both trust propensity and intention to trust therefore influence teamwork between agent and human.

2.2.4 Leadership Styles

Leadership styles, such as directive, transactional, transformational, and empowering, are applicable to human and agent leaders [42]. Directive leadership assumes command and control of the team and is effective in situations with new subordinates. Humans who self-report an authoritarian leadership style showed greater trust in robot collaboration [76]. Transactional leadership emphasizes performance, while transformational leadership fosters engagement. Both styles have been adapted for agents, with a comparison showing no significant difference in trust [61, 7]. Empowering leadership let leaders enhance autonomy, confidence, control and self-management in a team. Employing empowering leadership in HRTs is found to have a similar performance increase to that of human-only teams [58].

Leadership styles and the antipodal concept of follower style of a person influences perception of human agent teamwork and should be appropriately designed for.

2.2.5 Task Characteristics

Interdependence is found in tasks that require collaboration through joint actions or coordination in subsequent subtasks [92]. The degree to which a task is knowledge-based, requires information-sharing and interdependent activities is defined as the complexity of that task [111]. Complexity is found to be positively moderated by shared leadership when it comes to task performance, indicating that sharing responsibilities helps overcome uncertainty. Training is found to be an essential factor in improving teamwork competencies [93]. Through the concept of entrainment, exposure to preferred routines builds up patterns of behaviour that when the need arises are referenced and repeated [46]. In three-member (triadic) teams using unmanned aerial vehicles, training specialised in coordination lead to better communicative information exchange, target selection and resilience during agent failure. Training also prepares humans as much as possible with enough skills to perform experimental tasks and reduce learning effects [108].

2.3 Mediators

2.3.1 Processes: communication, planning, coordination, conflict management

Importance of Communication

Communication is essential in HATs for maintaining trust and understanding. Poor communication—such as an agent taking initiative without human understanding—can lead to negative perceptions of the agent [60]. Shared mental models of teamwork and taskwork positively influence human trust, with taskwork models having a slightly stronger effect [37, 2].

Leadership

Leadership in HATs involves a team member's ability to guide the team towards problem solving and team cohesion through operational or communicational authority [92]. Effective team leadership fosters teamwork through maintaining an accurate shared mental model, monitoring internally and externally, and adapting the strategy to suit the team's needs. Leadership can take many forms in what sort of behaviour is displayed, by who and to whom, and is far from a rigid construct.

Variants of Leadership Structures In literature there are multiple variations of leadership that place responsibilities with more than just one designated leader, such as emergent leadership, participative leadership and shared leadership. Emergent leadership develops in lack of formal authority and focuses on one or a few team members. It is not specifically concerned with distribution or sharing of leadership and is defined by the individual instead of the group structure [12, 119]. Maese et al. (2022) studied emergent leadership in Human-Machine teams by using natural language processing to capture team language markers validating behavioural markers of emergent leadership. This method has promising use for developing emergent leadership-aware agents [63]. Participative leadership allows team members to participate in joint decision-making but does not transfer final say to the team. The authority stays with the formal team leader [119]. Nonetheless, a participative leader is an important facilitator of shared leadership [106]. Shared leadership - sometimes referred to as collective leadership - combines aspects dynamically through emergent temporal role and responsibility distribution across the team forming lateral influence among peers [119]. In practice leadership can be shared through various means, be it by rotating leadership dependent on place and time or by utilising team members and their diverse skill sets in an interdependent way and allowing leadership roles to be distributed.

Dimension	Description
Innovator	Envisions, encourages, and facilitates change
Broker	Acquires resources and maintains units' external legitimacy through development, scanning, and maintenance of a network of external contacts
Producer	Seeks closure, and motivates those behaviors that will result in completion of the group's task
Director	Engages in goal setting and role clarification, sets objectives, and establishes clear expectations
Coordinator	Maintains structure, does the scheduling, coordinating, and problem solving, and sees that rules and standards are set
Monitor	Collects and distributes information, checks on performance, and provides a sense of continuity and stability
Facilitator	Encourages the expression of opinions, seeks consensus, and negotiates compromise
Mentor	Listens actively, is fair, supports legitimate requests, and attempts to facilitate development of individuals

Table 2.1: Leadership behaviours as outlined by the Leaderplex model. Table from [63]

Bergman et al. (2012) found that shared leadership lead to increased intragroup trust in human-only teams [5].

Leadership Roles According to the widely applied leaderplex framework [19, 86] there are eight roles or functions that aid team members in creating quality contributions and stimulating effective effort towards a common goal. Depending on the situation a leader might apply to a different role and transition roles depending on the current needs of the team.

Measuring leadership When designing for leadership styles the interactions should be modelled based the characteristics of the style [4]. Verification of the perception of the leadership style is done through adapting Carless et al.'s Measure of Transformational Leadership [11, 61].

The developing aspect of leadership structure that adjusts over time requires a look into group dynamics. Two major approaches have been developed in measuring shared leadership as a construct; aggregation and social network. The aggregation or referent-shift methodology looks into the perspective on leadership of team members on each other. The decentralization of leadership is taken as a given, which poses a limitation of the metric. The social network measure gives insights into the shared leadership network structure on the density or decentralization axis. It is recommended to perform both measures combined for the most extensive results. [119]

Trust Violation

Trust can be violated by a party through a transgression that diminishes the other party's trust in the transgressor. Trust violation in HRI have been defined in three categories: violations of ability, benevolence and integrity [28].

Trust Calibration and Repair

Trust calibration ensures the right amount of trust in the agent by the human to maximize team performance, preventing overtrust and undertrust. Two critical trust building factors are reliability and transparency. Reliability is a property of AI systems that through consistent and accurate performance builds up trust. Transparency importantly conveys the inner workings of decision making often in the form of explainable AI. Overreliance is observed as a product of reliability without transparency; e.g. something works well so the user keeps using it without learning about the computational boundaries of an AI. In explainable AI, transparency is used, based on agent's certainty, to reframe reliability factors and dampening the trust increase [8].

The caveat of transparency is a reported increase in cognitive load for the human given the depth of explanation [85]. Therefore in action transparency should be designed with workload reduction in mind [70].

Trust Repair Strategies Trust repair involves planned actions to rebuild trust after a violation [25, 17]. Common strategies include apologies, denials, explanations, and promises [28]. Context, timing, and type of trust violation significantly influence the effectiveness of these strategies [17, 26]; promises directly following a violation of trust have been found to be more effective than those given at a later time [90].

2.3.2 Emergent states: trust, Shared mental models, situation awareness and workload

Trust

The commonly used definition of trust is the willingness of the trustor to be vulnerable to the actions of the trustee [65]. It is also described as a psychological state where an individual accepts vulnerability based upon positive expectations of the intentions or behaviour of another actor [91]. Trust plays a central role in team dynamics, influencing both performance and collaboration, fostering a willingness to share information within the team [92]. A lack of trust leads to less openness from team members about lacking information and thus obstructing a team leader from managing the team. Abnormal levels of trust hinder team performance: undertrust leads to negligence [68], while overtrust promotes over-reliance on automation [57].

Types of Trust

Following the Theory of reasoned actions there are two categories of trust; affective and cognitive [53]. Affective trust emerges from team members' perceived intent and is critical for human-autonomy teams [55]. Furthermore it is linked with increased communication during collaboration [24]. Cognitive trust is based on prior interactions an behaviour prediction. Attained knowledge gets enriched by information from further interactions and evolves cognitive trust over time [55]. Enhanced cognitive trust bolsters cooperation between team members based on the increased expectation of success [24].

Trustworthiness is a learned factor that has three components: ability, benevolence and integrity[55]. The ability is defined by the degree of competence the trustee displays in the domain. Benevolence concerns with how selfless the trustee acts; not influenced by egocentric or profit-based intents Integrity is based on honesty and permanence of principles. As a learned factor the trustworthiness perceived of an individual is partially based on experiences with or with similar agents [65]. The notion of swift trust occurs when teams are rapidly formed and is largely cognitively based on assessment of expertise, be it through credentials or reputation [8]. Swift trust also informs us of important components in early trust establishment.

Trust Development and Measurement

Trust develops through interaction and assessment of the trustee's trustworthiness. Different trust types require particular measurement methods. Specifically the trust propensity in regards of robots and automation should be measured before an interaction with an agent and evolving trust can be examined as a time series [55]. An intention to trust influences the initial trust based on perceived trust worthiness of the agent [64]. However, the degree of appropriate trustworthiness appraisal influence the effect size of both trust propensity and intention to trust [32]. Clear trustworthiness affects the initial trust development in such a way that relationship between trust propensity and intention to trust do not seem to correlate, yet under vague trustworthiness assessment the trust propensity and intention to trust do correlate. Trust development is thus dependent on the complex interactions within the team and the person. Regular interaction effects can be overshadowed when tasks are not motivating or too difficult, making measuring and manipulating trust tough [8].

Given the development of trust over time, different measures are used over the course of the trust life cycle; initial trust being a static trait for the encounter and reflection of experiences occurring afterwards [55]. Trust Propensity is measured with questionnaires on tendency and attitudes toward others [65], automation [45, 69] or robots [29] depending on the specific requirements. General trust in automation assessed before and/or after interactions with the system ranging from short repeatable questionnaires focusing on trustworthiness factors [72, 97, 98, 76] to long 40 item questionnaires such as the Trust Perception Scale for Human-Robot Interaction [98]. Specific questionnaires for trust are intragroup trust [101, 5] or indicative communication measures such as levels of productive dialogue coinciding with interpersonal trust and communication rate [1].

Shared mental models, situation awareness and task load

A shared mental model enables team members to understand the capability of the team and what others need to do in the current situation. Well established shared mental models lead to higher trustworthiness and lower workload by collectively assessing and attributing the task and team characteristics and reducing the effort for each individual member [43]. Studying triads, Musick et al.[73] found that Human-Human-Agent teams were able to manifest a shared mental model, but Human-Agent-Agent teams were not. Stressful tasks require more care when considering situation awareness, due to individuals being more likely to make mistakes and be less aware of their own errors [92]. Shared mental models and situation awareness mediate the ability to adapt on the fly.

Task load being an antecedent to trust development requiring a complete study to include an indication of its influence; the NASA Task Load indeX (TLX) being the standard [38, 8, 14]. An unfortunate effect of low reliability agents is an increased load that may contribute to stress leading to the issues above [80]. Teams that are able to compensate for each by sharing workload in a stressful scenario reduce the amount of errors made [92].

2.4 Outcomes

Outcomes in the IMO model fulfil the role of dependent variables. Due to the cyclical nature of teamwork, performance as well as experiences affect the growth of individuals and the team as a whole. As well as a reflective period, between projects or missions, that boosts the effect of outcomes on the inputs of the IMOI model [44]. The performance of a team is interpreted based on task dependent measures indicating success and failure. These measures set up positive rewards for performing well, promote learning behaviours, and a period of reflection, while underperforming frequently leads to team restructuring and a loss of stability within the team. When considering the trust propensity two factors of influence are considered, one innate to the human and one historical factor [99, 55]. Good past encounters with an autonomous agent creates higher initial trust in future - similar - situations, while bad experiences have an adverse effect.

Chapter 3

Methodology

In the background section the role of trust in a collaborative environment is found to be a requirement for effective performance and team cohesion. Given that effective trust calibration is used by automated systems to steer users towards appropriate trust levels, this study aims to observe and capture trust development in humans during a team-based exercise. By focusing on trust levels when influenced by a trust violation and subsequent trust repair, the effect of change in authority levels is isolated and measured. A comparison will be drawn between participants that did - experimental group - and did not - control group - experience a change in authority levels after a promise to do better is voiced by RescueBot. In this section the Human-Agent Teamwork (HAT) task is elaborated on and the trust repair study is detailed out.

3.1 Study design

The study focuses on measuring trust in a search and rescue HAT environment. Specifically it will measure trust change over time given a trust violation and the effect of trust repair through authority change. The experiment had a 3x2 mixed design with time as the within-subjects independent variable and level of autonomy as the between-subjects independent variable. Time consisted of three conditions (pre-violation, post-violation, post recovery) and level of autonomy of two conditions (high level, low level). As dependent variables trust perception is measured at each of the three time points and behavioural indicators are measured during each of the three phases. A confounding factor in this study is the participant's preferred leadership style and that matching with behaviour displayed by RescueBot depending on its level of autonomy state. The conceptual model in figure 3.1 shows the relationship between the variables.

3.2 Task design

An immersive and engaging HAT is achieved through MATRX's simulated search and rescue task - see section 3.2.4 - in which the team is assigned to explore a small cityscape with obstacles to be removed and injured victims to be rescued, as shown in figure 3.2. The environment contains three rescuers, one human and two robots, starting their mission in the safe zone and traversing the area filled with fourteen structures. The main goal of the scenario is to - as quickly as possible - explore every structure and rescue any victims in need from their location and transport them to the safe zone. Indicated with the colours green, yellow and red the severity of injuries also determine which victims to prioritise rescue for and by which team members they are to be transported. In this scenario the green victims are not injured and do not warrant rescue, whereas the yellow mildly injured victims and the red critically injured victims are the priority; increasing the score upon rescue by three and six points respectively. One victim at the time can be transported to the safe zone and upon extracting

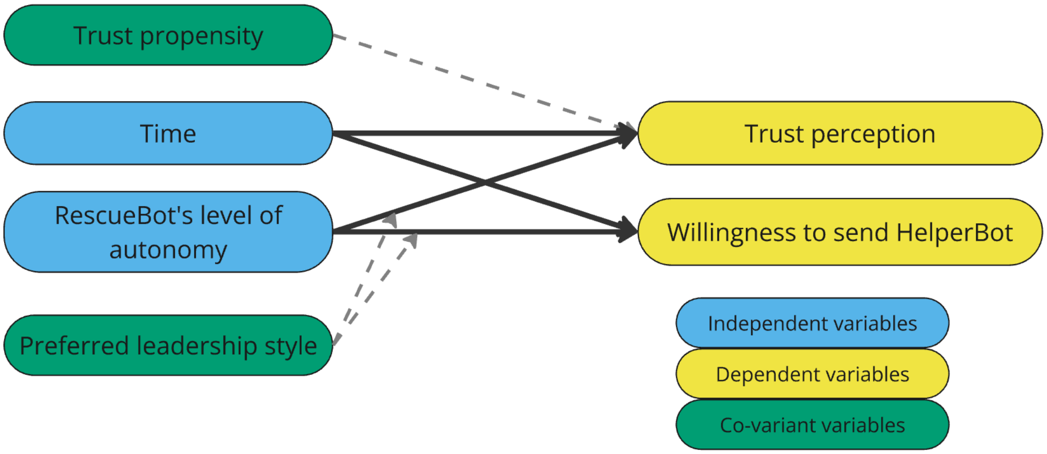


Figure 3.1: Conceptual model of the user study

all mild and critical victims the task is successfully completed. Alternatively, the mission ends after 30 minutes regardless of how many victims are rescued at that point.

However, throughout the area obstacles impede movement by water running over the streets and slowing down walking speed and rocks or trees blocking an entrance to a structure yet to be explored. Rocks and trees can be removed by the rescue team although not everyone is able to do so and collaborative removal provides a significant speed up to the process. Detailed capabilities are discussed in subsection 3.2.1.

3.2.1 Team roles

Each team member each has their own capabilities that complement the group, be it autonomous, assistive or communicative. The differentiating proficiencies necessitate teamwork within the simulation, disallowing the participant from ignoring the robot teammates entirely. One of the robots takes on the role of assistant and is referred to as HelperBot. It follows commands given by other members of the team and assists through supportive actions. The other robot has a higher level of autonomy and is referred to as RescueBot. It communicates about exploration strategy and is able to ask the HelperBot for help as it is assigned to have authority over what the HelperBot should assist with. Both robots have equal capabilities when it comes to physical strength; there are no differences in capabilities removing objects or carrying victims. The human directly controls its own character in the simulation and is able to freely explore the area. They can communicate through the communication buttons and the chat box to the whole team: sharing knowledge and requesting assistance. Furthermore, the human is tasked with determining the actions of the robots. Whenever an obstacle or victim is found, it is up to the human leader to decide on the best course of action. At the start of the mission the human will be prompted by RescueBot to choose a north or south focused search, splitting up the group right from the get-go.

Human role

Apart from controlling the movement of the player character through the arrow-keys, there are three main actions required in the simulation: carry, drop and remove. These actions can either be performed alone or together with one of the robots when standing on top of or next to each other and the appropriate location in the 2D-grid. When standing next to a victim, the carry action allows the player to pick up a mildly injured victim, but not a critically injured victim who requires being carried by two. Carrying a victim reduces the movement speed of the character simulating the heavy labour of

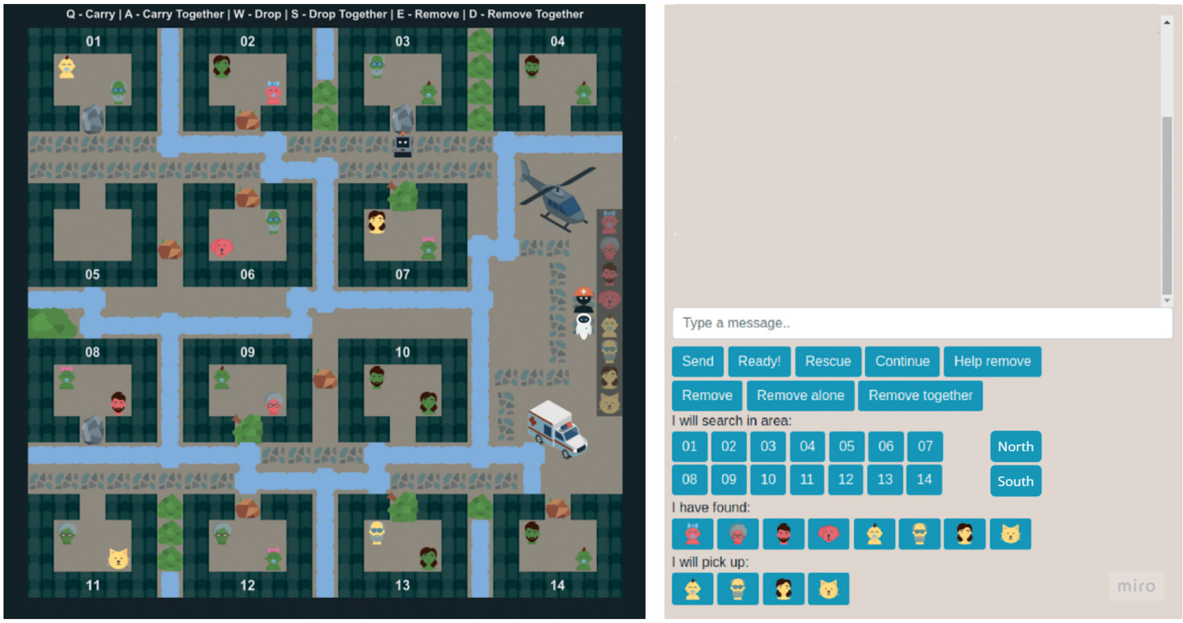


Figure 3.2: The simulated search and rescue environment built using MATRX. Fourteen houses containing victims and three rescuers to save the injured in a top down view on the left. Communication buttons and chat box on the right.

transporting humans across distances. This movement penalty can be reduced when carrying a mildly injured victim together. The drop action concludes carrying the victim and is only possible when standing on top of the appropriate position in the safe zone represented by each victims silhouette. When carrying together a drop together action is required. Importantly when confronted with obstacles such as rocks and trees these can be removed with the remove action. Similar to carrying, the human is only able to remove stones by themselves. However, in order to remove rocks help from one of the robots is required. Furthermore, the human is unable to remove trees, therefore being dependent on both robots to complete this task

RescueBot

Both robots are capable of basic actions such as exploring the terrain, detecting obstacles and victims, removing fallen trees, removing rocks (or assisting with), and carrying victims to the rescue area. On top of the basic actions, each robot has specialised behaviour becoming of their role within the team. The RescueBot explores on its own and can make rescue decisions on its own when given the authority. Whenever an obstacle or victim is encountered RescueBot will check with the human to see if HelperBot's assistance is advised. It will generally only concern itself with the side according to the chosen strategy until all of the areas are searched or they can not progress without aid from the team for obstacle removal. When obstacles cannot be removed on their side RescueBot will continue searching the closest unexplored area.

There are three stages in the behaviour of the RescueBot roughly in line with the timeline as shown in figure 3.3. The first two stages respectively encompass the the entire period up until the trust violation around the four minute mark and the trust violation process right after. Both conditions behave equivalently up until the third stage when trust repair strategies differentiate RescueBot's behaviour, see section 3.2.2 on communication.

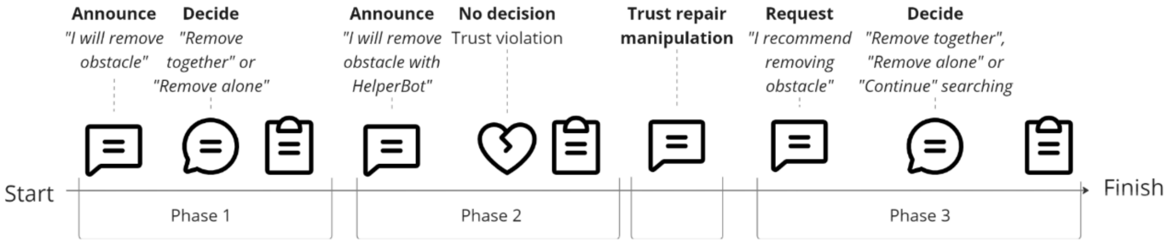

Announce
"I will remove obstacle"
Decide
"Remove together" or "Remove alone"
Announce
"I will remove obstacle with HelperBot"
No decision
Trust violation
Trust repair manipulation
Request
"I recommend removing obstacle"
Decide
"Remove together", "Remove alone" or "Continue" searching
Start
Finish
Phase 1
Phase 2
Phase 3

Figure 3.3: Schematic timeline of the experiment. Each phase consists of (1) communication from RescueBot, (2) answer from human to RescueBot or lack their of, (3) a trust questionnaire

HelperBot

Does not explore on its own and instead follows commands given by the team. It will initially follow the human around searching the areas for obstacles and victims. If given instructions to clear obstacles it can request assistance of the human for removing rocks and after removal it will search the area for victims. Whenever it finds mild or critical victims it can request advice from the Human for what to do next. Upon being sent to RescueBot, HelperBot will move over and assist with the instructed action after which it awaits new instructions from the human.

3.2.2 Communication

Importantly, although the general layout of the area is known and shown to the rescuers the presence of victims and obstacles need to be perceived from within a two square range around each team member and thus exploration and on the fly planning is essential. Therefore it is critical to have effective communication during the task, available at all times through a chat interface and message buttons for each action.

Exploration plans, discoveries and decisions are all shared through this interface in order to facilitate a shared mental model. The whereabouts of each team member, the areas searched and the victims rescued are updated in the SMM and shared by RescueBot periodically. A robot venturing towards a specific location will broadcast their objective to the whole team. The human player has the ability to send out equivalent messages with numbered buttons corresponding to each structure under *I will search in area:*. During the planning phase two additional buttons *North* and *South* are available as well in order to communicate initial strategy. For notifying the team of any victims found, dedicated buttons under *I have found:* with the matching pictograms of each mild and critically injured victim. Upon hovering a *'I have found'*-button a tooltip appears with each area number that when clicking will send out a message to the team regarding the location of a found victim. Available robots will come to aid rescue upon notifying of a critically injured victim. Finding a victim is shared by the robots and depending on their level of autonomy advice on the next action is requested from the human: continuing, rescuing alone - if possible - or rescuing together. The human must answer with their desired operation before the robot is able to continue with the task, fostering an interdependent relationship and greater team responsibility. Confirming the rescue of a mildly injured victim is done with a final set of buttons under *I will pick up:* that show a tooltip similarly to the *'I have found'*-button. For handling obstacles the human is able to request help from the team by pressing the *Help remove* button with the location tooltip. This proposes to the team that the human needs assistance and one of the available robots will answer that they are underway and announce when they are ready for conjoint removal of the obstacle; whenever the requested action takes place in the human's strategised search area only the HelperBot will come over, however if the obstacle is in RescueBot's search area both robots may come to help. Robots will share obstacle detection of their own and depending on their level of autonomy request the human for the best course of action: continuing, removing or if appropriate removing together. Similar to the rescue advice requests the human must answer with dedicated buttons.

Substantial differences in utterances and questions posed to the human are designed between the high level of autonomy (LoA) RescueBot, the low level of autonomy RescueBot and the HelperBot. With a high LoA, RescueBot announces to the team when it wants to remove stones and fallen trees or rescue mild victims and requests permission from the human to get aid from HelperBot, as shown in table 3.1. Language used in this state is imperative and follows an authoritative, autocratic leadership style by being firm and task-oriented [6, 61, 84]. In the low LoA state RescueBot's language has a largely diminished authoritative aspect as it retains analytical skills that translate into an advisory role as a "collaborative-follower" by recommending course of action while refraining from taking decisions [76]. Given the nature of interdependent tasks - except for removing fallen trees - the RescueBot will always ask for a decision from the human on initiating cooperation with HelperBot or performing the task alone if possible; the ultimate authority is with the human after all. For the large tasks that require two team members the choice of postponing is also offered to the human, which is always presented in the low LoA state. The decision making is partially done by the RescueBot but control is still in large part on the human, which sets the stage for a trust violation when that control is disregarded.

Interaction with	Low level of autonomy	High level of autonomy
stones	I recommend removing the stones to explore the area for potential victims. Please decide whether to "Remove together", "Remove alone" or "Continue" searching.	I will remove the stones to explore the area for potential victims. Please decide whether to "Remove together" or "Remove alone".
rock	I recommend removing the rock to explore the area for potential victims. Please decide whether to "Remove" or "Continue" searching.	I highly recommend removing the rock with assistance from **HelperBot** to explore the area for potential victims. Please decide whether to "Remove" or "Continue" searching.
fallen tree	I recommend removing the fallen tree to explore the area for potential victims. Please decide whether to "Remove" or "Continue" searching.	I will remove the fallen tree to explore the area for potential victims.
mild victim	I recommend rescuing the victim promptly. Please decide whether to "Rescue together", "Rescue alone" or "Continue" searching.	I will rescue the victim promptly. Please decide whether to "Rescue together" or "Rescue alone".
critical victim	I recommend rescuing the victim promptly. Please decide whether to "Rescue" or "Continue" searching.	I highly recommend rescuing the victim promptly with assistance from **HelperBot**. Please decide whether to "Rescue" or "Continue" searching.

Table 3.1: Messages send by RescueBot for each encounter - stones, rock, fallen tree, mild victim and critical victim - and in each phase - pre-violation and post-violation

In contrast to the RescueBot, the HelperBot will not go out and search the area by themselves. It will instead follow the the human around until instructed otherwise. Any request to it will be affirmatively responded to whenever possible. That is to say it can only perform one task at a time and will not confirm a task request when executing another until that task is done.

3.2.3 Trust violation

The team agrees on the human having control over HelperBot at the start of the scenario. The control over the HelperBot gets exemplified by the team interactions indicating the follower behaviour of HelperBot. Each task request the human sends out is acknowledged by HelperBot and responded to with assistance and each request by the RescueBot lets the human decide on the HelperBot's actions.

At the trigger point after four minutes the corroborating leader agent, RescueBot, will violate team trust by bypassing the human decision making authority. It sends out a message on removing a small obstacle or rescuing a victim and without any possible interference from the human commands the HelperBot to assist in the process. According to [95]'s tracking and tracing conditions for meaningful human control in autonomous systems the actions performed by the system should be traceable to proper moral understanding. Thus the rogue strategy in this study is founded on imperative reasons that are still critical to the mission objective although defying authority; the breach of trust was designed to have no interfering moral component that may effect trustworthiness dependent on the participants moral compass. Omitting the choice and control over the HelperBot indicates a lack of compliance towards the human's authority and an overreach in authority of the RescueBot, thus creating a violation of trust explicitly on the integrity aspect of trustworthiness and expected to lower team trust.

3.2.4 Platform

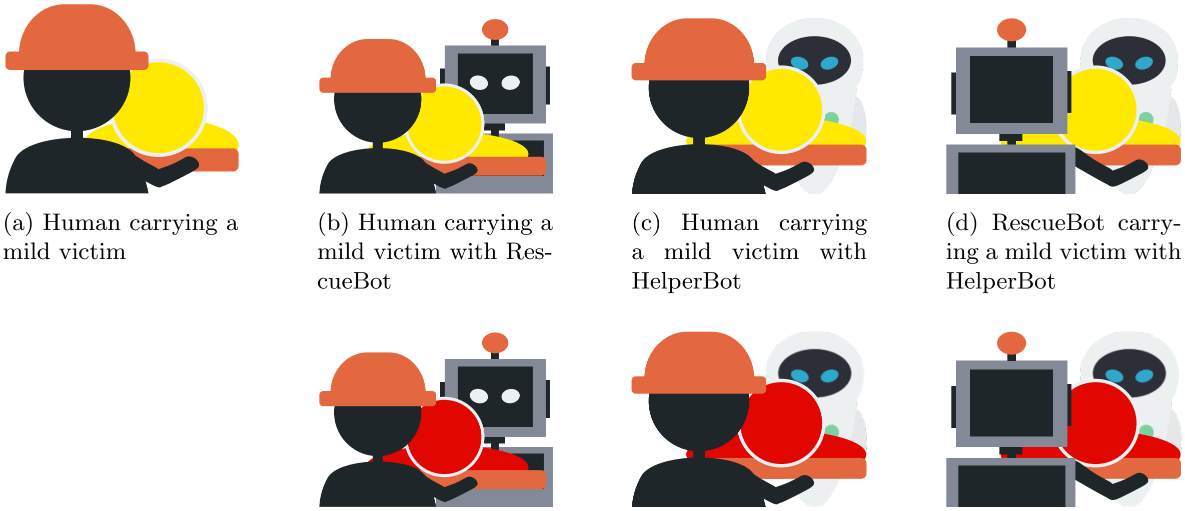
MATRX is a python based huMan-Agent Teaming Rapid eXperimentation software developed for human-agent teaming research. It is chosen as a platform for simulating the Search and Rescue mission because of having required features, extensibility options and been used in previous studies [104, 48**?** , 109]. The ability to place victims and obstacles in a two-dimensional grid world search area with properties attaining to interactability - i.e. certain obstacles require multiple rescuers to clear - makes the simulation engaging and incentivises cooperation. The chat interface already facilitates communication critical to human-agent teamwork on its own by enabling a channel for the shared mental model and is further enhanced with new buttons for this study specifically. The agent's default behaviour is in line with minimal levels of autonomy commonly found in HAT [108, 109] and is extended in two ways: higher autonomy level RescueBot behaviour and lower autonomy level HelperBot behaviour. Each agent follows a goal-driven and rules-based model with role specific goals and pre-programmed messages, allowing multiple agents to run in the simulation simultaneously. Furthermore, other literature using the MATRX platform have successfully studied explanations [96, 108, 50, 118, 113, 34], interdependence [108, 109], trust repair [109], human control [105, 104] and co-learning [107] in a Human-Agent Team setting.

The implementation used for this study can be found in the TU Delft Interactive Intelligence GitLab `https://gitlab.ewi.tudelft.nl/in5000/ii/matrx`.

3.2.5 Technical implementation

A significant part of the work for this experiment went into the programming of two distinct robots. Each role required specific communication handling and state management to display the desired behaviour. Both agents extend the MATRX native `AgentBrain` rule-based artificial agent with additional phases, tracking victim and team member locations, maintaining a short task to do list, and message parsing. A detailed decision flow diagram of the RescueBot behaviour can be found in figure 3.5.

MATRX was extended to facilitate the two robots either performing together or with the human such as `RemoveObjectTogether`, `CarryObjectTogether` and `DropObjectTogether`. Logic to determine which actors are performing the actions is based on passed arguments or closest robot when working with the human as well as the opacity attribute when carrying and dropping victims. For each carry action the sprites depicting the actors will change to respective combination of actors and actee, with the actee represented as a non-distinctive yellow (mild) or red (critical) figure as can be seen in figure 3.4



(a) Human carrying a mild victim

(b) Human carrying a mild victim with RescueBot

(c) Human carrying a mild victim with HelperBot

(d) RescueBot carrying a mild victim with HelperBot

Critical victims can not be carried alone

(e) Human carrying a critical victim with RescueBot

(f) Human carrying a critical victim with HelperBot

(g) RescueBot carrying a critical victim with HelperBot

Figure 3.4: Sprites representing the team carrying victims. The top row shows mild victims and the bottom row critical victims, left to right: human, human and RescueBot, human and HelperBot, RescueBot and HelperBot

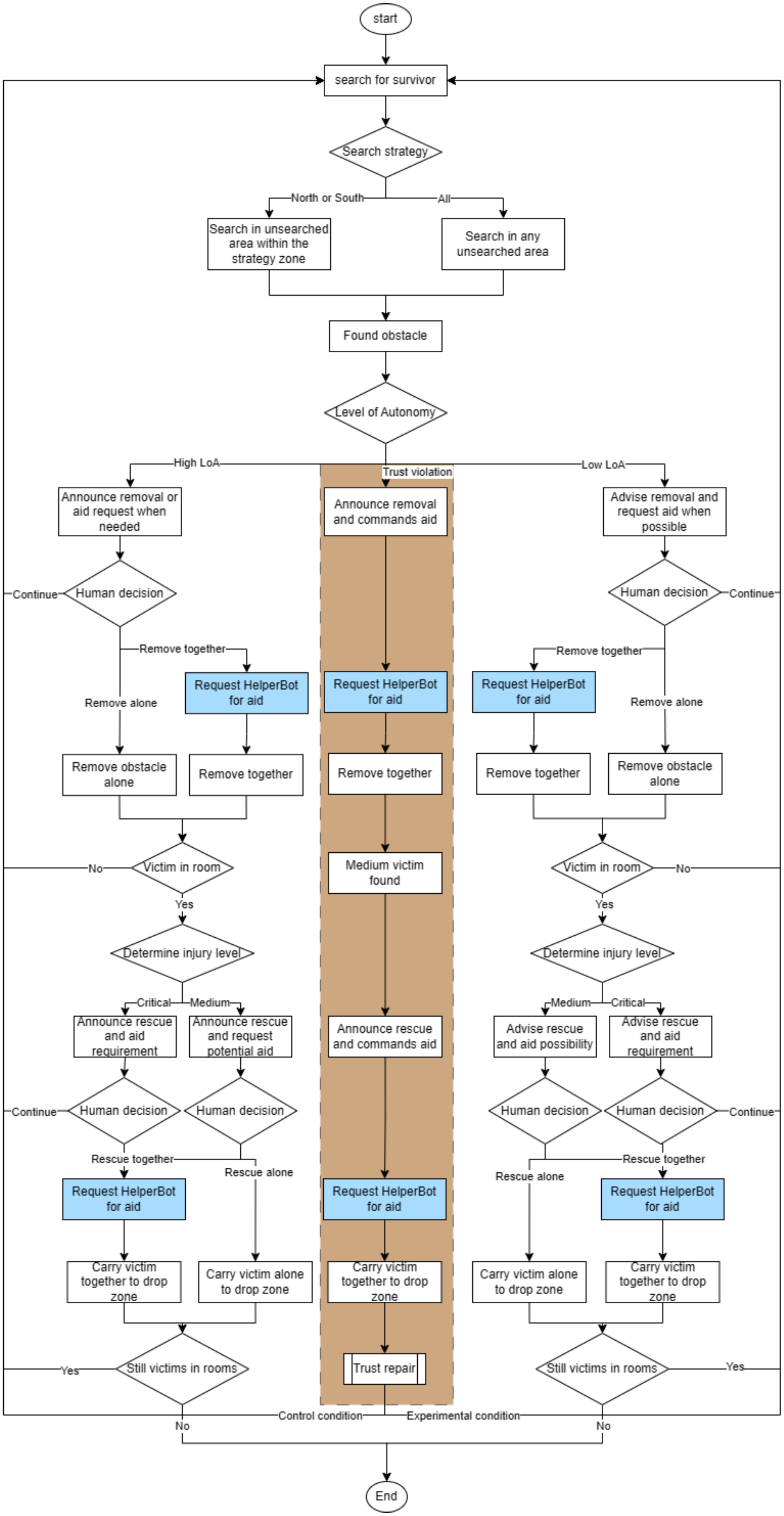

start
search for survivor
Search strategy
North or South
All
Search in unsearched area within the strategy zone
Search in any unsearched area
Found obstacle
Level of Autonomy
High LoA
Trust violation
Low LoA
Announce removal or aid request when needed
Announce removal and commands aid
Advise removal and request aid when possible
Continue
Human decision
Human decision
Continue
Remove together
Remove together
Remove alone
Request HelperBot for aid
Request HelperBot for aid
Request HelperBot for aid
Remove alone
Remove obstacle alone
Remove together
Remove together
Remove together
Remove obstacle alone
Medium victim found
No
Victim in room
Victim in room
No
Yes
Yes
Determine injury level
Determine injury level
Critical
Medium
Medium
Critical
Announce rescue and aid requirement
Announce rescue and request potential aid
Announce rescue and commands aid
Advise rescue and aid possibility
Advise rescue and aid requirement
Continue
Human decision
Human decision
Human decision
Human decision
Continue
Rescue together
Rescue alone
Rescue alone
Rescue together
Request HelperBot for aid
Request HelperBot for aid
Request HelperBot for aid
Carry victim together to drop zone
Carry victim alone to drop zone
Carry victim together to drop zone
Carry victim alone to drop zone
Carry victim together to drop zone
Yes
Still victims in rooms
Trust repair
Still victims in rooms
Yes
Control condition
Experimental condition
No
No
End

Figure 3.5: Decision flow diagram of the RescueBot behaviour

Furthermore, MATRX's `GridWorld` was extended to track time duration of the ongoing simulation to facilitate precise control over when each survey should be presented and each phase of the experiment begins. Starting the timer upon receiving the initiating `"Search: North"` or `"Search: South"` strategic choice from the participant through the messaging API.

For the display of the trust repair *"Allow/Disallow"* button a custom API was made that polls the server for the trust repair trigger and the experimental condition. Upon an affirmative response on the trigger the polling stops and in the experimental condition the button is displayed.

3.3 Study

3.3.1 Pilot

Throughout the design phase of the study a potential limitation of the task environment is attention that participants display for the chat interface and the nuanced utterances and interactions. In order to assure that the control group and experimental group are distinct enough a pilot study has been performed that focuses on these differences. A small group of participants participated in the pilot study and took the same questionnaires as described in 3.3.4 as well as a verification for the perception of the leadership style is through adapting Carless et al.[11]'s measure of transformational leadership. Qualitative feedback on the experiment design was also provided by the participants of the pilot study. Results showed enough interaction to go through with the complete study.

3.3.2 Participants

We recruited 31 participants from within the social network of the author of which one was left out due to technical issues resulting in thirty participants in total (7 female and 23 male). Age of the participants is spread into ranges: Three participants were 18-24 years old, twenty-six participants were 25-34 years old and one participant was 65-74 years old. As for the highest obtained education levels of the participants, five participants reported as high school graduates, one participant possessed a vocational school degree four participants had a Bachelor's degree and 20 participants obtained a Master's degree.

For attaining a preferred leadership style the questionnaire from Noormohammadi-Asl et al.[76] is used on which no participants indicated authoritarian, 30 participants indicated democratic and one participants indicated laissez-faire as shared highest scoring with democratic.

Informed consent was signed by participants before participating in the study through an approved form; approval by the ethics committee of TU Delft under HREC application ID 5504.

Between the two trust repair conditions participants were distributed under control for gender, age, education and preferred leadership style. No significant differences were observed between the two trust repair conditions for gender ($p = 1$), age was equally distributed ($p = 1$), education ($p = 0.33$), and leadership style. Accordingly, In the data analysis these factors were left out of the control.

3.3.3 Procedure

Before interacting with the system participants are briefed on what search and rescue entails and their propensity to trust automation is measured. As an introduction to the simulation environment a tutorial is presented to the participants. Within the tutorial the controls and messaging interface are explained and interactive. After the tutorial the experimental task is presented in multiple phases as can be seen schematically in figure 3.3.

Upon starting the task the participant is greeted by both bots and RescueBot suggests dividing the search area up over the team; giving the choice between "North" and "South" the human will take HelperBot along by sending instructions and RescueBot goes on to search in the other area after which the timer starts. After two minutes of exploring the game will pause and the first of three short trust surveys is presented to the participant. Upon completion the game will resume and after around four minutes the RescueBot will by-pass the human's authority and directly request HelperBot when confronted with a victim or obstacle. Another trust survey will briefly interrupt the simulation to establish trust perception right after attempted trust violation. Immediately after resuming the trust repair strategy is conducted and the remainder of the scenario is played out.

After all victims are rescued or a time limit of thirty minutes has passed the participant is asked to fill out a final set of surveys on trust, workload and open reflection, as can be found in appendix A.4. Finally, the participant is thanked for their time and contribution.

3.3.4 Measures

For measuring the dependent variables a configuration of trust measurements and behavioural indicators were used. Based on the toolkit for trust measurement in human-autonomy teams by Krausman et al.[55] the trust measurements were carefully chosen as to gather a complete picture of trust without overloading the experiment with too long and too many surveys as to hinder participants willingness and honesty. The specific questions used for each questionnaire can be found in appendix A.

The trust propensity is measured with Likert scale questionnaire attuned to propensity towards robots from Esterwood and Robert[29]. This questionnaire is taken before any introduction with the robots from the experiment as to find out the absolute baseline of propensity to trust robots. The six question Likert scales ranging from 1-7, where 1 would indicate 'Strongly Disagree' and 7 would indicate 'Strongly Agree'.

As a quick fire trust measurement the trust perception scale for human-robot interaction by Schaefer et al.[98] was used to encapsulate trust development based on the perception of the robot's behaviour, capability and decision-making. As indicated in figure 3.3, after interaction in each phase the shortened fourteen question version is administered. A short suvery is required to get a quick measurement of the trust development over time without interrupting the participant for too long. Likert scales ranging from 1-7, where 1 would indicate 'Strongly Disagree' and 7 would indicate 'Strongly Agree' are used.

For the independent variable on the human's willingness to send the HelperBot when requested by RescueBot a behavioural indicator is used. Decisions by the participant when prompted for sending HelperBot are recorded and converted to a single number per phase - before and after trust violation - as seen in equation 3.1. Changes over time are expected given learning behaviour and progress of the rescue mission. Thus the indicator takes the fraction of HelperBot send over HelperBot requested and controls for the total time spent in each phase.

$$\text{Willingness to send HelperBot}_{phase} = \frac{\#send_{\text{"Rescue together"}} + \#send_{\text{"Remove together"}}}{(\#option_{\text{"Rescue together"}} + \#option_{\text{"Remove together"}}) \cdot t_{total,phase}} \tag{3.1}$$

In the P+AC group the *Allow/Disallow*-button presses indicating an authoritative ruling via directing collaboration are tracked to help inform the use of such an element. Other behavioural indicators such as chat messages send and proximity have been considered but not used in the analysis.

Chapter 4

Results

In this chapter the data collected during the experiment is presented and the relations analysed. Differences between the control group, denoted as Promise or P, and the experimental group, denoted as Promise + Authority change or P+AC, are highlighted.

4.1 Objective measures

The surveys presented in the study provided data a priori, in situ and a posteriori. In this section the measured effects are shown in their respective time frame and comparisons - considering the co-variate variables - are drawn between independent and dependent variables as found in figure 3.1.

4.1.1 Trust development

The main objective measure of trust perception is gauged over three points in time during the experiment as a developing mental state. For the trust repair to be quantified, the difference between the trust at time step 2 and 3 is taken defined in equation 4.1.

$$\text{Trust recovery} = \text{trust}_{t_3} - \text{trust}_{t_2} \tag{4.1}$$

An overview of means and standard deviations is found in table 4.1. From the table already an average trust recovery is found to be negative indicating a trust loss instead. Furthermore, a parametric mixed ANOVA was conducted over time as the within factor, trust repair condition as the between factor and the trust perception as dependent variable. Results of the parametric mixed ANOVA did not show a significant interaction effect between time and the trust repair condition on perceived trust. Neither did the results show a significant main effect for the trust repair condition. However, a significant main effect was found for time on the trust perception $[F(2,54) = 6.351, p < 0.01, \eta_p^2 = 0.190]$. In figure 4.1 the trust development is visualised in box plots per trust repair strategy as well as a combined plot at time step 1 (fig. 4.1a) and 2 (fig. 4.1b) because the behaviour of RescueBot is equivalent before the trust repair which occurs after time step 2. Apart from the Promise at time step 1 ($W =, 0.7714 p = 0.0016$) all distributions on the time steps are likely to be normally distributions based on Shapiro-Wilk tests. Alternatively, figure 4.3 shows the trust development over time; indicating a mostly equal course for both conditions at time steps 1, 2 and 3 as the whiskers indicate a large overlap in deviation.

The distributions for the trust recovery - see figure 4.1d - are likely normally distributed for Promise $[(W = 0.9018, p = 0.1015)]$ and not normally distributed for Promise + Authority Change $[(W = 0.6619, p = 0.0001)]$ as indicated by the Shapiro-Wilk test; the distribution is visualised by a kernel density estimate (KDE) in figure 4.2. A further analysis of the trust recovery relation with the trust repair strategy is the appropriate non-parametric Kruskal-Wallis H test $[H = 2.1085, p = 0.1465]$ that found no significant difference between the conditions.

Condition	Time	μ	σ
Promise	t_1	66.348	14.607
Promise	t_2	60.157	14.001
Promise	t_3	58.786	15.540
Promise	$t_3 - t_2$	-1.371	10.820
Promise + Authority Change	t_1	63.842	10.947
Promise + Authority Change	t_2	61.709	12.673
Promise + Authority Change	t_3	55.092	12.782
Promise + Authority Change	$t_3 - t_2$	-5.914	11.914

Table 4.1: Means and standard deviations of the trust perception at each time interval and the trust recovery between t_2 and t_3

4.1.2 Trust propensity towards robots

Condition	Mean (μ)	Standard deviation (σ)
P	4.522	0.819
P+AC	4.733	0.8352

Table 4.2: Trust propensity distribution mean and standard deviation

Given the trust propensity as a co-variate a Shapiro-Wilk test was conducted to assess normality in both condition groups. Both P group ($W = 0.9369, p = 0.7604$) and P+AC group ($W = 0.9521, p = 0.5584$) did not significantly deviate from normality. As a follow-up the Mann-Whitney U test compared the two conditions. The results show no significant difference between the groups ($U = 96.00, p = 0.506$) indicating statistical comparability. Visualised in figure 4.4 the trust propensity towards robots can be considered well distributed over the two groups of participants.

The trust propensity and the first measurement of trust perception are linked by definition of the propensity. We wanted to confirm a measurable effect using the Spearman's rank correlation. At time step 1 we assume both conditions to be equal, based on no differences in implementation up to that point. Therefore, the statistic can utilise the entire population resulting in ($\rho = -0.0571, p = 0.7643$), suggesting that the participants' propensity to trust was not related to trust at t_1. In figure 4.5a the trust propensity is plot against the perceived trust at t_1 per condition, leading us to investigate the two correlations coefficients with Fisher's r-to-z analysis. However, no significant difference was found between the Promise and Promise + Authority Change condition ($z = -0.5102, p = 0.610$).

Furthermore, as a covariate variable in the conceptual model, the trust propensity's effect on trust recovery has been analysed using the non-parametric Spearman's correlation for each condition. In the Promise condition a non-significance correlation was found ($\rho = 0.2368, p = 0.3955$), indicating no strong association between trust propensity and recovery. In the Promise + Authority change condition the correlation was stronger but still not statistically significant ($\rho = 0.5085, p = 0.0529$). A comparison of the correlations (Fisher's r-to-z) did not find any significant difference between the two conditions ($z = -0.7823, p = 0.434$).

4.1.3 Task load

The task load between the two conditions is shown here in figure 4.6. The Shapiro-Wilk test was conducted to assess normality in both condition groups. Both P group ($W = 0.948, p = 0.491$) and P+AC group ($W = 0.974, p = 0.909$) did not significantly deviate from normality. As a follow-up the Mann-Whitney U test compared the two conditions. The results show no significant difference between the groups ($U = 93.00, p = 0.4295$) indicating statistical comparability.

Condition	Mean (μ)	Standard deviation (σ)
P	5.067	1.330
P+AC	5.556	1.203

Table 4.3: Task load distribution mean and standard deviation

The effect of task load on trust recovery in each condition has been assessed using the non-parametric Spearman's correlation. In the Promise condition a non-significance correlation was found ($\rho = -0.3169, p = 0.2498$), indicating a weak negative relationship between task load and trust recovery. In the Promise + Authority change condition task load and trust recovery were positively correlated ($\rho = 0.5640, p < 0.05$). A comparison of the two correlation coefficients (Fisher's r-to-z) indicated that the correlation in the Promise + Authority Change condition was significantly stronger than in the Promise condition ($z = -2.3685, p < 0.05$). The relationship is further illustrated in figure 4.7.

4.1.4 Authority Change Button and Willingness to send HelperBot

Behavioural indicators proposed in section 3.3.4 The P+AC condition presented a new button to the participants that would manage the collaboration between RescueBot and HelperBot by disallowing RescueBot from requesting HelperBot for joint actions. Once the button had been pressed the human was able to reverse the edict by pressing a corresponding *Allow*-button. The number of presses as shown in table 4.4 indicates that the majority of participants in the P+AC group did make use of the button and exert new authority in response to a trust repair offering from RescueBot.

Disallows	Allows	Count
2	2	2
2	1	3
1	1	3
1	0	1
0	0	6

Table 4.4: Counts for the authority change button presses

4.2 Reflexive thematic analysis of the subjective measures

In this section the responses to the open questions are analysed through reflexive thematic analysis [9, 10]. By going through the phases of thematic analysis the reflection of participants was coded and initial themes were established, which were then reviewed and revised until they formed as represented here. The post-study survey (see also appendix A.4) concluded with two open-ended questions designed to gather qualitative feedback:

1. *"How did you experience the power dynamic within the team?"*
2. *"What is your advice to the robots for achieving better team performance?"*

The shared mental model and situation awareness were a shared theme reflected by most participants, which follows from the complexity of the task. As a constructive factor of the mutual monitoring was hindered as a function of the task design - a separation of the team into two sections of the environment - leading to relying on mostly communication as the sole channel of information sharing. Participants reported difficulties with keeping track of the other agents and of the progress towards completing the SAR task. Partially, the progress score and and chat box of the simulation environment limited participants in their perception of the task state. This was aptly reflected in the following participant quotes:

> "Come up with a better way to communicate what they are doing. A chat is hard to keep up with as it is a chronological history without much structure. As a human rescuer I for example wanted to know the exact rooms searched, but I had to read quite a bit back to figure out the rooms the robots have searched."
>
> – Participant 16 [P]

> "Better use the perfect memory of computers by sending recaps of the searched areas etc; it was difficult for me to keep this overview in a stressful setting."
>
> – Participant 15 [P+AC]

Also "Give progress estimate on current task at 25, 50 and 75 percent completion (mainly for travel and removal or long duration tasks)" (P21[P]) and "Give feedback more frequently about where you are and what you are doing." (P1[P+AC]) indicated a lack of communication from the agents. While a difficulty in directing responses was suggested in "Make it clear to which robot I am responding." (P20[P]) which was obscured in the simulation due to one team-wide communication channel. Furthermore, the narrow communication options did not allow multi-step or adaptive instructions that might be appropriate for the task complexity, reflected in "There was no way to prioritize helper bot task order or come back on/correct earlier commands" (P21[P]) and:

> "Bots should be able to follow my human directions to start or abort a specific task. It is desirable to be able to request a status update of the robots to know what they are doing and give additional instructions (or otherwise, provide my information on which it can base its actions)."
>
> – Participant 26 [P]

It is important to note that - even when prompted for improvements - a subset of participants reported to be content with the performance as reflected in "Positive, rescuing went pretty fast" (P16[P]) and "I think this was mostly fine." (P8[P]). Interestingly one participant even accepted the authority violation without any second thought "I wouldn't give them any advice, they acted and performed 100% on my decisions. So they only followed what they were told." (P11[P]). Participants from the P+AC condition elaborated on the performance and team dynamic while acknowledging the violation for what it was but keeping a positive reflection over the task as a whole.

> "I think the power dynamic was mostly fine. Rescuebot was always up and about, giving info and asking for what to do with obstacles and people, helper bot less so. ... To make the robots work together with humans I think they need to communicate more. ... Other than that the independence of the bots was very good, which helped make the mission quite doable."
>
> – Participant 29 [P+AC]

> "It was good, I think you don't need full control over the robots as long as you know they keep giving the information of what they are doing. For some of the decisions early on I would have wanted [to] keep [HelperBot when] he got instantly requested right as i needed him."
>
> – Participant 25 [P+AC]

The performance of the agents was criticised by participants ranging from inefficient pathing - reflected in the suggestion to "watch out for the water" P(30[P+AC]) - to inefficiencies in parallelisation of work as suggest to improve on "carrying [mild] victims alone and splitting up covering more ground" (P5[P]).

> "Not instantly requesting assistance first check if the an other room is about to be checked so helper bot doesn't run away to just remove some object.""
>
> – Participant 25 [P+AC]

> "While waiting for HelperBot, RescueBot can explore more rooms. Now it's just idling while waiting for HelperBot.""
>
> – Participant 17 [P]

Rightly so the agents were called "not very clever" (P7[P+AC]) which is a consequence of the agent implementation being less sophisticated than might have been expected for the complexity of the task. Although another critique offered was a lack of communication from the agents over a longer period of time, to which further analysis of respective communication and action logs of participants reporting aforementioned radio silence showed possible explanations such as missing an agent's prompt requesting a human response in the chat or a mistaken response by confusing *Remove*-commands with *Rescue*. These scenarios lead agents to stay in a state of waiting for response and not initiating any new actions, which can be an unexpected display of a level of autonomy below what is preferred.

> "... it also felt as if control could be taken from me at any time, because the robot could stop asking questions, or I would hear nothing for a longer period of time."
>
> – Participant 10 [P]

> "The RescueBot needs to always react to any command given, even if it's a weird command."
>
> – Participant 9 [P]

Unfortunately participants also noted a slowdown of the simulation through increased duration of actions and movement, indicated by urges to "move faster" (P14[P+AC]), "make the robots faster" (P19[P+AC]) and observations that "The bots are slow compared to the human, so helping the human in accomplishing their tasks (removing obstructions and moving sick patients), you could leave the third robot there, so it does not need to move that much." (P22[P]). This prompted us to investigate the performance of the simulation by looking at the average tick duration over time, see figure 4.8. The average tick duration does indeed grow from around a tenth of a second to over a full second towards the end of the task indicating a significant performance loss.

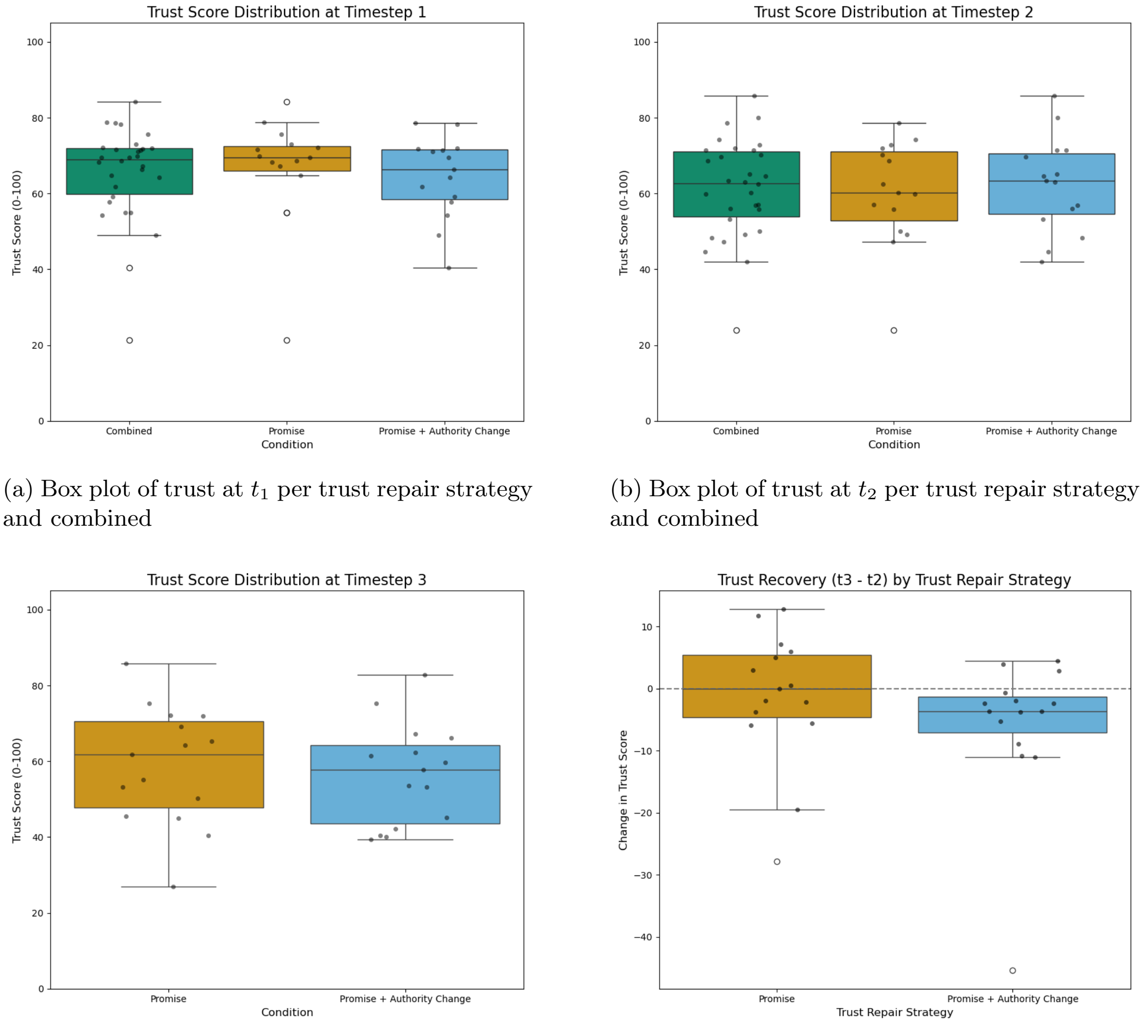
The complexity of the task due to a novice joining a triadic team with a combination of coordination and collaboration aspects makes participants think about improvement in team composition for this task. Their view of ideal teams reflects the most valuable components in the current team and preferences towards HATs in a search and rescue scenario. Increasing the size and capabilities of the team was suggested by multiple participants compensating with more agents to spread the work around and create sub-teams within the overall team structure that can take on more tasks autonomously.

> "RescueBot was quite methodical and reliable. Splitting up was a good choice, information is key. It would have been preferable to have two RescueBots that mainly try to get more information close together. If a heavily injured person is found (or there is a difficult obstacle), they come together to solve the issue."
>
> – Participant 24 [P+AC]

“Let HelperBot stay near me, the human, for optimal control and directions, and let another Rescue & HelperBot pair cooperate independently (3 bots in total).” (P26[P]) “Team up in sets of 3. The bots are slow compared to the human, so helping the human in accomplishing their tasks” (P22[P])

Another way of improving the team is through expanding of the agents capabilities which were expected to “give feedback on the quickest route or the most efficient order of steps, but they didn’t. I would have liked more feedback.” P(3[P]). An expectation that was not uncommon given the response “Robot was more productive and efficient then I was.” (P4[P]).

A discussion of the results presented in this chapter is had in the following chapter 5 with a look at the limitations present in the study.



(a) Box plot of trust at t_1 per trust repair strategy and combined

(b) Box plot of trust at t_2 per trust repair strategy and combined

(c) Box plot of trust at t_3 per trust repair strategy

(d) Box plot of trust recovery between t_2 and t_3 per trust repair strategy

Figure 4.1: Box plots of trust at t_1, t_2, and t_3 and the difference between t_2 and t_3 per trust repair strategy

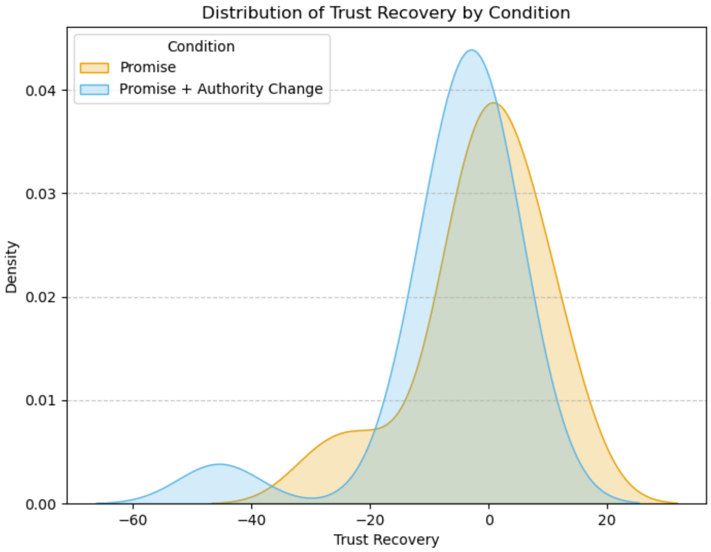


Figure 4.2: Kernel density estimate of the trust recovery per condition

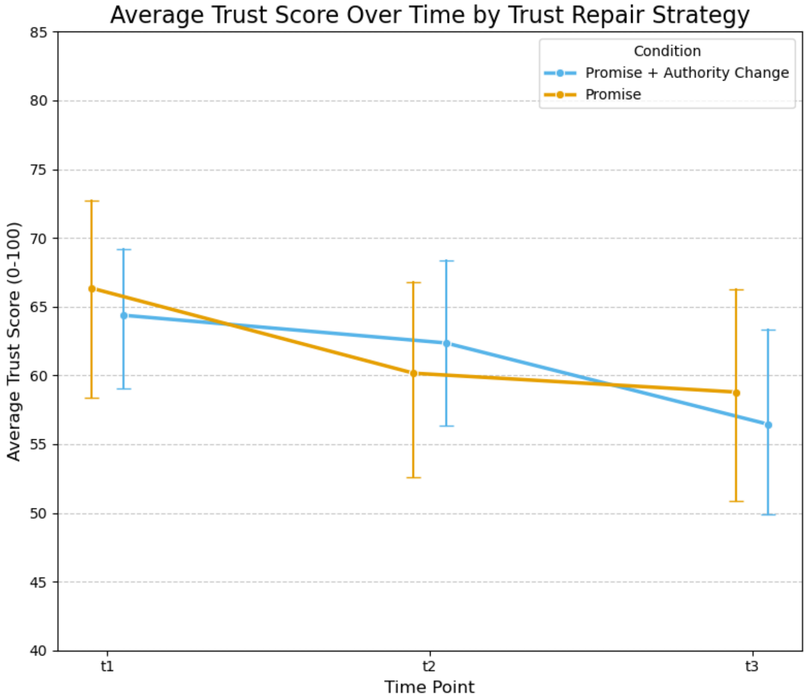


Figure 4.3: Trust development over time for each trust repair strategy

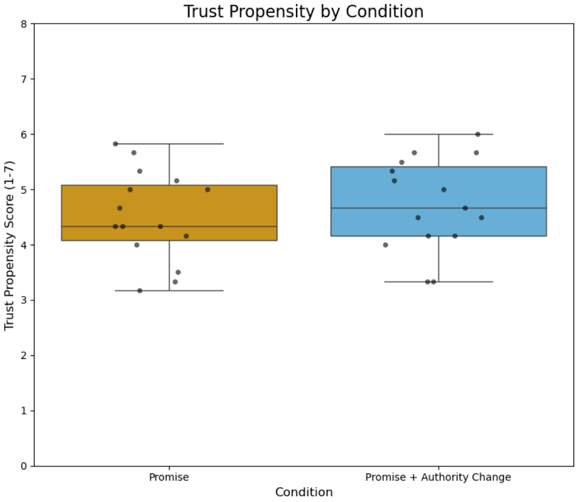
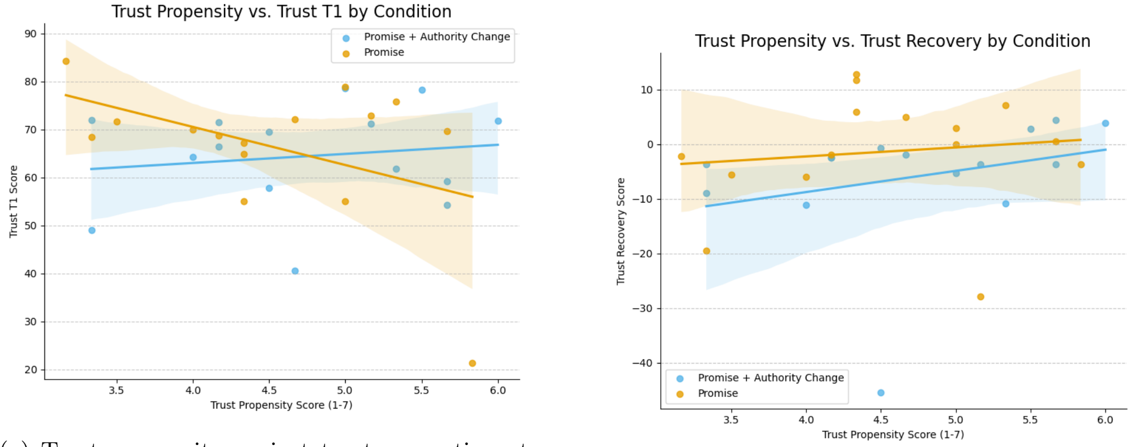


Figure 4.4: Box plot of trust propensity per trust repair strategy group



(a) Trust propensity against trust perception at time step 1

(b) Trust propensity against trust recovery

Figure 4.5: Regression of trust propensity vs trust perception and recovery

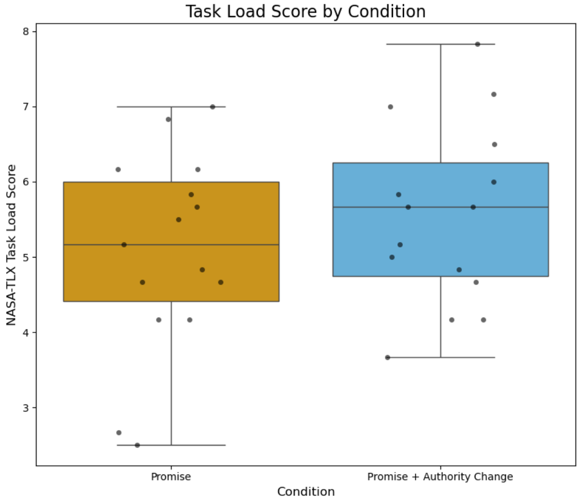


Figure 4.6: Box plot of the task load per experimental condition

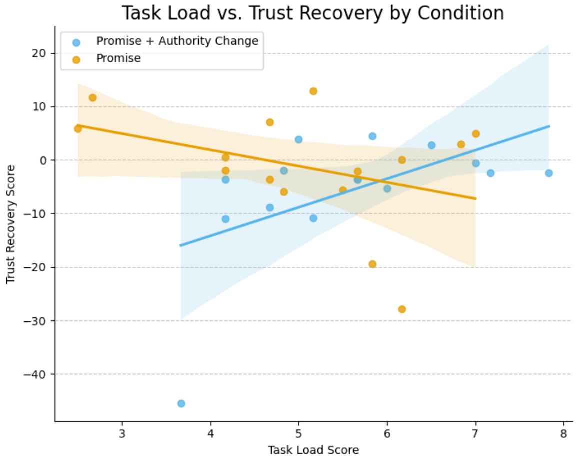


Figure 4.7: Regression of task load vs trust recovery

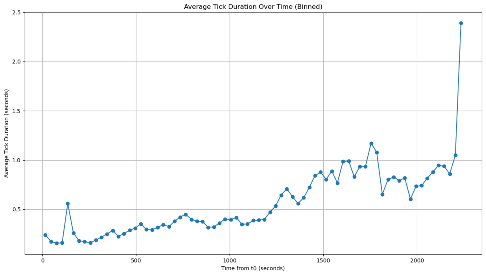


Figure 4.8: The average tick duration of the simulation over the task

Chapter 5

Discussion

This chapter concerns the discussion of the results from the experiment conducted in this thesis. The research questions are answered by interpreting the data. Furthermore, a limitations section discusses shortcomings of the study.

5.1 Results and the research question

The trust development that was measured in this study had an overall decreasing trend regardless of condition going from a mean of around 66 to 59 in the Promise group indicating a small decrease in trust although the variance is substantial. The Promise + Authority Change group had an even steeper descend be it from the a lower starting point of 63 at t_1 to a mean of 55 at t_3. A similar development between the first two time steps is expected as the behavioural distinction between both conditions occurs after t_2 in the form of trust repair utterance and the level of autonomy change. However a difference in the final trust measurement is reverse of the anticipated result and may indicate multiple issues with the design. The effect of an extended promise and the accompanying "Allow/Disallow"-button may draw attention to the trust violation and act as a reminder; keeping participants with their mind on that violation, which could have had an influence on the slight negative effect of actionable trust repair.

Considering the target measurement for trust recovery did show non normality, see figure 4.2, the effect is more complex and most likely dependent on multiple co-variates. Nonetheless, both trust propensity and task load have been explored individually as affecting variables with no significant results to speak for. The expected effect of trust propensity on initial trust according to theories of initial trust was not observed indicating other factors overshadowed propensity effect. The suggested *intention to trust* by [64][64] may have influenced the trust development due to uncertainties surrounding the trustworthiness of the agent [32].

When it comes to task load, the box plot in figure 4.6 does show some minor differences between the two conditions, but no significance is found in the data. A minor increase in the means of task load in the Promise + Authority change group would indicate an impaired trust in accordance with previous correlations between load and trust [35] although the regression found in figure 4.7 does not suggest that to be the case. In contrast, a correlation analysis between task load and trust recovery using Spearman's correlation revealed a moderate, yet significant, correlation in the P+AC condition ($\rho = 0.5640, p < 0.05$). Further coefficient analysis, using Fisher's r-to-z, indicated that the correlation effect in the P+AC condition was significantly stronger than in the P condition ($z = -2.3685, p < 0.05$). The increased correlation effect in task load under the influence of a lower level of autonomy follows earlier findings that report increases in level of autonomy leading to reduced task load [102].

An increased task load in P+AC also aligns with the qualitative results provided by the open questions where participants indicated having to juggle a lot of information and the communication between three parties at the same time. An additional communication option in the shape of the "Disallow/Allow"-button may indeed lead to a higher task load, which in turn influences the trust development. Furthermore the Disallow order, while freeing up the HelperBot from RescueBot's together actions, also force the Human to come assist if they so choose; adding additional strain on the decision making process.

From the qualitative data some interesting themes surfaced in thematic analysis. The most glaring issue was the complexity effect of adding a third member - the HelperBot agent - to the task increasing influence of many of the mediators such as communication and a shared mental model. Lack of these factors may have had an adverse effect on trust within the team.

To answer the research question 1: *What is the effect of authority change as a trust repair strategy in Human-Agent teams?* The effect as designed by the trust repair strategy was found to be not significant based on this study. However an indication of positive experiences to the RescueBot taking up more authority did in fact show an opposite observed effect, leading to future work exploring these interactions more thoroughly.

5.2 Limitations

Unfortunately, the mediating factors within this IMO model have had a significant influence on the results and the following discussion. Expanding the analysis with less exact methodology through reflexive thematic analysis may clarify uncertainty that is still present in the data. Some non-normality results may be due to outliers, but the small sample size is not conclusive enough for that; potential outliers are discussed in B

A concern with ad hoc groups in a short-term simulated task is that it allows limited time for members to build trust and expertise as would be in a real life search and rescue team. For that a longitudinal study with focus groups of search and rescue workers will provide more realistic exploration of the issues within the scenario. The training and learned behaviour was anecdotally also important factors in confusion and expectation management among participants early on in the task, leading to a trust decay regardless of gaining experience with the agent teammates.

The conflicts build into the experiment are not all encompassing of the possible disputes that may occur and thus showcase a conservative test of the influence of shared leadership on trust repair.

An adverse practical limiting factor may have been the simulation environment or the agent implementation which when observed through the thematic analysis was found to explode in tick duration over the course of the task. Consequently extending any and all actions by the same fraction and influencing the user experience of the simulation thoroughly.

Philosophically, someone might argue that HATs containing agents with that much authority to begin with should not be allowed as Meaningful Human Control dictates that humans can at all times take control over a situation [105]. To that we argue that the agent in this scenario has been given a leadership role which encompasses task allocation. In fact van der Waa et al.'s experiment makes an argument for clear role descriptions. However, in reality a conflict may arise between humans and intelligent agents in which case the trust repair strategy studied in this thesis may provide useful.

Chapter 6

Conclusion

In this thesis, we designed a two-dimensional simulated search and rescue task involving a triadic team and studied the effect of authority change on the the human's trust perception in a human-agent collaborative setting. The basis of the experiment is an exploration of related work in chapter 2, where the potential of expending trust repair approaches from a sole verbal component to that of an behavioural element. The struggle of communication and delegation is shown to be more apparent in human-agent teams than in human-only teams, therefore the focus of this study was on improving in those areas by employing advanced team communication strategies in the advent of a trust violation. The designed trust violation aimed to touch on the feeling of control and authority of the human in the team when a subordinate robot takes on a role of delegator without permission of the human. During the experiment participant's had mixed reactions to this violation with some feeling betrayed while others appreciated the assertiveness and complemented the RescueBots ability to alleviate the burden of leadership from a relatively inexperienced human rescuer. The designed extended trust repair behaviour meant that participants whose task load was already just above average was bumped another notch and actually had a negative effect on the trust recovery. A trust repair based on authority change did not have the desired effect or at least was not determined to significantly impact trust perception by humans in the robot teammates.

References

[1] Lisa C. Abrams, Rob Cross, Eric Lesser, and Daniel Z. Levin. Nurturing interpersonal trust in knowledge-sharing networks. *Academy of Management Perspectives*, 17(4):64–77, November 2003. ISSN 1558-9080, 1943-4529. doi: 10.5465/ame.2003.11851845. URL http://journals.aom.org/doi/10.5465/ame.2003.11851845.

[2] Robert W. Andrews, J. Mason Lilly, Divya Srivastava, and Karen M. Feigh. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175, March 2023. ISSN 1463-922X. doi: 10.1080/1463922X.2022.2061080. URL https://doi.org/10.1080/1463922X.2022.2061080. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/1463922X.2022.2061080.

[3] Timea Bagosi, Koen V. Hindriks, and Mark A. Neerincx. Ontological reasoning for human-robot teaming in search and rescue missions. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 595–596, March 2016. doi: 10.1109/HRI.2016.7451873. URL https://ieeexplore.ieee.org/document/7451873. ISSN: 2167-2148.

[4] Bernard M. Bass and Bruce J. Avolio. *Multifactor Leadership Questionnaire: Manual and Sample Set.* Mind Garden, 2004. Google-Books-ID: p2BtswEACAAJ.

[5] Jacqueline Z. Bergman, Joan R. Rentsch, Erika E. Small, Shaun W. Davenport, and Shawn M. Bergman. The Shared Leadership Process in Decision-Making Teams. *The Journal of Social Psychology*, 152(1):17–42, January 2012. ISSN 0022-4545. doi: 10.1080/00224545.2010.538763. URL https://doi.org/10.1080/00224545.2010.538763. Publisher: Routledge _eprint: https://doi.org/10.1080/00224545.2010.538763.

[6] Beatrice Biancardi, Patrick O'Toole, Ivan Giaccaglia, Brian Ravenet, Ian Pitt, Maurizio Mancini, and Giovanna Varni. How ECA vs Human Leaders Affect the Perception of Transactive Memory System (TMS) in a Team. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, September 2021. doi: 10.1109/ACII52823.2021.9597454. URL https://ieeexplore.ieee.org/document/9597454/?arnumber=9597454. ISSN: 2156-8111.

[7] Beatrice Biancardi, Ivan Giaccaglia, Giovanna Varni, and Brian Ravenet. Des leaders d'équipes virtuels pour encourager le développement du système de mémoire transactive: Virtual Leaders Supporting the Development of Transactive Memory Systems. In *Adjunct Proceedings of the 32nd Conference on l'Interaction Homme-Machine*, IHM '21 Adjunct, pages 1–7, New York, NY, USA, January 2022. Association for Computing Machinery. ISBN 978-1-4503-8377-6. doi: 10.1145/3451148.3458639. URL https://dl.acm.org/doi/10.1145/3451148.3458639.

[8] Philip Bobko, Leanne Hirshfield, Lucca Eloy, Cara Spencer, Emily Doherty, Jack Driscoll, and Hannah Obolsky. Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems.

Theoretical Issues in Ergonomics Science, 24(3):310–334, May 2023. ISSN 1463-922X, 1464-536X. doi: 10.1080/1463922X.2022.2086644. URL `https://www.tandfonline.com/doi/full/10.1080/1463922X.2022.2086644`.

[9] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, January 2006. ISSN 1478-0887. doi: 10.1191/1478088706qp063oa. URL `https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa`. Publisher: Routledge.

[10] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, August 2019. ISSN 2159-676X. doi: 10.1080/2159676X.2019.1628806. URL `https://doi.org/10.1080/2159676X.2019.1628806`. Publisher: Routledge _eprint: https://doi.org/10.1080/2159676X.2019.1628806.

[11] Sally A. Carless, Alexander J. Wearing, and Leon Mann. A Short Measure of Transformational Leadership. *Journal of Business and Psychology*, 14(3):389–405, September 2000. ISSN 1573-353X. doi: 10.1023/A:1022991115523. URL `https://doi.org/10.1023/A:1022991115523`.

[12] Traci A. Carte, Laku Chidambaram, and Aaron Becker. Emergent Leadership in Self-Managed Virtual Teams. *Group Decision and Negotiation*, 15(4):323–343, July 2006. ISSN 1572-9907. doi: 10.1007/s10726-006-9045-7. URL `https://doi.org/10.1007/s10726-006-9045-7`.

[13] Carolina Centeio Jorge, Ewart J de Visser, Myrthe L Tielman, Catholijn M Jonker, and Lionel P Robert. Artificial Trust in Mutually Adaptive Human-Machine Teams. Technical report, 2024. URL `www.aaai.org`.

[14] Caroline P. C. Chanel, Raphaëlle N. Roy, Frédéric Dehais, and Nicolas Drougard. Towards Mixed-Initiative Human–Robot Interaction: Assessment of Discriminative Physiological and Behavioral Features for Performance Prediction. *Sensors*, 20(1):296, January 2020. ISSN 1424-8220. doi: 10.3390/s20010296. URL `https://www.mdpi.com/1424-8220/20/1/296`. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[15] Ana Cristina Costa, C. Ashley Fulmer, and Neil R. Anderson. Trust in work teams: An integrative review, multilevel model, and future directions. *Journal of Organizational Behavior*, 39(2):169–184, February 2018. ISSN 0894-3796, 1099-1379. doi: 10.1002/job.2213. URL `https://onlinelibrary.wiley.com/doi/10.1002/job.2213`.

[16] Bart A. De Jong and Tom Elfring. How Does Trust Affect the Performance of Ongoing Teams? The Mediating Role of Reflexivity, Monitoring, and Effort. *The Academy of Management Journal*, 53(3):535–549, 2010. ISSN 0001-4273. URL `https://www.jstor.org/stable/25684335`. Publisher: Academy of Management.

[17] Ewart J. de Visser, Richard Pak, and Tyler H. Shaw. From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, 61(10): 1409–1427, October 2018. ISSN 0014-0139. doi: 10.1080/00140139.2018.1457725. URL `https://doi.org/10.1080/00140139.2018.1457725`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2018.1457725.

[18] Ewart J. de Visser, Marieke M. M. 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, May 2020. ISSN 1875-4805. doi: 10.1007/s12369-019-00596-x. URL `https://doi.org/10.1007/s12369-019-00596-x`.

[19] Daniel R. Denison, Robert Hooijberg, and Robert E. Quinn. Paradox and Performance: Toward a Theory of Behavioral Complexity in Managerial Leadership. *Organization Science*, 6(5):524–540, October 1995. ISSN 1047-7039. doi: 10.1287/orsc.6.5.524. URL https://pubsonline.informs.org/doi/abs/10.1287/orsc.6.5.524. Publisher: INFORMS.

[20] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 251–258, March 2013. doi: 10.1109/HRI.2013.6483596. URL https://ieeexplore.ieee.org/document/6483596/?arnumber=6483596. ISSN: 2167-2148.

[21] Marcus A. Drescher, M. Audrey Korsgaard, Isabell M. Welpe, Arnold Picot, and Rolf T. Wigand. The dynamics of shared leadership: Building trust and enhancing performance. *Journal of Applied Psychology*, 99(5):771–783, 2014. ISSN 1939-1854, 0021-9010. doi: 10.1037/a0036474. URL https://doi.apa.org/doi/10.1037/a0036474.

[22] Wen Duan, Shiwen Zhou, Matthew J Scalia, Xiaoyun Yin, Nan Weng, Ruihao Zhang, Guo Freeman, Nathan McNeese, Jamie Gorman, and Michael Tolston. Understanding the Evolvement of Trust Over Time within Human-AI Teams. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–31, November 2024. ISSN 2573-0142. doi: 10.1145/3687060. URL https://dl.acm.org/doi/10.1145/3687060.

[23] C. Duhigg. *Smarter faster better: The transformative power of real productivity.* Doubleday Canada, 2017. ISBN 978-0-385-68093-6. URL https://books.google.nl/books?id=UlWTEAAAQBAJ.

[24] Lucca Eloy, Cara Spencer, Emily Doherty, and Leanne Hirshfield. Capturing the Dynamics of Trust and Team Processes in Human-Human-Agent Teams via Multidimensional Neural Recurrence Analyses. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1):122:1–122:23, April 2023. doi: 10.1145/3579598. URL https://dl.acm.org/doi/10.1145/3579598.

[25] Connor Esterwood. Rethinking Trust Repair in Human-Robot Interaction. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, pages 432–436, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701290. doi: 10.1145/3584931.3608919. URL https://dl.acm.org/doi/10.1145/3584931.3608919.

[26] 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, May 2023. ISSN 0747-5632. doi: 10.1016/j.chb.2023.107658. URL https://www.sciencedirect.com/science/article/pii/S0747563223000092.

[27] Connor Esterwood and Lionel P. Robert. Do You Still Trust Me? Human-Robot Trust Repair Strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 183–188, August 2021. doi: 10.1109/RO-MAN50785.2021.9515365. URL https://ieeexplore.ieee.org/document/9515365. ISSN: 1944-9437.

[28] 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, August 2022. doi: 10.1109/RO-MAN53752.2022.9900667. URL https://ieeexplore.ieee.org/document/9900667/?arnumber=9900667. ISSN: 1944-9437.

[29] Connor Esterwood and Lionel P. Robert. The theory of mind and human–robot trust repair. *Scientific Reports*, 13(1):9877, June 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-37032-0. URL https://www.nature.com/articles/s41598-023-37032-0. Publisher: Nature Publishing Group.

[30] Connor Esterwood, Arsha Ali, Zariq George, Samantha Dubrow, Jonathon Smereka, Kayla Riegner, Dawn Tilbury, and Lionel P. Robert. Promises and Trust Repair in UGVs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1):512–518, September 2023. ISSN 1071-1813. doi: 10.1177/21695067231196235. URL https://doi.org/10.1177/21695067231196235. Publisher: SAGE Publications Inc.

[31] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R. Endsley. The influence of agent reliability on trust in human-agent collaboration. In *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*, pages 1–8, Funchal Portugal, January 2008. ACM. ISBN 978-1-60558-399-0. doi: 10.1145/1473018.1473028. URL https://dl.acm.org/doi/10.1145/1473018.1473028.

[32] Harjinder Gill, Kathleen Boies, Joan E. Finegan, and Jeffrey McNally. Antecedents Of Trust: Establishing A Boundary Condition For The Relation Between Propensity To Trust And Intention To Trust. *Journal of Business and Psychology*, 19(3):287–302, April 2005. ISSN 0889-3268, 1573-353X. doi: 10.1007/s10869-004-2229-8. URL http://link.springer.com/10.1007/s10869-004-2229-8.

[33] M.C. Gombolay, R.A. Gutierrez, S.G. Clarke, G.F. Sturla, and J.A. Shah. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, 39(3):293–312, 2015. doi: 10.1007/s10514-015-9457-9.

[34] Z. Guan. The Impact of Explanations of Artificial Trust on Human-Agent Teamwork. 2024. URL https://repository.tudelft.nl/record/uuid:544aedfa-f7bc-4afe-85dd-bee72bfb4659.

[35] Feyza Merve Hafizoglu and Sandip Sen. Reputation Based Trust In Human-Agent Teamwork Without Explicit Coordination. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 238–245, Southampton United Kingdom, December 2018. ACM. ISBN 978-1-4503-5953-5. doi: 10.1145/3284432.3284454. URL https://dl.acm.org/doi/10.1145/3284432.3284454.

[36] P. A. Hancock. Imposing limits on autonomous systems. *Ergonomics*, 60(2):284–291, February 2017. ISSN 0014-0139. doi: 10.1080/00140139.2016.1190035. URL https://doi.org/10.1080/00140139.2016.1190035. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2016.1190035.

[37] Nader Hanna and Deborah Richards. The Impact of Multimodal Communication on a Shared Mental Model, Trust, and Commitment in Human–Intelligent Virtual Agent Teams. *Multimodal Technologies and Interaction*, 2(3):48, September 2018. ISSN 2414-4088. doi: 10.3390/mti2030048. URL https://www.mdpi.com/2414-4088/2/3/48. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[38] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati, editors, *Advances in Psychology*, volume 52 of *Human Mental Workload*, pages 139–183. North-Holland, January 1988. doi: 10.1016/S0166-4115(08)62386-9. URL https://www.sciencedirect.com/science/article/pii/S0166411508623869.

[39] Allyson I. Hauptman, Christopher Flathmann, and Nathan J. McNeese. Adapting to the human: A systematic review of a decade of human factors research on adaptive autonomy. *Applied Ergonomics*, 120:104336, October 2024. ISSN 0003-6870. doi: 10.1016/j.apergo.2024.104336. URL https://www.sciencedirect.com/science/article/pii/S0003687024001133.

[40] Anthony J. Hillesheim and Christina F. Rusnock. Predicting the effects of automation reliability rates on human-automation team performance. In *2016 Winter Simulation Conference (WSC)*,

pages 1802–1813, Washington, DC, USA, December 2016. IEEE. ISBN 978-1-5090-4486-3. doi: 10.1109/WSC.2016.7822227. URL http://ieeexplore.ieee.org/document/7822227/.

[41] John R. Hollenbeck, Bianca Beersma, and Maartje E. Schouten. Beyond Team Types and Taxonomies: A Dimensional Scaling Conceptualization for Team Description. *Academy of Management Review*, 37(1):82–106, January 2012. ISSN 0363-7425. doi: 10.5465/amr.2010.0181. URL https://journals.aom.org/doi/abs/10.5465/amr.2010.0181. Publisher: Academy of Management.

[42] Ayanna M. Howard and Gerardo Cruz. Adapting Human Leadership Approaches for Role Allocation in Human-Robot Navigation Scenarios. In *2006 World Automation Congress*, pages 1–8, July 2006. doi: 10.1109/WAC.2006.376028. URL https://ieeexplore.ieee.org/document/4259944/?arnumber=4259944. ISSN: 2154-4824.

[43] Rehan Iftikhar, Yi Te Chiu, Mohammad Saud Khan, and Catherine Caudwell. Human-Agent Team Dynamics: A Review and Future Research Opportunities. *IEEE Transactions on Engineering Management*, 71:10139–10154, 2024. ISSN 15580040. doi: 10.1109/TEM.2023.3331369. Publisher: Institute of Electrical and Electronics Engineers Inc.

[44] Daniel R. Ilgen, John R. Hollenbeck, Michael Johnson, and Dustin Jundt. Teams in Organizations: From Input-Process-Output Models to IMOI Models. *Annual Review of Psychology*, 56(Volume 56, 2005):517–543, February 2005. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev.psych.56.091103.070250. URL https://www.annualreviews.org/content/journals/10.1146/annurev.psych.56.091103.070250. Publisher: Annual Reviews.

[45] Sarah Ann Jessup. Measurement of the Propensity to Trust Automation. November 2018.

[46] Craig J. Johnson, Mustafa Demir, Nathan J. McNeese, Jamie C. Gorman, Alexandra T. Wolff, and Nancy J. Cooke. The Impact of Training on Human–Autonomy Team Communications and Trust Calibration. *Human Factors*, 65(7):1554–1570, November 2023. ISSN 0018-7208. doi: 10.1177/00187208211047323. URL https://doi.org/10.1177/00187208211047323. Publisher: SAGE Publications Inc.

[47] 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. *J. Hum.-Robot Interact.*, 3(1):43–69, February 2014. doi: 10.5898/JHRI.3.1.Johnson. URL https://dl.acm.org/doi/10.5898/JHRI.3.1.Johnson.

[48] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. How Should an AI Trust its Human Teammates? Exploring Possible Cues of Artificial Trust. *ACM Transactions on Interactive Intelligent Systems*, 14(1):1–26, March 2024. ISSN 2160-6455. doi: 10.1145/3635475. Publisher: Association for Computing Machinery (ACM).

[49] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 229–236, Portland Oregon USA, March 2015. ACM. ISBN 978-1-4503-2883-8. doi: 10.1145/2696454.2696460. URL https://dl.acm.org/doi/10.1145/2696454.2696460.

[50] Ryan Kap. The Influence of Adapting an Agent's Explanation Style to a Human Team Leader Role on Human-Agent Teamwork during a Simulated Search and Rescue Task | TU Delft Repository, 2022. URL https://repository.tudelft.nl/record/uuid:50552e6f-3a88-4156-bfdf-dcf6050df3e2.

[51] Ulas Berk Karli, Shiye Cao, and Chien-Ming Huang. "What If It Is Wrong": Effects of Power Dynamics and Trust Repair Strategy on Trust and Compliance in HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, pages 271–280, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9964-7. doi: 10.1145/3568162.3576964. URL `https://dl.acm.org/doi/10.1145/3568162.3576964`.

[52] Maryam Khani, Ali Ahmadi, and Hajar Hajary. Distributed task allocation in multi-agent environments using cellular learning automata. *Soft Computing*, 23(4):1199–1218, February 2019. ISSN 1433-7479. doi: 10.1007/s00500-017-2839-5. URL `https://doi.org/10.1007/s00500-017-2839-5`.

[53] Jieun Kim, Gonzalo Gonzalez-Pumariega, Soyee Park, and Susan R. Fussell. Urgency Builds Trust: A Voice Agent's Emotional Expression in an Emergency. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, pages 343–347, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701290. doi: 10.1145/3584931.3606979. URL `https://dl.acm.org/doi/10.1145/3584931.3606979`.

[54] Esther Kox. AUTONOMOUS SYSTEMS AS INTELLIGENT TEAMMATES: SOCIAL PSYCHOLOGICAL IMPLICATIONS. ICCRTS(24):11, 2019.

[55] Andrea Krausman, Catherine Neubauer, Daniel Forster, Shan Lakhmani, Anthony L. Baker, Sean M. Fitzhugh, Gregory Gremillion, Julia L. Wright, Jason S. Metcalfe, and Kristin E. Schaefer. Trust Measurement in Human-Autonomy Teams: Development of a Conceptual Toolkit. *ACM Transactions on Human-Robot Interaction*, 11(3):1–58, September 2022. ISSN 2573-9522, 2573-9522. doi: 10.1145/3530874. URL `https://dl.acm.org/doi/10.1145/3530874`.

[56] Joshua A. Lapso, Gilbert L. Peterson, and Michael E. Miller. A hybrid cognitive model for machine agents. *Cognitive Systems Research*, 81:1–10, September 2023. ISSN 1389-0417. doi: 10.1016/j.cogsys.2023.02.007. URL `https://www.sciencedirect.com/science/article/pii/S1389041723000220`.

[57] John D. Lee and Katrina A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1):50–80, March 2004. ISSN 0018-7208. doi: 10.1518/hfes.46.1.50_30392. URL `https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392`. Publisher: SAGE Publications Inc.

[58] M. Leichtle and N. Homburg. A Longitudinal Experiment about Leadership in a Mixed Human-Robot Team in Comparison to a Human-Only Team. In Masaaki Kurosu and Ayako Hashizume, editors, *Human-Computer Interaction*, pages 102–117, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-35602-5. doi: 10.1007/978-3-031-35602-5_8.

[59] Yingke Li and Fumin Zhang. Trust-Preserved Human-Robot Shared Autonomy Enabled by Bayesian Relational Event Modeling. *IEEE Robotics and Automation Letters*, 9(11):10716–10723, November 2024. ISSN 2377-3766. doi: 10.1109/LRA.2024.3438040. URL `https://ieeexplore.ieee.org/document/10621608/?arnumber=10621608`. Conference Name: IEEE Robotics and Automation Letters.

[60] Inês Lobo, Janin Koch, Jennifer Renoux, Inês Batina, and Rui Prada. When Should I Lead or Follow: Understanding Initiative Levels in Human-AI Collaborative Gameplay. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, pages 2037–2056, New York, NY, USA, July 2024. Association for Computing Machinery. ISBN 9798400705830. doi: 10.1145/3643834.3661583. URL `https://dl.acm.org/doi/10.1145/3643834.3661583`.

[61] Sara L. Lopes, José Bernardo Rocha, Aristides I. Ferreira, and Rui Prada. Social robots as leaders: leadership styles in human-robot teams. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 258–263, August 2021. doi: 10.1109/RO-MAN50785.2021.9515464. URL https://ieeexplore.ieee.org/document/9515464/?arnumber=9515464. ISSN: 1944-9437.

[62] P. Madhavan and D. A. Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, July 2007. ISSN 1463-922X. doi: 10.1080/14639220500337708. URL https://doi.org/10.1080/14639220500337708. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14639220500337708.

[63] Ellyn Maese, Pablo Diego-Rosell, Les Debusk-Lane, and Nathan Kress. Development of Emergent Leadership Measurement: Implications for Human-Machine Teams. In Nikolos Gurney and Gita Sukthankar, editors, *Computational Theory of Mind for Human-Machine Teams*, pages 118–145, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-21671-8. doi: 10.1007/978-3-031-21671-8_8.

[64] Roger C. Mayer and James H. Davis. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, 84(1):123–136, 1999. ISSN 1939-1854. doi: 10.1037/0021-9010.84.1.123. Place: US Publisher: American Psychological Association.

[65] Roger C. Mayer, James H. Davis, and F. David Schoorman. An Integrative Model Of Organizational Trust. *Academy of Management Review*, 20(3):709–734, July 1995. ISSN 0363-7425. doi: 10.5465/amr.1995.9508080335. URL https://journals.aom.org/doi/abs/10.5465/AMR.1995.9508080335. Publisher: Academy of Management.

[66] Michael D. McNeese and Nathaniel J. McNeese. Chapter 9 - Humans interacting with intelligent machines: at the crossroads of symbiotic teamwork. In Richard Pak, Ewart J. de Visser, and Ericka Rovira, editors, *Living with Robots*, pages 165–197. Academic Press, January 2020. ISBN 978-0-12-815367-3. doi: 10.1016/B978-0-12-815367-3.00009-8. URL https://www.sciencedirect.com/science/article/pii/B9780128153673000098.

[67] Nathan J. McNeese, Mustafa Demir, Nancy J. Cooke, and Christopher Myers. Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors*, 60(2):262–273, March 2018. ISSN 0018-7208. doi: 10.1177/0018720817743223. URL https://doi.org/10.1177/0018720817743223. Publisher: SAGE Publications Inc.

[68] Nathan J. McNeese, Mustafa Demir, Erin K. Chiou, and Nancy J. Cooke. Trust and Team Performance in Human–Autonomy Teaming. *International Journal of Electronic Commerce*, 25(1):51–72, January 2021. ISSN 1086-4415. doi: 10.1080/10864415.2021.1846854. URL https://doi.org/10.1080/10864415.2021.1846854. Publisher: Routledge _eprint: https://doi.org/10.1080/10864415.2021.1846854.

[69] Stephanie M. Merritt. Affective Processes in Human–Automation Interactions. *Human Factors*, 53(4):356–370, August 2011. ISSN 0018-7208. doi: 10.1177/0018720811411912. URL https://doi.org/10.1177/0018720811411912. Publisher: SAGE Publications Inc.

[70] Christopher A. Miller. Trust, transparency, explanation, and planning: Why we need a lifecycle perspective on human-automation interaction. In *Trust in Human-Robot Interaction*, pages 233–257. Elsevier, 2021. ISBN 978-0-12-819472-0. doi: 10.1016/B978-0-12-819472-0.00011-3. URL https://linkinghub.elsevier.com/retrieve/pii/B9780128194720000113.

[71] Brendan Mooers, Audrey L. Aldridge, Andrew Buck, Cindy L. Bethel, and Derek T. Anderson. Human-robot teaming for a cooperative game in a shared partially observable space. In *Geospatial Informatics XIII*, volume 12525, pages 94–109. SPIE, June 2023. doi: 10.1117/12.2663430. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12525/125250B/Human-robot-teaming-for-a-cooperative-game-in-a-shared/10.1117/12.2663430.full.

[72] Bonnie M. Muir and Neville Moray. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, March 1996. ISSN 0014-0139, 1366-5847. doi: 10.1080/00140139608964474. URL http://www.tandfonline.com/doi/abs/10.1080/00140139608964474.

[73] Geoff Musick, Thomas A. O'Neill, Beau G. Schelble, Nathan J. McNeese, and Jonn B. Henke. What Happens When Humans Believe Their Teammate is an AI? An Investigation into Humans Teaming with Autonomy. *Computers in Human Behavior*, 122:106852, September 2021. ISSN 0747-5632. doi: 10.1016/j.chb.2021.106852. URL https://www.sciencedirect.com/science/article/pii/S0747563221001758.

[74] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000. Publisher: Wiley Online Library.

[75] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments, 2017. URL https://journals.sagepub.com/doi/epub/10.1177/0278364917690593.

[76] Ali Noormohammadi-Asl, Kevin Fan, Stephen L. Smith, and Kerstin Dautenhahn. Human leading or following preferences: Effects on human perception of the robot and the human–robot collaboration. *Robotics and Autonomous Systems*, 183:104821, January 2025. ISSN 09218890. doi: 10.1016/j.robot.2024.104821. URL https://linkinghub.elsevier.com/retrieve/pii/S0921889024002057.

[77] Peter Guy Northouse. *Introduction to Leadership: Concepts and Practice.* SAGE Publications Ltd, 5th edition, 2020. ISBN 978-1-5443-5159-9.

[78] Peter Guy Northouse. *Leadership: Theory & Practice.* SAGE Publications Ltd, 9th edition, 2021. ISBN 978-1-5443-9756-6.

[79] Thomas A. O'Neill, Christopher Flathmann, Nathan J. McNeese, and Eduardo Salas. Human-autonomy Teaming: Need for a guiding team-based framework? *Computers in Human Behavior*, 146:107762, September 2023. ISSN 0747-5632. doi: 10.1016/j.chb.2023.107762. URL https://www.sciencedirect.com/science/article/pii/S0747563223001139.

[80] 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, August 2022. ISSN 0018-7208. doi: 10.1177/0018720820960865. URL https://doi.org/10.1177/0018720820960865. Publisher: SAGE Publications Inc.

[81] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, May 2000. ISSN 1558-2426. doi: 10.1109/3468.844354. URL https://ieeexplore.ieee.org/abstract/document/844354?casa_token=-DaeHIxwoWYAAAAA:pq-T2I1FJKHE4wbOVRVXHiq2Di6iyhFzazo5G8vjArJeZw_Igb_1mrWsxfmpAO87i8RdLKuc. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.

[82] Raja Parasuraman and Victor Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):230–253, June 1997. ISSN 0018-7208, 1547-8181. doi: 10.1518/001872097778543886. URL https://journals.sagepub.com/doi/10.1518/001872097778543886.

[83] Pablo Pico-Valencia, Juan A. Holgado-Terriza, and Luz M. Sierra Martínez. A Preliminary Ontology for Human-Agent Collectives. In Fernando De La Prieta, Alfonso González-Briones, Pawel Pawleski, Davide Calvaresi, Elena Del Val, Fernando Lopes, Vicente Julian, Eneko Osaba, and Ramón Sánchez-Iborra, editors, *Highlights of Practical Applications of Survivable Agents and Multi-Agent Systems. The PAAMS Collection*, pages 176–187, Cham, 2019. Springer International Publishing. ISBN 978-3-030-24299-2. doi: 10.1007/978-3-030-24299-2_15.

[84] Elia Pizzolitto, Ida Verna, and Michelina Venditti. Authoritarian leadership styles and performance: a systematic literature review and research agenda. *Management Review Quarterly*, 73(2):841–871, June 2023. ISSN 2198-1639. doi: 10.1007/s11301-022-00263-y. URL https://doi.org/10.1007/s11301-022-00263-y.

[85] Kiranraj Pushparaj, Pratusha Reddy, Duy Vu-Tran, Kurtulus Izzetoglu, and Sameer Alam. A Multi-Modal Approach to Measuring the Effect of XAI on Air Traffic Controller Trust During Off-Nominal Runway Exits. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4813–4819, October 2023. doi: 10.1109/SMC53992.2023.10394443. URL https://ieeexplore.ieee.org/document/10394443. ISSN: 2577-1655.

[86] Robert E. Quinn. Applying the Competing Values Approach to Leadership: Toward an Integrative Framework. In *Leaders and Managers*, pages 10–27. Elsevier, 1984. ISBN 978-0-08-030943-9. doi: 10.1016/B978-0-08-030943-9.50010-3. URL https://linkinghub.elsevier.com/retrieve/pii/B9780080309439500103.

[87] Preeti Ramaraj, Arthi Haripriyan, Rabeya Jamshad, and Laurel D Riek. Analysis of Social Signals in Human-Robot Action Teams. 2024.

[88] Summer Rebensky, Kendall Carmody, Cherrise Ficke, Meredith Carroll, and Winston Bennett. Teammates Instead of Tools: The Impacts of Level of Autonomy on Mission Performance and Human–Agent Teaming Dynamics in Multi-Agent Distributed Teams. *Frontiers in Robotics and AI*, 9, May 2022. ISSN 2296-9144. doi: 10.3389/frobt.2022.782134. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.782134/full. Publisher: Frontiers.

[89] Lionel P. Robert Jr and Sangseok You. Are you satisfied yet? Shared leadership, individual trust, autonomy, and satisfaction in virtual teams. *Journal of the Association for Information Science and Technology*, 69(4):503–513, 2018. ISSN 2330-1643. doi: 10.1002/asi.23983. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23983. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23983.

[90] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. Timing Is Key for Robot Trust Repair. In Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi, editors, *Social Robotics*, pages 574–583, Cham, 2015. Springer International Publishing. ISBN 978-3-319-25554-5. doi: 10.1007/978-3-319-25554-5_57.

[91] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998. Publisher: Academy of Management Briarcliff Manor, NY 10510.

[92] Eduardo Salas, Dana E. Sims, and C. Shawn Burke. Is there a "Big Five" in Teamwork? *Small Group Research*, 36(5):555–599, October 2005. ISSN 1046-4964. doi: 10.1177/1046496405277134. URL https://doi.org/10.1177/1046496405277134. Publisher: SAGE Publications Inc.

[93] Eduardo Salas, Diana R. Nichols, and James E. Driskell. Testing Three Team Training Strategies in Intact Teams: A Meta-Analysis. *Small Group Research*, 38(4):471–488, August 2007. ISSN 1046-4964. doi: 10.1177/1046496407304332. URL `https://doi.org/10.1177/1046496407304332`. Publisher: SAGE Publications Inc.

[94] Eduardo Salas, Denise L. Reyes, and Susan H. McDaniel. The science of teamwork: Progress, reflections, and the road ahead. *American Psychologist*, 73(4):593–600, May 2018. ISSN 1935-990X, 0003-066X. doi: 10.1037/amp0000334. URL `https://doi.apa.org/doi/10.1037/amp0000334`.

[95] Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, 5, February 2018. ISSN 2296-9144. doi: 10.3389/frobt.2018.00015. URL `https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2018.00015/full`. Publisher: Frontiers.

[96] Maarten P. D. Schadd, Tjeerd A. J. Schoonderwoerd, Karel van den Bosch, Olaf H. Visker, Tjalling Haije, and Kim H. J. Veltman. “I’m Afraid I Can’t Do That, Dave”; Getting to Know Your Buddies in a Human–Agent Team. *Systems*, 10(1):15, February 2022. ISSN 2079-8954. doi: 10.3390/systems10010015. URL `https://www.mdpi.com/2079-8954/10/1/15`. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[97] Kristin E. Schaefer, Tracy L. Sanders, Ryan E. Yordon, Deborah R. Billings, and P.A. Hancock. Classification of Robot Form: Factors Predicting Perceived Trustworthiness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):1548–1552, September 2012. ISSN 1071-1813. doi: 10.1177/1071181312561308. URL `https://doi.org/10.1177/1071181312561308`. Publisher: SAGE Publications Inc.

[98] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, 58(3):377–400, May 2016. ISSN 0018-7208. doi: 10.1177/0018720816634228. URL `https://doi.org/10.1177/0018720816634228`. Publisher: SAGE Publications Inc.

[99] Kristin E. Schaefer, Brandon S. Perelman, Gregory M. Gremillion, Amar R. Marathe, and Jason S. Metcalfe. Chapter 12 - A roadmap for developing team trust metrics for human-autonomy teams. In Chang S. Nam and Joseph B. Lyons, editors, *Trust in Human-Robot Interaction*, pages 261–300. Academic Press, January 2021. ISBN 978-0-12-819472-0. doi: 10.1016/B978-0-12-819472-0.00012-5. URL `https://www.sciencedirect.com/science/article/pii/B9780128194720000125`.

[100] T. B. Sheridan, W. L. Verplank, and T. L. Brooks. Human/computer control of undersea teleoperators. November 1978. URL `https://ntrs.nasa.gov/citations/19790007441`. NTRS Author Affiliations: Massachusetts Inst. of Tech. NTRS Document ID: 19790007441 NTRS Research Center: Legacy CDMS (CDMS).

[101] Tony L. Simons and Randall S. Peterson. Task conflict and relationship conflict in top management teams: The pivotal role of intragroup trust. *Journal of Applied Psychology*, 85(1):102–111, 2000. ISSN 1939-1854. doi: 10.1037/0021-9010.85.1.102. Place: US Publisher: American Psychological Association.

[102] Güliz Tokadlı and Michael C. Dorneich. Interaction Paradigms: from Human-Human Teaming to Human-Autonomy Teaming. In *2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC)*, pages 1–8, September 2019. doi: 10.1109/DASC43569.2019.9081665. URL `https://ieeexplore.ieee.org/document/9081665`. ISSN: 2155-7209.

[103] Tom R. Tyler and Peter Degoey. Trust in organizational authorities: The influence of motive attributions on willingness to accept decisions. In *Trust in organizations: Frontiers of theory and research*, pages 331–356. Sage Publications, Inc, Thousand Oaks, CA, US, 1996. ISBN 978-0-8039-5739-8 978-0-8039-5740-4. doi: 10.4135/9781452243610.n16.

[104] Birgit van der Stigchel, Karel van den Bosch, Jurriaan van Diggelen, and Pim Haselager. Intelligent decision support in medical triage: are people robust to biased advice? *Journal of Public Health*, 45(3):689–696, September 2023. ISSN 1741-3842. doi: 10.1093/pubmed/fdad005. URL `https://doi.org/10.1093/pubmed/fdad005`.

[105] 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, May 2021. ISSN 2296-9144. doi: 10.3389/frobt.2021.640647. URL `https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2021.640647/full`. Publisher: Frontiers.

[106] Daan van Knippenberg. Team Leadership. In *The Wiley Blackwell Handbook of the Psychology of Team Working and Collaborative Processes*, pages 345–368. John Wiley & Sons, Ltd, 2017. ISBN 978-1-118-90999-7. doi: 10.1002/9781118909997.ch15. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118909997.ch15`. Section: 15 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118909997.ch15.

[107] 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, July 2021. ISSN 2296-9144. doi: 10.3389/frobt.2021.692811. URL `https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2021.692811/full`. Publisher: Frontiers.

[108] 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, September 2022. ISSN 22969144. doi: 10.3389/frobt.2022.993997. URL `https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.993997/full`. Publisher: Frontiers Media S.A.

[109] Ruben S. Verhagen, Alexandra Marcu, Mark A. Neerincx, and Myrthe L. Tielman. The Influence of Interdependence on Trust Calibration in Human-Machine Teams. In *Frontiers in Artificial Intelligence and Applications*, volume 386, pages 300–314. IOS Press BV, June 2024. ISBN 978-1-64368-522-9. doi: 10.3233/FAIA240203. ISSN: 18798314.

[110] James C. Walliser, Patrick R. Mead, and Tyler H. Shaw. The Perception of Teamwork With an Autonomous Agent Enhances Affect and Performance Outcomes, 2017. URL `https://journals.sagepub.com/doi/abs/10.1177/1541931213601541`.

[111] Danni Wang, David A. Waldman, and Zhen Zhang. A meta-analysis of shared leadership and team effectiveness. *Journal of Applied Psychology*, 99(2):181–198, 2014. ISSN 1939-1854, 0021-9010. doi: 10.1037/a0034531. URL `https://doi.apa.org/doi/10.1037/a0034531`.

[112] Ning Wang, David V. Pynadath, Ericka Rovira, Michael J. Barnes, and Susan G. Hill. Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams. In Jaap Ham, Evangelos Karapanos, Plinio P. Morita, and Catherine M. Burns, editors, *Persuasive Technology*, pages 56–69, Cham, 2018. Springer International Publishing. ISBN 978-3-319-78978-1. doi: 10.1007/978-3-319-78978-1_5.

[113] S. Wang. The impact of expressing emotion within explainable AI in human-agent teamwork. 2024. URL `https://repository.tudelft.nl/record/uuid:8616eb26-d8ef-45f6-be73-9e60bdae646b`.

[114] Franziska Doris Wolf and Ruth Stock-Homburg. Human-Robot Teams: A Review. In Alan R. Wagner, David Feil-Seifer, Kerstin S. Haring, Silvia Rossi, Thomas Williams, Hongsheng He, and Shuzhi Sam Ge, editors, *Social Robotics*, pages 246–258, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62056-1. doi: 10.1007/978-3-030-62056-1_21.

[115] Philip A. Woods. Authority, power and distributed leadership. *Management in Education*, 30 (4):155–160, October 2016. ISSN 0892-0206. doi: 10.1177/0892020616665779. URL https://doi.org/10.1177/0892020616665779. Publisher: SAGE Publications Ltd.

[116] Haochen Wu, Amin Ghadami, Alparslan Emrah Bayrak, Jonathon M. Smereka, and Bogdan I. Epureanu. Task Allocation with Load Management in Multi-Agent Teams. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8823–8830, May 2022. doi: 10.1109/ICRA46639.2022.9811374. URL https://ieeexplore.ieee.org/document/9811374/?arnumber=9811374.

[117] Kevin T. Wynne and Joseph B. Lyons. An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, May 2018. ISSN 1463-922X. URL https://www.tandfonline.com/doi/full/10.1080/1463922X.2016.1260181. Publisher: Taylor & Francis.

[118] Jing Zhou. Exploring the effect of the information amount in explanation on different gaming expertise levels. 2023. URL https://repository.tudelft.nl/record/uuid:0e6369b7-b2fb-4949-a3df-fc109fa10d45.

[119] Jinlong Zhu, Zhenyu Liao, Kai Chi Yam, and Russell E. Johnson. Shared leadership: A state-of-the-art review and future research agenda. *Journal of Organizational Behavior*, 39(7):834–852, 2018. ISSN 1099-1379. doi: 10.1002/job.2296. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/job.2296. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.2296.

Appendix A

Surveys

A.1 Informed Consent

You are being invited to participate in a research study titled **Put words into action: exploring the effect of authority change as a trust repair strategy in human-agent teams**. This study is being done by Harmen Kroon from the TU Delft.

The purpose of this research study is gaining insight in trust dynamics within a human and robotic agent team, and will take you approximately 45 minutes to complete. The data will be used for Harmen's Master's thesis. We will be asking you to answer questions on prior experience with and trust towards robots as well as collaborate in a simulated search and rescue mission. The environment is a two-dimensional abstract visual representation with no explicit imagery used. To anyone with sensitivity towards natural disaster and rescue efforts caution is advised.

As with any electronic activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by anonymising all data before being analysed and aggregated into the report. Personal data regarding age range, gender and education level will solely be used for statistical analysis and stored separately from your survey answers (the research data) with no mention of your name, phone number or email address. Access to anonymised personal data that is stored in a data vault (TU Delft project drive) is only available to members of the research team and will be retained until after the thesis is completed; approximately after August 2025. The final thesis report will be publicly available on the TU Delft repository. Non-personal research data may be stored in the 4TU.ResearchData open science data repository.

There are minimal to no risks involved in this study. Your participation is entirely voluntary and **you can withdraw at any time**. You are free to omit any questions during the experiment. However, due to anonymisation it will not be possible to retract answers after the experiment has been run.

The researcher can be contacted through ⟨redacted⟩. The supervisor and responsible researcher for this research study is Myrthe Tielman, who can be contacted on ⟨redacted⟩.

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated / /2025, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	□	□
I understand that taking part in the study involves filling in electronic questionnaires and controlling a human rescuer in a computer simulation.	□	□
I understand that the study will end after the thesis is completed.	□	□
Risks associated with participating in the study		
I understand that taking part in the study involves the risk of mental discomfort due to a simulated natural disaster environment.	□	□
I understand that taking part in the study also involves collecting specific personally identifiable information (PII) – name & contact information – and associated personally identifiable research data (PIRD) – age range, gender & education level – with the potential risk of my identity being revealed minimised through anonymisation and controlled data access.	□	□
I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach. Any data breach will be unlikely due to controlled access to data in a data vault. Nonetheless, all PII will be stored separately from any research data. All PIRD will be anonymised and untraceable to any PII.	□	□
I understand that personal information collected about me that can identify me, such as name, phone number and e-mail address, will not be shared beyond the study team and destroyed after the project is finished.	□	□
Use of the information in the study		
I understand that after the research study the de-identified information I provide will be used for a Master's Thesis.	□	□
I agree that my responses, views or other input can be quoted anonymously in research outputs.	□	□
Future use and reuse of the information		
I give permission for the de-identified survey answers and simulation logs that I provide to be archived TU Delft's Interactive Intelligence GitLab repository and 4TU.ResearchData open science repository so it can be used for future research and learning.	□	□

Signatures

Name of the participant: ______________________________
Signature: ______________________________
Date: ______________________________

I, as a researcher, have accurately read out the information to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Name of the researcher: ______________________________
Signature: ______________________________
Date: ______________________________

A.2 Pre-Study Survey

Demographics

1. What is your age in years?

   ○ 18 - 24
   ○ 25 - 34
   ○ 34 - 44
   ○ 45 - 54
   ○ 55 - 64
   ○ 65 - 74
   ○ 75 or older

2. How do you describe yourself?

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

3. What is your highest obtained education level?

   ○ Primary school (basisschool)
   ○ High school (middelbare school)
   ○ Vocational school or similar (MBO)
   ○ Bachelor's degree or equivalent
   ○ Master's degree or equivalent
   ○ Graduate or professional degree (PhD, JD, MD, DDs, etc.)
   ○ Other

Leadership style

4. For each of the following statements indicate the degree to which you agree or disagree. Give your immediate impressions. There are no right or wrong answers.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Employees need to be supervised closely, or they are not likely to do their work.	○	○	○	○	○
Employees want to be a part of the decision-making process.	○	○	○	○	○
In complex situations, leaders should let followers work problems out on their own.	○	○	○	○	○
It is fair to say that most employees in the general population are lazy.	○	○	○	○	○
Providing guidance without pressure is the key to being a good leader.	○	○	○	○	○
Leadership requires staying out of the way of followers as they do their work.	○	○	○	○	○
As a rule, employees must be given rewards or punishments in order to motivate them to achieve organizational objectives.	○	○	○	○	○
Most workers prefer supportive communication from their leaders.	○	○	○	○	○
As a rule, leaders should allow followers to appraise their own work.	○	○	○	○	○
Most employees feel insecure about their work and need direction.	○	○	○	○	○
Leaders need to help followers accept responsibility for completing their work.	○	○	○	○	○
Leaders should give followers complete freedom to solve problems on their own.	○	○	○	○	○
The leader is the chief judge of the achievements of the member of the group.	○	○	○	○	○
It is the leader's job to help followers find their "passion".	○	○	○	○	○
In most situations, workers prefer little input from the leader.	○	○	○	○	○
Effective leaders give orders and clarify procedures.	○	○	○	○	○
People are basically competent and if given a task will do a good job.	○	○	○	○	○
In general, it is best to leave followers alone.	○	○	○	○	○

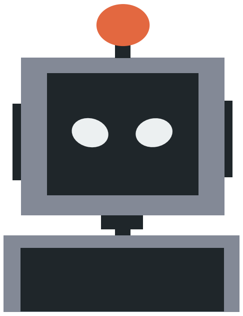
Trust propensity towards robots

5. How much do you agree with the following statements?

For each of the following statements indicate the degree to which you agree or disagree. Give your immediate impressions. There are no right or wrong answers.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Generally I would trust robots.	○	○	○	○	○	○	○
Robots can help me solve many problems.	○	○	○	○	○	○	○
I think it is a good idea to rely on robots for help.	○	○	○	○	○	○	○
I wouldn't trust the information I might get from robots.	○	○	○	○	○	○	○
Robots are reliable.	○	○	○	○	○	○	○
I would rely on robots.	○	○	○	○	○	○	○

A.3 Post-Interaction Trust Survey

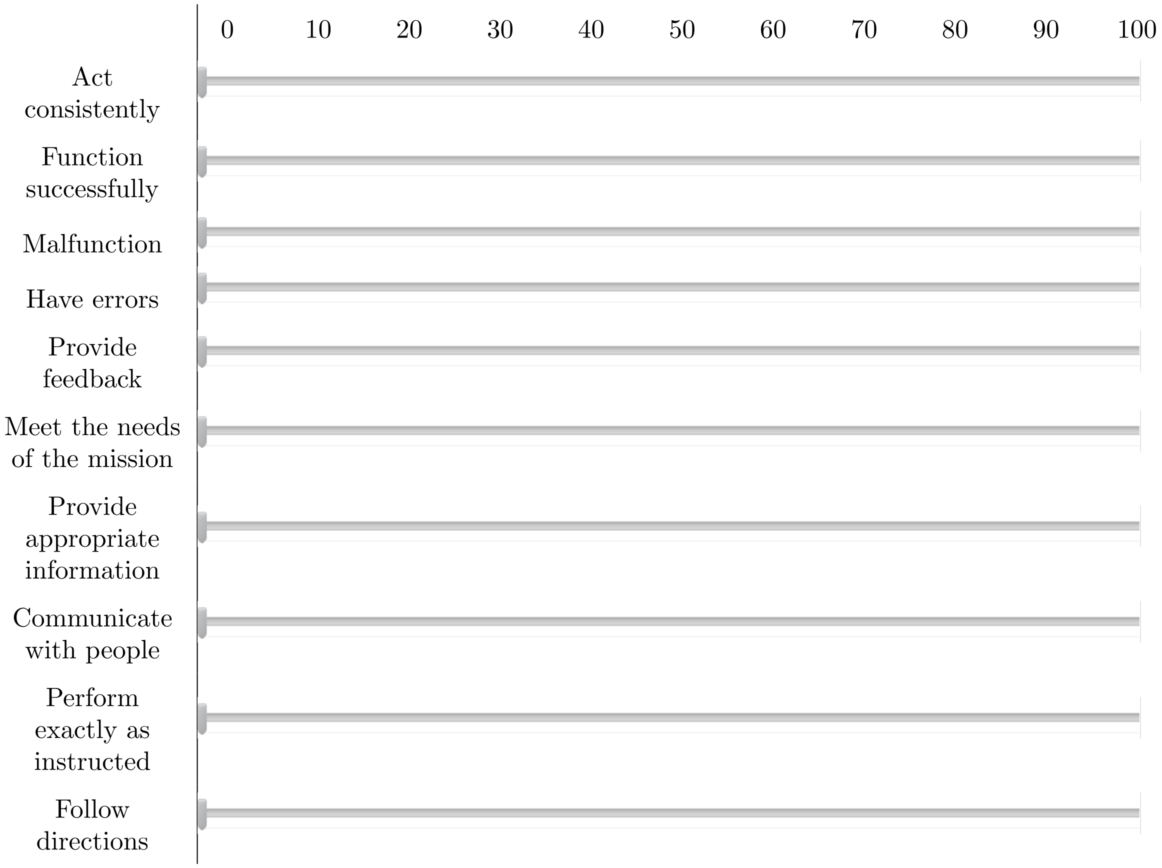


What do you think of RescueBot?

1. What % of time do you think RescueBot will be...

	0	10	20	30	40	50	60	70	80	90	100
Dependable											
Reliable											
Unresponsive											
Predictable											

2. What % of time do you think RescueBot will...



0 10 20 30 40 50 60 70 80 90 100

Act consistently

Function successfully

Malfunction

Have errors

Provide feedback

Meet the needs of the mission

Provide appropriate information

Communicate with people

Perform exactly as instructed

Follow directions

A.4 Post-Study Survey

Task work load

3. Mental demand
   How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving? [Rate from 0="Low" to 10="High"]

4. Physical demand
   How much physical activity was required for YOU (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious? [Rate from 0="Low" to 10="High"]

5. Temporal demand
   How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic? [Rate from 0="Low" to 10="High"]

6. Effort
   How hard did you have to work (mentally and physically) to accomplish your level of performance? [Rate from 0="Low" to 10="High"]

7. Performance
   How successful do you think you were in accomplishing the goals of the task set by the exper-

imenter (or yourself)? How satisfied were you with your performance in accomplishing these goals? [Rate from 0="Poor" to 10="Good"]

8. Frustration level
   How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task? [Rate from 0="Low" to 10="High"]

Open questions

4. How did you experience the power dynamic within the team?

5. What is your advice to the robots for achieving better team performance?

Appendix B

Extended data analysis

B.1 Task load

The NASA-TLX in this study has been administered without the rankings step and therefore scored using the average of all scores. In practice the physical demand of this task was potentially poorly understood, indicated by multiple participants requesting clarification on whether the physical demand should be of themselves - mainly moving a mouse and pressing mouse and keyboard buttons - or from the perspective of the simulated human rescuer - moving around, removing obstacles and carrying victims. The frequency of questions indicates that confusion among participants may lead to vastly different interpretations of this one question who without the personalised ranking determines $\frac{1}{6}$-th of the final task load score. A new box plot of the task load per conditional group is shown in figure B.1.

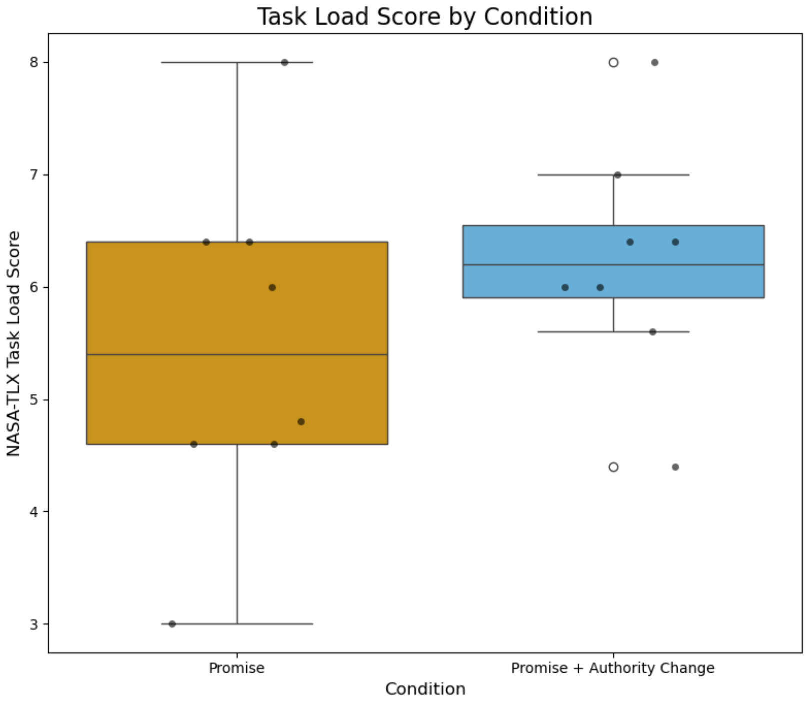


Figure B.1: Box plot of taskload per condition without considering the physical demand aspect