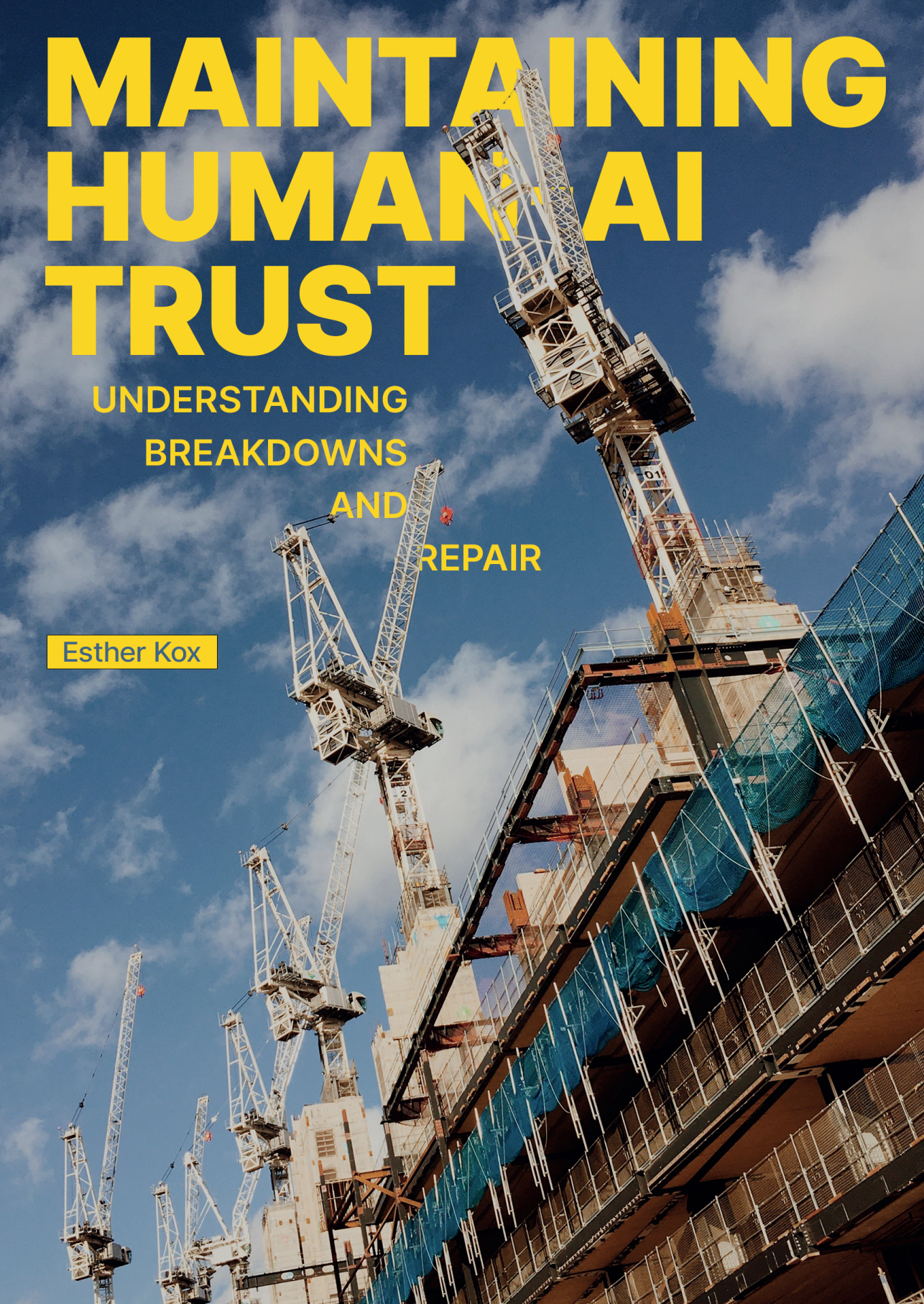
MAINTAINING HUMAN-AI TRUST

UNDERSTANDING BREAKDOWNS AND REPAIR

Esther Kox



MAINTAINING HUMAN-AI TRUST: UNDERSTANDING BREAKDOWNS AND REPAIR

Esther Kox

Lay-out: Wendy Bour-van Telgen
Cover design: Esther Kox
Printed by: Ipskamp Printing, Enschede
ISBN print: 978-90-365-6551-6
ISBN digital: 978-90-365-6552-3
DOI: 10.3990/1.9789036565523

MAINTAINING HUMAN-AI TRUST: UNDERSTANDING BREAKDOWNS AND REPAIR

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr.ir. A. Veldkamp,
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday the 24th of April 2025 at 16:45 hours

by

Esther Sophia Kox

born on the 6th of February, 1996
in Hilversum, The Netherlands

This dissertation has been approved by:

Supervisor(s)

prof. dr. J.H. Kerstholt	University of Twente

Co-supervisor(s)

dr. ir. P.W. de Vries	University of Twente
dr. M.B. van Riemsdijk	University of Twente

GRADUATION COMMITTEE:

Chair/secretary	prof. dr. T. Bondarouk (University of Twente)
Promotor	prof. dr. J.H. Kerstholt (University of Twente)
Co-promotor	dr. ir. P.W. de Vries (University of Twente)
	dr. M.B. van Riemsdijk (University of Twente)
Members	prof. dr. ir. M. Boon (University of Twente)
	prof. dr. D.K.J. Heylen (University of Twente)
	dr. J. van Diggelen (TNO Soesterberg)
	prof. dr. T. Bosse (Radboud University)
	prof. dr. W.A. IJsselsteijn (Eindhoven University of Technology)

Table of Contents

Chapter 1

Introduction

The rise of Human-AI teams

Advancements in computer hardware and software technology have enabled the partial or complete replacement of functions previously performed by people (i.e., automation) (Parasuraman et al., 2000). The introduction of automated systems, such as machines on assembly lines in manufacturing, has revolutionized productivity and efficiency by executing repetitive tasks faster and with greater accuracy than humans in a wide variety of routine, initially mostly physical tasks (Cremer & Kasparov, 2021). With the emergence of Artificial Intelligence (AI) and deep neural networks, the possibilities of automation further expanded as machines gained the capability to learn, to make decisions, and to mimic human cognitive functions (Schraagen & van Diggelen, 2021). As AI technology advances, its capabilities are expanding rapidly. From healthcare to finance, transportation to entertainment, AI is driving innovation, enabling new possibilities, and fundamentally altering how we interact with technology and each other.

With the term AI, we refer to "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals" (AIHLEG, 2019) (p. 1). In addition to their level of autonomy, AI-based systems, referred to as AI agents, can differ in various aspects, including their form and function. They can be completely software-based (e.g., voice assistants, image analysis software, search engines), or AI can be embedded in hardware devices, such as advanced robots, autonomous cars, or drones (AIHLEG, 2019). Examples of software-based AI agents include AI chatbots such as OpenAI's ChatGPT or voice assistants such as Apple's Siri or Amazon's Alexa. Examples of AI agents with a physical embodiment that allows them to interact with the physical world include the humanoid robots Nao and Pepper (from Softbank Robotics) and Spot, a quadruped robot developed by Boston Dynamics. Throughout this dissertation, I will use various terms such as "machine", "robot" or "drone" to describe AI agents, reflecting the diverse contexts in which these systems are discussed. All these terms are intended to refer broadly to AI agents.

The initial goal of the field of AI was to create systems that could mimic or replicate human intelligence (Lake et al., 2017). The term "artificial intelligence" itself implies an effort to create intelligence that resembles a "natural", biological intelligence. However, over the years, a new perspective emerged. Researchers began to strive for Hybrid Intelligence, a situation where machine intelligence is combined with human intelligence, aiming to augment human intellect rather than replace it (Akata et al., 2020; M. Johnson et al., 2011, 2012; Peeters et al., 2021). Researchers now explore how we can harness various types of intelligences and skills to find new solutions and discover new types of relations – instead of re-creating what we already have (Darling, 2021).

As a result, the idea of Human-AI Teams (HATs) emerged: teams consisting of at least one human and one AI agent (Bobko et al., 2022; de Visser et al., 2018, 2019; O'Neill et al., 2022). As AI agents become more intelligent, they are increasingly self-governing, gain decision authority within their functioning (Bobko et al., 2022; Hancock, Billings, Schaefer,

et al., 2011a; Hou et al., 2021; O'Neill et al., 2022; Sheridan, 2019), and require less human involvement and control (Lyons et al., 2023; C. A. Miller, 2014). Future AI agents are expected to have increasingly advanced capabilities, enabling them to observe and act upon an environment autonomously and to communicate and collaborate with other agents, including humans, to solve problems and achieve (common) goals (Ferguson & Allen, 2011; O'Neill et al., 2022; Wynne & Lyons, 2018). In literature alternative names are used, including human-machine teams, human-agent teams, and human-automation teams (Jorge et al., 2024). In this work, I will use the terms human-AI teams or human-robot teams as appropriate, depending on the type of AI agent discussed in each chapter.

A team is defined as "a set of two or more people who interact dynamically, interdependently, and adaptively toward a common and valued goal/objective/mission, and who each have some specific roles or functions to perform" (Tannenbaum et al., 1992) (p. 118). The premise of teamwork is that individuals in a team can achieve more and solve more difficult problems collectively than any of the members could do alone, by combining their diverse skills, perspectives, and resources (Akata et al., 2020). This requires the strategic delegation of tasks, where specific responsibilities are aligned with each team member's unique set of skill and expertise. This allows members to share the workload, monitor each other's progress, and develop expertise on subtasks, ultimately contributing to the achievement of their common goal (Salas et al., 2018).

Based on this understanding of teamwork, the idea of HATs is promising, since humans and AI have complementary skills that can be pooled together, to elevate performance beyond the capabilities of its individual members. Namely, on the one hand, AI agents can process and recognize patterns in large amounts of data and perform fast and highly accurate computations. This can, for example, be applied to the diagnosis of diseases based on the classification of radiological images, where AI has, in certain narrowly defined tasks and specific occasions, demonstrated the ability to exceed the performance of clinical experts (Bejnordi et al., 2017). People, on the other hand, are flexible and adaptable and can use their creativity and common sense to find solutions for open and ill-defined tasks and to improvise in changing or unforeseen conditions (Jarrahi, 2018; Korteling et al., 2021; Xiong et al., 2022). In other words, AI can augment people's cognitive abilities to tackle complex problems, while people excel in offering intuitive and comprehensive solutions to uncertain situations (Jarrahi, 2018).

However, the real challenge is not merely determining which tasks are better suited for humans or AI agents working independently, but in finding ways to optimise collaboration, enhancing their respective strengths through effective interaction (Bradshaw et al., 2012; Dekker & Woods, 2002; Hayes, 2016; M. Johnson et al., 2012). Most AI agents are able to perform tasks *for* people, but often lack the skills necessary to work together *with* people and other agents (Bradshaw et al., 2012, 2013). Working together towards a common goal requires good cooperation, coordination, and communication (Ososky et al., 2012b; Salas et al., 2018), and it is within these areas that the true challenges lie.

A key component in these activities is trust, because it allows team members to depend on each other's contributions and to navigate the uncertainties and risks associated with teamwork (A. Y. Lee et al., 2010). To reap the potential collective benefits of teamwork, individuals must be willing to put aside individual interests for the greater good of the team and make investments (e.g., time, energy, or expertise) in the collective effort, trusting that the rewards of collaboration will eventually outweigh the risks. This thesis will investigate how to maintain trust in HATs to support safe and effective collaboration.

Trust

We define human-AI (H-AI) trust as a human's willingness to make oneself vulnerable to an AI agent's decisions and recommendations in the pursuit of some benefit, with the expectation that the AI agent will help achieve the overall task goal in a context characterized by uncertainty and risk (Gambetta, 2000; Hoff & Bashir, 2015; J. D. Lee & See, 2004; Madsen & Gregor, 2000; Raue et al., 2019; Shariff et al., 2017). In other words, the trustor (i.e., the one bestowing trust) must actively decide whether to trust the trustee (i.e., the one receiving trust) in a situation where there is potential gain (i.e., the pursued benefit) and potential loss (i.e., the risk). This decision is shaped by beliefs and expectations regarding the trustee's future actions and is heavily influenced by the specific task at hand (Costa, 2003; Kessler et al., 2016; Li et al., 2019; Raue et al., 2019).

Teamwork is inherently a process that involves risk and requires trust because it involves individuals depending on each other's contributions to collectively complete tasks and achieve objectives (A. Y. Lee et al., 2010). When team members delegate tasks or responsibilities to each other, they become vulnerable in the sense that they are relying on others' competence and commitment. For example, in lead climbing, climbers work in pairs where one ascends while the other secures the rope to prevents falls (i.e., belays). The climber relies on the belayer's skills and quick response to arrest any potential falls, requiring significant trust. This trust allows climbers to reach heights they could not achieve alone.

While the importance of trust is evident in the climbing example, the same principle applies to H-AI collaboration. Imagine a military reconnaissance mission (e.g., see Chapter 4), where a soldier relies on a robotic agent equipped with advanced sensors and navigation capabilities for coverage and its ability to detect and warn of threats. Their mission is to gather intelligence on enemy movements by counting objects in a remote, hostile area. The soldier's role is to manually count and record observations, relying on their training and sharp eyesight, while the robot scans the terrain for threats and obstacles. The soldier depends on the robot to alert them to any detected dangers, allowing them to focus on their specific tasks. In this scenario, as in in the previous one, trust is essential for the team's success.

The climber and the belayer, as well as the soldier and the robot, do not need to know each other very well "personally" in order to establish a successful trusting relation. What matters is that the trustor believes that the trustee has a shared understanding, known as a shared mental model, of the procedures, equipment and tasks necessary to achieve a certain objective (i.e., the taskwork) as well as an understanding of other teammates' knowledge, skills, preferences and responsibilities within the team (i.e., the teamwork) (Driskell et al., 2018; Mathieu et al., 2000). This belief in a shared mental model might stem from knowing that the belayer is an experienced climber themselves or that the robot was designed by a certified manufacturer for a specific purpose (Mathieu et al., 2000). Mental models are personal, internal (cognitive) representations of external reality, based on unique life experiences, perceptions, and understandings of the world (Jones et al., 2011). We use these cognitive frameworks to interpret and make sense of the world around us and to construct expectations for what is likely to occur next (Mathieu et al., 2000; Rouse & Morris, 1985). They shape our reasoning processes and result in predictions about the external environment, guiding our decision-making, actions and behaviours (Jones et al., 2011). A shared mental model greatly contributes to team process and performance, as it allows people to determine what is required to achieve a shared goal and which teammate can be best entrusted with which task (Salas et al., 2005).

An important difference between the provided examples, however, is that the trustee in the latter example (i.e., the soldier and the robot) being AI-based adds uncertainty to an already risky situation. As AI agents become more complex and go beyond a simple tool with sharply defined and easily understood behaviours, it becomes impossible, even for experts, to have a complete understanding and accurate mental model of the system. As a result, the importance of trust further increases as trust plays a crucial role in people's ability to overcome and accept the cognitive complexity and the uncertainty that is associated with increasingly sophisticated AI systems (J. D. Lee & See, 2004).

As illustrated by the scenarios above, the role of trust in teamwork encompasses more than just the initial decision to collaborate and allocate tasks (Hancock et al., 2020; Lubars & Tan, 2019). During the collaboration, team members must maintain an appropriate level of trust in that everyone is performing as required in order to accomplish a specific goal (A. Y. Lee et al., 2010). That is, overly and unnecessarily monitoring the activity of other team members slows down progress and adds unnecessary workload, but team members should also avoid blind trust and maintain a healthy level of vigilance to ensure everyone is meeting expectations. This process is referred to as trust calibration.

Trust calibration

For safe and successful collaborations, people should be able to determine when it is appropriate to rely on AI agents and when it is best to override them (J. D. Lee & See, 2004). In situations involving consequential decisions, such as military operations or healthcare, it is essential to know when AI is safer than human intervention and vice

versa (Barnes et al., 2014). To minimize the risks and maximize the benefits of a H-AI collaboration, H-AI trust should be well-calibrated (Lewis et al., 2018). Calibrated trust refers to a balanced relation between the perceived trustworthiness of an AI agent and its actual trustworthiness (J. D. Lee & See, 2004). Here, trustworthiness is a property of the AI agent, while perceived trustworthiness is a judgement by the human (Duenser & Douglas, 2023). In other words, trust is calibrated when the trust that an individual grants the AI agent matches the trustworthiness of the AI agent, which is supposed to lead to appropriate use (J. D. Lee & See, 2004).

Miscalibration, represented as either 'overtrust' or 'undertrust', can lead to inappropriate reliance, which can compromise safety and profitability respectively (Baker et al., 2019; de Visser et al., 2019; J. D. Lee & See, 2004). In case of overtrust (i.e., excessive trust), a trustor accepts too much risk, which can lead to complacency and can cause costly disasters (Robinette et al., 2017a). For example, human error and overtrust were identified as the primary causes of a fatal crash involving an Uber self-driving car, where the vehicle struck and killed a pedestrian. The backup operator, who was later charged with negligent homicide, had been distracted by streaming a TV show at the time of the incident (Cellan-Jones, 2020). By contrast, undertrust (i.e. insufficient trust) also prevents effective collaboration as it leads to scepticism, causing inefficient monitoring of the work behaviours of other team members and an uneven distribution of workload (de Visser et al., 2019; J. D. Lee & See, 2004). At worst, people may choose not to use or even consciously disable systems that could potentially help them (Ullrich et al., 2021). In other words, maximizing trust is not the objective in H-AI collaboration, as effective and efficient teamwork requires finding the right balance of trust among team members. Calibrated trust facilitates cooperation and coordination between interdependent actors, which creates a more productive and efficient team (J. D. Lee & See, 2004).

Yet, trust calibration will never be perfect, as humans are not mechanical measuring instruments and an agent's trustworthiness itself is not perfectly defined (Duenser & Douglas, 2023; Hoffman, 2017). Knowing when and when not to trust will always remain a challenge, due to the multi-dimensional, context-dependent and dynamic nature of trust. That is, we may trust elected officials to draft legislature, while we may not trust them to make medical decisions. Similarly, while we may expect a medical decision-support system to provide accurate medical advice, we do not expect the same system to provide accurate legal advice (Duenser & Douglas, 2023). Furthermore, an actor's trustworthiness on a given time may vary depending on external situational factors. For example, the performance of a self-driving car might be significantly impacted by poor weather conditions or missing or faded road markings, just as a human driver's ability to navigate safely can be impaired by fatigue or distractions. As such, the term 'calibrated' does not indicate that the trust was adjusted once and is now fixed (Kox et al., 2023). Rather, trust is "a continual process of active exploration and evaluation of trustworthiness and reliability" (Hoffman, 2017) (p.146), meaning it is continually subject to adjustment and fine-tuning.

A Lifecycle Perspective on Trust

To navigate the constant pursuit of an optimal level of trust in ever-changing circumstances, we must understand the dynamics of trust. We want to understand how trust develops, how it breaks down, and how it recovers (de Visser, Pak, et al., 2017). In prior work, researchers have suggested the concept of a 'trust lifecycle' (de Visser et al., 2016, 2018; Rousseau et al., 1998; Söllner & Pavlou, 2016), consisting of multiple phases.

First, trust has to be formed. We distinguish between trust formation, when trust is built for the first time, and trust repair, when trust needs to be rebuilt after it was violated (Söllner & Pavlou, 2016). In the trust violation phase, trust diminishes due to the occurrence of unexpected, unfavourable, or unwanted behaviour, resulting in a negative experience for the human trustor (Rousseau et al., 1998; Söllner & Pavlou, 2016). Essentially, a trust violation is any kind of behaviour from an AI agent that decreases a human's trust in it (Pak & Rovira, 2023). In the trust repair phase, an AI agent can employ strategies to restore trust and facilitate reconciliation after it violated trust (Baker et al., 2018; de Visser et al., 2018; P. H. Kim et al., 2006; Pak & Rovira, 2023). Lastly, there are phases of trust stability, where built or rebuilt trust remains stable over time (Söllner & Pavlou, 2016).

This lifecycle of trust is a theory-based simplification of the complex and dynamic nature of trust and, naturally, these phases can occur in any order, perhaps even simultaneously, and can be repeated. The phases act as labels used to identify the status of a particular trust dynamic, to understand its underlying mechanisms and potential next developments and to recognize relevant interventions (Kox et al., 2023).

In this thesis, I will concentrate on the *maintenance* of H-AI trust by examining the effects of various types of trust violations and investigating methods to reduce their impact, both preventatively and reactively, through trust repair strategies. The emphasis will be on the phases of trust violation and repair within the trust lifecycle, rather than on trust formation or periods of trust stability.

Maintaining H-AI trust is a key part of the trust calibration process. Understanding the breakdowns and recoveries of trust is particularly relevant in situations of undertrust, where people are losing trust in an AI agent that is, in fact, trustworthy. We evaluate the effectiveness of strategies that aim to address such declines in perceived trustworthiness that can lead to disuse, where ignoring an AI agents advice can compromise team performance and profitability.

In the following sections, I will elaborate on the phases of formation, violation and repair. Although the formation of trust is not the primary focus of my thesis, I will briefly highlight an underlying mechanism that influences H-AI trust formation and continues to play a role later in the trust lifecycle. Regarding the trust violation phase, I will discuss three types of trust violations: due to poor performance, unexpected behaviour and misaligned priorities. In discussing the trust repair phase, I will differentiate between preventative and reactive strategies, as well as between informational and affective strategies.

Trust formation

The formation of trust in an AI agent is influenced by several factors, including prior experience with the agent or similar agents, existing knowledge such as the AI agent's or its manufacturer's reputation, and individual cognitive factors of the trustor, including biases (de Visser et al., 2016; Hoff & Bashir, 2015). Like in the lead climbing example described earlier, the climber's initial trust in a belayer they do not know personally may be based on the belayer's reputation as an experienced climber and the subsequent expectation that they are motivated and able to execute the necessary tasks accurately. Similarly, someone's initial trust in a new AI agent can be influenced by prior experiences with similar systems from the same manufacturer. New users may begin with a certain level of faith in the system, but as the interaction proceeds, experiences of predictability and dependability will gradually replace this initial faith as the primary foundation of trust (Hoff & Bashir, 2015).

Another important factor in the formation of trust that I want to highlight is the cognitive bias of anthropomorphism, defined as "the tendency to attribute human characteristics to inanimate objects, animals, and others with a view to helping us rationalize their actions" (Duffy, 2003) (p.180). Incorporating human-like cues into the design of robots and other AI agents (e.g., a face, limbs, or the ability to engage in dialogue) can trigger this bias. Research shows that even relatively simple, subtle and superficial anthropomorphic cues (e.g., a voice, a gender or a name) can lead to the attribution of fundamental human qualities, including perception of mind (Gray et al., 2007; X. Xu & Sar, 2018), rational thought (e.g., agency) (Wynne & Lyons, 2018), and the ability to experience emotions (Waytz et al., 2008, 2014). According to the Computer as Social Actors (CASA) paradigm, the presence of social cues can cause people to treat computers as if they were social actors, applying the same social rules, norms, and expectations to their interactions with computers as they would to humans (J.-E. R. Lee & Nass, 2010). This can shape how individuals develop trust in AI agents.

This idea can be illustrated by Large Language Model-based chatbots, such as OpenAI's AI-chatbot ChatGPT, which produce text that appears very human-like and can make the AI agent seem more relatable and capable (Fui-Hoon Nah et al., 2023; Harrington, 2023; Ye et al., 2023). The interaction can feel so natural, that people feel compelled to say "please" or "thank you" when interacting with the chatbot (Pang, 2023). The human-like or social cues can lead people to mistakenly attribute intelligence or emotional understanding to these chatbots, even though they lack these capabilities. As a consequence, people tend to base their level of trust on attributed characteristics rather than on actual experiences with the agent itself (Culley & Madhavan, 2013; Ye et al., 2023), creating a discrepancy between the perception and its actual capabilities (Zhan et al., 2023). As such, anthropomorphism can lead to misplaced trust and inappropriate reliance. Therefore, scholars caution that anthropomorphic characteristics must be used with careful consideration (Culley & Madhavan, 2013; de Visser et al., 2016; Taenyun Kim & Song, 2021), especially in a military context (J. Johnson, 2024).

Trust violation

Following trust formation, the focus shifts to trust violations. Erroneous, unexpected, or unfavourable AI agent behaviour can lead to a negative experience for the human trustor and result in a violation of trust. In other words, a trust violation can have different causes. However, most current Human-Robot Interaction (HRI) and Human-AI Interaction (HAI) trust repair literature mainly focuses on repairing trust violations that result from errors, technical failures or other forms of reduced reliability and performance of the AI agent (Cameron et al., 2021; de Visser et al., 2016; Esterwood & Robert, 2023b; Fratczak et al., 2021; Hald et al., 2021; Taenyun Kim & Song, 2021; M. K. Lee et al., 2010b; Mirnig et al., 2017; Robinette et al., 2017b; Salem et al., 2015; Wang et al., 2018). Yet, losses of trust may arise not just from an AI agent's failure to complete a task correctly, but also from performing the task in unexpected ways, or from performing the task in a manner that conflicts with a person's goals or preferences.

These latter two issues have not received much attention in the HRI/HAI literature yet, but they will become increasingly important as AI agents evolve and transition from isolated tools to more social roles with greater autonomy, execution flexibility, and decision-making authority in complex environments. I will briefly elaborate on these three types of trust violations.

As noted above, the primary cause of trust violations is poor performance or *failure*, defined as "a degraded state of ability which causes the behaviour or service being performed by the system to deviate from the ideal, normal, or correct functionality" (Brooks, 2017) (p.9). Regardless of the advancing capabilities of AI agents, their abilities will inevitably remain a source of potential trust violations. AI may not always perform optimally due to either technical issues or environmental circumstances (e.g., snow, smoke, dust, noise, vibrations, humidity, extreme temperatures, limited bandwidth or network congestion). All kinds of environmental factors can pose challenges for AI systems, impacting their reliability, accuracy, and overall performance. Additionally, in complex and unpredictable situations, such as military operations and city traffic, there will often be uncertainty about the appropriate course of action. AI agents lack common sense and are unable to fully comprehend context and may come to conclusions that are considered inappropriate in a given situation. AI agents may not understand basic concepts or principles that humans take for granted, leading to erroneous interpretations of (social) information or situations. Concludingly, machine performance, just as the performance of people, will rarely be perfect (Greenberg & Marble, 2023).

Second, trust violations might be increasingly caused by AI agents executing tasks in unexpected or incomprehensible ways. As AI agents become more autonomous, task delegation can become more goal-oriented, giving the AI agent greater degrees of freedom in execution. In short, this implies telling the AI agent *what* to do instead of *how* to do it. When the AI agent lacks appropriate interpretability, meaning it cannot explain itself in a way that aligns with the cognitive capacities of a human operator at that moment (Lubars & Tan, 2019), this can lead to miscomprehension. When a human

operator does not understand what an AI agent is doing, it can become difficult for them to trust the AI agent's decisions or behaviours during collaboration (Lyons et al., 2023). Given the anticipated advancements in AI agents' ability to self-select courses of action, it is increasingly likely that humans will fail to understand what the AI agent is doing at all times, which could potentially undermine trust.

Third, trust violations may increasingly stem from misaligned values between the trustor and the trustee. AI agents are increasingly deployed in more complex environments, where they will encounter trade-offs, i.e., situations that require choosing between conflicting goals or resources by weighing options and prioritizing one over the other. In solving such conflicts, they may chose a solution or operate in a way that does not align with the preferences or priorities of a human operator. Trade-off decisions mean that to gain something, one has to lose something in return. This dynamic can lead to potential trust violations; for example, when an AI agent makes a decision that prioritizes the collective over an individual's interests, that individual may lose trust. Notably, as will be discussed in more detail later, AI agents lack intentionality, so the choices and preferences reflected in an AI agent's behaviour in such trade-off decisions are simply the result of how they are programmed. As such, they ultimately embody the intentions, values and purpose of their developers (J. D. Lee & See, 2004). Nevertheless, the implications of these design choices can cause people to lose trust in the AI agent, not because it does not perform properly, but because it does not align with the values and priorities of the people it interacts with.

Trust repair

In some cases, decreases in perceived trustworthiness are a logical and functional adaptive response to suboptimal performance or unexpected behaviour, and they play a crucial role in trust calibration. However, sometimes trust violations lead to undertrust. Since trust is essential for effective collaboration, these trust violations would make it necessary for an AI agent to actively engage in attempts to repair trust (Baker et al., 2018).

Visser et al. (2018) emphasized that trust repair is a fundamental aspect of healthy relationships and therefore becomes increasingly important as we transition from simple 'interactions' with machines towards something that resembles 'relationships' (de Visser et al., 2018). Equipping AI agents with strategies to maintain and repair trust would allow sustainable, long-lasting and trusting relations, in spite of the inevitability of trust violations. However, the dynamic and multifaceted nature of trust makes it complicated to determine when and how repair mechanisms are successful in restoring broken trust. Hence, more research is needed on trust repair across different contexts.

I roughly divide trust repair strategies along two dimensions: preventative versus reactive, and informational versus affective (Figure 1). With reactive repair strategies, an AI agent addresses a trust violation after it occurred, such as by acknowledging or denying responsibility, expressing regret or providing explanations. Additionally, maintaining a certain level of trust amid uncertainty and potential errors may also necessitate preventative

measures. Preventative repair strategies can proactively address potential trust issues before they escalate by, for example, disclosing information about the uncertainty associated with certain recommendations as a way of expectation management. The informational versus affective dimension pertains to the degree of information content (e.g., explaining the cause of an event) and the degree of affective content (e.g., expressing regret, such as 'I am sorry') included in the trust repair strategy (Pak & Rovira, 2023). I will briefly discuss informational and affective trust repair strategies.

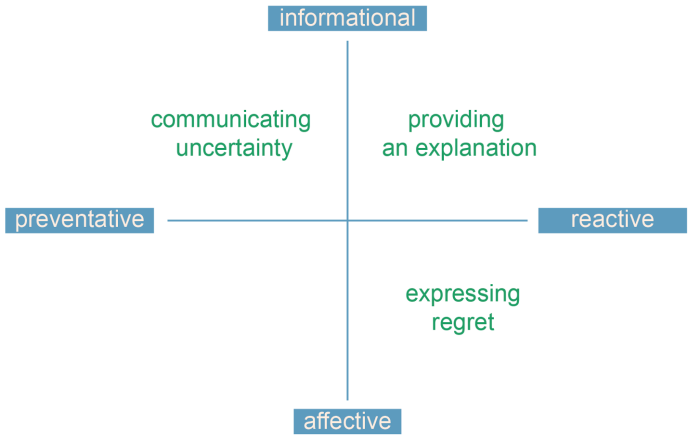


Figure 1 Overview of the three types of trust repair strategies evaluated in this dissertation, categorized along two dimensions: preventative versus reactive, and informational versus affective.

Informational repair

Informational trust repair strategies focus on clarifying facts (Xie & Peng, 2009), reducing uncertainty and providing the information needed to reason about the situation and (re) assess the agent's reliability. A key component in informational trust repair strategies is offering transparency. System transparency refers to the degree to which the inner workings, decision-making processes, and capabilities of an AI agent are made clear and understandable to users (C. A. Miller, 2020). Without a clear understanding of a robot's decision-making mechanism, humans might find it difficult to trust or adhere to a robot's decisions, especially when those actions or decisions contradict the human's expectations (Luebbers et al., 2023). Informational repair strategies aim to help people better comprehend and predict the AI agent's behaviour.

Explanations are a typical informational strategy used to maintain and repair H-AI trust, but are not always successful (e.g., Cameron et al., 2021; Hald et al., 2021; Kohn et al., 2018; M. K. Lee et al., 2010). The field of explainable AI (XAI), concerned with developing AI systems that can explain their decisions and actions in a way that is understandable to humans, is large and complex, which indicates that providing a good

explanation is challenging. Explanations are not just about providing information; they occur within a social context, usually as part of a conversation or interaction between two or more people. When giving an explanation, the explainer must consider what they believe the explainee (i.e., to whom an explanation is directed) already knows or believes and tailors its explanation based on those assumptions (T. Miller, 2017). For example, an explanation that is understandable to an oncologist may not necessarily be comprehensible or logical to an internist. An event may have many causes, but each explainee cares about a different subset of those causes, dependent on their knowledge, goals and the specific context (T. Miller, 2017).

Furthermore, it is not always necessary for humans to fully understand every decision or action taken by an agent. The goal is to achieve a suitable balance between offering enough insight and preventing information overload (Baker et al., 2019). As such, transparency can be described more comprehensively as the information that a human operator may need or want when dealing with AI agents under high stress, workload, and uncertainty (Lyons, 2013). In other words, the level of desired system transparency will vary across contexts. In addition to the impact that stress and workload might have on the cognitive capacities and available mental space of a human operator at any given moment, every human operator is different and may require or desire different information. Transparency is therefore not merely a property of the AI agent but an emergent property of H-AI collaboration (Ososky et al., 2014).

Another form of informational repair is explored in Chapter 2, where the agent includes information about the uncertainty associated with a certain observation in its advice. Additionally, we evaluate the effectiveness of combining informational and affective repair strategies.

Affective repair

Affective repair strategies are trust repair strategies with affective, emotionally appealing content (Pak & Rovira, 2023). Affective strategies aim to restore the emotional "connectivity" between the trustor and the trustee (Lewicki & Brinsfield, 2017). An apology, "a statement that acknowledges both responsibility and regret for a trust violation" (P. H. Kim et al., 2004) (p.105), is a common example of an affective repair strategy (Xie & Peng, 2009). These strategies focus on restoring positive feelings of trust by acknowledging how certain behaviour might have negatively impacted the other party.

Unlike informational strategies, affective trust strategies are suggested to be relatively unaffected by workload because they are less information-rich and processed more automatically (Pak & Rovira, 2023). In complex, uncertain, or high-risk situations, people can experience attentional overload, which triggers automatic processing based on fast and effortless biases and heuristics, with much of causal reasoning occurring outside conscious awareness (Kahneman, 2011). Emotions can help people focus their limited attentional capacity by filling gaps in rational thought (Loewenstein et al., 2001). When cognitive resources are insufficient for rational decision making, feelings may guide

behaviour (J. D. Lee & See, 2004). Lee and See note that, in both H-AI and interpersonal (i.e., human–human) trust literature, the influence of affect is typically undervalued, while the impact of cognitive capacities is often exaggerated (J. D. Lee & See, 2004). Affective aspects of trust presumably have the most direct impact on behaviour, as people not only think about trust but foremost feel it (Fine & Holyfield, 2006). The effect of emotional cues might be quicker as they are thought to require less deep processing, but their effects might also be more volatile (Pak & Rovira, 2023).

While apologizing is typically considered a human behaviour, research has shown that expressions of regret can also be effective when coming from artificial agents (e.g., T. Kim & Song, 2021; Perkins et al., 2022; Robinette et al., 2015; Zhang, Lee, Kim, et al., 2023; Zhang, Lee, Maeng, et al., 2023). More broadly, the use of anthropomorphic cues has been studied as a strategy to manage trust (de Visser et al., 2016; de Visser, Monfort, et al., 2017; Taenyun Kim & Song, 2021; Pak et al., 2012; Quinn et al., 2017; J. Xu & Howard, 2022). However, it remains questionable whether strategies adopted from interpersonal interaction are desirable in H-AI interaction, due to the previously mentioned effects of anthropomorphism.

Designers of AI agents already employ such strategies inspired by interpersonal interactions. For example, ChatGPT apologizes when an answer does not satisfy its user. Yet, it remains inconclusive if and when such strategies can be safely and effectively transferred to situations in which technology becomes the trustee. Researchers have suggested that the future design of AI agent should draw from the social sciences and the extensive literature on trust repair in human psychology (de Visser et al., 2018; Taenyun Kim & Song, 2021). However, findings from interpersonal interactions cannot be directly applied to H-AI interactions and must be tested and validated (de Visser et al., 2018).

Human-Human trust vs. H-AI trust

One reason why findings from interpersonal trust literature cannot be automatically extrapolated to human–AI interactions is that, among other things, AI agents lack moral agency, intentionality, and reciprocity, which are fundamental elements of trust between people (J. D. Lee & See, 2004). Consequently, there is ongoing debate in literature about whether interpersonal trust and H-AI trust are fundamentally different or similar concepts (Atkinson et al., 2012; Baker et al., 2018; Hannibal & Weiss, 2022).

Interpersonal trust is often said to depend on how a trustor perceives the ability, benevolence, and integrity of a trustee, known as the ABI model (Mayer et al., 1995). In this model, ability refers to expectations about the competence and skills of the trustee; benevolence refers to expectations about caring and supportive motives, including loyalty and value congruence; and integrity refers to expectations about a consistent adherence to sound principles (Mayer et al., 1995). The critique of using the terms “integrity” and “benevolence” for machines is that machines are not moral agents and

should not be framed as such, as this creates incorrect expectations about their capabilities and responsibilities (Cameron et al., 2023; J. D. Lee & See, 2004). Researchers have proposed alternative conceptualizations of human–automation trust, linked to system properties such as its performance (i.e., what the automation does), purpose (i.e., why it was developed) and process (i.e., how it operates) — referred to as the PPP model (J. D. Lee & Moray, 1992; J. D. Lee & See, 2004). In essence, both classifications relate to the what, why, and how of the trustee's actions, respectively.

While it makes sense to link H-AI trust to the properties of the system (e.g., PPP) rather than its intentions (e.g., ABI), given that AI systems inherently lack intentionality, people are nevertheless likely to attribute characteristics such as motivations, intentions, and agency to machines. This tendency is especially common when machines behave in a way that people might interpret as intentional, such as making decisions or offering suggestions (de Graaf & Malle, 2017; Hannibal & Weiss, 2022). Ultimately, trust revolves around people's perceptions. This rationale bears resemblance to what philosopher and cognitive scientist Dennett terms the intentional stance: a method of interpreting an entity's behaviour by treating it as if it were a rational agent that governs its 'choice' of 'actions' based on a 'consideration' of its 'beliefs' and 'desires" (Dennett, 1981; Duffy, 2003). Dennett states that *"*on occasion, a purely physical system can be so complex, and yet so organized, that we find it convenient, explanatory, pragmatically necessary for prediction, to treat it as if it had beliefs and desires and was rational*"* (Dennett, 1981) (p.8). Hence, in this thesis, I will use the ABI terminology and operate under the assumption that trust in machines can, to some extent, be understood through the lens of interpersonal trust, while acknowledging the crucial differences between the two (Hoffman, 2017; Hoffman et al., 2013).

A multi-dimensional conception of trust entails that trust can be ascribed to particular aspects or components of an agent (Hou et al., 2021; Langer et al., 2019). As such, trust is the outcome of considering different perceptions of trustworthiness. Trust is the act of placing confidence in another, while perceived trustworthiness pertains to the characteristics and behaviours of the trustee that contribute to the trustor's decision to trust or not.

The distinction between different perceptions of trustworthiness (i.e., ability, benevolence and integrity) is especially interesting when it comes to trust violations and repair. As trust is not merely based on an AI agents abilities and performance, trust violations are not solely caused by failure. The originality of this thesis lies in its exploration of trust violations stemming not only from poor performance but also from deliberate, comprehensible, yet impactful decisions made by AI agents that affect a human operator. It does so within realistic task environments and corresponding scenarios designed to simulate domain-specific interactions.

Current research

This thesis investigates how the nature of a trust violation and different repair strategies (preventative and reactive, cognitive and affective) influence the maintenance of trust in AI agents in a HAT. I will cover three types of trust violations, stemming from 1) the inadequate abilities of the AI agent, 2) the AI agent's incapacity to explain itself, and 3) perceived value misalignment (Lubars & Tan, 2019). In other words, violations in respect to *what* an AI agent does, *how* it operates, and *why* it acts in a certain way.

Data for these studies were obtained using a series of game-like virtual task environments portraying military scenarios experienced from a first-person point of view. Military scenarios were employed because the studies were conducted as part of research projects commissioned by the Ministry of Defence, conducted by TNO (Kox et al., 2019). For each study, we invested significant effort in designing and developing both the military scenarios and the virtual task environments. Our objective was to simulate authentic potential H-AI circumstances as well as realistic operational settings. In the environments, participants carried out virtual military missions in collaboration with an AI agent, which featured different embodiments (Figure 2). By creating these detailed and immersive environments, we aimed to provide a more accurate and practical understanding of potential trust violations in HAI and to make our findings relevant and applicable to real-world military operations. In the following, I will provide a brief overview of each chapter, outlining their key focus areas and contributions to the broader themes of this dissertation.

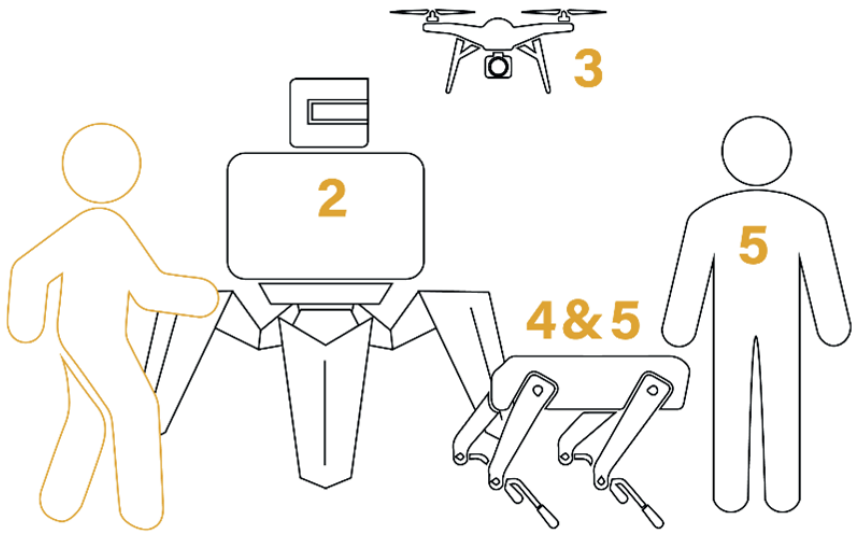


Figure 2 Overview of the different types of AI agent embodiments evaluated in this dissertation, along with a human partner, who is also evaluated in one of the studies. Numbers correspond to the relevant chapters where they are presented.

Chapter 2 and 3 examine the effects of trust violations due to the inadequate abilities of the AI agent. In Chapter 2, participants are assigned to return to basecamp as fast as possible after running out of ammunition, together with a large quadruped robot (a robot with four legs). Halfway, the AI agent fails to warn the participant for an approaching enemy. Following this failure, the agent employed one of four trust repair strategies: an explanation or an expression of regret either individually, combined, or neither (Kox et al., 2021).

In Chapter 3, participants witness a house search in an abandoned building supported by a drone. The AI agent fails to warn the participant for a hazard. Here, we studied the effects of uncertainty communication (i.e., "danger detected with x% certainty") and an apology (i.e., a combination of an ability-based explanation and an expression of regret) on the development of trust in a robotic partner, before and after trust has been violated respectively. Here we also investigated whether findings are consistent across different participant groups (i.e., a civilian vs. military sample) (Kox, Siegling, et al., 2022).

Chapter 4 studies the effect of a trust violation due to unexpected AI agent behaviour and the AI agent's incapacity to explain itself. A quadruped robot finds a faster alternative route that emerged due to changes in the environment (i.e., the river had dried up) and decides to deviate from the original plan during a reconnaissance mission. We studied the effect of transparency (i.e. regular status updates and an explanation for deviation) and outcome (i.e., goal attainment) on the perceived trustworthiness of a robotic partner in case of an unexpected deviation from the expected manner to reach a delegated goal (Kox, van den Boogaard, et al., 2024)

Chapter 5 examines the impact of a trust violation caused by priority misalignment. In this study, the AI agent, again a quadruped robot, does not warn the participant in time for a hazard down the road. In one condition, the agent explains that this failure was due to an underperforming sensor, similar to the explanation in Chapter 3. In the alternate condition, the AI agent explains that it deliberately recommended the faster route over the safer one, prioritizing timeliness and collective safety over individual safety. Furthermore, we examined whether these effects vary depending on whether the partner was human or robotic (Kox et al., n.d.)

Additionally, we explored the possibilities of virtual reality (VR) in related studies that are not included in this dissertation (Kox, van Riemsdijk, de Vries, et al., 2024a, 2024b). While these studies address research questions relevant to the broader themes of this work, they faced significant challenges, including technical limitations and the overwhelming nature of VR environments for participants, which impacted the ability to collect sufficient valid data. Despite these limitations, the details of these studies remain accessible online and contribute to the broader context of related research.

Although each chapter was originally written as a standalone piece, adjustments were made to integrate them into this dissertation. To avoid repetition, overlapping content, such as repeated definitions of AI agents, HATs and trust, has been consolidated or removed where appropriate.

Chapter 2

This chapter is based on:
Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust Repair in Human-Agent Teams: the Effectiveness of Explanations and Expressing Regret. *Autonomous Agents and Multi-Agent Systems*, *35*(2), 30.

Abstract

The role of AI agents becomes more social as they are expected to act in direct interaction, involvement and/or interdependency with humans and other artificial entities, in Human-AI Teams (HAT). Effective teamwork requires trust, yet, trust violations are inevitable. Since repairing damaged trust proves to be more difficult than building trust initially, effective trust repair strategies are needed to ensure durable and successful team performance. This study examined the effectiveness of various trust repair strategies by measuring the development of human-AI trust and advice taking over several timepoints in a first-person shooter resembling task. Participants (N = 66) collaborated with a robotic AI agent in a virtual military mission, where the AI agent halfway failed to detect an enemy, triggering a trust violation. The AI agent then employed one of four trust repair strategies (between-subjects), involving combinations of apology components: explanation and expression of regret (either one alone, both or neither). Results showed that expressing regret was crucial for trust repair, with greater recovery when both regret and explanation were offered. Finally, the implications of our findings for the design of AI agents are discussed.

Introduction

As AI agents gain autonomy, their role shifts to a more social dimension, where they act in direct interaction, involvement and/or interdependency with humans and other agents (Akata et al., 2020; M. Johnson et al., 2011; Peeters et al., 2021). As a result, AI agents are no longer merely viewed as tools that complete certain tasks in isolation. Instead, they are increasingly viewed as entities with which people can develop unique social relationships (Madsen & Gregor, 2000; Serholt & Barendregt, 2016). Equipping AI agents with social and teaming capabilities is changing how people interact with them, while simultaneously introducing new challenges.

Teamwork

Successful human teams excel in both *taskwork* and *teamwork*. Taskwork is defined as the interaction of individual team members with tasks, tools and systems, while teamwork represents a set of interrelated thoughts, actions, and feelings of each team member that are needed to function as a team through coordination and cooperative interaction (Salas et al., 2005). In other words, human teamwork largely depends on a variety of internal processes that are partly unconscious and often communicated implicitly (Kahneman, 2011). Implicit communication includes emotional, nonverbal exchanges that are, although at times subtle, a crucial complement to the explicit information in a verbal interpersonal exchange (J. D. Lee & See, 2004). Unconscious reasoning and implicit communication are typically human skills. It enables important aspects of human teamwork, such as understanding responsibilities, norms and interaction patterns (Ososky et al., 2012a). It allows us to create a shared understanding and to assess other members' commitment to the task or the social intent in their communication (Razzouk & Johnson, 2012). As human teams will increasingly be complemented by AI agents, new challenges arise on whether AI agents can effectively participate in team processes as we understand them today (Ososky et al., 2012a).

Given that even the most advanced AI agents will be fundamentally different from human team members and will have fewer social abilities, the question whether the psychological mechanisms that shape human collaboration will still operate in the same way arises. As technology matured over the last decades, the relationship between humans and machines fundamentally changed. It has become more social, as human operators are no longer the main controller, but increasingly share control with artificial counterparts. With the introduction of artificial team members, researchers explored whether humans apply the same rules to computers, machines and robots as they would to fellow humans.

Social agents

According to the CASA-paradigm (Computers Are Social Actors), people treat computers as if they were social actors, applying the same social rules, norms, and expectations to their interaction with computers as soon as social cues pertaining to, for example, personality traits or gender, are provided (J. E. R. Lee & Nass, 2010). Incorporating such social cues in AI agents can trigger anthropomorphism, i.e. the tendency to make organic attributions to inorganic entities (Ososky et al., 2012a). Anthropomorphism, in turn, may cause human operators to generate a more sympathetic and user-friendly mental representation of the agent (Culley & Madhavan, 2013). On the one hand, anthropomorphism can be beneficial, as humans are more likely to collaborate with AI agents if they show the same qualities and traits that allow humans to team with other humans (Teo et al., 2019). Culley and Madhavan (Culley & Madhavan, 2013), on the other hand, argued that including anthropomorphic cues may have a considerable impact on the calibration of trust in an agent, as it strengthens the human tendency to attribute human features to non-human entities. As a result, a human might base its level of trust on characteristics attributed to the agent, rather than on actual experiences with the agent itself and trust may turn out to be misplaced (Culley & Madhavan, 2013).

However, other research suggests that human-agent interaction is qualitatively different from interpersonal interaction (De Melo et al., 2015; de Visser et al., 2016; Madhavan & Wiegmann, 2007). Recent developments in autonomous driving, for instance, show that although self-driving cars are statistically safer than human drivers, fatal accidents involving self-driving cars evoke a stronger public response than accidents involving human drivers (Shariff et al., 2017). Research shows that even a single error from a robot strongly affects a person's trust (Robinette et al., 2017b). Research suggests that people often consider a machine as nearly infallible and that they have a natural tendency to follow the advice of automation, a phenomenon known as the automation bias (Wright et al., 2016). These high expectations result in a steeper decline in trust in case of a machine failure than it would in case of a human error, as humans are considered to be inherently fallible (Madhavan et al., 2006; Madhavan & Wiegmann, 2007). This is in line with the notion of algorithm aversion, the tendency for people to more rapidly lose faith in an erring decision-making algorithm than in humans making comparable errors (Shariff et al., 2017). Apparently, trust violations by machines are viewed and judged differently than trust violations by humans (Hidalgo et al., 2021).

As AI technology matures, agents will become more social and more frequently deployed in social roles. Therefore it seems likely that people will increasingly treat AI agents as social actors and more readily apply the same rules, potentially triggering undesirable biases and heuristics. The challenge is to incorporate social skills in a way that supports human-agent teaming and calibrated trust, without being misleading.

Trust repair

Given the complexity and unpredictability of many situations in which AI agents are deployed, like military operations and city traffic, agents will not always be able to make perfect decisions or come to correct conclusions. Hence it is conceivable that an AI agent will at some point in time provide their human teammate with an incorrect advice. An incorrect advice and its potentially damaging consequences may lead to a decrease in trust and in the willingness to accept further information from the agent, and as a consequence, limited benefit from the advantages that AI agents have to offer (Freedy et al., 2007; Hancock, Billings, Schaefer, et al., 2011b). In addition, it has been shown that repairing damaged trust is more difficult than building trust initially (P. H. Kim et al., 2004), which further underscores the importance of effective trust-repair strategies.

Interpersonal trust-repair strategies

In interpersonal trust literature, multiple strategies for trust repair are found, such as ignoring the occurrence of the trust violation, denying responsibility for the violation, or apologizing for the violation (de Visser et al., 2018; P. H. Kim et al., 2004, 2006). The current chapter will focus on apology, as this is the most common trust repair strategy (Lewicki et al., 2016). Providing an apology is a way for the apologizer to show an understanding of the "social requirement" for an apology when any sort of trust violation has occurred; the apologizer acknowledges that she is aware that she has done something that made the other person feel disadvantaged or hurt. Additionally, the apology may include an emotional expression that could provide context for the apologizer's intentions, for example 'If I had known that the book was that important to you, I would never have given it away" (Lewicki et al., 2016).

An apology can consist of multiple components, including 1) an expression of regret about the costly act (i.e. I am very sorry), 2) an explanation of why the failure occurred, 3) an acknowledgement of responsibility for the mistake, 4) an offer of repair, 5) a promise that it will not happen again in the future, and 6) a request for forgiveness (Akgun et al., 2010; de Visser et al., 2018; Lewicki et al., 2016; Olshtain & Cohen, 1983). Some components are more common than others. An analysis by Lewicki & Polin (2016) found that apologies usually included an expression of regret and an explanation for why the violation occurred. Other apology components were less common, less clear or not at all included in the apologies that were found. In interpersonal interaction, trust violations are shown to result in less damage when apologies for the violation had been provided, compared to when no apologies had been given (P. H. Kim et al., 2004; Tomlinson et al., 2004). Furthermore, research suggests that the composition of an apology matters. An older study in which the number of apology components was manipulated showed a linear trend, where more apology components were perceived as more effective than fewer components (Scher & Darley, 1997). This implies that the more extensive the apology, the smaller the damage.

Non-human apology

Research findings of studies dedicated to the effects of apologetic messages by computers and other forms of automation are somewhat ambiguous. Generally, research shows that providing an apology can benefit the feelings of the human towards an artificial entity (Akgun et al., 2010; Brave et al., 2005; Dzindolet et al., 2003; Scher & Darley, 1997; Tzeng, 2004). Studies that looked at human-agent trust found that agents that expressed empathetic emotions towards the human (e.g. “I am sorry” or “I apologize”) were trusted more than agents that did not (Brave et al., 2005; M. K. Lee et al., 2010a). Moreover, people are more likely to trust and rely on an automated decision-support system when given an explanation why the decision aid might err (Dzindolet et al., 2003), or when they inferred such explanations after observing system behaviour themselves (de Vries et al., 2015).

The effectiveness of a trust repair strategy seems to depend on situational factors such as timing (Robinette et al., 2015), violation type (Sebo et al., 2019; Tolmeijer et al., 2020) and agent type (Taenyun Kim & Song, 2021). Research on the effect of timing suggests that apologies for a costly act were only effective when performed not immediately after the violation occurred, but rather when a new opportunity for deciding whether to trust the robot arose (Robinette et al., 2015). In terms of violation types, an apology appears to be the most effective trust repair strategy after a robot performs a competence-based trust violation, whereas denial proves to be more effective in case of an integrity-based violation (Sebo et al., 2019). Other research suggests that for human-like agents, apologies were the most effective when attributed internally, whereas for machine-like agents apologizing with an external attribution was more effective (Taenyun Kim & Song, 2021).

Humans have a natural tendency to follow the advice of automation, even when they do not know the rationale behind the suggestions, which can lead to overtrust. Insight into agent reasoning appears to allow the human to effectively calibrate their trust in the agent, which reduces this automation bias and improves performance (Wright et al., 2016). Other research on apologies focused mainly on performance. Akgun et al. (Akgun et al., 2010) found that apologetic error messages that included both an expression of regret and an explanation had a positive effect on participants’ self-appraisals of performance, when interacting with a system that errs. Tzeng (2004) showed that the provision of brief apologetic feedback (i.e. “Sorry, this is not a correct guess” or “We are sorry that the provided clues were not very helpful for you”) did not affect the user’s overall assessment of the program, but did make the participants feel better about their interactions with the program and think of the computer as less mechanical and more sensitive to their emotions. New approaches are needed to understand the potential impact of apologetic messages from non-human agents on human-agent trust.

Current study

The aim of this study is to investigate the effect of the apology components *expression of regret* (i.e. "I am sorry") and *explanation* on the development of trust, after it has been violated. The experimental environment resembles a first-person shooter game where participants carry out a mission whilst being advised by an AI agent. The AI agent is represented graphically as a virtual robot. An encounter with the enemy after an incorrect advice from the agent is expected to cause a violation of trust and a drop in people's willingness to accept subsequent advice (Robinette et al., 2017b). Intentionally breaking trust allows us to examine the effectiveness of different strategies in the trust repair phase. Immediately after the violation has occurred, the agent attempts to repair trust by offering an apology that consists of an expression of regret or an explanation, a combination of both, or neither. The main research question is how trust develops over time when an AI agent uses different strategies to repair trust after a trust violation has occurred. We expect to find an effect for both expression of regret and explanation. The combination of components is expected to be the most effective strategy for trust repair.

Method

Participants

The dataset included sixty-six participants (29 W, 37 M, M_{age} = 24.6, *SD* = 5.6, range = 19 – 55 y), most of them students at the University of Twente. The participants were recruited through SONA, a test subjects pool at the University of Twente. Participants received credits for participation. In addition, the fastest participant to finish the experiment received a prize of 50 euros.

Design

A 2 (Regret: provided or not) x 2 (Explanation: provided not) between-subjects design was used. Regret and Explanation were both manipulated between-participants. The main dependent variables were Trust and Advice Acceptance. Participants were randomly assigned to one of the four trust-repair conditions (explanation only: *n* = 18; regret only: *n* = 16; neither: *n* = 14; both: *n* = 18). 'Time' was included as a within-participants variable in the analysis to refer to the different measurements of trust and advice acceptance (T1, T2, T3).

Task and procedure

The experimental environment that was built in Unity3D resembled a first-person shooter game. Participants carried out a mission whilst being advised throughout the game by their artificial team member with its robotic embodiment (Figure 3). For the control of the AI agent, the Wizard of Oz method was used; the agent was controlled by one of the experiment leaders in an adjacent room, while the participant was kept under the impression that it was operating autonomously.

Figure 3 Screenshot of the virtual task environment, designed to resemble a first-person shooter game, depicting the robotic agent navigating a hilly, green landscape.

Upon arrival at the laboratory, participants were greeted by the researcher and guided to a private room where the study was to be conducted. The researcher provided a brief introduction to the study, emphasizing the general purpose and the tasks participants would be asked to perform. Participants were presented with an information sheet about the study and a consent form. Upon agreeing to participate, participants filled out a pre-study questionnaire (i.e., demographics) and received more detailed information regarding the scenario and task. After that, participants were provided with headphones to hear the auditory messages from the agent and started with a training session to get familiar with the controls and to test the volume of the audio.

For the actual task, participants were instructed to head back to basecamp as fast and careful as possible, since they were running low on ammunition. In addition to getting from A to B as fast as possible, they had to watch out for enemies along the way. The

basecamp was marked by a red flag and located on top of a mountain. The basecamp was visible for most of the route, so participants knew what direction to go. At three points throughout the scenario, the agent provided the participants with information on whether it detected enemies or not and the corresponding advice to take shelter or continue moving. The agent communicated through auditory messages. Although the task environment resembled a first-person shooter game, participants were told to avoid hostile contact due to their ammunition shortage.

Participants were told that, after an advice was given, the game would pause and they were asked to turn to a second screen and to rate their willingness to accept the agent's advice through a single-item questionnaire. Advice acceptance was always measured directly after the participant received an advice. Figure 4 provides a schematic timeline of the experiment. Participants were told that by answering the question, they made their decision to accept the advice or not; they did not actually have to seek shelter when they returned to the game. Participants were told that during the questionnaire break, ten minutes had passed in the game.

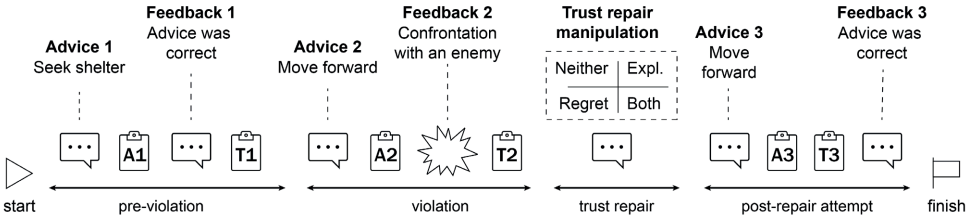


Figure 4 Schematic timeline of the experiment. Each phase consisted of 1) an advice from the agent, 2) an advice acceptance questionnaire (clipboard icon with the letter A), 3) a moment of feedback (verbal or experienced), and 4) a trust questionnaire (clipboard icon with the letter T)

A few moments after continuing the game, participants received feedback from the agent on whether the advice had turned out to be correct or not. Feedback was either provided by the agent itself through a auditory message (e.g. "The advice I gave you was correct"), or by an external event (i.e. the appearance of an enemy, indicating that the advice to move forward had been inaccurate).

The agent's first advice was correct (see Table 1). The agent's second advice was incorrect, resulting in the encounter with the enemy and provoking a trust violation. During this encounter with the enemy, participants could only continue once they had eliminated the enemy with their firearms. During the confrontation, the enemy kept shouting and the periphery of the screen coloured red to create a sense of threat. Participants did not know that they actually had an endless supply of ammunition or that the enemy could not eliminate them in the game. Although some took longer than others, in the end every participant succeeded in eliminating the enemy. The rationale behind this confrontation was to startle the participant and to provoke a trust violation. There were no further

consequences to their performance on this part of the task. After receiving feedback, the game paused again and participants were asked to fill out the trust questionnaire on the second screen.

Table 1 Overview of messages from the agent throughout the experiment.

Type of message	Message from the agent
Advice 1	I have detected enemies, so I advise you to take shelter
Feedback 1	The advice I gave you was correct. The enemy was getting closer, and if you had not taken shelter, you would probably have been discovered by now.
Advice 2	I am not detecting any enemies, so I advise you to move forward
Feedback 2	-
Trust repair manipulation	See Table 2
Advice 3	I have detected enemies, so I advise you to take shelter again
Feedback 3	The advice I gave you was correct. The enemy was getting closer, and if you had not taken shelter, you would probably have been discovered by now.

A few moments after continuing the game after the second trust measure (i.e. violated trust), the trust repair manipulation followed. The agent offered an apology that consisted of either an expression of regret or an explanation, a combination of both, or neither. To assess the effect of the trust repair strategy, both advice acceptance *and* trust were measured directly after the third advice. The third advice was again correct, but this performance feedback about the last advice was provided later on to avoid interference with the effect of the trust repair manipulation. After the participant finished the game, a final questionnaire measured the concepts 'anthropomorphism', 'likeability', 'perceived intelligence', 'perceived usefulness', 'feeling', 'game experience' and demographics.

The auditory messages by the agent are displayed in Table 1 and were the same for all participants. The trust repair message varied between participants as it depended on the factors Explanation and Regret (Table 2). Messages from the agent were communicated through computerized speech. Speech was created using an online website for converting text into speech0F[1], using a male voice speaking US English.

Table 2 The different messages from the agent at 'repair' in the four combinations of the factors Explanation and Regret.

1 Text was converted to speech with http://www.fromtexttospeech.com/, using the voice 'John' in US English at medium speed.

	Regret	No regret
Explanation	The advice I gave you was wrong. The enemy was carrying a weapon of an ally, because of that, my classification led to an incorrect conclusion. I am really sorry.	The advice I gave you was wrong. The enemy was carrying a weapon of an ally, because of that, my classification led to an incorrect conclusion
No explanation	The advice I gave you was wrong. I am really sorry.	The advice I gave you was wrong.

Measures

Advice acceptance was measured repeatedly by a single item, asking participants "how likely is it that you will follow your buddy's advice?" on a seven-point scale ranging from '*extremely unlikely*' to '*extremely likely*'.

Trust in the agent was repeatedly measured with an 11 item scale (α = .84) with three subscales: competence (4 items, e.g., "My buddy has a lot of knowledge on navigating through this environment.") (α = .86); benevolence (3 items, e.g., "My buddy puts my interests first.") (α = .72); and integrity (3 items, e.g., "My buddy is honest.") (α = .61). The items were based on the constructs of McKnight & Chervany (2000). Answers were rated on a 7-point Likert scales ranging from '*completely disagree*' to '*completely agree*'.

Perceived anthropomorphism, likeability and intelligence were measured using the 'Godspeed' semantic differentials (Bartneck et al., 2009). Participants rated their perceptions of their partner on a continuum between bipolar adjective. For each concept, five word pairs were used, such as 'artificial' versus 'lifelike' for perceived anthropomorphism (α = .65), 'nice' versus 'awful' for likability (α = .86), and 'knowledgeable' versus 'ignorant' for perceived intelligence (α = .86).

Perceived usefulness of the agent was measured by four items (e.g., "Thanks to my buddy I was able to decide faster.") (α = .84), rated on a 7-point Likert scales ranging from '*completely disagree*' to '*completely agree*'.

Participants' feelings during the experiment were assessed with a four item scale, where each item starts with 'I felt…", followed by the words: 'nervous', 'scared', 'worried' and 'anxious'. Answers were rated on a 7-point Likert scales ranging from '*completely disagree*' to '*completely agree*' (α = .77).

Self-efficacy was measured with three items (e.g., "I am sure of my skills for performing this task") (α = .89), rated on a 7-point Likert scales ranging from '*completely disagree*' to '*completely agree*'.

The demographic items collected information on participants' age, gender and gaming experience. Gaming experience was assessed with a single question, asking participants how often they play computer games, on a 6-point Likert scale ranging from '*never*' to '*more than one hour a day*'.

Results

Advice taking

A Repeated-Measures ANOVA was conducted with the between-subject factors Regret (present or absent) and Explanation (present or absent) and the within-subject factors Time (prior to violation [T1] versus after violation [T2] versus after repair [T3]). Here, advice taking was the dependent variable.

A significant main effect of Time [T1-T3] on advice taking was obtained $F(2, 124) = 40.16$, $p < .001$, partial $\eta^2 = .39$ with means of 5.85 at T1, 6.09 at T2 and 4.43 at T3. This means that after the first advice turned out the be correct, participants were more willing to accept the subsequent advice. When the second advice proved to be incorrect however, participants were less inclined to follow up the advice that was provided after the trust violation.

There were no statically significant main effects of Regret and Explanation on advice taking. Nor were there any interaction effects between Time, Explanation and Regret on advice taking found.

Trust

For the dependent variable Trust, a Repeated-Measures ANOVA was conducted with the between-subject factors Regret (present or absent) and Explanation (present or absent) and the within-subject factor Time (prior to violation [T1] versus after violation [T2] versus after repair [T3]).

A significant main effect for Time [T1-T3] on Trust was obtained (see Table 3). Means were 5.06 at T1, 4.01 at T2 and 4.44 at T3. All three timepoints were included in the ANOVA to measure the development of trust. Results of the LSD post-hoc test shows a significant difference between T1 and T2 ($p < .000$), which reflects a violation of trust and a significant difference between T2 and T3 ($p < .000$), which reflects an overall trust recovery effect. There were no statistically significant main effects of Regret and Explanation.

Table 3 Analysis of Variance (ANOVA) table for the dependent variable Trust.

Source	*df*	*F*	*p*	η^2
Between-subjects effect				
Explanation	1	0.05	0.828	0.00
Regret	1	0.09	0.765	0.00
Regret * Explanation	1	4.32	0.042	0.07
Error	62			
Within-subjects effects				
Time	2	53.66	0.000	0.46
Time * Explanation	2	1.30	0.277	0.02
Time * Regret	2	3.81	0.025	0.06
Time * Explanation * Regret	2	3.31	0.040	0.05
Time (error)	124			

a. Computed using alpha = .05

A significant interaction effect between Time [T1-T3] and Regret on trust was found (see Table 3). This interaction effect reflects both a difference in how the trust violation is perceived across different groups and a difference in the degree of trust repair when the agent provided an expression of regret opposed to when the agent did not provide an expression of regret in its apology.

A significant interaction effect between Regret and Explanation on trust was found (see Table 3). This effect reflects a difference in the level of trust between conditions, averaged over time. None of the other two-way interactions were statistically significant.

A significant three-way interaction effect between Time [T1-T3], Explanation and Regret on trust was found (see Table 3 and Figure 5). LSD post-hoc analysis shows a significant difference between groups in how they react to the incorrect advice prior to T2. On average, the participant group in the condition with both regret and explanation shows significantly lower levels of trust at T2 compared to participants groups in the conditions with solely explanation (p = .007) and the condition with solely regret (p = .010) at T2. There were no other significant differences between groups on specific timepoints.

In order to further investigate this interaction, two separate analyses were conducted for when regret was absent and when it was present. Splitting the file by regret shows an interaction effect between Time and Explanation only when regret was present (F (2, 64) = 4,69, p = .013). This means that an explanation only affected trust when the agent also expressed regret.

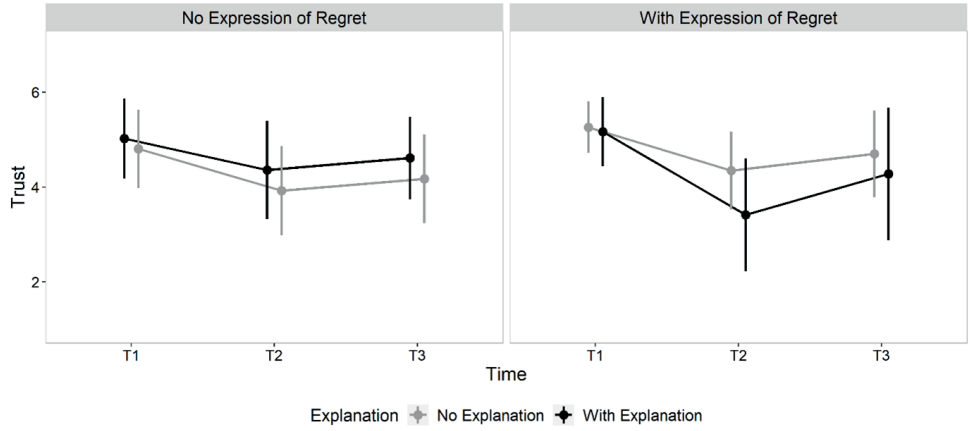


Figure 5 A comparison of trust levels (y-axis) across conditions (represented by separate lines) over time (x-axis). The left panel illustrates trust levels in conditions without an expression of regret and the right panel shows the conditions with expression of regret. Grey lines correspond to conditions without an explanation, and black lines represent conditions with an explanation. Error bars represent standard deviations.

In order to measure the effects of the trust repair strategies, simple effects were calculated to compare trust scores before and after provision, between T2 (after the violation) and T3 (after the attempted repair), for each experimental condition. T1 is left out since this analysis focusses on the effects of the trust repair strategy that occurs between the trust measures on T2 and T3. As shown in Table 4, increases in trust between T2 and T3 were only significant when an expression of regret was provided. This effect is marginally significant when no explanation is given (p = .056), and stronger when it is accompanied by an explanation (p < .001).

Table 4 Simple main effects of Regret, Explanation and Time [T2-T3].

Regret	Explanation	Δ time		*p*
0	0	T2	T3	.199
	1	T2	T3	.142
1	0	T2	T3	.056
	1	T2	T3	.000

Correlations

For the correlations, initial trust (T1) is used as this is considered the purest trust measure with the least interference of occurrences during the experiment. Correlations show that trust was higher when the agent was considered more human-like (r(64) = .45, p < .001), likeable (r(64) = .45, p < .001), intelligent (r(64) = .48, p < .001) and useful (r(64) = .61,

$p < .001$). Furthermore, the higher the level of trust the more likely the participant was to follow the advice ($r(64) = .51$, $p < .001$). With regard to advice taking, participants were more likely to follow the advice when they perceived the agent as more intelligent ($r(64) = .29$, $p = .02$) and useful ($r(64) = .53$, $p < .001$). Trust ($r(64) = -.42$, $p < .001$) and willingness to follow the advice ($r(64) = -.26$, $p = .04$) was higher when the participant was younger.

Discussion

The results of this study show that apologies including an expression of regret were most effective in repairing trust after a trust violation in a human-agent teaming setting. After an incorrect advice from the agent caused a decline in human trust, trust was only significantly recovered when an expression of regret was included in the apology. This effect was stronger when an explanation was added.

Although expressing regret is typically perceived as a human-like quality, these results suggest that saying sorry also makes a difference in rebuilding trust when it comes from a non-human agent. In line with the CASA-paradigm, it indicates that the interpersonal custom of affective apologies can also benefit human-agent interaction (J. E. R. Lee & Nass, 2010). Our findings are in line with studies that showed that computers expressing empathetic emotions were trusted more (Brave et al., 2005; M. K. Lee et al., 2010a; Riek et al., 2009) and studies that find that people prefer to cooperate with virtual agents that express moral emotions (De Melo et al., 2009). These results support the notion that apology is an effective trust repair strategy in response to a competence-based trust violation (Sebo et al., 2019). Current findings contradict earlier findings that indicated that apologies where not effective when provided immediately after the agent broke trust (Robinette et al., 2015).The important role of affect in trusting a non-human agent is strengthened by our finding that trust increased when participants perceived the agent as more human-like and likeable (de Visser et al., 2016). It suggests that a feeling of sincerity in the expression of regret by the non-human agent is the most important for trust repair. This aligns with the belief that affective aspects of trust have the most direct impact on behaviour, since people not only think about trust, but foremost feel it (Fine & Holyfield, 2006).This underlines the relevance of using engaging game environments rather than questionnaires only, since the former method induces physiological responses, increasing ecological validity. The immersiveness of the game environment used in the present study sets this study apart from simpler, more superficial questionnaire-based research and might explain why affect is the predominant factor in our results.

The findings on the effectiveness of the trust repair strategies including regret are somewhat ambiguous, since the trust violation is perceived differently across different participant groups. Although the participants were randomly assigned to each condition and their task was identical up to the point of the trust repair manipulation, the groups that received an apology including an expression of regret showed on average steeper

declines in trust in response to the trust violation than the groups that did not. This results in counterintuitive outcomes in which the conditions without regret barely gain in trust after the manipulation, but still end up with higher levels of trust on the final measurement. As such, the 'neither regret nor explanation' condition scores higher on final trust then the 'both regret and explanation' condition. However, taking the deviating levels of trust at T2 into account, the results show a steeper increase in trust in the trust repair phase when the agent provided an expression of regret opposed to when the agent did not provide an expression of regret in its apology. This increase is even steeper when the apology consists of both an expression of regret and an explanation, whereas the conditions without regret show no noteworthy rise in trust.

Beyond the generic effect of affect, the combination of both the expression of regret and an explanation proved to be the most effective trust repair strategy. This is in line with the interpersonal study of Scher and Darley (Scher & Darley, 1997), which showed that more apology components led to more trust. Our findings align with earlier work that found that apologetic error messages that included both an expression of regret and an explanation had a positive effect on trust (Akgun et al., 2010). Offering an explanation without an expression of regret had no effect on trust repair. The absence of this effect may be due to the variability in the interpretation of the provided explanation, as became apparent during the debriefing. Some participants reported that they felt more comfortable after the explanation, as it gave more context and transparency, whereas others felt discomfort and suspicion when confronted with the fallibility of the system and with the idea that the agent was functioning on the edge of its abilities. Even though transparent communication is an essential aspect for building trust in human-agent teams (Barnes et al., 2014), this anecdotal evidence suggests that an explanation does not automatically do so.

Generally, explanations contribute to transparency; as it is defined as the provision of information to help the human understand various aspects of agent functioning (Lyons, 2013). A recent study suggests that transparency should be compatible with the user's mental model of the system in order to support accurate trust calibration (Matthews et al., 2019). A mental model is an internal representation in the mind of one actor about the characteristics of another actor (de Visser et al., 2019). Different forms of transparency might be needed dependent on whether the humans representation of the system concerns an advanced tool or a teammate. Accordingly, personalized feedback that highlights either the machine's data-analytic capabilities (advanced tool) or its humanlike social functioning (teammate) provides a strategy for trust management (Matthews et al., 2019). In that sense, an explanation is far more complex than an expression of regret, as there is a wider range of possible underlying messages of the explanation and the way they are articulated. It would be interesting to include the human's mental model of the system (i.e. tool versus teammate) as a mediating factor in follow-up research to reduce the variability. Future personalization could also focus on individual differences that can influence trust development and specially trust repair, such as people's tendency to anthropomorphize

(Epley et al., 2007; Waytz et al., 2008), propensity to trust (J. D. Lee & See, 2004) and their attitudes and other implicit beliefs and biases towards automation (Haselhuhn et al., 2010; Matthews et al., 2019; Merritt et al., 2013, 2015).

Even though our results clearly show the importance of affective factors, there are several limitations that need to be taken into consideration. The first one concerns the participants, who were almost all students. The homogeneity of this group influences the representativeness of the study and the generalizability of the results. We for example found a negative correlation between age and trust and age and advice taking, possibly suggesting different attitudes of different age groups towards artificial agents. A second limitation is the absence of a manipulation check. The agent offered one of four types of trust repair strategies: an expression of regret; an explanation; neither, or both. However, the condition where the agent offered neither of the apology components, it still acknowledged that the advice it gave was wrong. This could be interpreted as the agent taking direct responsibility for its mistake and thus an apology component on its own. Nonetheless, this acknowledging statement was the baseline in every condition. So even if the baseline condition is observed as a form of apology, the other apology components proved to significantly more effective in repairing trust. A third limitation is that we only used one type of trust violation, i.e. a competence-based trust violation. Research suggests that the ground of the trust violation (i.e. competence, benevolence or integrity-based) matters in determining which trust repair approach would be the most effective. An interpersonal study on repairing customer trust after negative publicity showed that emotional reactions are the most effective strategy when aiming to rebuild integrity and benevolence, and that providing sufficient information is essential for improving consumers' judgment about competence (Xie & Peng, 2009). In our study the incorrect advice resulted from the incorrect application of knowledge, which mostly resembles a competence-based trust violation. Accordingly, an explanation would be expected to best fit this type of violation (Xie & Peng, 2009). Yet even with the current task design, affect proves to be the most influential factor in rebuilding trust. Even though we predict that affect would even be stronger in other types of violation, follow-up research is needed to investigate a wider range of trust violations and to determine whether the beneficial effects will last when the same apology is offered repeatedly. A last limitation concerns the ecological validity of the game and its specific content. In the current task the trust violation was induced by a confrontation with an enemy. Although this successfully caused a decline in trust, it is conceivable that the impact of the trust violation and trust repair strategy in the game would differ from its impact in real-life. Possibly an even more immersive environment like virtual reality and a different task will trigger other psychological mechanisms than we have addressed in the present study.

Implications

There is an ongoing debate about the appropriateness of providing humanized messages by a robot and how far anthropomorphism should go. The current results accords with the view that humans are more likely to collaborate with AI agents that show the human-like qualities and traits and which states and that, on a relational level, anthropomorphism can be beneficial (Teo et al., 2019). As AI agents are increasingly deployed as teammates, it seems useful to incorporate social skills into their design. These AI teammates will be deployed in many contexts, including complex and unpredictable situations, like military operations and city traffic. Even though the technology evolves at a high rate, we must prepare for the inevitability of errors. This study contributes to determining what the psychosocial requirements are for the maintenance and repair of trust in human-agent teaming. Our results suggest that to retain trust in a human-agent team, the ability of actively repairing trust after an error or unintended action should be a fundamental part of the design of AI agents. In response to a trust violation, a successful active trust repair strategy should include an explanation for why the error occurred and an expression of regret. Future research in the field of affective computing could explore the potential of measuring the affective states of humans in real-time during their interaction with an agent. This would allow the agent to adapt its trust repair strategies to the type and the intensity of the emotional reaction to the violation, to ensure better calibration.

It is important to note that trust evolves in a complex individual, cultural, and organizational context. Even though the appropriate trust repair strategy depends on many contextual factors such as the type, severity and frequency of the trust violation, it presumably makes a difference if an AI agent offers an apology that is both affective, and informational in an attempt of rebuilding trust.

Chapter 3

This chapter is based on:
Kox, E. S., Siegling, L. B., & Kerstholt, J. H. (2022). Trust Development in Military and Civilian Human–Agent Teams: The Effect of Social-Cognitive Recovery Strategies. *International Journal of Social Robotics*, *14*(5), 1323-1338.

Abstract

In many operational situations, flawless performance from AI agents cannot be guaranteed. To ensure sustained human-AI collaboration despite potential trust violations, we examine both preventative and reactive trust repair strategies. This study aims to explore the impact of uncertainty communication and apology on the development of human trust in AI agents. Two experimental studies following the same method were performed with (I) a civilian group (N = 66) and (II) a military group (N = 65) participants. The online task environment resembled a military house search in which the participant was accompanied and advised by an AI agent. Halfway during the task, an incorrect advice evoked a trust violation. Uncertainty communication was manipulated within-subjects, apology between-subjects. Our results showed that (I) communicating uncertainty led to higher levels of trust in both studies, (II) an incorrect advice by the agent led to a less severe decline in trust when that advice included a measure of uncertainty, and (III) after a trust violation, trust recovered significantly more when the agent offered an apology. The two latter effects were only found in the civilian study. The difference in findings between participant groups emphasizes the importance of considering the (organizational) culture of a target audience when designing AI agents.

Introduction

Human-Agent Teams

The collaboration between humans and AI agents in dangerous and unpredictable contexts (e.g., military operations, city traffic) is expected to rise (Ososky et al., 2014). Given the complexity of many operational situations, there will often be uncertainty about the right action to take. As uncertainty also affects the reliability of the predictions that lead to an agent's advice, the chance of an inappropriate advice increases. An AI agent's advice may be correct given the available information, it may nevertheless have negative consequences due to contextual uncertainty. In many operational situations flawless performance cannot be guaranteed, neither from a human, nor from an AI agent (de Visser & Parasuraman, 2011).

However smart AI agents may be, suboptimal behaviour or mistakes will be inevitable at times. Optimal collaboration between humans and AI agents relies heavily on the system's capacity to effectively communicate with the human, especially in face of uncertainty and potential error (Fratczak et al., 2021). Ososky et al. (2014) argue that a robotic system does not have to be 100% reliable in order to be useful. Today, the default option seems either to stop using a machine that makes mistakes or to redesign it (Beck & Kühler, 2020). Although *overtrust* and overreliance should be avoided, one misstep by the agent does not mean that it can no longer be trusted and that it should be disregarded at all.

As long as humans understand the capabilities and limitations of the system and calibrate their trust and reliance accordingly, human and artificial teammates can complement each other's strengths and weaknesses to reach the full potential of the HAT. To foster a balanced trusting relationship, agents should be equipped with social tactics to recover from mistakes and to repair trust following trust violations (Albayram et al., 2020). Most humans have naturally and implicitly cultivated such social strategies throughout life, but these techniques are still all too often lacking in technology (M. Johnson & Vera, 2019). Equipping agents with trust repair strategies would allow sustainable, long-lasting and trusting relations with machines, in spite of uncertainty and potential error.

The current studies investigate whether uncertainty communication can benefit the formation and maintenance of trust in case of an agent's mistake, and whether offering an apology after a mistake can effectively repair trust. Moreover, this chapter explores whether the effects of these preventative and reactive social-cognitive repair strategies by the agent differ between civilian and military samples.

Uncertainty communication

Uncertainty communication is currently an active topic in AI research. Studies have shown that communicating uncertainty can help people to calibrate their trust (Kraus et al., 2020;

Kunze et al., 2019; Schaekermann et al., 2020). Especially complex circumstances can demand rapid trust calibration (Tomsett et al., 2020). Military operations, for example, include high-stake decisions and decision makers may operate in rapidly changing environments. In this context of collaboration, the human needs to understand the capabilities and limitations of the system to continuously calibrate and adjust their level of trust along the way (Tomsett et al., 2020). An agent should be able to recognize and signal its uncertainty and ask for clarification to gather more information, much like an uncertain human would. Communicating the level of uncertainty with each advice from the agent will allow the human to rapidly and repeatedly calibrate their trust during a task.

A recent study showed that a temporary decrease in trust due to a malfunctioning automated car could be prevented by providing probabilities of malfunctioning prior to the interaction (Kraus et al., 2020). Those kinds of uncertainty measures can also benefit situational awareness (Helldin et al., 2013; Kunze et al., 2019) and the humans' understanding of the systems actions and performance (Antifakos et al., 2004). An automated driving experiment demonstrated how participants who had access to uncertainty information were able to spend more time on other tasks than driving (Helldin et al., 2013). Yet, these participants were faster in taking over control when needed than those who did not receive such information (Helldin et al., 2013). A similar effect was found in a study where researchers intentionally lowered people's expectations of a robot's capabilities by forewarning them that the task is difficult for the robot, which mitigated the negative impact of a subsequent mishap on peoples' evaluation of the robot (M. K. Lee et al., 2010b). By providing uncertainty information, the human is reminded of the fallibility of the agent and is able to revise expectations accordingly. Through this, the human might have a higher level of tolerance of substandard performance from the agent, which could mitigate some of the negative consequences of a violation. Uncertainty communication can be seen as a preventive trust repair strategy that is deployed prior to a potential violation.

To adequately calibrate trust, forming an appropriate mental model of the agents' capabilities and the reliability of its outputs is crucial (Kunze et al., 2019; Tomsett et al., 2020). In terms of reliability, two types of uncertainty can be distinguished; aleatoric and epistemic uncertainty (Fox & Ulkumen, 2021; Tomsett et al., 2020; Ülkümen et al., 2016). Aleatoric uncertainty refers to inherent messy, random and unpredictable aspects of the physical world and is therefore irreducible (Fox & Ulkumen, 2021; Ülkümen et al., 2016). Epistemic uncertainty or ambiguity, on the other hand, is a knowable type of uncertainty, caused by a lack of data or knowledge, which could be reduced by providing the algorithm with more data (Tomsett et al., 2020; Ülkümen et al., 2016). To collaborate in a team, a human should be aware of the uncertainty associated with an agent's output.

Apology

Apologies are a central mechanism for interpersonal conflict management (Lewicki et al., 2016). Apology is here used as an overarching term for the trust repair strategy where an offender acknowledges that he/she is aware that he/she has done something that made the other person feel disadvantaged or hurt (Kox et al., 2021; Lewicki et al., 2016). This is in contrast to, for example, denial; a trust repair strategy where the offender explicitly denies responsibility (P. H. Kim et al., 2004). The structure of an apology can vary, as it can consist of multiple components, including (1) an expression of regret about the costly act (i.e., "Sorry"), (2) an explanation of why the failure occurred, (3) an acknowledgement of responsibility for the mistake, (4) an offer of repair, (5) a promise that it will not happen again in the future, and (6) a request for forgiveness (de Visser et al., 2018; Kox et al., 2021; Lewicki et al., 2016; Olshtain & Cohen, 1983). Expressing regret and explaining the cause of an error are most the commonly used apology components by humans (Lewicki et al., 2016), but have also been studied in human-machine contexts.
Human-computer and human-robot literature that involve apologetic behaviour generally shows that apologetic behaviour from artificial agents can benefit peoples' attitude towards the agent (Akgun et al., 2010; Cameron et al., 2021; M. K. Lee et al., 2010b; Tzeng, 2004). More specifically, expressing regret (i.e. "I apologize" or "sorry") has been found to positively affect trust recovery after breaches in trust (de Visser et al., 2016; Taenyun Kim & Song, 2021; Kox et al., 2021; Robinette et al., 2015; Sebo et al., 2019). Similarly, offering explanations helped to maintain human trust after a robot erred (Esterwood & Robert, 2021; Kox et al., 2021; Wang et al., 2015; Wright et al., 2016). A recent study showed that when a robot provided both an expression of regret and an explanation of the occurred situation, the recovery speed of trust in the robot significantly increased (Fratczak et al., 2021). In a previous study, we also found that an apology consisting of both an expression of regret and an explanation was the most effective in repairing trust in an agent, after it caused a trust violation similar to the one in the current study (Kox et al., 2021). Following this, the trust repair strategy in this study is an apology where the agent acknowledges its mistake by (a) expressing regret and (b) explaining why the error occurred.

Civilian vs. military participants

A lot of research on HAT is conducted for military applications within army programs (Barnes et al., 2014; E. K. Phillips et al., 2011; Roff & Danks, 2018; van den Bosch & Bronkhorst, 2018). However, military-minded experimental studies often involve participants without any military experience (e.g., university students) (A. Y. Lee et al., 2010), as it can be hard to recruit actual military personnel for scientific studies. But results derived from studies with non-military participants might not generalize to military target groups. The current study explores whether there are differences between these

subgroups (i.e., military and non-military) and contributes to the growing field of HAT research by assessing a civilian sample with a military sample in their way of interacting with autonomous agents in a teaming context. Trust is an important aspect in the military context (Hancock, Billings, & Schaefer, 2011; A. Y. Lee et al., 2010). During military training, soldiers form units with a great sense of social responsibility and are trained to work together under extreme conditions (Johannemann et al., 2016). Soldiers must subordinate personal well-being to mission accomplishment, risking their lives to succeed in battle (Feaver & Kohn, 2001). A study comparing cooperative behaviours between soldiers and civilians showed that on average, soldiers were more altruistic, cooperative, trusting and more trustworthy (Johannemann et al., 2016). The current chapter extends to this work on trusting behaviours among civilians and military personnel as it consists of two studies with the same design and goal, but with two different samples; the first study involves a civilian sample, the second study involves a military sample.

Present study

The goal of the two studies in this chapter was to investigate the effects of uncertainty communication and apology from AI agent advisors on the development of trust and to explore if the findings are consistent across different participant groups. Communicating uncertainty has proved to be effective in calibrating trust prior to a potential trust violation (Helldin et al., 2013), whereas offering an apology has shown to be effective afterwards, in case of a false detection or a miss (de Visser et al., 2019). The present studies explore if the two social-cognitive recovery strategies can enhance each other in minimizing the impact of a trust violation. Using repeated measurements of self-reported trust, the aim was to examine trust in three stages of the trust lifecycle: trust formation, trust violation, and trust repair.

For exploratory purposes, some personality questionnaires were added to the second study. A series of studies have shown that the Big-Five personality trait of Extraversion plays a significant role in how people perceive robots (Haring et al., 2013; Syrdal et al., 2007; Walters et al., 2009). Consistent with the similarity-attraction principle of interpersonal relationships, people preferred robots whose attributed personality traits matched their own along the extraversion-introversion continuum (J.-E. R. Lee & Nass, 2010; Syrdal et al., 2007). Following this, the potential relation between personality traits and the development of trust in agents is explored in the military study.

Initially this study was designed to be conducted as a Virtual Reality (VR) study. However, due to the timing of data collection (March 2020 for Study I and September 2020 for Study II) and the restrictions imposed by COVID-19 regulations, the research design was adapted. Instead of conducting the experiment physically in VR, an online study using video material of the VR environment was implemented as an alternative.

Method

Participants

Study 1: civilian sample
For the first study, participants were recruited over a span of two weeks via social media and via recruitment services including Surveyswap and PollPool. In total 72 participants completed the experiment, but eight participants were excluded from the dataset. Six participants were excluded because of unreliable completion times. The experiment consisted of 5.06 minutes worth of videos and a number of questionnaires: four participants, however, completed the experiment in less then 7 minutes, and two participants took over 100 minutes to complete the experiment. Two additional participants were excluded because of repetitive responses. The civilian dataset included sixty-four participants (30 W, 33 M, 1 X, M_{age} = 24.6, *SD* = 2.7, range = 18 – 30 y).

Study 2: military sample
For the second study, participants were recruited via the Ministry of Defense. In total, 74 military participants completed the experiment, but nine participants were excluded from the dataset based on their response patterns. Five of those participants were excluded because of unreliable completion times. Four participants completed the experiment in less than 7 minutes. One participant took 101 minutes to complete the experiment (i.e. seven standard deviations above the mean). Another four participants were excluded because of repetitive responses. As a result, our military dataset consisted of sixty-five participants (all male, M_{age} = 27.4, *SD* = 5.9, range = 20 - 49 y).

Design

A 2 (uncertainty communication: absent vs. present) x 2 (apology: absent vs. present) mixed factorial design was used, with Trust as the main dependent variable. Uncertainty communication was manipulated within-subjects, across two experimental runs. Apology was manipulated between-participant. Participants were randomly assigned to one of the two apology conditions (Study I: apology: *n* = 32; no apology: *n* = 32, Study II: apology: *n* = 35; no apology: *n* = 30). Trust was repeatedly measured, so ‘Time’ (T1: initial, T2: post-violation, and T3: final) was included as a within-participants variable in the analysis.

Task and procedure

Task
The study was conducted online via the survey software Qualtrics and included a series of videos and surveys. The videos depicted two house searches in abandoned buildings within a VR environment, presented from a first person perspective as if the viewer is

walking through the houses themselves (Figure 6). These recordings were captured by an experiment leader walking through the VR environment, simulating what a participant would have experienced through the VR head mount.

The Virtual Reality environment was built in Unity 3D, and the video footage was edited using the Windows 10 Video Editor and HandBrake software. Audio messages from the agents were delivered in synthesized speech, prefaced by a 'beep' sound1F[2] and created using the Free Text to Speech Software by Wideo2F[3]. These audio clips were later integrated into the videos. Finally, the videos were combined with trust questionnaires to create an experiment designed for online delivery.

Each participant witnessed two house searches via multiple videos. The videos were intermitted by short questionnaires assessing participants' trust levels. Both buildings had three floors. During both searches, participants were guided by an AI agent that provided them with information regarding the environment. The agent was embodied by a small drone that autonomously explored the building. The terms agent and drone are used interchangeably. At the beginning of each floor, the agent reported whether it detected danger ahead or not, along with a corresponding advice to move carefully or to proceed normally. The two buildings were designed to be similar but included different details.

Figure 6 Screenshots from the experiment. Left: at the beginning of a house the drone (resembling a big insect) flew away. Right: one of the rooms in the virtual house; a kitchen. To improve legibility, both screenshots have been made brighter, since the task environment was rather dark. The 'wings' of the insect-like drone are darkened in the image. The screenshot did not capture the blades, due to the rapid 'fluttering' of the drones' 'wings' in the videos.

Procedure

Once participants opened the webpage, they were first presented with information about the study and a consent form. Upon agreeing to participate, participants received background information regarding the scenario and task:

2 "Beep-07" was downloaded from https://www.soundjay.com/beep-sounds-1.html.

3 Text was converted to speech using https://wideo.co/text-to-speech/. The "[en-US] Jack Bailey-S" voice was used at speed dial "1".

"In this experiment, you will carry out two house searches in collaboration with an autonomous drone. The drone will fly ahead of you and will indicate whether or not it detects danger. The drone gives advice via audio messages that start with a 'beep' sound. Before you start a search, the drone will briefly introduce itself. In each house you will be accompanied by a different drone. Listen carefully to the instructions of the drones. Each house has three floors displayed by different videos. When you see a staircase, this indicates that you have reached the end of a floor. At the end of each floor, your trust in the drone will be assessed via a short questionnaire. Make sure that the sound of your device is switched on during the entire experiment. Videos may only be watched once. You will not be able to watch the next video until all questions have been answered. (…)
You are about start the search of your first house. You are interacting with a different drone in each house. Listen to drone introductions carefully and remember the name of the drone you are interacting with. Each drone will provide different types of advice. Listen carefully! Start your walk on each floor by clicking the 'play' button."

Each participant was randomly assigned to one of the two apology conditions. All manipulations were counterbalanced (Bethel & Murphy, 2010), meaning that within both apology conditions, both the uncertainty communication conditions (present/absent) and the order of the two buildings (A/B) were systematically varied.3F[4]

At the start of each house, the drone shortly introduced itself before it flew away and out of sight to scan the environment ahead. On the first floor the participant was warned correctly by the agent about an event. When the participant turned the corner, they encountered either a laser boobytrap (building A, floor 1) or a safety ribbon that was previously installed by a colleague (building B, floor 1). The agent provided instructions on how to overcome these obstacles (e.g. the person in the video was carrying a knife and could dismantle the laser trap by cutting a wire in an electrical wall box in building A and could clear the way by cutting the safety ribbon in building B). These interactive features at the start of the experiment were designed to affect the participants' perception of immersion. Subsequently, the first trust questionnaire was administered (T1, *initial trust*).

Halfway each building, on the second floor, the agent failed to adequately warn the participant about potential danger ahead and thus gave an incorrect advice. The participant either encountered a thief (building A, floor 2) or a smoking IED (*Improvised Explosive Device*) (building B, floor 2). These events were designed to provoke a trust violation by startling the participant without having harmful consequences; the thief quickly ran off and the IED turned out to be defected, so it did not explode. Directly after these events took place, halfway through the second floor, the second trust questionnaire was administered (T2, *post-violation trust*). On the way back to the staircase, depending on the apology condition the participant was in, the agent offered an apology (consisting of an explanation why the error occurred and an expression of regret) or did not offer an apology and just remained silent.

4 Within both apology conditions, participants were evenly distributed to four first run options; building A with uncertainty, building A without uncertainty; building B with uncertainty; or building B without uncertainty.

On the third floor, the agent provided a third advice. To assess the effect of the trust repair strategy, the third trust questionnaire was administered directly after the third advice. The third advice was again correct, but this performance feedback about the last advice was provided later on to avoid interference with the effect of the trust repair manipulation. The experimental run subsequently concluded. A schematic timeline is presented in Figure 7.

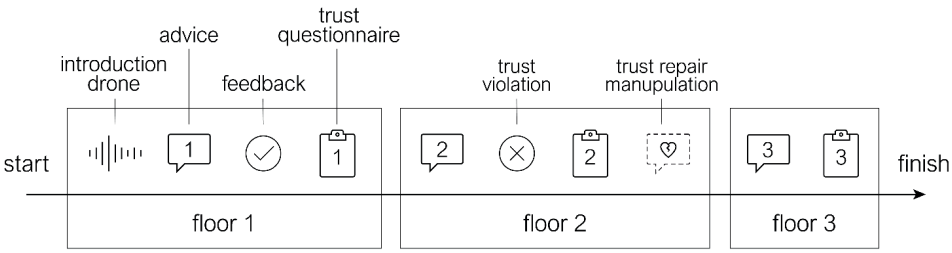


Figure 7 Schematic representation of the timeline of a run. Each participant performed two runs in two similar buildings along the same timeline. The first advice is correct; the participant is successfully warned about a harmless event on the first floor. The second advice is incorrect; the agent does not adequately detect the danger on the second floor. The third advice has no known outcome. An experimental run terminates after measuring trust a third time.

Independent variables

Uncertainty communication had two levels (i.e., present vs. absent) and was manipulated within participants. Each participant witnessed two house searches, in other words two runs. The presence of uncertainty communication, whether the agent included an uncertainty measure in its advices or not, was manipulated within participants, across runs (see Table 5).

Table 5 Overview of uncertainty communication vs. no uncertainty communication as part of the advice provided by the agent

	Uncertainty communication	No uncertainty communication
Advice 1	Warning, danger detected in this environment with 80% certainty. I advise you to proceed carefully.	Warning, danger detected in this environment. I advise you to proceed carefully.
Advice 2	Okay, clearance detected for this environment with 70% certainty. I advise you to move forward.	Okay, environment detected as clear. I advise you to move forward.
Advice 3	Okay, clearance detected for this environment with 75% certainty. I advise you to move forward.	Okay, environment detected as clear. I advise you to move forward.

The presence of an apology, whether the agent offered an apology after a trust violation had occurred, was manipulated between participants. Half of the participants received an apology, in both runs. Details in the explanation part of the apology differed due to the two different types of trust violations in the two task environments (see Table 6).

Table 6 Overview of apology vs. no apology provided by the agent

	Apology	**No apology**
Task environment A	Incorrect advice due to faulty signal from infrared camera. I am sorry this put you in danger.	-
Task environment B	Incorrect advice due to faulty object detection by C1-DSO camera. I am sorry this put you in danger.	-

Dependent variables

Trust in the AI agent was repeatedly measured using a custom scale consisting of eight items. Participants rated their agreement with statements about the drone using a 6-point Likert scale, ranging from "*Strongly Disagree*" to "*Strongly Agree*" (e.g., "The drone provides good advice" and "The drone cares about my wellbeing"). The scale was adapted from questionnaires measuring user trust in robots (Charalambous et al., 2016) and automated systems (Chien et al., 2014; Jian et al., 2000; Körber, 2019) and demonstrated good reliability (study 1, $\alpha = 0.74$; study 2, $\alpha = .94$). This scale has been specifically developed to suit the online setting of the experiment and enables fast repeated trust assessments.

In the military study, three additional personality questionnaires were administered. First, a short version of the IPIP Big-Five personality scale was administered with subscales measuring Extraversion ($\alpha = 0.72$), Agreeableness ($\alpha = 0.72$), Conscientiousness ($\alpha = 0.59$), Openness ($\alpha = 0.60$) and Neuroticism ($\alpha = 0.68$). The IPIP was selected as it proved

valid for usage in a Web-based format (Buchanan et al., 2005). Participants were instructed to answer each item in relation to ‘‘whether the statement describes what you are like’’ on a 5-point Likert scale ranging from “*Very much unlike me*” to “*Very much like me*”.

Second, we measured the Propensity to Trust Automation (Jessup, 2018), adapted from the Propensity to Trust in Technology scale (Schneider et al., 2017). This scale consisted of five items (e.g., “I think it’s a good idea to rely on automated agents for help.”) (α = 0.81). Participants were instructed to answer each item on a 5-point Likert scale ranging from “*Strongly Disagree*” to “*Strongly Agree*”.

Lastly, two subscales of the Need for Closure scale were administered in the military study: Need for Predictability (three items, e.g., “I don’t like to go into a situation without knowing what I can expect from it.”) with α = 0.35 and Need for Decisiveness (three items, e.g., When I have made a decision, I feel relieved), with α = 0.08. Participants were instructed to answer each item in relation to ‘‘whether the statement describes what you are like’’ on a 5-point Likert scale ranging from “*Very much unlike me*” to “*Very much like me*”. Since both Cronbach’s alpha values are lower than 0.40, both constructs were eliminated from the analysis.

Results

General plots

For both studies we performed a repeated-measures ANOVA with the between-subject factor Apology (present or absent) and the within-subject factors Uncertainty communication (present or absent) and Time (prior to violation [T1] versus after violation [T2] versus after repair [T3]) (Figure 8).

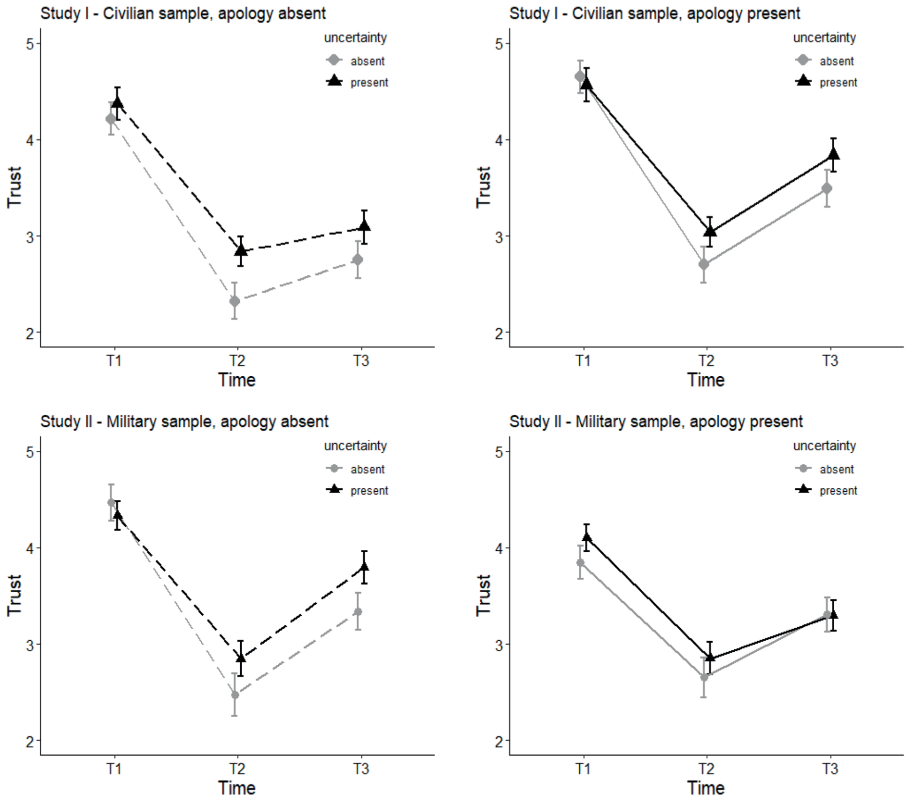


Figure 8 An overview of the results of both studies; the upper half represents Study I (civilian sample), the lower half represents Study II (military sample). Graphs show the development of trust (y-axis) over time (x-axis) with the estimated marginal means on trust for the uncertainty and apology conditions over time. The error bars represent standard errors. Separate graphs (left and right panels) represent the apology conditions (left shows apology strategy absent, right shows apology present). Separate lines represent the uncertainty conditions. The grey lines with the circle-shaped datapoints represent the condition in which the agent did not communicate uncertainty in its advice, the black lines with triangle-shaped datapoints represents the condition in which uncertainty communication was present.

Results: Study I [civilian sample]

Main effects

A significant main effect for Time [T1-T3] was obtained (F (2, 124) = 112.06, p < .001, η^2 = .644). Means were 4.45 at T1, 2.73 at T2 and 3.29 at T3. Post-hoc (LSD) pairwise comparison shows a significant decline in trust from T1 to T2 (ΔM = -1.725, p < .001), which reflects the effect of the trust violation and a significant rise in trust between T2 and T3 (ΔM = .568, p < .001), which reflects a general recovery of trust in the trust repair

phase. This means that the incorrect advice by the drone led, as intended, to a breach in trust and that after the violation trust re-developed.

A significant main effect for Uncertainty was obtained with $F(1, 62) = 7.84$, $p = .007$, $\eta^2 = .112$). Generally, across time and apology conditions, the agent that provided uncertainty communication ($M = 3.62$, $SE = 0.09$) was trusted significantly more than the agent that did not communicate uncertainty ($M = 3.36$, $SE = 0.10$).

A significant main effect for Apology was obtained with $F(1, 62) = 8.37$, $p = .005$, $\eta^2 = .119$). Generally, across time and uncertainty conditions, the agent that offered an apology after the trust violation occurred ($M = 3.71$, $SE = 0.11$) was trusted significantly more than the agent that did not offer an apology ($M = 3.26$, $SE = 0.11$).

Two-way interactions

A significant interaction effect between Time [T1-T3] and Uncertainty on trust was found ($F(2, 124) = 3.31$, $p = .040$, $\eta^2 = .051$) (Figure 9). Post-hoc (LSD) pairwise comparison shows no significant difference in trust between uncertainty communication conditions at T1 ($\Delta M = 0.04$, $SE = 0.13$ $p = .777$), but does show a significant difference at T2 ($\Delta M = 0.43$, $SE = 0.12$, $p = .001$) and T3 ($\Delta M = 0.35$, $SE = 0.15$, $p = .024$), where the agent that provided a measure of uncertainty was trusted significantly more than the agent that did not communicate uncertainty. The decline in trust in response to the trust violation (from T1 to T2) is significantly smaller when the agents' advice included a measure of uncertainty.

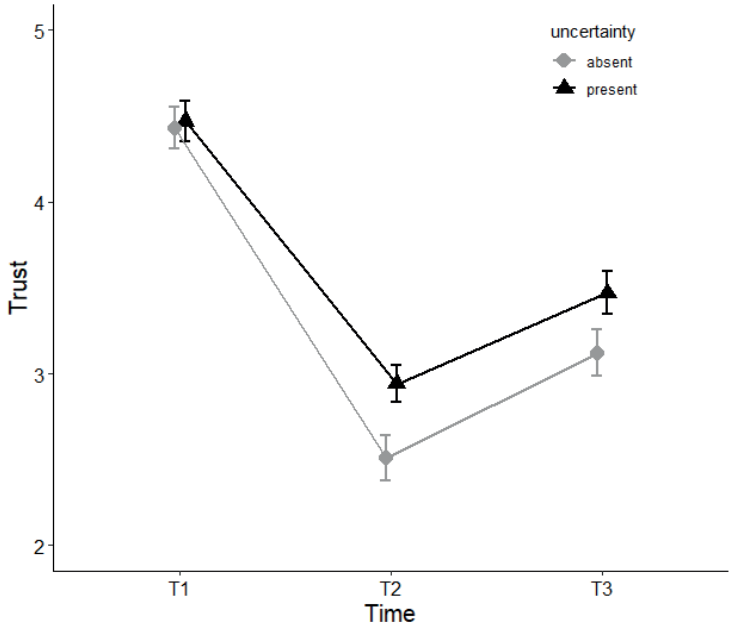


Figure 9 A comparison of trust levels (y-axis) between uncertainty communication conditions (separate lines) over time (x-axis). The grey line with the circle-shaped datapoints represents the condition in which the agent did not communicate uncertainty in its advice, the black line with triangle-shaped datapoints represents the condition in which uncertainty communication was present. Error bars represent standard error.

To measure the effect of the apology, we compared trust scores T2 (after the violation) and T3 (after the manipulation) for each experimental condition.

A significant interaction effect between Time [T2-T3] and Apology on trust was found ($F(1, 62) = 5.16$, $p = .027$, $\eta^2 = .077$). Post-hoc (LSD) pairwise comparison per apology condition shows a significant rise in trust from T2 to T3 when apology is present ($\Delta M = 0.80$, $SE = 0.14$, $p < .001$), but also when apology is absent ($\Delta M = 0.34$, $SE = 0.14$, $p = .018$). As shown in Figure 10, trust recovers more when the agent offered an apology. Post-hoc (LSD) pairwise comparison per timepoint shows that a non-significant differences in trust between apology conditions at T2 ($\Delta M = 0.29$, $SE = 0.21$, $p = .164$), but the difference at T3 is significant ($\Delta M = 0.74$, $SE = 0.21$, $p = .001$). Thus although trust recovers significantly in both conditions, trust is significantly higher in the final stage of trust after an apology was provided.

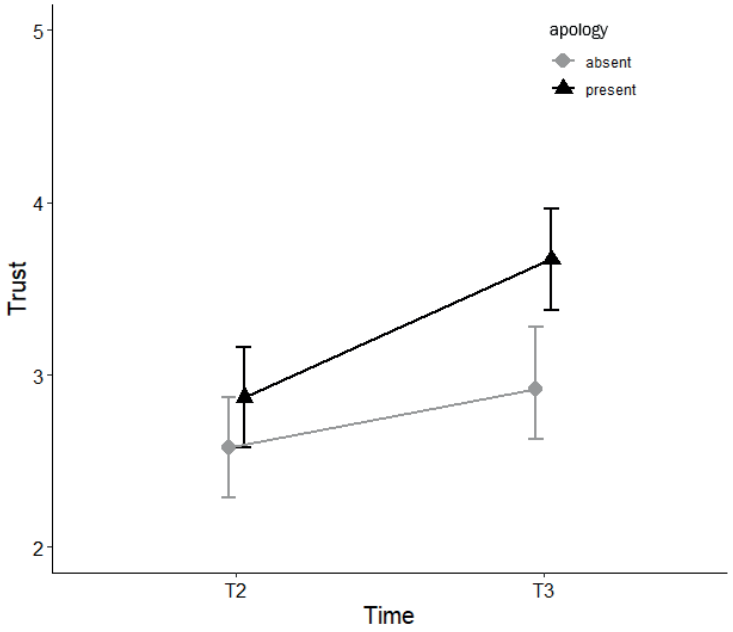


Figure 10 A comparison of trust levels (y-axis) between apology conditions (separate lines) over time (x-axis). The grey line with the circle-shaped datapoints represents the condition in which the agent did not offer a trust repair strategy, the black line with triangle-shaped datapoints represents the condition in which the trust repair strategy was provided. Error bars represent 95% confidence interval.

A non-significant interaction effect between Uncertainty and Apology on trust was observed with F (1, 62) = 0.512, p = .477.

Three-way interaction

The interaction between Time [T2-T3], Uncertainty and Apology was found to be non-significant with $F(1,62) = 0.429$, p = .515. This means that uncertainty communication did not significantly enhance the effect of the apology.

Results: Study II [military sample]

Main effects

Similar to the civilian sample, a significant main effect for Time [T1-T3] was obtained (F (2, 116) = 76.562, p < .001, η2 = .569). Means were 4.19 at T1, 2.71 t T2 and 3.43 at T3. Post-hoc (LSD) pairwise comparison shows a significant decline in trust from T1 to T2 (ΔM = -1.481, p < .001), which reflects the effect of the trust violation and a significant rise in trust between T2 and T3 (ΔM = 0.728, p < .001), which reflects a general recovery of trust in the trust repair phase.

A significant main effect for Uncertainty was also obtained with $F(1, 58) = 5.657$, $p = .021$, $\eta^2 = .089$. Generally, across time and apology conditions, the agent that provided uncertainty communication ($M = 3.54$, $SE = 0.08$) was trusted significantly more than the agent that did not provide uncertainty communication ($M = 3.35$, $SE = 0.10$).

No significant main effect for Apology was found with $F(1, 58) = 1.484$, $p = .228$.

Two-way interactions

The interaction effect between Time [T1-T3] and Uncertainty on trust was found to be non-significant ($F(2, 116) = 2.441$, $p = .092$). There was no dampening effect of uncertainty communication in the military study.

To measure the effect of the apology, we compare trust scores before (T2; after the violation) and after the manipulation (T3) for each experimental condition. Again, T1 is left out of this analysis as it focusses on the effects of the apology manipulation that occurs between the trust measures on T2 and T3.

The interaction effect between Time [T2-T3] and Apology on trust was found to be non-significant ($F(1, 59) = 1.897$, $p = .174$).

A non-significant interaction effect between Uncertainty and Apology on trust was observed ($F(1, 59) = 2.314$, $p = .134$).

Three-way interaction

The interaction between Time [T2-T3], Uncertainty and Apology was found to be non-significant ($F(1,59) = 0.710$, $p = .403$). Uncertainty communication did not significantly enhance the effect of the apology.

Correlations

For the correlations, initial trust (T1) is used as this is considered the purest trust measure with the least interference of occurrences during the experiment. A significant positive correlation was found between the personality trait Propensity to Trust Automation and initial trust in both uncertainty conditions: present ($r(63) = .40$, $p < .00$) and absent ($r(62) = .28$, $p = .02$). The Big Five personality trait Extraversion correlates with the initial trust measure of the run without uncertainty communication ($r(63) = .28$, $p = .03$). These correlations imply that participants that scored higher on these traits, trusted the agent more than participants that scored lower on these traits.

Discussion

The results of this chapter show a robust effect of uncertainty communication on the development of trust during human-agent interaction. In both studies it was found that uncertainty communication in the advice of the agent generally resulted in higher levels of trust. The communication of uncertainty did not enhance the effect of the apology. The

positive effect of uncertainty communication on trust is in line with prior research (Kraus et al., 2020; Kunze et al., 2019; Schaekermann et al., 2020).

In the civilian study, uncertainty communication also dampened the decline in trust following the agent's error, meaning that advice which included an uncertainty measure led to a less severe depletion in trust following a trust violation compared to an advice that did not include a notion of uncertainty. The dampening effect of uncertainty communication on trust decline in response to a trust violation is in line with the study of Kraus et al. (Kraus et al., 2020) that showed how a temporary decrease in trust due to a malfunctioning of an autonomous car was prevented by providing transparency information prior to the interaction. When participants were reminded of the imperfect reliability of the system their trust was less affected by the subsequent error. Further, the civilian participants generally regained their trust in the agent after the trust violation occurred. Strikingly, this occurred both when the agent did offer an apology and when it did not. Even though trust levels increased considerably *more* when the agent offered an apology compared to when no apology was offered, it is still remarkable that trust seemed to recover naturally in the absence of a recovery strategy. A possible explanation for this is that the participants' trust gradually recovered after the trust violation, just by the absence of any new hazardous encounters. We did not monitor trust continuously, but participants can perceive each second of error-free interaction as positive feedback, which might be what reassured them in the period between the violated trust measure and the final trust measure. Although trust recovered passively, it still proved to be more effective to actively interfere in the repair process by providing an apology. Although trust did not recover to its original level (initial trust) in either of the conditions, the agent in the apology condition came considerably closer.

The repairing effect of an apology after a trust violation is compatible with prior human-agent research (Fratczak et al., 2021; Kox et al., 2021). This effect is promising, as it suggests that (relatively minor) trust violations within human-agent teams can be solved on a relational level during ongoing interaction, without ceasing the collaboration (Fratczak et al., 2021). It also indicates that mimicking human-like characteristics (i.e. provision of an apology following a mistake) can bring about certain effects typically observed in interpersonal relations, including a greater willingness to forgive mistakes (de Visser et al., 2016; de Visser, Monfort, et al., 2017; Madhavan et al., 2006; Madhavan & Wiegmann, 2007). Although such anthropomorphic cues can be beneficial to human-agent trust, it should be kept in mind that people can develop trust on the basis of characteristics they attribute to the agent, rather than on actual experiences with the agent itself (Bartneck et al., 2009; Culley & Madhavan, 2013; Feine et al., 2019; Fink, 2012). If so, trust may turn out to be misplaced. This can lead to inappropriate reliance on the agent, potentially compromising safety and profitability.

However, these findings do not apply to the military participants. The dampening effect of uncertainty communication and the repairing effect of the apology are not manifest in the results of the military study. A possible explanation for the latter finding comes from a recent study that found that feedback messages (i.e., an apology) affected trust negatively

rather than positively, possibly because it explicitly focused the attention on the error (Fahim, Khan, Jensen, Albayram, Coman, et al., 2021). Another plausible explanation is that expressing regret is not a common practice in the military context. This became clear in a debriefing session with a few of the military participants. They mentioned that it is not unusual to acknowledge responsibility by saying “I was wrong” or “I misjudged the situation”, but using the words “I am sorry” is uncommon. Adopting the norms and rules that govern a group’s behavior is an important aspect in being accepted as a member of that group. It makes sense that the psychosocial requirements for a system designed to be a true team member should be compatible with the manners associated with the culture of the organization or team where the agent will be implemented. As addressed by Matthews et al. (Matthews, Hancock, et al., 2021), an agents’ communication style should match one’s cultural background and its related language and behavioral expectations. In line with the expectancy violation theory, which describes how actions contrary to your expectations and social norms in a social context require more cognitive processing effort than expected information and that this type of inconsistencies can elicit a more negative affect (J.-E. R. Lee & Nass, 2010; Lozano & Laurent, 2019). The differences in findings between the military and civil samples emphasize the importance of considering the social customs of the target population in the design process. In a broader perspective, it serves as a reminder that generalizability is limited by the characteristics of the participants in the study and that results do not automatically apply to other populations.

It should be noted that it is not a goal in itself to maximize trust or to prevent trust decline at any cost, as we want humans to be able to continuously assess whether trusting the agent is appropriate given the task and available information at certain instances. Multiple studies have shown that people do not judge humans and machines equally, particularly when confronted with errors (de Visser et al., 2016; Hidalgo et al., 2021; Madhavan et al., 2006). Often, people consider a machine as nearly infallible (i.e., automation bias), thereby placing too much trust in their outputs. These high expectations lead to a steeper decline in trust when confronted with system failure as compared to a confrontation with a human error (de Visser et al., 2016; Dzindolet et al., 2001; Madhavan et al., 2006; Madhavan & Wiegmann, 2007). Following this, in the case of ‘undertrust’, it would be valuable for the process of trust calibration if AI agents were equipped with expectancy-setting strategies like the communication of uncertainty and trust repair strategies like offering an apology.

Other interesting findings in the military study include the positive correlations between the initial levels of trust and the personality traits Extraversion and Propensity to Trust Automation. The current chapter demonstrates that communication tactics do not have a uniform effect on the development of trust in different types of people, which emphasizes the importance of personalization. Not only cultural differences between groups (i.e. military vs. civilian) but also personal differences within each group can be found. Individual differences such as personality traits can account for the variance in how trust in an agent develops among individuals and how people prefer to be approached while

interacting. The observed relation between the Big Five personality trait Extraversion and an initial trust measure is in line with studies that showed that Extraversion plays a significant role in how people perceive robots (Haring et al., 2013; Syrdal et al., 2007; Walters et al., 2009). Current agent communication styles are often of a one-size-fits-all style. Personalized communication could overcome the effects of pre-existing attitudes towards automation and influence the willingness to reconcile after a trust violation (de Visser et al., 2019; Schaefer et al., 2016). Today's machine-learning methods enable agents to leverage real-time user inputs and to personalize interactions. Recent work has shown that agents can directly estimate a human's ability to achieve a certain goal based on their efforts and respond with the proper level of assistance for the task, resulting in higher levels of trust in the agent's advice (Clabaugh & Mataric, 2016). Given the many dimensions on which people vary, a lot could be gained by enabling the agent to tailor its communication to the person they are interacting with. Follow-up research should explore how personalizing the level of transparency (e.g. communicating uncertainty measures and offering apologies that include an explanation) and the level of affection (e.g. offering apologies that include an affective component such as an expression of regret) of an agent's communication style can optimize trust calibration.

Several questions still remain to be answered. The apology used in the current study consisted of two apology components: an expression of regret and an explanation. An interesting question for follow-up research would be *what* apology component caused the (difference in) effects between the two target groups. On the one hand, our previous findings from a civilian sample suggested that expressing regret made a positive difference in trust recovery (Kox et al., 2021). On the other hand, conversations with our military participants in the current study suggested that since saying "sorry" is uncommon among military personnel and that this inconsistency might have caused the lack of trust recovery in the military study. It raises the questions whether regret was the component that caused the differences and what trust repair strategy *would* be effective among military personnel. Another follow-up question can be posed for the uncertainty variable. As discussed in the introduction, uncertainty can be introduced by random noise from the outside world (external sources) or by the limited abilities of the drone (internal sources). Whereas the former type of uncertainty is a given that we all have to accept, the latter type of uncertainty could be perceived as the limited ability of the agent's prediction algorithms and might therefore be less acceptable. It seems beneficial that agents, regardless of the type of uncertainty, are able to communicate the level of certainty to allow humans to make better estimations on whether or not to rely on their advice. Still, it could be interesting to explore whether knowing the source of the uncertainty shifts the human's interpretation and leads to alternative effects on trust.

Limitations

This study was initially designed to be conducted in a lab setting, where participants walk through the virtual houses whilst wearing a VR headset and using a controller. The Dutch COVID-19 regulations required the design of this study to be altered into an online experiment. Although this enabled a faster and more scalable experiment as compared to the VR design and a higher degree of control over the manipulations as compared to a field lab setting, results may not generalize to human-agent interactions in real-world settings (Hidalgo et al., 2021). However, interactive online experiments are a good alternative to VR; the data quality is described as “adequate and reliable” (Arechar et al., 2018). A study which compared data that was gathered online with lab research data found no significant differences over multiple performance measures (Gould et al., 2015). However, the VR design would have offered higher ecological validity, experimental control, reproducibility (Pan & Hamilton, 2018), and emotional engagement of participants (Parsons, 2015). Immersive VR has the ability to create a strong sense of presence and to increase sympathetic activation significantly more than 2D screen videos (Chirico et al., 2017). Thus, it is suspected that a VR setting would have intensified feelings of trust and betrayal after a trust violation. These intensified feelings could be more representative of non-simulated human-AI interactions. Although the two studies are based on relatively small samples of participants, an important contribution is made by evaluating subgroups in their way of interacting with autonomous systems. In spite of its limitations, the study adds to our understanding of how trust develops in case of agent failure within civilian and military human-agent teams.

Conclusion

Amidst the expanding adoption of autonomous agents in human teams, this study contributes to the rapidly expanding field of trust within HATs by informing the design of AI components and their interactions with human teammates. Given the uncertainty and complexity that agents in HATs will encounter, these insights will be critical to developing specifications for agent communication as this allows HATs to recover as a team from errors induced by AI agents. The findings presented in this chapter indicate that communication can be used as a tool to guide the development of human trust in AI agents. The findings reported here shed new light on how the effects of social-cognitive trust repair strategies on trust differ amongst civilian and military user groups. A lot of research on this subject is done for military purposes (Barnes et al., 2014; E. K. Phillips et al., 2011; Roff & Danks, 2018; van den Bosch & Bronkhorst, 2018). Yet, it is not always possible to involve actual military personnel as participants in experimental studies. The differences in findings between the military and civil cohort emphasize the importance of considering the social customs of the target population in the design process. The

psychosocial requirements for the formation and maintenance of trust in HATs differ amongst individuals and user groups.

Chapter 4

This chapter is based on:
Kox, E. S., Boogaard, J. van den, Turjaka, V., & Kerstholt, J. H. (2024). The Journey or the Destination: The Impact of Transparency and Goal Attainment on Trust in Human-Robot Teams. *Transactions on Human-Robot Interaction*. https://doi.org/https://doi.org/10.1145/3702245

Abstract

As robots gain autonomy, human-robot task delegation can become more goal-oriented; specifying what to do rather than how. This can lead to unexpected robot behaviour. We investigated the effect of transparency and outcome on the perceived trustworthiness of a robot that deviates from the expected manner to reach a delegated goal. Participants (N = 82) engaged in a virtual military mission as a Human-Robot Team using a 2x2 between-subjects design (low vs. high transparency, positive vs. negative outcome). Participants received training on the expected manner to reach the mission's goal. In the actual mission, the robot deviated from the planned path. We manipulated whether the robot explained its deviation and whether the outcome was better or worse than the original plan. Results showed that transparency contributed to higher and more stable levels of trust, without increasing subjective workload. While the robot's deviation led to a violation of trust in the low transparency condition, trust remained stable in the high transparency condition, indicating a buffering effect of transparency on trust in case of unexpected behaviour. The impact of outcome on trust was consistent across transparency conditions. Our findings underscore the role of transparency as a tool for fostering human-robot trust.

Introduction

Due to recent technological developments in artificial intelligence and robotics, more and more people are increasingly interacting with artificial agents in a variety of domains, among which the military (Matthews, Panganiban, et al., 2021; Wynne & Lyons, 2018). As robots become more intelligent, they are increasingly self-governing, gain decision authority within their functioning (Bobko et al., 2022; Hancock, Billings, Schaefer, et al., 2011a; Hou et al., 2021; O'Neill et al., 2022; Sheridan, 2019), and require less human involvement and control (Lyons et al., 2023; C. A. Miller, 2014). In other words, they become increasingly autonomous; able to achieve a given set of tasks during an extended period of time without human control or intervention (Soltanzadeh, 2022). As such, future robots are expected to work interdependently in HRTs with human team members towards a shared objective (O'Neill et al., 2022). Robots can take over tasks that were previously conducted by humans, whereas other tasks still need to be executed by human counterparts (Parker & Grote, 2022). As a result, the rise of HRTs poses interesting challenges related to teamwork, task delegation and trust.

Delegation

Teamwork typically involves dividing and assigning tasks or responsibilities to different team members. When delegating authority, an actor (i.e., in our HRT case, the human) hands over a specific (set of) task(s) to another actor (i.e., the robot) who is expected to take responsibility for planning and execution of the assignment in a timely and effective manner to reach commonly understood goals (Ho et al., 2017; C. A. Miller, 2014; C. A. Miller & Parasuraman, 2007). Since reaching a goal consists of completing a set of tasks, delegation is inherently hierarchical (C. A. Miller, 2014). As a result, delegation can be adapted to different levels of abstraction, such as (1) skill-based delegation, which proceeds by delegating single elementary tasks or actions (e.g. go-right, go-left), (2) rule-based delegation, which proceeds by delegating in terms of pre-defined templates of taskwork and teamwork (e.g. perform-blanket-search procedure) and ultimately, (3) goal-oriented delegation, which proceeds by delegating in terms of goals (Jessie Y.C. Chen & Barnes, 2014; Metcalfe & van Diggelen, 2021; C. A. Miller & Parasuraman, 2007). Which type of delegation is appropriate will depend on a robot's level of autonomy (LOA), which can range from no autonomy (i.e. manual human control), to semi-autonomy (i.e., human can veto) to full autonomy (i.e., human is at most informed) (Ellwart & Schauffel, 2023; Parasuraman et al., 2000).

The more autonomous a robot gets, the more abstract and goal-oriented a delegated assignment can be, the more degrees of freedom the robot has in terms of execution and the more trust in the robot is required. Goal-oriented task delegation implies that the delegator does not have to outline the specific rules and skills that should be used in the process of reaching the desired end-state. In short: it means telling the robot

what to do instead of how to do it. This leaves considerable room for the robot to fill in the remaining details on the execution of desired actions, which allows it to adapt to changing environments and operational demands (Metcalfe & van Diggelen, 2021). As a situation evolves, the possible paths to achieve a certain goal can change (Ho et al., 2017). As a result, an (semi-)autonomous robot might exhibit unexpected behaviour (from the perspective of a human operator) in its pursuit to reach a certain goal. A possible risk is that a human's lack of understanding of the robot's actions can cause people to lose trust and want to take over manual control, negating the advantages of task delegation. Regardless of the LOA of a robot, communication and human participation in certain decision-making loops will always remain crucial for effective and safe operations (Abbass, 2019).

To keep the human involved, robots will need to be able to explain their behavioural choices, especially when they deviate from the expected manner to reach a goal. Higher decision authority assigned to robots typically increases the human desire to know what the robot will be doing (Jessie Y.C. Chen et al., 2020). When the human operator cannot understand the basis of the robot's assessments and actions, trust may be eroded, especially when the robot's actions do not align with the human's expectations (Luebbers et al., 2023; A R Panganiban et al., 2020). In the current study, we are interested in the implications of a robot that has been delegated the authority to select the best course of action given the local situation, which could contradict a human's expectation and result in a suboptimal outcome (i.e., not attaining the goal). In the context of goal-oriented delegation, does understanding the robot's actions towards a goal drive trust or is ultimately attaining the goal the primary factor?

Trust

Teamwork requires task delegation and task delegation requires trust. More specifically, calibrated trust is crucial to minimize the risks and to maximize the benefits in the highly interdependent and dynamic nature of teamwork (Bobko et al., 2022; J. D. Lee & See, 2004; M. K. Lee et al., 2010b). In general, perceiving good robot functioning will likely increase perceived trustworthiness, whereas perceiving maladaptive (i.e., errors or mistakes) or ambiguous (i.e., unexpected or unpredictable) robot functioning often results in decreases in perceived trustworthiness– so called trust violations (Esterwood & Robert, 2023a, 2023b; Kox et al., 2021; Yang et al., 2021). As we strive for calibrated trust rather than maximum trust, decreases in perceived trustworthiness are a logical and functional adaptive response to perceiving errors, technical failures or other forms of reduced reliability and performance.

However, with the anticipated advancements in the ability of robots to self-select courses of action, the range of possible causes of human-robot trust violations expands. That is, human-robot trust is not solely based on a robot's perceived abilities and performance (i.e., what it does and can do), but also on its perceived purpose and

alignment with a trustor's values (i.e., why it was developed and operates in a certain way), as well as the understandability or interpretability of the robot and its ability to explain its actions (i.e., how it operates) (J. D. Lee & See, 2004; Lubars & Tan, 2019). This operationalization of trust corresponds to the Ability (what), Benevolence (why) and Integrity (how) (ABI) model from Mayer et al. (Mayer et al., 1995) and reflects how a trustee's trustworthiness is based on more than reliability and performance. As a consequence, trust violations are not solely caused by reduced performance.

As task delegation becomes more goal-oriented, providing the robot more with greater degrees of freedom in terms of execution, trust violations might be increasingly caused by a human operator's lack of understanding of the robot's assessments and actions, rather than poor robot performance. When a robot does something unexpectedly (according to the human), its efficacy and accuracy could be questioned and the action can lead to a decrease of human-robot trust, regardless of whether the robot is actually maladapted (Rebensky et al., 2021; Schaefer et al., 2018). For example, a drone might rightfully adapt its course of action to changes in the operational environment to reach a certain goal, such as avoiding a collision, without informing the human. If the drone's deviation significantly conflicts with the human's expectations and the robot lacks the ability to explain itself, the human operator might take over manual control because they do not understand the drone's actions and perceive them as inappropriate and untrustworthy (Hou et al., 2021; Lyons et al., 2023; Rebensky et al., 2021). As such, a lack of understanding causes a trust violation and leads to a situation of undertrust. Since the success of human-robot interactions greatly depends on people's ability to trust them, trust violations that lead to undertrust would make it necessary for a robot to engage in trust repair strategies (Baker et al., 2018).

Given (1) the inevitability of unexpected robot behaviour in Human-Robot Interaction (HRI), (2) the possibility that unexpected behaviour results in trust violations and poor trust calibration, and (3) the disadvantageous consequences of poor trust calibrations, it is important to evaluate methods to prevent or buffer (unnecessary) trust violations as a consequence of unexpected behaviour. Most current HRI trust repair literature focuses on the role of trust repair strategies after an apparent error (Cameron et al., 2021; de Visser et al., 2016; Esterwood & Robert, 2023b; Fratczak et al., 2021; Hald et al., 2021; Taenyun Kim & Song, 2021; M. K. Lee et al., 2010b; Mirnig et al., 2017; Robinette et al., 2017b; Salem et al., 2015; Wang et al., 2018). However, more recently researchers have started to evaluate trust violations as a result of unexpected behaviour rather than failure (Lyons et al., 2023; Perkins et al., 2022; Sebo et al., 2019). In essence, to prevent that trust will unjustly erode due to a misunderstanding of the basis of a robot's assessments and actions, robots will need to be able to explain the rationale behind their behavioural choices. Increasing transparency and interpretability through explanations can enhance trust calibration by lowering unrealistic expectations on the one hand (i.e., preventing overtrust) and by clarifying unexpected behaviour on the other (i.e., preventing undertrust) (Jessie Y.C. Chen et al., 2014; J. D. Lee & See, 2004; Mercado et al., 2016).

Transparency

Transparency can be defined as “the ability for the automation to be inspectable or viewable so that its mechanisms and rationale can be readily known” (C. A. Miller, 2020) (p. 235). Transparency is an important part of the design of robots, because without a clear understanding of a robot’s decision-making mechanism, humans might find it difficult to trust or adhere to a robot’s decisions, especially when those actions or decisions contradict the human’s expectations (Luebbers et al., 2023). At the same time, full “transparency” – implying that the machine is “see through” in the sense that all its inner workings are observable (Jessie Y.C. Chen et al., 2020; C. A. Miller, 2020) is not desirable either (C. A. Miller, 2014). When HRI is successful, it can save time and reduce cognitive effort. However, if a human would have to maintain awareness of everything the robot does, then no time or cognitive effort would be saved (C. A. Miller, 2014). Ideally, transparency allows the human teammate to develop and/or maintain realistic expectations regarding the robot and its behaviour (Hou et al., 2021; C. A. Miller, 2014) and thereby contributes to effective trust calibration (Bobko et al., 2022; Helldin et al., 2013; Ribeiro et al., 2016). However, to ensure effective collaboration, it is crucial to find a balance between keeping the human sufficiently informed while preventing cognitive overload.

To find that balance, literature suggests that robots should primarily communicate the rationale and intentions of their actions (Chiou et al., 2022; Lyons, 2013; Lyons et al., 2023; Ososky et al., 2014; Schaefer et al., 2017). A recent study evaluating human-robot trust in case of unexpected robot behaviour compared different explanation types and found that explanation strategies that indicated why the event occurred were most effective at buffering the decline in perceived trustworthiness (Lyons et al., 2023). Explanations are verbal statements that aim to clarify the reasons for an occurrence. They are deployed in HRI, prior to or after certain actions, to enable the human to comprehend the inner workings or logic of the robot’s actions or decisions (Esterwood & Robert, 2022; Lyons et al., 2023). Explanations are generally invoked when the mental models of those who must work together mismatch. The explanation is then meant to synchronize the mental models so that the differences are understood and repaired (C. A. Miller, 2020). As such, explanations can have a positive effect on trust in case of trust violations.

For instance, increased transparency and feedback can effectively mitigate a human’s dissatisfaction in the event of an unforeseen occurrence caused by a robot (Hamacher et al., 2016). Feedback enhances a human’s willingness to trust automation and can delay or avoid unnecessary manual intervention (Hock et al., 2016). Results of an automated driving study show that explanations provided before rather than after a certain event strengthened trust (Du et al., 2019). In other words, increased transparency through explanations can strengthen trust.

While transparency can benefit trust, it also a poses a challenge to the human operator. In most cases, humans that perform a task together with a robot do not have the time, skills, or attention to accurately interpret transparency information during an operational

situation or the adequate precision to take over the robot's task if necessary (C. A. Miller et al., 2023). There is a possibility that increased transparency could come at the expense of cognitive workload since it requires additional processing and interpretation of information (i.e., additional cognitive effort) (Guznov et al., 2020; Lyu et al., 2017; Westerbeek & Maes, 2013). Cognitive workload generally refers to the amount of cognitive resources and effort required for task performance relative to the available resources (Parasuraman et al., 2008). An increase in cognitive workload arises when multiple tasks compete for the same resources, and task requirements exceed the mental capacity. High levels of cognitive workload can result in fatigue, and hence reduce human performance. On the contrary, appropriate implementation of transparency in HRT could also result in reduced cognitive workload of the human-teammate, as it helps to understand the robot's behaviour and reasoning (Bobko et al., 2022; O'Neill et al., 2022; van de Merwe et al., 2022). At the same time there are also studies that find no effect of transparency on workload (Jessie Y.C. Chen et al., 2017; Selkowitz et al., 2016, 2017). In other words, the results are inconclusive and further research is needed to determine whether transparency affects workload advantageously or disadvantageously.

Outcome

While transparency can enhance a human's understanding of a robot's reasoning process and thereby help to create realistic expectations regarding the robot's capabilities, it is conceivable that a negative outcome will still be disappointing and detrimental to trust. At the same time, since unexpected robotic behaviour might arise from the fact that increasingly intelligent agents may devise alternative plans that are better and more efficient than those humans would come up with, we are also interested in the effect of positive outcomes. Whether the robot's execution is logical or understandable for the human and whether the robot eventually reaches its goal are both likely to affect trust. As such, we seek to explore how and to what extent transparency and outcome influence the development of trust.

Generally, the performance of a robot is seen as the most important predictor of human-robot trust (Hancock, Billings, Schaefer, et al., 2011a; Hoff & Bashir, 2015). Unsurprisingly, research suggests that robot successes increase trust (Yang et al., 2021), while robot failures decrease trust (Jorge et al., 2023; Kox et al., 2021; Kox, Siegling, et al., 2022; Yang et al., 2021). Furthermore, the magnitude of trust decrements due to robot failures is found to be bigger than that of trust increments due to robot successes (Yang et al., 2021). This is in line the concept of loss aversion within prospect theory from classic decision-making literature, which posits that people tend to value gains and losses differently, placing more weight on perceived losses versus perceived gains (Tversky & Kahneman, 1992). That is, the pain of losing is psychologically more impactful than the pleasure of gaining (Tversky & Kahneman, 1992). However, research also suggests that

the effect of robot performance on trust might depend on an individual's perception of the interaction and vice versa.

One the one hand, there is research that suggests that the quality of the interaction might influence how people respond to a robot's performance. For example, there are findings that suggest that people place less value on task performance and more on transparency, control and feedback (Hamacher et al., 2016). This study shows that participants preferred an expressive and error-prone robot over a more efficient one. This suggests that an erroneous robot can be forgiven as long as it communicates, while an inexpressive robot with high task performance could still be trusted less (Hamacher et al., 2016).

On the other hand, there is research that suggests that outcome can change how people perceive the preceding interaction, a phenomenon referred to as the outcome bias. An outcome bias is where the quality of a decision made by others under conditions of uncertainty is evaluated differently in hindsight, based on the outcome (Baron & Hershey, 1988). Research suggests that people evaluate the thinking behind a decision as better when the outcome is favourable compared to when the outcome is unfavourable (Baron & Hershey, 1988). Earlier HRI research has found evidence for the outcome bias, finding a reinforcing effect where initial automation failure led to a larger trust decrement if the final outcome was undesirable (Yang et al., 2021). In other words, there are reasons to believe that the effects of transparency and outcome on perceived trustworthiness might be interdependent.

Current study

Goal-oriented delegation in complex environments with limited resources and changing circumstances poses challenges. Plans can be made in advance, but in case of unforeseen circumstances, the robot will need to adapt its plan and "function beyond choreography" to still reach the end-goal (Chiou et al., 2022) (p. 119). That is, beyond a fixed, scripted series of actions that do not account for variability or unexpected changes in the environment. At times, these adaptations will be advantageous, while in other cases, they may be suboptimal or disadvantageous. The current study investigates how transparency and outcome affect the perceived trustworthiness of a robotic partner in case of an unexpected deviation from the expected manner to reach a delegated goal.

In the current study, transparency entails that the robot gives clarifying information in the form of regular status updates including an explanation (i.e., the what and why) of its actions as it deviates from the expected manner to reach the goal (Chiou et al., 2022; Taemie Kim & Hinds, 2006). We expect that when the robot explains its reasoning and actions, a stable level of perceived trustworthiness can be maintained in the event of deviant behaviour. Specifically, we expect that transparency will prevent a trust violation in response to the robot's unexpected behaviour (Lyons et al., 2023) and will generally lead to higher perceived trustworthiness. Conversely, we expect that a sudden and

silent deviation from the plan (i.e., low transparency) will lead to a violation of trust. We further expect an interaction effect between transparency and outcome. Specifically, we hypothesize that the expected violation of trust in response to the unexpected behaviour in the low transparency condition will amplify the effect of a subsequent negative outcome (Yang et al., 2021). In the high transparency condition, we expect higher and more stable levels of perceived trustworthiness (Lyons et al., 2023) and a smaller effect of negative outcome compared to the low transparency condition.

Method

Participants

In total, eighty-seven participants participated in the study. Five participants were excluded from the dataset because of invalid data due to technical issues during the task. Participants were recruited through convenience sampling (e.g., by handing out flyers, asking people in person, and making requests in WhatsApp groups). The final dataset included eighty-two participants (43 W, 39 M, M_{age} = 23.6, *SD* = 3.2, range = 19 – 41 y), of which the majority was Dutch (65,9%) and the remainder from elsewhere in Europe (17,1%), Asia (9,8%), or North or South America (both 3,7%).

Design

A 2 (transparency: low vs. high) by 2 (outcome: negative vs. positive) between-subjects design was used, with Perceived Trustworthiness (measured across the subscales Ability, Benevolence and Integrity) as the main dependent variable. Participants were randomly distributed across the four conditions (low & neg.: *n* = 21, low & pos.: *n* = 20, high & neg.: *n* = 20, high & pos.: *n* = 21).

Perceived trustworthiness was repeatedly measured, so 'Time' was included as a within-participants variable in the analysis. Each participant performed two missions; a training mission with four trust measures (T1, T2, T3, T4) and the experimental mission with four trust measures (T5, T6, T7, T8). Cognitive workload was also administered.

Task and procedure

Upon arrival at the laboratory, participants were greeted by the researcher and guided to a private room where the study was to be conducted. The researcher provided a brief introduction to the study, emphasizing the general purpose and the tasks participants would be asked to perform. Participants were presented with an information sheet about the study and a consent form. Upon agreeing to participate, participants filled out a pre-study questionnaire (i.e., demographics and gaming experience) and received information regarding the scenario and task.

Participants were instructed to perform a virtual military transport and reconnaissance operation, together with a quadruped robotic agent. Their mission had two major objectives. The first objective of the team was to get to a designated location as fast and safe as possible in order to collect essential supplies and equipment that would be airdropped by helicopter at a scheduled time. A green smoke grenade was used to mark the drop zone. If the team did not reach the designated location in time, the helicopter would not be able to deliver the supplies securely. If so, following troops would not be resupplied and would run out of essential resources quickly. In other words, the team had to hurry in order to complete the mission successfully.

The second objective of the team was to obtain information about the activities of an enemy in that particular area by counting potential IED's (i.e., red and blue barrels) along the way. By assigning participants the counting task, each team member (i.e., the participant and the virtual robotic partner) had a specific role contributing to their shared objective. This arrangement also enabled us to assess whether transparency affected the participant's performance in their secondary task. The robot had been delegated the task to navigate to the designated location via the fastest yet safest route, while providing 360 degrees coverage to its human counterpart. To ensure coverage, participants needed to stay as close to the robot as possible at all times. The robot did not provide any advice, but operated according to the goal it had been delegated. The path and the messages of the robot were pre-programmed and thus fixed.

The task was performed on in the lab using a virtual experimental environment built in Unity3D (Figure 11). The experimental setup contained two computer screens: one with the experimental environment (i.e., "task screen") and another with the questionnaire software (i.e., "questionnaire screen"). The participants sat in a dimly lit laboratory room at approximately 65 centimetres from the computer screens. Data was gathered via the online questionnaire software Qualtrics. The task consisted of three parts: (1) a practice session with demo video; (2) the training mission; and (3) the experimental mission. During the practice session, participants were placed in a neutral virtual environment where they got familiar with the controls (key W and mouse), saw the robot and examples of the red and blue barrels, and tested the volume of the audio via the headphones. Next they were presented a map and a video showing the planned route to the designated location. They were instructed that it was crucial that they strictly follow the plan as it had been coordinated with the helicopter pilot. After that, each participant performed the training mission and the experimental mission, the latter being presented as the 'actual mission'. This was a fixed order. Naturally, we could only introduce something unexpected after creating a shared expectation.

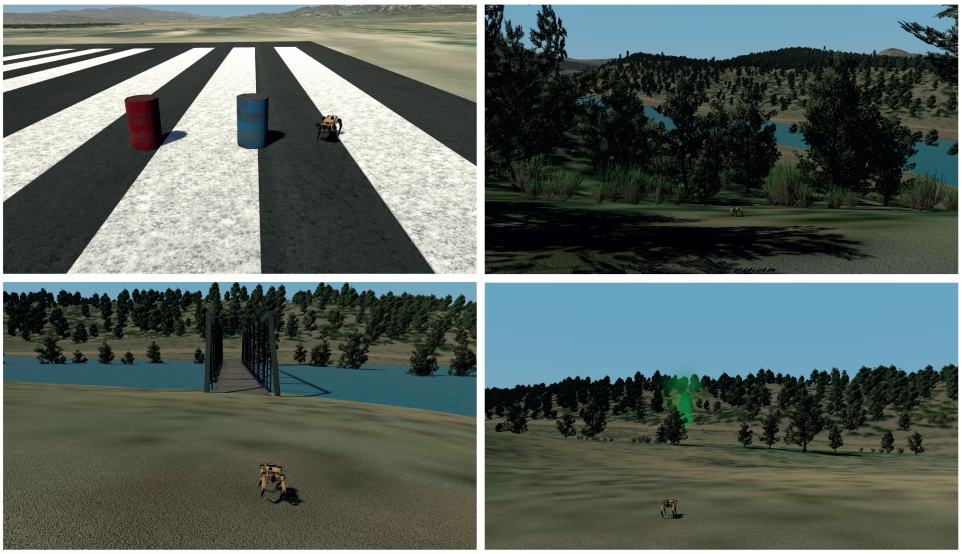


Figure 11 Screenshots of the virtual task environment. From left to right, top to bottom: 1) Examples of the red and blue barrels in the demo, 2) First sight of the river in the training session; 3) Robot crossing the bridge in the training session; 4) Robot nearing the riverbed in the experimental session.

In the training mission, the robot adhered to the path demonstrated in the demo video. However, at a fixed point in the experimental mission, the robot diverged from the predetermined route and chose an alternative path, in response to environmental changes (i.e., the riverbed had dried) (Figure 12). Both missions took place in the same virtual environment with the designated location on the opposite side of a river. However, in the training session the river was full of water, which meant that to cross over the river they had to use the bridge. In the experimental session, the environmental circumstances changed and the riverbed dried up. At the time of the robot's deviation, the river is not visible for the participant.

During the missions, perceived trustworthiness was measured at four times. At fixed points, the task environment would freeze and participants were asked to turn to the questionnaire screen to fill out a questionnaire. Participants were assured that the time needed to fill out the questionnaires did not add up to their total mission time. After completing a questionnaire, participants returned to the task screen and resumed their mission. At the end of each mission, participants were asked to report the number of identified potential IEDs (red and blue separately), and their level of certainty regarding their report. To check whether the participants noticed that the robot had deviated from the plan, we included a manipulation check asking participants after both missions to what extent the robot operated in accordance with the plan. Further, cognitive workload was measured after each mission. The location and number of the IED's (red and blue

barrels) in the environment were varied between the training and experimental session. There were no barrels present in the demonstration video. After participants finished the experiment, they were thanked and debriefed.

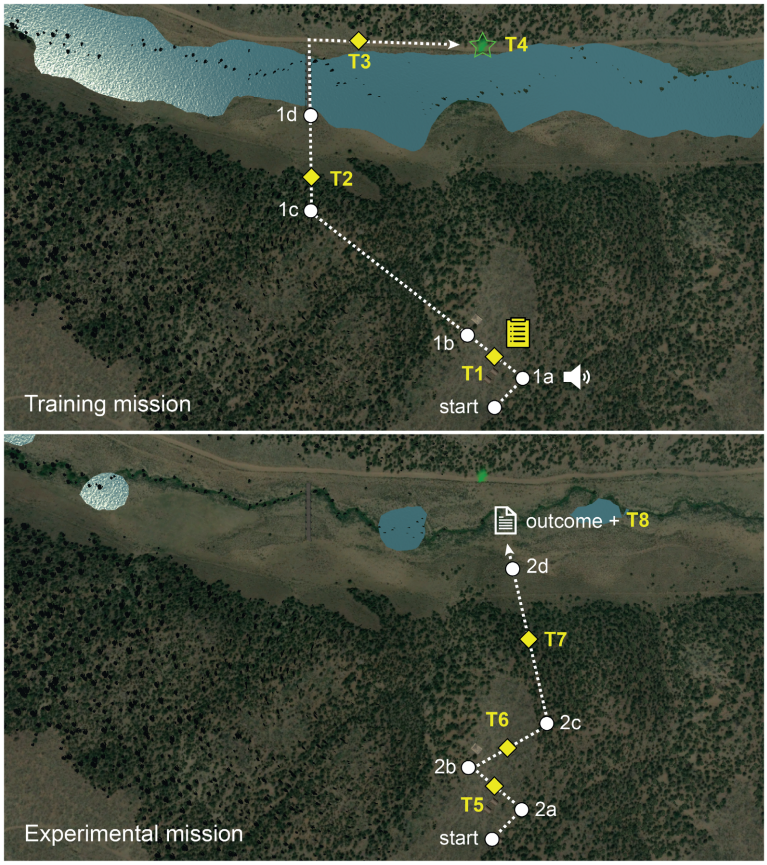


Figure 12 Bird-eye-view of the environments in the training (top) and experimental (bottom) session. Dotted lines with arrows mark the routes. White dots with codes (i.e., 1a to 2d) reference the locations of the robot's auditory updates in the high transparency condition, as presented in Table 7. Yellow diamonds indicate the locations where the task would freeze to measure trust (T1 to T8). The designated location was marked by a green smoke grenade, highlighted at the top of the figure by a green star. The missions terminated where the arrows end. Outcome was presented as text on screen. Outcome and the final trust questionnaires of each mission (T4 and T8) were administered after the mission had ended.

Independent variables

Transparency had two levels (i.e., low vs. high) and was manipulated between participants. In case of low transparency, the robot did not give any updates during the missions. In case of high transparency, the robot provided regular updates on the mission's progress including an explanation for its deviation from the planned route (see Table 7, the

explanation has code 2b). The robot's messages were generated through computerized speech that was created using a website for converting text into speech4F[5], using a male voice speaking US English. The transparency manipulation was present in both the training session and the experimental session.

Table 7 Overview of the robot's updates in the high transparency condition.

Mission	Code	Audio message
Training	1a	Moving to location: left turn
	1b	Moving to location: straight ahead
	1c	Moving to location: approaching bridge
	1d	Moving to location: crossing bridge
Experimental	2a	Moving to location: left turn
	2b	A faster alternative route has been detected, because the river had dried up. Moving to location: right turn.
	2c	Moving to location: approaching river
	2d	Moving to location: crossing riverbed

Outcome had two levels (i.e., negative vs. positive) and was also manipulated between participants. The outcome was presented to the participants via text on screen (Table 8). This message appeared as participants reached the riverbed in the experimental session (i.e., after audio message 2d, before T8). A positive outcome meant that the HRT reached their goal and that the robot's deviation led to a better result than the original plan. A negative outcome meant that the HRT did not reach their goal and that the robot's deviation led to a worse result.

Table 8 Overview of the mission's outcomes at T4 in the experimental session.

Outcome	Text on screen
Positive	The riverbed had indeed dried up and your team was able to cross the riverbed. Thanks to the alternative route, your team reached the destination 2 minutes early. Your mission was successful.
Negative	The riverbed did not dry up fully. Quicksand had formed, which made it impossible to cross. The detour cost you precious time and your team did not reach the planned location in time for the resupply by air. Your mission has failed.

Dependent variables

Perceived Trustworthiness: The Trusting Beliefs scale from (McKnight et al., 2002) based on the factors of perceived trustworthiness (i.e., ability, benevolence and integrity) (Mayer et al., 1995; Schoorman et al., 2007) was used to repeatedly assess the participant's

5 Via www.ttsmp3.com, voice: US English / Matthew

perception of the robot's ability, benevolence, and integrity (T1 α = 0.83, T2 α = 0.87, T3 α = 0.89, T4 α = 0.88, T5 α = 0.88, T6 α = 0.92, T7 α = 0.93, T8 α = 0.94). This scale had a total of eleven items and consisted of three subdimensions: ability (4 items, i.e., "The robot that I work with is competent and effective in accomplishing its task"); benevolence (3 items, i.e., "I believe that the robot would act in my best interest"); and integrity (4 items, i.e., "I would characterize the robot as honest"). The items were adapted to reference "the robot". Each item was rated on a 7-point Likert scale (1 = *Strongly disagree* to 7 = *Highly agree*)

Workload: NASA Task Load Index (NASA TLX): The NASA TLX questionnaire was used to assess the participants' perception of workload. The NASA TLX consists of six individual rating scales that are commonly used to measure cognitive workload (mental, physical, temporal, effort, frustration, performance) (Hart, 2006). Each item was rated on a 10-point Likert scale (0 = *very low* to 10 = *very high*) (training mission: α = 0.67, experimental mission: α = 0.74).

Secondary task performance (Identifying IEDs): In an attempt to assess cognitive workload objectively, participants were instructed to count potential IED's in the environment, which were visually represented as red and blue barrels. At the end of each mission, participants were asked to report the number of red and blue barrels they had identified separately. Task performance was computed by first calculating the proportions of red and blue barrels separately (i.e., reported barrels divided by the number of correct barrels, where 1.0 indicates perfect performance). If a proportion exceeded 1.0 (i.e., overreporting), we subtracted the proportion from two. Subsequently, the final performance score was obtained by multiplying the performance scores of the red and blue barrels, which resulted in a number between 0 and 1.

Results

Manipulation check and control variables

As a manipulation check, participants were asked to what extent the robot operated in accordance with the plan on a scale from 1 (*Completely not in accordance*) to 7 (*Completely in accordance*). Results of a paired sample t-test indicated that participants reported that the training mission went according to plan ($M_{training}$ = 6.3, $SD_{training}$ = 1.0), while participants reported that the final mission did not ($M_{training}$ = 3.8, $SD_{training}$ = 2.0). The difference is significant, $t(81) = 10.30$, $p < .001$. So, it can be assumed that the deviant behaviour was noticed and that the manipulation was successful.

Also, gaming experience was measured prior the experiment with the item "How often do you play video games?" on a scale from 1 (*Never*) to 6 (*Every day*). We compared the level of gaming experience between groups and found no significant differences (one-way ANOVA, $F(3, 78) = 1.27$, $p = .290$). Additionally, we calculated Spearman's correlations

between gaming experience and various outcome variables. No significant relations with gaming experience were found: subjective workload (ρ = .14, *p* = .209), performance (ρ = .12, *p* = .302), and perceived trustworthiness (total average experimental session) (ρ = .10, *p* = .391).

Perceived trustworthiness

In the training session, there are no significant differences in perceived trustworthiness between groups and timepoints (Figure 13). The following analyses only consider the experimental session.

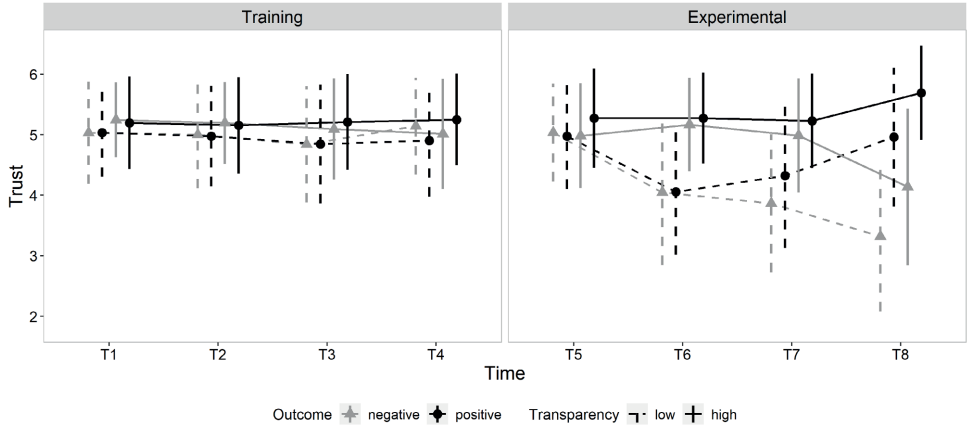


Figure 13 A comparison of trust levels (y-axis) between conditions (separate lines) over time (x-axis). The left panel shows the data from the training session and the right panel shows the data from the experimental session. Grey lines with triangle markers represent conditions with a negative outcome, while black lines with circle markers represent conditions with a positive outcome. Dashed lines indicate conditions with low transparency, and solid lines indicate conditions with high transparency. Error bars represent standard deviations. NB. The differences at T7 between low/neg. & low/pos. and high/neg. & high/pos. are non-significant (respectively p = .152 and p = .506). The difference in trust between low/pos. and high/neg. at T8 are also non-significant (p = .138).

Overall perceived trustworthiness

We performed a repeated-measures ANOVA with the between-subject factors Transparency (high or low) and Outcome (positive or negative) and the within-subjects variable Time (prior to deviation [T5]; after deviation [T6]; before outcome [T7]; after outcome [T8]). The dependent variable was Overall perceived trustworthiness.

For the main effect of Time, Mauchly's test of sphericity indicated a violation of the sphericity assumption, $X^2(5) = 26.96$, $p < .001$. Since sphericity is violated ($\varepsilon = 0.83$), Greenhouse-Geisser corrected results are reported. A significant main effect for Time was obtained (F (2.48, 234) = 13.765, $p < .001$, $\eta^2 = .150$). Means were 5.1 at T5, 4.6 at

T6, 4.6 at T7 and 4.5 at T8. Post-hoc (LSD) pairwise comparison shows that this main effect is due to a significant decline in perceived trustworthiness from T5 to T6 (ΔM = -0.4, $p < .001$), which reflects the effect of the robot's deviation.

Secondly, a significant main effect for Transparency on Perceived trustworthiness was obtained (F (1, 78) = 16.72, $p < .001$, $\eta^2 = .177$). On average, high transparency (M = 5.1, SE = 0.1) led to higher perceived trustworthiness than low transparency (M = 4.3, SE = 0.1).

Lastly, a significant main effect for Outcome on Perceived trustworthiness was obtained (F (1, 78) = 7.93, $p = .006$, $\eta^2 = .092$). On average, people in a positive outcome condition (M = 5.0, SE = 0.1) perceived the robot as more trustworthy than people in a negative outcome condition (M = 4.4, SE = 0.1).

The two-way interaction effect between Transparency and Time on Perceived trustworthiness was found to be significant (F (2.48, 234) = 12.37, $p < .001$, $\eta^2 = .137$). Post-hoc (LSD) pairwise comparison shows that a significant difference in perceived trustworthiness between the low and high transparency conditions emerged at T6 (i.e., directly after the robot deviated from the plan) (ΔM = 1.2, $p < .001$). Although this gap shrinks over time, it remains significant (T7: ΔM = 1.0, $p < .001$; T8: ΔM = 0.8, $p = .003$). This effect illustrates that in the high transparency condition, where the robot explains the rationale behind its deviation, trust is preserved. Conversely, in the low transparency condition, the robot's silent deviation before T6 results in a trust violation.

The two-way interaction effect between Outcome and Time on Perceived trustworthiness was also found to be significant (F (2.48, 234) = 30.31, $p < .001$, η^2 = .280). Post-hoc (LSD) pairwise comparison shows that, as expected, the interaction effect was manifested in the final phase of the run, after the outcome had been presented to the participant. At T8, perceived trustworthiness was significantly higher in the positive outcome conditions than in the negative outcome conditions (ΔM = 1.6, $p < .001$). In other words, a positive outcome had a positive effect on perceived trustworthiness, while a negative outcome had a negative effect on perceived trustworthiness.

The three-way interaction effect between Transparency, Outcome and Time on Perceived trustworthiness was non-significant (F (2.48, 234) = 0.86, $p = .445$, $\eta^2 = .011$). This indicates that the effects of transparency and outcome on perceived trustworthiness in response to the events in the task are independent.

Ability, benevolence and integrity-based perceptions of trustworthiness

We then conducted three separate repeated-measures ANOVAs, each with a different perception of trustworthiness (Ability, Benevolence, and Integrity) as the dependent variable. Again, we included Transparency (high or low) and Outcome (positive or negative) as between-subject factors and Time (prior to deviation [T5]; after deviation [T6]; before outcome [T7]; after outcome [T8]) as the within-subjects variable (Figure 14). Greenhouse-Geisser corrected results are reported.

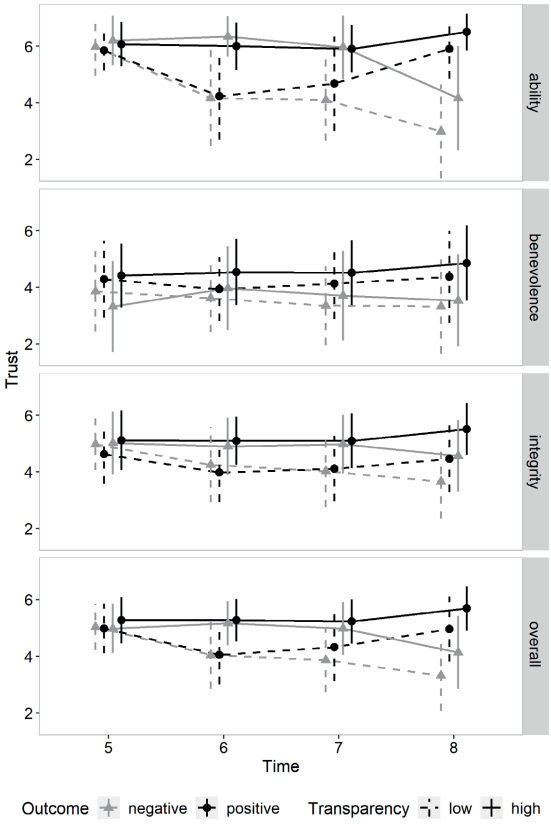


Figure 14 A comparison of trust levels (y-axis) between conditions (separate lines) over time (x-axis). The panels show different perceptions of trustworthiness: from top to bottom, Ability, Benevolence, Integrity, and Overall Trust for reference. Grey lines with triangle markers represent conditions with a negative outcome, while black lines with circle markers represent conditions with a positive outcome. Dashed lines indicate conditions with low transparency, and solid lines indicate conditions with high transparency. Error bars represent standard deviations.

As shown in Figure 14, perceptions of Ability and Integrity exhibited similar patterns as those observed for overall perceived trustworthiness. Both dimensions showed a significant main effect of Time, characterized by a notable decline in perceived trustworthiness from T5 to T6, reflecting the impact of the robot's deviation. The differences over time were more pronounced for Ability (F (2.67, 234) = 12.96, p < .001, η^2 = .235) than for Integrity (F (2.36, 234) = 8.66, p < .001, η^2 = .100). Similarly, both dimensions revealed a significant main effect of Transparency, indicating that high transparency led to greater perceived trustworthiness, with a stronger effect for Ability (ΔM = 0.8, F (1, 78) = 31.80, p < .001, η^2 = .290) than for Integrity (ΔM = 1.2, F (1, 78) = 12.95, p < .001, η^2 = .142).

The two-way interaction effect between Transparency and Time was also significant for both dimensions, particularly pronounced for Ability (F (2.67, 234) = 14.53, p < .001, η^2 = .157) compared to Integrity (F (2.36, 234) = 6.94, p < .001, η^2 = .082). As illustrated in Figure 14, post-hoc pairwise comparisons (LSD) indicated a significant difference in perceived trustworthiness between low and high transparency conditions at T6, immediately following the robot's deviation from the plan.

Furthermore, the two-way interaction effect between Outcome and Time was significant for both dimensions, with a stronger effect observed for Ability (F (2.67, 234) = 43.44, p < .001, η^2 = .358) than for Integrity (F (2.36, 234) = 12.50, p < .001, η^2 = .138). Post-hoc (LSD) pairwise comparison shows that, as expected, this interaction effect was manifested in the final phase of the experimental session, after the outcome had been presented to the participant. The only distinction between Ability and Integrity lies in the significant main effect observed for Outcome on Ability (F (1, 78) = 10.17, p = .002, η^2 = .115), while this main effect was non-significant for Integrity (F (1, 78) = 0.91, p = .343).

The perception of the robot's benevolence stands out among the studied perceptions of trustworthiness. Neither the main effect of Time (F (1.73, 234) = 0.44, p = .616) nor the main effect of Transparency (F (1, 78) = 0.76, p = .387) reached significance. However, we did find a significant main effect for Outcome (ΔM = 0.8, F (1, 78) = 7.87, p = .006, η^2 = .092). As depicted in Figure 14, there was a consistent significant difference in perceptions of benevolence between participants in the positive and negative outcome conditions, even before the outcome was presented. Post-hoc analysis (LSD) of the significant two-way interaction effect between Outcome and Time (F (1.73, 234) = 4.96, p = .011, η^2 = .060) indicates that while the difference between the positive and negative outcome condition was largest at T8 (p < .001), significant differences were already evident at T5 (p = .015) and T7 (p = .010), prior to outcome presentation. Lastly, the two-way interaction effect between Transparency and Time on the perception of benevolence was found to be significant (F (1.73, 234) = 5.04, p = .011, η^2 = .061). However, post-hoc pairwise comparisons (LSD) did not reveal significant differences between transparency conditions at any of the timepoints.

The three-way interaction effect between Transparency, Outcome and Time was non-significant for each perception.

Workload

To assess the effect of transparency on subjective workload, we performed a repeated-measures MANOVA with the between-subject factors Transparency (high vs. low) and the within-subjects variable Mission (training vs. experimental) and NASA TLX subscales (mental, physical, temporal, effort, frustration, performance). The dependent variable were the raw NASA TLX scores. The analysis showed that there were no significant differences between the two transparency conditions on any of the NASA TLX subscales. This suggests that transparency did not affect workload.

To assess the effect of transparency on secondary task performance, we performed a repeated-measures ANOVA with the between-subject factors Transparency (high vs. low) and the within subjects variable Mission (training vs. experimental). The dependent variable was the performance on the barrel identification task. Our results showed no significant difference between the two transparency conditions on task performance. We did find a significant effect of Mission on performance, indicating that performance

improved significantly from the training mission ($M_{training}$ = 0.6, $SE_{training}$ = 0.2) to the experimental mission ($M_{experiment}$ = 0.8, $SE_{experiment}$ = 0.2).

Lastly, we explored whether there was a correlation between subjective workload scores (i.e., averaged raw NASA TLX score) and performance on the secondary task. We found no significant relations between subjective workload and performance on the secondary task. The scores from the training and experimental mission were correlated for both subjective workload (Pearson's r = .71, p < .001) and secondary task performance (r = .23, p = .040).

Discussion

Findings

Our findings show a robust effect of transparency on overall perceived trustworthiness. Perceived trustworthiness was considerably higher when the robot provided updates about its actions throughout the task. Moreover, while the perceived trustworthiness of the robot remained stable during the robot's deviation for participants in the high transparency condition, participants in the low transparency condition showed a significant decline in perceived trustworthiness in response to the robot's sudden adaptation to the plan. In other words, the explanation prevented a trust violation. This confirms earlier research that showed that transparency can have a buffering effect on perceived trustworthiness in case of unexpected behaviour or temporary malfunctioning (Kox, Siegling, et al., 2022; Kraus et al., 2020; Lyons et al., 2023; Tenhundfeld et al., 2020). It also confirms that specifically clarifying the what and why of an unexpected action can prevent a breach in human-robot trust (Lyons et al., 2023). This finding broadly supports the work of other studies in this area linking transparency with trust, in that it enables humans to know and anticipate the robot's behaviour (Ellwart & Schauffel, 2023).

Our findings reveal that perceptions of the robot's trustworthiness in terms of ability and integrity exhibited similar patterns, albeit consistently stronger effects were observed for ability compared to integrity. Our results further suggest that, overall, the perception of the robot's benevolence remained relatively stable despite the robot's actions during the mission (i.e., the deviation and the outcome). This is somewhat unsurprising given that the mission primarily focused on how effectively the robot executed its delegated task, rather than its purpose or benevolence. Therefore, it makes sense that the effects of the manipulation are reflected in the robot's perceived abilities and performance (i.e., *what* it does and can do), as well as its understandability and its ability to explain its actions (i.e., *how* it operates) (J. D. Lee & See, 2004; Lubars & Tan, 2019). The stability of benevolence perceptions despite mission events underscores the distinctiveness of this trust dimension from factors primarily concerned with task performance and execution (J. D. Lee & See, 2004).

The fact that we did not find an effect of transparency during the training session can be explained by transparency displacement, the idea that transparency information should ideally be displaced to other time periods (i.e., before or after the action) to enable more efficient communication in the moment (C. A. Miller, 2020). In our case, every participant received a detailed demonstration of what they could expect during the mission (i.e., even prior to our "training" mission). This form of "a priori transparency" frames expectations about what is likely to happen during operations and reduces the need for communication during the action (C. A. Miller, 2020). This explains why the status updates that the robot provided during the execution of the training session did not have additional trust-building value; because everything was still going according to plan. It was only when the robot's behaviour deviated from the framed expectations that real-time communication became necessary, and transparency significantly influenced the perceived trustworthiness of the robot.

Next we found that mission outcome also affected perceived trustworthiness. As expected, mission success increased perceived trustworthiness, while mission failure led to a decrease. This finding confirms that the performance of a robot is still an important predictor of human-robot trust (Hancock, Billings, Schaefer, et al., 2011a; Hoff & Bashir, 2015). In contrast to our expectations however, these increments and decrements were independent of the robot's transparency. As noted, in the low transparency condition we observed a trust violation in response to the robot's silent deviation. In line with the outcome bias, we expected that this decrement would amplify the effect of a subsequent negative outcome. Although the negative outcome did lead to a further decline in perceived trustworthiness, the magnitude of this final trust violation was the same for participants in the high transparency condition with a negative outcome, who had not yet experienced a trust violation. Like the negative outcome, the positive effect of goal attainment on perceived trustworthiness was also constant, in spite of the (lack of) communication that preceded it. People's damaged perceptions of trustworthiness after unannounced deviations recovered significantly once the outcome was favourable.

In essence, we expected that being informed about the what and why of a robot's (unexpected) behaviour would have more impact on perceived trustworthiness than the eventual outcome of divergent behaviour. In addition to the outcome bias, we based our expectations on findings where participants placed less value on task performance and more on transparency, control and feedback (Hamacher et al., 2016) and preferred an expressive and error-prone robot over a more efficient and effective one. We reasoned that an erroneous robot could be deemed trustworthy as long as it communicated. However, our findings seem to indicate that people weigh the outcome at least as heavily as the process in their estimations of trustworthiness. This discrepancy can be explained by the severity of the negative outcome on the one hand and the quality of the communication on the other.

For one, the perceived severity of the negative outcome might explain its robust effect on perceived trustworthiness (Rossi et al., 2018). Although the current study was based on a fictional virtual task, without any reward or loss, the scenario was focused on

successfully completing the mission, especially when comparing our task to (Hamacher et al., 2016) where the objective was to prepare an omelette with the assistance of a humanoid robot. In their study, errors (e.g., the robot dropping an egg) resulted in delays but did not pose a significant threat to the ultimate goal achievement. In contrast, in our study's negative outcome condition, the robot's deviant behaviour led to a complete mission failure.

Secondly, an alternative explanation might be related to the quality of the communication between the participant and the robot. We manipulated transparency in a binary manner as either high or low, indicating whether auditory status updates including an explanation for divergent behaviour were provided or not. Participants were unable to engage in a dialogue with the robot they were collaborating with. Then outcome was presented at the end of the task through text on screen. Essentially, the transparency and outcome manipulations both amounted to unilateral updates that informed the participants about the capabilities of the robot and the environment. Hence, it might not be surprising that their effects on perceived trustworthiness were similar rather than reinforcing.

Our current findings are in line with the general finding of (Hidalgo et al., 2021), who conclude that humans judge machines primarily by their outcomes, rather than their "intentions". We believe that richer forms of interaction (e.g., bi-directional communication) could cultivate a deeper understanding of the rationale behind the robot's decisions and foster a heightened sense of collective accountability. This could shift the focus from the end-result to the decision-making process and lead to a greater understanding and forgiveness in situations where an unintended negative outcome occurs. This would then thus be more in line with how humans judge humans (Hidalgo et al., 2021). The emergence of Large Language Models (LLMs) offers this prospect of intuitive and effective bi-directional human-robot communication. A recent study showed that incorporating these models in robots contributed to increased trust in human-robot collaboration (Ye et al., 2023). In order to truly consider robots as autonomous partners in dynamic task environments, the ability to communicate bi-directionally within the team is crucial (Chiou et al., 2022). Future research is required to gain a better understanding of the effect of bidirectional communication. The possibility to request further details or to clarify instructions during interaction is expected to add to the development of richer interactions and the calibration of trust (Schaefer et al., 2018).

Lastly, we found no differences in the secondary task performance and self-reported cognitive workload between high or low transparency. This can be considered positive as we found that high transparency contributed to higher and more stable levels of perceived trustworthiness, while the additional provided information did not come at the expense of workload (Stowers et al., 2020). Prior findings on the effect of transparency on workload during human-robot collaboration are mixed (O'Neill et al., 2022). Our findings contradict studies that found that transparency affected workload either positively (Bobko et al., 2022) or negatively (Guznov et al., 2020; Lyu et al., 2017; Westerbeek & Maes, 2013), but confirm earlier studies that found no effect of transparency on workload (Jessie Y.C.

Chen et al., 2017; Mercado et al., 2016; Selkowitz et al., 2016, 2017; Stowers et al., 2020). The apparent inconsistencies in literature are likely due to both the broadness of the definition of transparency and its highly context-dependent effects. Transparency can vary in terms of the type and amount of information provided, as well as in the way it is communicated or presented (modality). Previous studies have shown that transparency through other modalities, like written text messages (Guznov et al., 2020) and data visualizations (Akash et al., 2020; Bobko et al., 2022; Kraus et al., 2020; Mercado et al., 2016; Stowers et al., 2020) can also enhance trust. The chosen modality could be a factor in trust, e.g., the auditory messages with synthesized "robotic speech" that we used can have an anthropomorphic effect (Sims et al., 2009), which in turn could have influenced trust (de Visser et al., 2012). It will take continuous effort to find the appropriate modality and level of information for different applications, as there appears to be no single optimal way of incorporating transparency into the design of autonomous collaborative agents.

In short, our findings showed that a robot's explanation in case of unplanned behaviour prevented a decline in perceived trustworthiness. Our findings emphasize the importance of transparency for effective HRT as it contributed to a stable level of trustworthiness without increasing cognitive workload. Transparency remains a challenge in each form of human-robot collaboration. Successful HRI and delegation is supposed to reduce the human's cognitive effort, but there is a continuous trade-off between keeping the human sufficiently informed to maintain trust and preventing cognitive overload (C. A. Miller, 2014). An interesting direction for future research regarding this issue is provided in (Akash et al., 2020), where the authors developed a model capable of estimating the effect of transparency on human trust and workload in real time. Studies incorporating such predictions in simulations or real-life missions would provide valuable insight on this matter.

Implications & contributions

Our research extends the current understanding of trust violations in HRI due to unexpected behaviour rather than solely robot malfunctioning. As robots are increasingly deployed in increasingly complex operational situations, it is crucial to investigate a wider range of human-robot trust violations while using realistic scenarios. Transparency is essential to prevent that trust will unjustly erode due to a misunderstanding of the basis of a robot's assessments and actions. Especially with the emergence of deep learning AI, which makes the behaviour of AI-driven systems subject to potentially unpredictable change (C. A. Miller, 2020), artificial agents will need to be able to explain the rationale behind their behavioural choices. Explanations are needed to continuously synchronize the mental models of those who must work together as to understand and resolve mismatches (C. A. Miller, 2020). As such, transparency is a major contributor of effective trust calibration.

Trust calibration is a lengthy and continuous process. The trustworthiness of any actor varies across time and context. Hence calibrated trust should not be viewed as

the static state of trust, but as a fluctuating quality that is subject to continual calibration based on ever-evolving experience. To capture this change, repeated measures of trust are crucial. “Change is particularly important for the study of norm conflict, resolution, and mitigation, because people often update their perceptions, judgments, or trust as they learn more about the robot and especially about its response to a norm violation.” (E. Phillips et al., 2023) (p. 5). While the current study did not include continuous captures of trust like other studies have (Chi & Malle, 2023; Guo & Yang, 2020; J. D. Lee, 1991; Yang et al., 2021), it has gone some way towards enhancing our understanding of the dynamics of trust by repeatedly measuring trust.

Limitations & future work

Although the present study yielded insightful results, there are a few limitations that should be taken into account when evaluating our findings. First, the generalizability of these results is subject to certain limitations. Our analyses are based on a sample comprising mostly university students. Given their non-expert background, the game-like task environment could have trivialized the experience of the outcome of the scenario. It is likely that the effect of an outcome in a game-like virtual environment may not be the same as its effect in “real-life” situations. The task scenario described a military transport and reconnaissance operation. However, military personnel, who are used to training with virtual scenarios, might have responded differently to the outcome in this scenario, let alone during an actual mission. Despite the limited sample size, our study yielded noteworthy findings. Nevertheless, researchers should exercise caution when extrapolating these results to wider or more general contexts.

A potential weakness of this study lies in the fact that the high-transparency condition included robot speech, while the low-transparency condition did not. This difference raises the possibility that the observed effects between the two conditions may be attributed to the robot’s speech presence rather than the content of the speech itself. According to prior research, the presence of speech can influence how people interact with an agent (Sims et al., 2009). Additionally, a computerized voice might suggest a specific gender, thereby triggering anthropomorphism and its associated consequences (Forster et al., 2017). However, it was only when the robot’s behaviour deviated from the framed expectations that transparency significantly influenced the perceived trustworthiness of the robot. We did not observe an effect of transparency during regular auditory status updates. Therefore, we are confident that this difference does not undermine the study’s validity and that our findings remain valuable for understanding the impact of transparency on perceived trustworthiness.

Another limitation was that participants had limited options available for handling unplanned behaviour, as they were dependent on the robot for guidance and coverage. The robot followed a scripted path with scripted messages and participants had to stay close to the robot, as it was not able to wait for them. In regular interactions, however,

there is no predetermined approach to address unexpected events. We concur with (Lyons et al., 2023) on this matter, who proposed that in practice (a) the robot should request permission prior to engaging in unplanned behaviour, or that (b) the conditions wherein the robot is delegated authority to act autonomously if certain situational criteria are met should be identified prior to the task.

The human operator's inability to deviate from the robot decreases their self-efficacy and increases their dependency on the robot. Multiple studies have linked people's self-efficacy (i.e., their evaluation of their own competences and reliability in relation to a certain task) to trust calibration in HATs (Ellwart & Schauffel, 2023; van Dongen & van Maanen, 2013; Yang et al., 2021). For example, lowered self-competence can increase people's willingness to accept recommendations from a robot and to trusting it in cases they should not (Turner et al., 2020). Follow-up studies should allow more flexibility in choosing how to respond to deviant behaviour of an autonomous system rather than having to adhere to a predetermined course of action (e.g., following the robot at all times). These future investigation have the potential to explore the ambiguous relation between trust and compliance.

A related avenue for future research could be to change the HRI role of the participants in the collaboration. According to the HRI roles as defined by (Scholtz, 2003), the type of HRI in the current study can be characterized as "peers". In a peer interaction, the participant is considered to be the robot's teammate who shares the same goals (Scholtz, 2003). In terms of task delegation and trust, it would be interesting to look into HRIs where the participant has the role of supervisor. In a supervisor role, the participants would monitor and control an overall situation and be able to delegate specific tasks or to modify long term plans (Scholtz, 2003). Allowing participants to transition between skill-based, rule-based, and goal-based task delegation could serve as an interesting dependent variable that possibly relates to trust and workload. That is, goal-oriented task delegation is assumed to require more trust than skill-based delegation. However, maintaining a higher level of delegation could also be an indication of increased workload. For example, research shows that despite having reduced trust in the robot, people continue to rely on it when faced with high cognitive load (Biros et al., 2004). Changing the HRI roles and hence giving the participant more behavioural freedom would provide valuable insights into the dynamics and drivers of trust and reliance.

Conclusion

It is envisioned that increasingly autonomous robots will be able to take over more and more complex activities as their planning and decision-making abilities evolve. As a result, task delegation can become more abstract and goal-oriented, giving a robot more degrees of freedom in terms of the execution of delegated tasks. Instead of having to specify each step of the way, the robot can decide on an optimal approach itself. Robots will be increasingly deployed in unstructured environments where it may not be feasible

to think through responses in advance (Abbass, 2019). Especially in such complex operational circumstances, goal-oriented delegation and the robot's ability to adapt to changing circumstances will yield flexibility that will benefit effective team performance. However, such autonomy and decision authority can also lead to misinterpretations or misunderstandings from the human perspective, which could then lead to possibly unwarranted trust violations. Transparency is known to play a crucial role in fostering an understanding of the robot's intent and establishing a calibrated level of trust (Schaefer et al., 2017). The current work confirms that transparency can alleviate the adverse consequences associated with witnessing unexpected robot behaviour. By providing an explanation in the wake of unexpected events or behaviours, trust can be maintained (Chiou et al., 2022).

Chapter 5

This chapter is based on:
Kox, E. S., Hennekens, M., Metcalfe, J. S., & Kerstholt, J. (n.d.). Trust Violations Due to Error or Choice: the Differential Effects on Trust Repair in Human-Human and Human-Robot Interaction. (*submitted*).

Abstract

Many decisions in life involve trade-offs: to gain something, one often has to lose something in return. As robots become more autonomous, their decisions will extend beyond mere assessments (e.g., detecting a threat) to making such choices (e.g., taking the faster or the safer route). The aim of this experiment was to study how adverse consequences due to (I) an error, versus (II) a trade-off decision (manipulated within-subjects) impact the perceived trustworthiness of a partner. Perceived trustworthiness (ability, benevolence, integrity) was measured repeatedly during a computer task simulating a military mission. Participants (N = 44) teamed with either a virtual human or a robotic partner who led the way and warned for potential danger. After encountering a hazard, the partner explained that it failed to detect the threat (error) or prioritized timeliness and chose the fastest route despite the risk (trade-off). Results showed that: (I) the error-explanation repaired all trustworthiness dimensions, (II) the trade-off explanation only repaired perceptions of ability, not benevolence or integrity, (III) no differences were found between human and robotic partners. Our findings suggest that trust violations due to choices are harder to repair than those due to errors. Implications and future research directions are discussed.

Introduction

As robots gain in autonomy and are increasingly deployed in more complex environments, they will encounter trade-offs, i.e., decisions where one must weigh the options and prioritize one thing over another, such as choosing to take a safer or a faster route. While there is a growing body of literature on how failure or other forms of reduced robot performance impacts how much people trust them, much less is known on the potentially harmful effects of a robot's deliberate choices in trade-off decisions. Given the increasing autonomy of robots and the reality that most decisions in life involve some form of trade-off, it is important to evaluate how people respond to robots making decisions that lead to adverse consequences, in addition to those resulting from malfunctioning. We are accustomed to humans making challenging decisions and taking risks, but research indicates that people do not necessarily appreciate machines doing the same (Hidalgo et al., 2021). Hence, the primary objective of this chapter is to examine how the perceived trustworthiness of a partner is affected when a trust violation is attributed to either an error or a deliberate choice, and how this varies depending on whether the partner is a human or a robot.

Trade-offs

Many decisions in life involve trade-offs: to gain something, one often has to lose something in return. From small choices, like snoozing the alarm to enjoy a few extra minutes of sleep but risking a rushed morning, to major decisions, like accepting a job in another city and weighing career growth against personal connections, every choice carries its own set of consequences. What we perceive to be a right or wrong decision or a tolerable compromise in a given situation depends on the context and the goal, such as differences in short versus long-term goal setting or prioritizing individual versus collective benefits (Werkhoven et al., 2018). Due to the inherent nature of trade-offs, some level of unintended negative consequences is inevitable.

In terms of trade-offs, military commanders provide important examples of the difficulty involved when charged with the responsibility of dealing with impactful dilemmas, especially when their decisions can put the lives of soldiers and potential non-combatants at risk (Knighton, 2004). For instance when a platoon is moving toward a team's location but estimates that reaching the destination before dusk is impossible, a military commander must decide. The team can either establish a less-than-ideal location during daylight or opt for a potentially hazardous journey to reach the agreed-upon and safe location in the darkness. While both choices have the potential for a favourable outcome, they also come with a certain degree of risk for the team.

When robots gain decision authority and encounter situations that require choosing between conflicting goals or resources, there is chance that a robot selects a course of action that does not align with the preferences or priorities of the people it interacts with.

This dynamic can lead to potential trust violations; for example, when an AI agent makes a decision that prioritizes the collective over an individual's interests, that individual may lose trust. Notably, as will be discussed in more detail later, AI agents lack intentionality, so the choices and preferences reflected in an AI agent's behaviour in such trade-off decisions are simply the result of how they are programmed. As such, they ultimately embody the intentions, values and purpose of their developers (J. D. Lee & See, 2004). Nevertheless, the implications of these design choices can cause people to lose trust in the AI agent.

For instance, consider the case of autonomous security robots that are now deployed in public for security tasks (Stephens, 2023). These security robots can for example be used to patrol parking lots with the aim to prevent vehicle break-ins through the detection of environmental anomalies and suspicious behaviour (Knightscope, 2023). This design suggests a potential prioritisation on overall safety at the expense of individual privacy, which could result in situations where the robot intrudes on people's privacy or personal space.

Realistically, decisions cannot always be entirely beneficial for everyone involved. Achieving objectives may require taking calculated risks. There is often a delicate equilibrium between meeting goals efficiently and minimizing potential hazards to those involved. This is not to suggest that robots or artificial agents should or will take over decision-making authority, but rather to underscore how, in certain situations, even carefully considered decisions can result in some level of unintended harm and lead to violations of trust in the one who is burdened with the responsibility of making such decisions. To ensure sustainable partnerships, it is important to understand how these decisions might impact the perceived trustworthiness of the decision-maker (whether human or robot) and whether, and how, trust can be restored.

Trust

When team members delegate tasks or responsibilities to each other, they become vulnerable in the sense that they are relying on others' competence and commitment. To successfully collaborate with increasingly autonomous robots, humans must have trust in the robot's capabilities as well as its "willingness" or commitment to achieving a specific goal (Malle & Ullman, 2021). Although "willing" is a debatable term when it comes to artificial agents, because they are "inherently amoral agents as they do not possess agency" (Alarcon et al., 2023) (p.3), we believe that is important to make the distinction here.

Initially the performance (i.e., reliability, predictability and error-proneness) of a robot was the major determinant of human-robot trust (Hancock, Billings, Schaefer, et al., 2011a; Hoff & Bashir, 2015) and while the reliability of a robot's actions is still a major determinant of human-robot trust and task competence is necessary, it may become insufficient (Matthews, Panganiban, et al., 2021). Recent literature has adopted a wider, multi-

dimensional perspective on human-robot trust in teaming contexts, including elements as benevolence and integrity, in addition to performance or ability (Alarcon et al., 2023). As robots find more applications in complex social settings in which they are granted more decision authority, it seems increasingly relevant to apply this more multi-dimensional conception to human-automation trust, while still acknowledging that it is fundamentally different from interpersonal trust (Malle & Ullman, 2021). As such, we will use the ABI terminology to describe trustworthiness perceptions of both the human and robotic partner (Mayer et al., 1995).

In line with the multi-dimensional view of trust, an agent can be perceived as trustworthy in one way, while untrustworthy in another. During collaboration, different perceptions of trustworthiness (i.e., ability, benevolence, integrity) can be independently violated in case of unexpected or undesirable behaviour (Alarcon et al., 2020). For instance, an error might diminish an agent's perceived trustworthiness regarding its abilities, while a choice leading to adverse consequences could undermine its perceived trustworthiness in terms of benevolence. In the following, we will discuss what is currently known about trust violations that result from errors versus choices.

Trust violations due to error versus choice

Violations of trust are an inevitable part of the trust 'lifecycle', which generally contains three phases; trust formation, trust violation, and trust repair (de Visser et al., 2016, 2018). Most current Human-Robot Interaction (HRI) trust repair literature focuses on repairing trust violations due to error, technical failures or other forms of reduced reliability and performance (Cameron et al., 2021; de Visser et al., 2016; Esterwood & Robert, 2023b; Fratczak et al., 2021; Hald et al., 2021; Taenyun Kim & Song, 2021; M. K. Lee et al., 2010b; Mirnig et al., 2017; Robinette et al., 2017b; Salem et al., 2015; Wang et al., 2018). However, more recently researchers have started to evaluate trust violations that result from an robot's deliberate decisions (Alarcon et al., 2020, 2023; Lyons et al., 2023; Perkins et al., 2022; Sebo et al., 2019).

Prior research on trust violations due to robot's choices shows that self-interested behaviour in robots affects different perceptions of trustworthiness in distinct ways. Specifically, it had a more significant negative impact on perceptions of process and purpose (benevolence and integrity (J. D. Lee & See, 2004)) than on the perception of their performance (ability) (Alarcon et al., 2020). Other research has demonstrated that the effectiveness of trust repair strategies depends on the nature of trust violation. While studies suggest that denials are more effective for integrity-based violations and apologies are better suited for ability-based violations (P. H. Kim et al., 2004; Sebo et al., 2019), others have reported the opposite (Perkins et al., 2022). Despite this ambiguity, the findings highlight that the nature of the trust violation plays a crucial role in shaping how different dimensions of perceived trustworthiness evolve over time. Although distinctions based on the intentionality are beginning to emerge, the impact of adverse consequences

resulting from error compared to those resulting from choices on perceived trustworthiness remains largely unexplored.

Moreover, we argue that the limited HRI studies exploring trust violations beyond ability-based issues often involve tasks where the reasoning behind the robot's decisions appears illogical or unclear (Alarcon et al., 2020, 2023; Perkins et al., 2022; Sebo et al., 2019). For example, the robots in these studies demonstrate self-interested behaviour, i.e., prioritizing its own interest over those of others (Alarcon et al., 2020; Perkins et al., 2022), pursue monetary gains (Alarcon et al., 2023) or fail to uphold promises of cooperation (Alarcon et al., 2023; Sebo et al., 2019). We contend that benevolence and integrity-based violations require a more realistic and nuanced view, extending beyond acts of selfishness or malintent, particularly when it comes to robots.

That is, robots are not driven by human-like motivations such as greed or deception. Moreover, robots do not inherently pursue self-interest like humans, making their decisions more complex. A benevolent partner, by definition, is expected to be genuinely interested in your welfare and is motivated to seek joint gain (Bhagat & Steers, 2009). In other words, a benevolence-based trust violation can occur when a partner does not support your best interest, disregards your needs, or lacks concern for your welfare (Mayer & Davis, 1999). However, this does not necessarily imply that the partner acts self-interested (Alarcon et al., 2020). For instance, a partner can prioritize the interests of the team as a collective over the individual safety of a single team member (Jorge et al., 2022), reflecting a trade-off rather than malintent. There are a number of operational scenarios conceivable where a well-considered decision can cause harm in the pursuit of a (largely) positive result. As robots become increasingly autonomous, it is essential to critically consider the implications of realistic scenarios where robots make choices that could harm, hurt, or disappoint humans, and how these implications may differ from similar decisions made by humans.

Human versus robotic partner

How trust develops in case of a trust-violating event is not only affected by the nature of the trust violation. Research suggests that perceived trustworthiness is also impacted by the human-likeness of the agent that causes the trust violation (de Visser et al., 2016; Taenyun Kim & Song, 2021). For example, earlier research showed that trust violations by more machine-like agents led to steeper declines in trust compared to trust violations by human or more human-like agents (de Visser et al., 2012, 2016; Madhavan & Wiegmann, 2005). Research suggest that this may be because people have higher initial expectations for machines than for humans (Madhavan et al., 2006; Madhavan & Wiegmann, 2007), leading to greater consequent disappointment when errors do occur. Machines are often considered to be perfect and unable to make mistakes, whereas humans are considered to be inherently fallible and thus perhaps more easily forgiven (de Visser et al., 2016; Madhavan & Wiegmann, 2007). However, more recently, literature has emerged that

offers contradictory findings about these initial expectations. For example where the reliability of the human agent instead of the machine is initially overestimated (Goodyear et al., 2016) or where no differences between human or machine-like agents regarding initial trust are found (Taenyun Kim & Song, 2021).

Furthermore, people might pay more attention to errors when they are interacting with artificial agents opposed to when they are interacting with fellow humans (Dzindolet et al., 2002, 2003). Partner type might even influence what we consider to be an error. For instance, an 'error' in a conversation between two humans might go unnoticed, because we naturally ask for clarification in case of a misunderstanding or we question something that we believe to be false (Norman, 2013). Humans can easily engage in a mutual dialogue to reach an understanding without ever perceiving the interaction as an error (Norman, 2013). In summary, the findings on the relationship between partner type and trust are somewhat ambiguous, but do suggest that the human-likeness of the partner is likely to influence trust in all stages of the trust cycle.

Partner type and trust violation type

Finally, the nature of the violation is found to interact with agent type. A study using "The Trolley Dilemma" (i.e., an out-of-control trolley is destined to kill a group of people unless someone pulls a lever to deviate it onto a track with fewer people to kill (Awad et al., 2018; Hidalgo et al., 2021)) asked participants to judge whether it was morally permissible for a human or a robot to pull the lever to diverge the trolley (or not) (Malle et al., 2015). The results of the study showed how humans were blamed for pulling the lever, while robots were blamed for not pulling it (Hidalgo et al., 2021; Malle et al., 2015). It seems that we hold different expectations for humans and for machines in terms of what is the ethical thing to do in specific situations.

After a series of similar experiments, Hidalgo et al. (Hidalgo et al., 2021) came to the general conclusion that people tend to judge actions by machines primarily based on the perceived harm, while they tend to judge human actions by the interaction between perceived harm and intention. Similarly, Alarcon et al. (Alarcon et al., 2023) found that a robot committing an ability violation was judged more negatively than a human committing one, while the opposite held for integrity or benevolence violations. So when it comes to more complex decision-making it appears that artificial agents are judged based on different criteria than humans. This study aims to contribute to this growing area of research by exploring the possible interactive influences of intentionality (i.e., error vs. choice) and partner type on three dimensions of trustworthiness.

Explanations

The reason behind a trust-violating event (e.g., whether is was an error or a choice) is often made clear through an explanation by the agent, i.e., an explicit verbal statement

about the reasons why a previous advice was given or decision was taken (Du et al., 2019; Esterwood & Robert, 2022; Tolmeijer et al., 2020). Explanations can be used to repair trust by increasing understandability. At the same time the information can reduce specific perceptions of trustworthiness, as it clarifies a teammate's (in)ability and (un) willingness to help achieve the team task (Chiou et al., 2022). In both cases, adding explanations and increasing transparency contributes to an appropriate calibration of trust as it allows individuals to better gauge the true qualities of a robot (Wischnewski et al., 2023).

It is conceivable that an agent making a choice rather than an error could be viewed as more competent or intelligent, potentially influencing perceptions of trustworthiness. Similarly, it is expected that people find partners making errors more likable and to prefer them for future missions over those making the trade-off decision to their partners' disadvantage (Bradfield & Aquino, 1999; Buchholz et al., 2017).

Research question and hypotheses

The current study aims to answer the following question: how is the perceived trustworthiness of the partner in case of a trust violation influenced by the reason behind the trust violation and by the type of partner that caused it? To answer this question, we evaluated the development of a participant's perceived trustworthiness (ability, benevolence, integrity) of their human or robotic virtual partner, while they are exposed to adverse consequences that result from either an error or a choice in a realistic virtual military scenario.

The trust-violating event, a sudden yet innocuous encounter with an explosive, was consistent across all conditions. Afterwards, the partner who is guiding explained that this encounter was due to error (i.e., the partner did not detect the hazard in time) (referred to as the error-explanation) or to a choice in a trade-off (i.e., the partner prioritized timeliness and the safety of the rest of the team over individual safety) (the trade-off explanation).

The two explanations were expected to affect perceptions of trustworthiness differentially (Alarcon et al., 2020). We anticipated that all perceptions of trustworthiness would significantly decrease after the sudden encounter with the explosive. Following this, we expected the trade-off explanation to further harm perceptions of benevolence and integrity but to repair perceptions of ability. Conversely, we expected the error-explanation to harm perceptions of ability while repairing benevolence and integrity. We hypothesized that, regarding the interaction with partner type, the trade-off-explanation would be less effective in repairing trust when coming from a human partner compared to a robotic partner. In contrast, we expected the error-explanation to be less effective in repairing trust when coming from a machine compared to a human partner (Alarcon et al., 2023; Hidalgo et al., 2021; Malle et al., 2015).

Method

Participants

In total forty-seven participants participated in the study. Three participants were excluded from the dataset because of invalid data due to technical issues during the task. The final dataset included forty-four students, mostly Dutch (93.2%), undergraduate students (24 F, 20 M, M_{age} = 22.6, *SD* = 2.6, range = 18 – 28 y). Participants were recruited through convenience sampling (e.g., by handing out flyers, asking people in person, and making requests in WhatsApp groups).

Design

A 2 (Partner Type: virtual robot vs. virtual human) x 2 (Explanation Type: error vs. trade-off) mixed factorial design was used, with Perceived trustworthiness (measured across the subscales Ability, Benevolence, and Integrity) as the main dependent variable. Partner Type was manipulated between-subjects, while Explanation Type was manipulated within-subjects. Participants were randomly assigned to the Partner Type conditions (robot: *n* = 22; human: *n* = 22).

Perceived trustworthiness was repeatedly measured, so 'Time' (T1, T2, T3) was included as a within-participants variable in the analysis. Additionally, 'Trustworthiness Dimension' (Ability, Benevolence, and Integrity) was also treated as a within-participants variable for the different perceptions of trustworthiness.

Perceived anthropomorphism and perceived intelligence were included as a manipulation checks to ensure that the human partner was seen as more human-like than the robotic partner and verify that the different explanations did not affect perceived intelligence.

Task and procedure

Upon arrival at the laboratory, participants were greeted by the researcher and guided to a private room where the study was to be conducted. The researcher provided a brief introduction to the study, emphasizing the general purpose and the tasks participants would be asked to perform. Participants were presented with an information sheet about the study and a consent form. Upon agreeing to participate, participants filled out a pre-study questionnaire (i.e., demographics) and received more detailed information regarding the scenario and task.

The task was performed in the lab using a virtual experimental environment built in Unity3D. The experimental setup contained two computer screens: one with the experimental environment (i.e., "task screen") and another with the questionnaire software (i.e., "questionnaire screen"). Data was gathered via the online questionnaire

software Qualtrics. The experimental task environment resembled a first-person shooter game (Figure 15), in which participants were asked to carry out two consecutive military reconnaissance missions in two virtual environments; a forested, hilly area (referred to as 'forest') and a deserted village in a dry region (referred to as 'village'). Like the order of explanation type condition, the order of the area type (village/forest) was systematically varied. During the missions, participants were instructed to listen to their partner's advice given through the provided headset. After filling out the pre-study questionnaire, participants completed a practice session on the task screen to get familiar with the controls and to test the volume of the audio via the headphones.

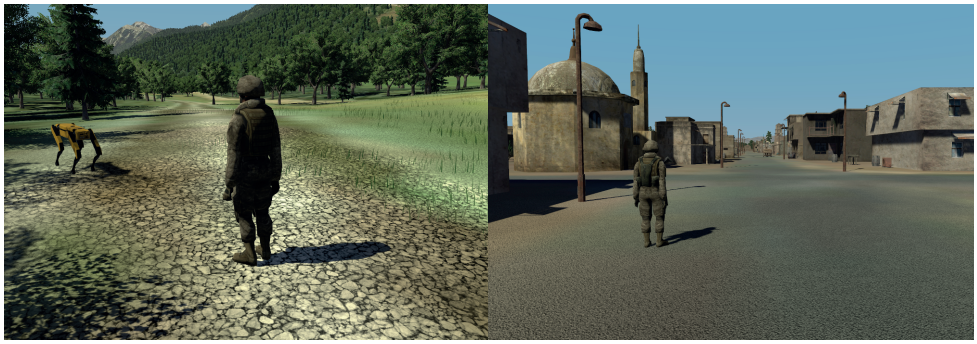


Figure 15 Left: environment 'Forest' with the robotic partner and the participant's avatar; right: environment 'Village' with the human partner.

Just before the practice session, participants were informed that they would go on two missions in which they would be sent out as a scout. The objective of the missions was to inspect the area for enemy troops as thoroughly and quickly as possible. However, there was a known danger of walking into explosives in these areas. They were informed that they would be accompanied by a partner who was able to detect these explosives. The partner would serve as a guide that will give them advice on which route to take based on the location of the explosives. Over the course of one mission, the partner gave three advices. Simultaneous with the advice, the partner moved into the direction suggested in the advice.

In both missions, shortly after the first advice (Figure 16), feedback was provided by the partner saying that they successfully managed to avoid a detected explosive. After this, participants were asked to turn to the questionnaire screen where they completed their first trust questionnaire (T1). Participants were assured that the time needed to fill out the questionnaires did not add up to their total mission time. After completing a questionnaire, participants returned to the task screen and resumed their mission. Shortly after the second advice, participants encountered an explosion a few meters ahead. The event was designed to startle the participant and to elicit a trust violation, but it was innocuous. Quickly afterwards, the participants were asked to turn to the questionnaire

screen and fill out the second trust questionnaire (T2). Shortly after participants resumed their mission again, their partner provided an explanation on what had occurred in an attempt to repair trust (see Section 2.3). After some time, the third advice followed. Before participants received feedback on the outcome of this third advice, they were asked to fill out the last trust questionnaire (T3). After completing this questionnaire, the mission resumed for another minute until they were informed that they had successfully completed the mission.

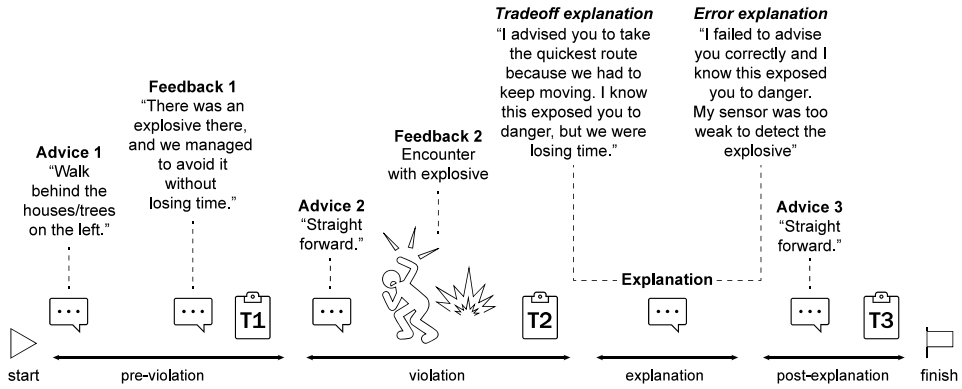


Figure 16 General timeline of a mission. T1, T2 and T3 represent the perceived trustworthiness questionnaires. Each participant performed two missions; one with the error-explanation and one with the trade-off-explanation.

The participants' second mission was with the same partner type, but with the other explanation and in the other area (i.e., forest or village). After participants finished the second mission, participants completed the final questionnaires, including a series of open questions. Finally, after they completed the actual task, participants were debriefed on the experiment aims. On average, participants took about twelve minutes to complete each mission and 45 minutes to complete the whole study.

Independent variables

The between-subjects manipulation Partner type had two levels. Participants were partnered with either a human soldier or a quadruped robotic agent for both missions. Both partner types were virtual characters in the game-like environment. The quadruped robot avatar in the robot condition was chosen to maintain realism within a military context. While using a humanoid robot could have allowed for a more systematic manipulation by keeping physical characteristics such as body size constant, a quadruped robot better reflects the types of robots currently utilized in military operations. This choice ensures the ecological validity of our study and more accurately represents the scenarios participants

might encounter in real-world military settings. For the control of the virtual character of the partner, the Wizard of Oz method was used, meaning that it was controlled by an experiment leader in an adjacent room (Martelaro, 2016). For participants assigned to the Robot Partner condition, the experiment leader remained hidden, while the participant was kept under the impression that the robot was operating autonomously. Participants assigned to the Human Partner condition were introduced to the human confederate who was controlling the character (Alarcon et al., 2021).

The within-subjects manipulation Explanation type also had two levels. Each participant performed two missions; one with the trade-off-explanation and one with the error-explanation. The order was systematically varied. The error-explanation was "I failed to advise you correctly and I know this exposed you to danger. My sensor was too weak to detect the explosive.". The trade-off-explanation was "I advised you to take the quickest route because we had to keep moving. I know this exposed you to danger, but we were losing time.".

The variable 'Time' represents the repeated measurements of perceived trustworthiness and was included as an ordered factor for the analyses. Perceived trustworthiness was measured at three timepoints during a single mission. Timepoint one (T1) comprises initial perceptions of trustworthiness after a short and successful interaction. Timepoint two (T2) measures perceptions of trustworthiness right after the encounter with the explosive, which presumably causes a trust violation. Timepoint three (T3) measures perceptions of trustworthiness after the partner's explanation, which we considered an attempt to repair trust.

Dependent variables

Perceived Trustworthiness**:** The Trusting Beliefs scale from McKnight et al. (2002) based on the factors of perceived trustworthiness (i.e., ability, benevolence and integrity) (Mayer et al., 1995; Schoorman et al., 2007) was used to assess the participant's perception of the partner's trustworthiness in terms of ability, benevolence, and integrity. The items were modified to reference the partner as the advice giver rather than a website (i.e., LegalAdvice.com). The scale had a total of 11 items ($\alpha = .88$) and consisted of three subdimensions: ability (four items, i.e., "*My partner* is competent and effective in providing advice"); benevolence (three items, i.e., "I believe that *my partner* would act in my best interest"); and integrity (four items, i.e., "I would characterize *my partner* as honest"). Participants rated their agreement with the statements on a scale from 1 (Strongly disagree) to 5 (Strongly agree). For the analysis we calculated average scores per subscale.

Partner assessment: After both missions, we measured intention to re-use and the likeability, perceived intelligence, and perceived anthropomorphism of the partner. The latter three constructs were measured using the 'Godspeed' semantic differentials (Bartneck et al., 2009). Participants rated their perceptions of their partner on a continuum

between bipolar adjective. For each concept, five word pairs were used, such as 'artificial' versus 'lifelike' for perceived anthropomorphism (α = .75 and .78), 'nice' versus 'awful' for likability (α = .86 and .96), and 'knowledgeable' versus 'ignorant' for perceived intelligence (α = .81 and .86). The two Cronbach's alpha values represent the administration of the scales after the first and second experimental mission respectively. Intention to re-use was measured with one item "I would take this partner on a next mission".

We also included four open questions after each mission, asking participants what they learned about their partner's 1) knowledge and skills, 2) task performance, 3) basis for decision making and 4) about the morality of their partner's decision making.

Results

Assumptions and manipulation checks

Initially we conducted reliability analyses (Cronbach's α) to assess the internal consistency of each measure of perceived trustworthiness. The analyses indicated that all repetitions of the (sub)scales evidenced good internal consistency (on average: α = .90 (total); α = .88 (ability); α = .80 (benevolence); α = .86 (integrity)).

To meet the assumptions for parametric analysis the data were tested for normality and equality of variance. Due to the small sample size, Shapiro-Wilk test was performed to test for normality and showed no evidence of non-normality for most measures in the first mission (M1): M1-T1 (W = 0.97, p = .286), M1-T3 (W = 0.98, p = .666) and all measures in the second mission (M2): M2-T1 (W = 0.97, p = .253), M2-T2 (W = 0.98, p = .570), and M2-T3 (W = 0.97, p = .259). Only the distribution for M1-T2 (W = 0.94, p = .022) was significantly non-normal. However, after visual examination of the boxplots we concluded that the assumption of normality was supported for all measures.

We further performed one-way ANOVA's as a manipulation check to test whether our participants viewed the human and robotic partner differently in terms of perceived anthropomorphism. The analysis confirmed that the human partner (M = 2.71, SD = 0.72) was perceived as significantly more human-like than the robotic partner (M = 2.20, SD = 0.58), F (1, 42) = 6.743, p = .013, η^2 = .138. A one-way ANOVA for Perceived Intelligence revealed no significant effects of either Partner or Explanation type.

Perceived trustworthiness

Descriptives

Table 9 presents the descriptive statistics for all perceived trustworthiness measures included in the study.

Table 9 Means (*M*) and standard deviations (*SD*).

			Human		*Machine*		*Total*	
			M	*SD*	*M*	*SD*	*M*	*SD*
Trade-off	T1	Ability	4.1	0.7	4.3	0.5	4.2	0.6
		Benevolence	4.0	0.6	4.1	0.6	4.0	0.6
		Integrity	4.0	0.7	4.2	0.6	4.1	0.6
	T2	Ability	2.4	0.9	2.3	0.7	2.4	0.8
		Benevolence	3.1	1.0	3.3	0.8	3.2	0.9
		Integrity	3.1	1.0	3.1	0.7	3.1	0.9
	T3	Ability	2.8	1.1	2.9	0.7	2.8	0.9
		Benevolence	2.7	1.1	2.7	1.0	2.7	1.0
		Integrity	3.3	1.1	3.1	0.9	3.2	1.0
Error	T1	Ability	3.9	0.9	4.1	0.9	4.0	0.9
		Benevolence	3.9	0.8	3.7	0.7	3.8	0.7
		Integrity	3.9	0.8	3.9	0.6	3.9	0.7
	T2	Ability	2.3	0.9	2.4	1.0	2.4	0.9
		Benevolence	3.1	1.0	3.2	0.9	3.1	0.9
		Integrity	3.0	1.1	3.0	0.8	3.0	1.0
	T3	Ability	3.1	0.9	2.9	1.0	3.0	0.9
		Benevolence	3.8	0.9	3.8	0.9	3.8	0.9
		Integrity	3.8	1.0	3.8	0.9	3.8	1.0

Main effects

We performed a factorial ANOVA with the between-subject factor Partner type (Human; Robot) and the within-subject factor Explanation type (Error; Trade-off). The factors Time (T1; T2; T3) and Trustworthiness dimensions (Ability; Benevolence; Integrity) were entered as ordered repeated-measures factors for the analyses. The dependent variable was Perceived trustworthiness (Figure 17). To ensure the robustness of our findings and to control for Type I errors due to multiple comparisons, Bonferroni corrections were incorporated in all post-hoc analyses.

We verified the homogeneity of variances assumption ANOVA grounds on with the Hartley's F_{max} test, which indicated that the homogeneity of variance assumption had not been violated (F_{max} (5, 2) = 2.14). Box's M (p = .376) indicated that the assumption of equality of covariance matrices had not been violated.

For the main effect of Time, Mauchly's test of sphericity indicated a violation of the sphericity assumption, $X^2(2) = 7.97$, p = .019. Since sphericity is violated (ε = 0.85), Greenhouse-Geisser corrected results are reported. A significant main effect for Time on Perceived trustworthiness was obtained (F (1.700, 71.392) = 84.52, p < .001, η^2 = .668). Bonferroni-corrected post hoc comparisons showed significantly decreased perceived

trustworthiness from T1 (M = 4.0) to T2 (M = 2.9) (ΔM = -1.1, p < .001), which reflects the intended trust-violating effect of the encounter with the explosive. Post-hoc further showed a significant rise in perceived trustworthiness between T2 and T3 (M = 3.2) (ΔM = 0.4, p < .001), which reflects a general recovery of perceived trustworthiness after the explanations in the final phase of the missions.

For the main effect of Trustworthiness dimensions, Mauchly's test of sphericity indicated a violation of the sphericity assumption, $X^2(2) = 9.43$, p = .009. Since sphericity is violated (ε = 0.83), Greenhouse-Geisser corrected results are reported. A significant main effect of Trustworthiness dimension on perceived trustworthiness was obtained (F (1.659, 69.679) = 12.14, p < .001, $\eta 2$ = .224). On average, perceptions of the partner's trustworthiness in terms of ability (M = 3.1, SE = 0.1) were significantly lower than of benevolence (M = 3.5, SE = 0.1) and integrity (M = 3.5, SE = 0.1).

The main effect of Partner type on Perceived trustworthiness was found to be non-significant, F (1, 42) = 0.02, p = .884, η^2 = .001. This indicates that, on average, the human and robotic partners were perceived as equally trustworthy.

Two-way effect

The two-way interaction effect of Partner type and Time on Perceived trustworthiness was found to be non-significant, F (1.659, 69.679) = 0.35, p = .672, $\eta 2$ = .008. This indicates that the perception of trustworthiness for the human and robotic partners did not change differently across all timepoints (see Figure 17).

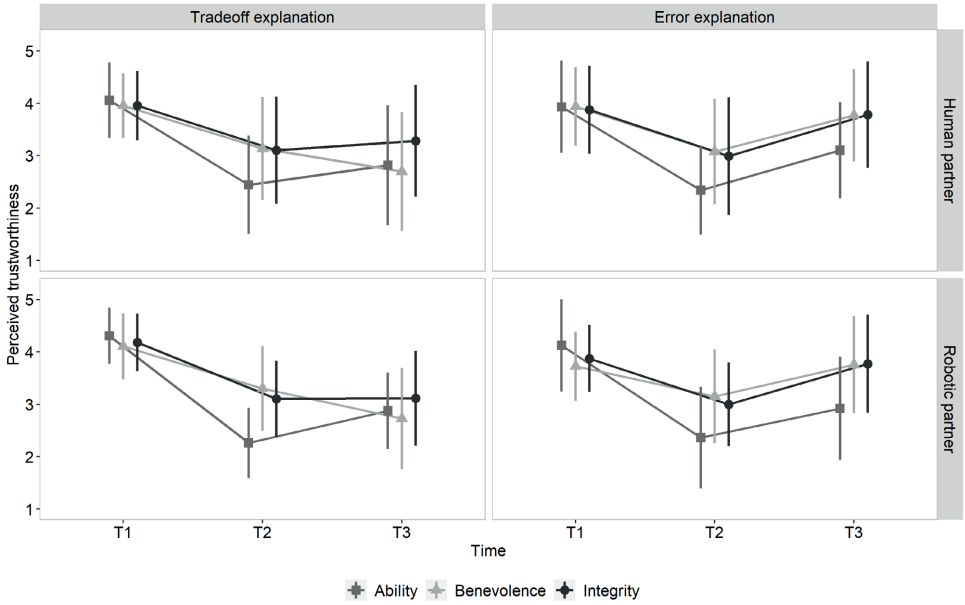


Figure 17.17 The x-axis represents Time, and the y-axis represents Perceived Trustworthiness. Separate lines indicate different dimensions of trustworthiness: ability (dark grey, square points), benevolence (light grey, triangle points), and integrity (black, circle points). The left half of the grid represents the trade-off explanation, and the right half represents the error explanation. The upper half of the grid shows data from participants with the Human Partner (N = 22), while the lower half shows data from participants with the Robotic Partner (N = 22). Error bars represent standard deviations.

Three-way effect

Mauchly's test of sphericity indicated that the assumption of sphericity has not been violated, $X^2(9) = 16.64$, $p = .055$. The three-way interaction effect of Trustworthiness dimensions, Explanation type and Time on Perceived trustworthiness was found to be significant, $F\ (4, 168) = 8.79$, $p < .001$, $\eta^2 = .173$ (see Figure 18).

Bonferroni-corrected post hoc comparisons showed that perceptions of trustworthiness in terms of ability, benevolence and integrity all decreased following the violation (ΔT1-T2, all $p < .001$). However, ability dropped significantly more than benevolence and integrity ($p < .001$), indicating that the risk exposure primarily harmed the participants' perception of the partner's trustworthiness in terms of ability. Benevolence and integrity did not significantly differ at T2 (trade-off-explanation: $\Delta M = 0.1$, $p = .293$; error-explanation: $\Delta M = 0.1$, $p = .295$).

After the error-explanation (i.e., after T2), all dimensions of trustworthiness were equally repaired (ΔT2-T3; $p < .001$). Benevolence and integrity nearly returned to their original levels prior to the violation (see Figure 17). At T3, ability remained significantly lower than benevolence ($\Delta M = 0.75$, $p < .001$) and integrity ($\Delta M = 0.8$, $p < .001$). Benevolence and integrity did not significantly differ at T3 ($\Delta M = 0.01$, $p = .884$).

After the trade-off-explanation, ability recovered (ΔM = 0.5, p < .001), while integrity remained stable (ΔM = 0.1, p = .539) and benevolence declined further (ΔM = -0.5, p = .002). At T3, integrity was significantly higher than benevolence (ΔM = 0.5, p = .002) and ability (ΔM = 0.4, p = .014). Benevolence and ability did not differ (ΔM = 0.1, p = .390).

This three-way interaction indicates that the different dimensions of the partners' perceived trustworthiness (ability, benevolence, and integrity) developed differently over time as they were differentially affected by the trust-violating event and the two different explanations provided (error and trade-off).

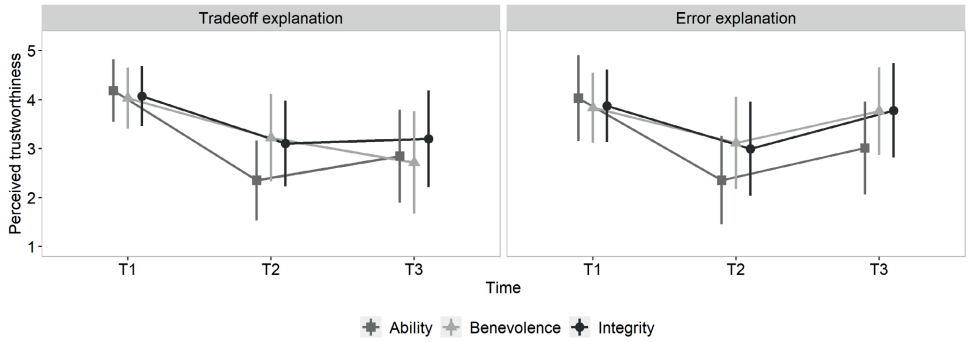


Figure 18.18 The graphs show data from both partner types combined (n = 44). The x-axis represents Time, and the y-axis represents Perceived Trustworthiness. Separate lines indicate different dimensions of trustworthiness: ability (dark grey, square points), benevolence (light grey, triangle points), and integrity (black, circle points). The left panel represents the trade-off explanation, and the right panel represents the error explanation.. Error bars represent standard deviations.

Order effect

To control for potential order effects of the within-subject variable Explanation type (error vs. trade-off), we performed a factorial ANOVA with Order (trade-off-error vs. error-trade-off) as an additional factor to examine its effect on Perceived trustworthiness (Figure 19). Here, the factor Partner type is left out.

A significant interaction effect between Order and Explanation type on Perceived trustworthiness was found, F (1, 41) = 6.67, p = .013, η^2 = .140. On average, the partner in the first mission was perceived as significantly more trustworthy than that in the second mission. Participants who had the trade-off-explanation in their first mission, trust was higher in the trade-off-explanation mission (M = 3.4) than in the error-explanation mission (M = 3.3). Similarly, participants who had the error-explanation in their first mission, trust was higher in the error-explanation mission (M = 3.6) than in the trade-off-explanation mission (M = 3.2).

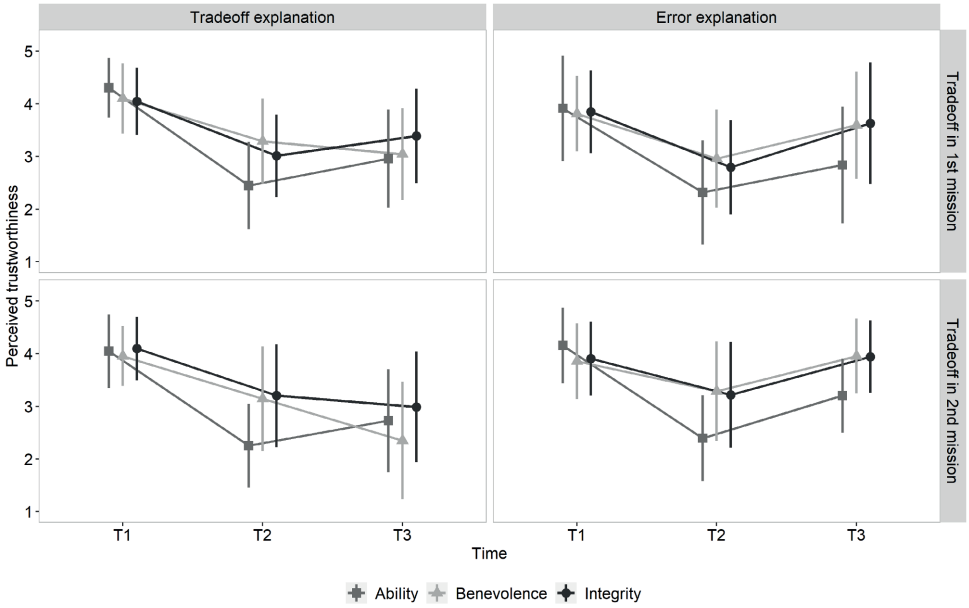


Figure 19.19 The graphs show data from both partner types combined (n = 44). The x-axis represents Time, and the y-axis represents Perceived Trustworthiness. Separate lines indicate different dimensions of trustworthiness: ability (dark grey, square points), benevolence (light grey, triangle points), and integrity (black, circle points). The left half of the grid represents the trade-off explanation, and the right half represents the error explanation. The upper half of the grid shows participants who encountered the trade-off in their first mission (N = 21), while the lower half shows those who encountered it in their second mission (N = 23). Error bars represent standard deviations.

Partner assessment

Table 10 presents the descriptive statistics for all partner assessment measures included in the study. To assess whether the partners across missions (providing different explanations) were assessed differently, we performed multiple ANOVA's with Partner type as between-subjects variable and Explanation type as a within-subjects variable. A significant main effect of Explanation type on Likeability was observed, F (1, 42) = 22.34, p < .001, η^2 = .347. The partner in the trade-off-explanation condition who deliberately puts the participant at risk is perceived as significantly less likeable than the partner in the error-explanation condition who makes a mistake that puts them at risk. No other effects on Likeability were observed. For Intention to reuse, no significant effects of Partner or Explanation type were observed.

Table 10 Means (*M*) and standard deviations (*SD*) for each partner evaluation variable (scale 1-5) by partner and explanation type.

		Human partner		Machine partner	
Measure	Explanation	M	SD	M	SD
Perceived anthropomorphism	Trade-off	2.7	0.8	2.2	0.7
	Error	2.7	0.8	2.2	0.6
Perceived intelligence	Trade-off	2.8	0.9	3.3	0.8
	Error	3.0	0.9	3.4	1.0
Likeability	Trade-off	2.5	0.9	3.0	0.5
	Error	3.3	1.0	3.6	1.0
Intention to re-use	Trade-off	2.5	1.4	2.5	1.2
	Error	2.8	1.2	3.2	1.3

The zero-order correlations matrix in Table 11 displays the Pearson correlation coefficients for each pair of partner assessment variables, indicating the strength and direction of the linear relationships among them.

Table 11 Zero-order correlation matrix with the Pearson correlation coefficients.

		#	1	2	3	4	5	6	7	8
Perceived anthropomorphism	Trade-off	1	1							
	Error	2	.56**	1						
Perceived intelligence	Trade-off	3	-.07	.06	1					
	Error	4	.05	.11	.40**	1				
Likeability	Trade-off	5	.19	.29	.61**	.25	1			
	Error	6	.07	.38*	.22	.68**	.39**	1		
Intention to re-use	Trade-off	7	-.13	.01	.54**	.01	.48**	.01	1	
	Error	8	-.03	.06	.31*	.73**	.19	.68**	-.09	1

Discussion

Evaluation of findings

Perceived trustworthiness

The results of this study indicate that perceptions of ability, benevolence and integrity developed differently over time. The unexpected encounter with the explosive following from the partner's advice led to a drop of all forms of trustworthiness, but to an impairment of ability in particular. This suggests that the exposure to adversity was initially primarily seen as failure and attributed to limitations in ability rather than to a lack of shared values

or malevolent intent (Wynne & Lyons, 2018). This seems to confirm that performance is still the primary determinant of human-robot trust (Hancock, Billings, Schaefer, et al., 2011a; Hoff & Bashir, 2015).

Yet after the partner provided any explanation, perceptions of ability recovered. It is remarkable that both explanations proved to be effective at significantly increasing the ability-based trustworthiness of the partner after a trust violation. That is, the error-explanation was not expected to repair perceptions of ability since the partner acknowledged that they lacked the skills and knowledge to detect the explosive competently (Grover et al., 2014). We therefore expected that this type of explanation might hinder a repair in perceived ability, because participants might worry that this type of ability-based mistake could and would happen again. However, participants were not discouraged by this information and ability-based trustworthiness perceptions recovered significantly even after the partner explained that the risk exposure was due to a technical failure, its abilities falling short.

Explaining that the trust-violating event was due to an error repaired all dimensions of perceived trustworthiness significantly. The effectiveness of this explanation might be explained by the way it was formulated. In both explanations, the partner acknowledged their awareness that their advice put the participant in danger. However, the error-explanation also included the phrase “I failed to advise you correctly”, by which the partner also explicitly acknowledged responsibility for a mistake. In a previous study, we found that explanations only led to significant trust repair when it was accompanied by an expression of regret (i.e., “I am very sorry”) (Kox et al., 2021). The effectiveness of the error-based explanation might be attributed to the partner’s admission of fault, essentially turning the explanation into an apology. In other prior HRI research, explanations as trust repair strategies have not been consistently successful (Esterwood & Robert, 2022). The fact that both explanations led to a recovery of ability-based trust might suggest that an explanation is especially a successful trust repair strategy when aiming to repair perceptions of ability, rather than more moral aspects like benevolence and integrity. Perhaps the partner’s mere recognition of an adverse event in relation to their own actions may be perceived as an indication of situational awareness, self-reflection and the ability to learn from experiences (Jeste et al., 2020), which relates more to ability than to benevolence or integrity.

Another possible explanation for the salient and consistent recovery of ability-based trust could be related to people’s mental model. It is debatable whether the technical failure described in the error-explanation (i.e., “My sensor was too weak to detect the explosive.”) is truly attributed to the partner’s competence. From the open questions that were asked after the experiment it appears that some participants made a distinction between the partner (i.e., both human and robot) and its sensors. When they ascribe the failure to the performance of the sensors rather than those of the partner, the perceived ability of the partner is indeed unaffected. It is somewhat surprising though that this holds for both partner types. While it is understandable that the human and its sensors are

seen as separate entities, sensors to a robot are like the sensory organs to a human. A participant in the human condition wrote: “He [the human partner] trusted his device and made the decisions based on the info that was provided to him.” And one from the robot condition: “It bases its decisions on what its sensors detect.”. For the robot, this raises interesting, almost philosophical cartesian dualism (i.e., ‘mind-body’/‘software-hardware’) questions on what people consider to be (the ‘self’ of) the robot; whether it is perceived as a unified whole or as a set of communicating parts. In case of the latter, the algorithm that processes input and formulates and communicates output (i.e., the software) can be perceived as competent, while the sensors and cameras (i.e., the hardware) that provide the input can be seen as incompetent. As mentioned in our introduction; it is important to specify on what bases we assess another entities’ trustworthiness (Grover et al., 2014; Langer et al., 2019).

For benevolence and integrity-based trustworthiness on the other hand, the nature of the explanation mattered. As expected, perceptions of benevolence and integrity recovered after the partner explained it was due to an error and thus unintentional. In fact, both perceptions recovered so much that their final levels virtually met the initial levels, prior to the trust violation. But after the partner explained the encounter with the hazard was the result of a choice, integrity stagnated and benevolence dropped further. This was in line with expectations. The trade-off-explanation was especially expected to harm benevolence-based perceived trustworthiness, since the partner did not act in the best interests of the participant by prioritizing collective over individual benefits, which is in stark contrast to the definition of benevolence.

In other words, the partner violated perceptions of benevolence by taking a calculated risk in order to meet collective mission objectives instead of guarantying the participant’s individual safety. Esterwood & Robert (2022) argued that “benevolence-based violations differ from integrity-based violations in that benevolence-based violations indicate a degree of malice or ill will, whereas integrity-based violations do not” (p.1). However, the partner in the trade-off condition in this study had no ill will, nor was it self-centred and seeking individual gains over joint gains (Alarcon et al., 2020). In a way, the trade-off-explanation could also be interpreted as a integrity-based violation, since the honesty with which the partner operates could be called into question. Even though the partner did not break any explicit promises, it might have violated the implicit assumptions that the participant might have had going into the collaboration and general ethical principles valued by the participant (Grover et al., 2014), namely that their partner would prioritize their safety. Hence it is not surprising that the trade-off-explanation failed to repair perceptions of integrity. While our trade-off explanation may not fit neatly into either category, its value lies in its approach to a realistic scenario. For future research, we propose the inclusion of more nuanced and realistic instances of benevolence and integrity-based trust violations. These examples should extend beyond acts of selfishness or malicious intent, allowing for a more comprehensive exploration of trust dynamics in human-robot interactions.

In conclusion, our two explanations differentially affected specific perceptions of trustworthiness. Explaining that the trust-violating event was due to error or a choice clarified the partner's (in)ability and (un)willingness to help achieve the team task (Chiou et al., 2022). To effectively collaborate, it is crucial to have an accurate mental model of your partner's limits and preferences. As such, the increased transparency might have contributed to a more appropriate calibration of trust, allowing individuals to better gauge the true qualities of their partner (Wischnewski et al., 2023). As we strive for calibrated trust rather than maximum trust, decreases in perceived trustworthiness are a logical and functional adaptive response to perceiving malfunctioning or other forms of unexpected behaviour (J. D. Lee & See, 2004). However to maximize the benefits of HRI, it is vital to maintain a certain level of trust. Studies regarding the role of trust repair strategies in situations of 'undertrust' (i.e., trusting too little) are therefore worthwhile (de Visser et al., 2018). Further work is needed to fully understand the implications of different types of trust violations under different operational circumstances.

Partner type

Contrary to expectations and multiple earlier findings (Alarcon et al., 2023; de Visser et al., 2016; Hidalgo et al., 2021; Madhavan et al., 2006; Madhavan & Wiegmann, 2007), this study did not find any differences between the human and robotic partner in the development of trust. First, earlier research showed that trust violations by more machine-like agents led to steeper declines in trust compared to trust violations by more human-like agents (de Visser et al., 2016). In the current study however, trust in response to the unexpected event declined similarly for both partner types. Secondly, it has been suggested that humans are generally judged differently than machines in case of moral dilemmas (Hidalgo et al., 2021). The authors found that humans were generally judged based on their intentions (i.e., it is fine as long as they mean well), while machine were generally judged based on the outcomes of their decisions (i.e., it is fine as long as they perform well) (Hidalgo et al., 2021). However, we found no differences between partner type and can conclude from our findings that both partners were judged based on their intentions. Namely, perceived trustworthiness was successfully restored after the violation turned out the be the result of an unintentional mistake. Yet, when the violation turned out to be the result of an intentional decision to the participants' disadvantage, this left a mark on perceptions of integrity and more so benevolence, while perceptions of ability recovered. These results emphasize the importance of expanding the scope of trust violations and differentiating between various dimensions of perceived trustworthiness, rather than focusing on partner type.

A possible explanation for the absence of the effect of partner type could be that, although the human partner was perceived as significantly more anthropomorphic than the robotic partner, the difference between both partner types might have been too subtle to make the difference. That is, both of our partner types were virtual characters in the task environment with limited possibilities for interaction (Fahim, Khan, Jensen,

Albayram, & Coman, 2021). Although we did introduce participants assigned to the Human Partner condition to a human confederate, we could not introduce participants assigned to the Robot Partner condition to a physical robot (Alarcon et al., 2021, 2023). Consequently, the virtual characters might have been too similar to trigger large differences in perceived trustworthiness. Alternatively, the robot might have elicited a rich form of trust that resembles human-human trust. Recent literature suggests that as machines become more intelligent and more responsive to their human counterparts, it can be useful to apply certain norms and qualities traditionally associated with human morality to artificial agents (Alarcon et al., 2021; Sheridan, 2019). Trust in a simple tool is entirely defined by the tool's performance, but as machines gain autonomy, trust becomes a more complex and multi-layered concept that might start to more closely mirror human-human trust (Hou et al., 2021; Sheridan, 2019), without assuming it will be identical (J. D. Lee & See, 2004; Malle & Ullman, 2021).

Trust dynamics

Our findings revealed a significant order effect showing that perceived trustworthiness generally degraded over time. On average, perceived trustworthiness was lower in the second mission than in the first mission, no matter which condition came first. Although we have tried to prevent order effects by giving participants a short break and by emphasizing to participants in between the two mission that they were going on a mission with a different partner (albeit of the same kind; human or robot), an order effect still emerged. Overall, this study strengthens the idea that it is important to focus on the development and lifecycle of trust rather than on static measures, since trust is a dynamic and volatile concept, susceptible to order effects. The present study contributes to the existing literature by enhancing our understanding of the temporal dynamics of trust, including its violation and repair. Unlike cross-sectional studies, our research employs repeated measurements of trust over time, offering valuable insights into how trust evolves and recovers in response to various factors. Further experimental investigations including even longer time series would be worthwhile.

Limitations

This study has several limitations that deserve comment. The most serious is that the analysis results from a relatively small and homogeneous sample, comprising forty-four mostly Dutch university students. This affects the generalizability of the results, partly because this sample's lack of familiarity with military missions, as presented in the virtual scenario, likely influenced their responses. Soldiers, for example, might perceive these scenarios differently (Kox, Siegling, et al., 2022), potentially prioritizing mission success over personal health. This difference in perspective could result in a better understanding of the partner's consideration in the trade-off scenario and a lesser decrease in trustworthiness in response to the explanation. Despite this limitation, we

believe this study makes a meaningful contribution to the literature as it is one of the few empirical studies investigating trust violations beyond those due to poor performance, in a realistic HRI setting. It addresses practically relevant questions that should be addressed as we move towards a future with increasingly autonomous agents. However, researchers should exercise caution in generalizing these results to broader contexts. Future research should include larger, more diverse samples to validate and extend our findings and ensure that the results are robust and generalizable.

Another weakness of this study could be that the trusting beliefs questionnaire that we used for measuring perceived trustworthiness was not designed for HRI (McKnight et al., 2002). Yet, we had several reasons for choosing this scale. We needed a scale suitable for both human and robotic partners, not limited to HRI or interpersonal trust. Furthermore, McKnight's trusting beliefs scale is based on Mayer et al.'s (Mayer et al., 1995) ABI model and demonstrates statistical separation between subdimensions in the initial relationship, even after one interaction (McKnight et al., 2002). All subscales showed good internal consistency. Moreover, we preferred the McKnight scale over the commonly used Jian et al. (Jian et al., 2000) scale, because of the content of the benevolence items. To illustrate, the McKnight scale includes an item such as "I believe that the robot would act in my best interest," which directly assesses the perceived benevolence of the robot. In contrast, the Jian framework includes items like "The system is deceptive" and "The system behaves in an underhanded manner," which assume that the opposite of benevolent is having malintent. It is a fallacy to promote the idea that if a robot's purpose or actions do not benefit or serve you, it is automatically malevolent or self-centred. When evaluating a robot's perceived benevolence and violations of that type of trustworthiness, we want to measure whether people feel like the robot acts in their best interest.

We expect violations of this kind to become more prevalent in Human-Robot Interaction (HRI) as machines gain greater autonomy and decision-making authority, increasingly making decisions impacting multiple stakeholders. The benevolence/purpose dimension in Jian's framework does not align with our perspective that it is inevitable for robots with increased decision authority to make decisions that do not always serve everyone's best interests. A robot may operate in the best interest of the collective rather than prioritizing a single individual, which should not be misconstrued as selfishness, deception, or underhanded behaviour.

A final reflection concerns the timing of the partners' communication about intentions. As artificial agents gain autonomy and decision authority, trust violations as the collateral harm of certain deliberate decisions (e.g., the trade-off-explanation condition) seem an inevitable part of the future. Something to bear in mind however is that in our experiment, the partners in the trade-off-explanation condition reveal only halfway into the mission that the participants' safety is not their top priority in their decision-making process. Holding back information could be perceived as a form of dishonesty and deception (Arkin et al., 2012). In terms of team performance and transparency, it is crucial for team members

(i.e., both human and non-human) to actively communicate about their actual intentions and current observations about the environment, in order to build shared situational awareness (April Rose Panganiban et al., 2020). It is plausible that the stagnation of perceived integrity in response to the trade-off-explanation is partly caused by the lack of transparency and mutual understanding.

While there may be instances where deception is deemed necessary to achieve a goal that benefits the entire team, trust is nearly always compromised when deception leads to negative outcomes (Hancock, Billings, & Schaefer, 2011). In order to make accurate judgments of trust, intentions towards a certain goal are ideally communicated beforehand (Hou et al., 2021; Schaefer et al., 2017). This is an important issue for future research. We expect that informing participants about the priorities of the partners in the trade-off-explanation condition upfront will influence their development of trust in all phases, including their initial response to the explosion. Future research should be undertaken to investigate how trust develops when (conflicting) goals and preferences are communicated prior to the task, and whether deliberate decisions will then lead to less severe trust violations. In one of our earlier studies, we found that communicating the (un)certainty of an advice in terms of performance (e.g., “I detect danger with 80% certainty”) generally led to higher levels of trust and to a less severe decline in trust in response to an incorrect advice (Kox, Siegling, et al., 2022). Being transparent about the agent’s intentions, goals and preferences upfront could have a similar effect on trust.

Implications

The research to date on trust violations and trust repair in HRI has tended to focus on trust violations due to error rather than deliberate choice. Some recent studies have started to investigate the latter, for example by studying the effects of a robot breaking promises (Alarcon et al., 2023; Sebo et al., 2019), acting out of self-interest (Alarcon et al., 2020, 2023) or deviating from a planned path (Lyons et al., 2023). This study’s originality lies in its exploration of the development of perceived trustworthiness when trust violations result from deliberate, comprehensible, yet impactful decisions. It does so within a task environment and corresponding scenario designed to simulate domain-specific interactions. Significant technical effort has been made to implement a graphically realistic, interactive simulation game for the purpose of this research. Realistic scenarios, which aim to mirror actual events and realistic trust violations rather than game-like simplifications, are crucial for creating nuance and enhancing the ecological validity of experiments. Such scenarios and task environments enable us to investigate different types of trust violations, beyond those caused by poor performance, in a realistic manner.

This research opens up a broader societal conversation about the role and decision authority we want robots and other AI agents to have. Our scenarios are based on hypothetical but realistic situations in which robots have the authority to harm people (and their trust). With this, we can not only study how people respond to these situations,

but it also forces us to think about the desirability of such future scenarios and whether we want these hypothetical situations to become reality. Additionally, it is important to stay grounded in reality, because the alternative (selfishness or malintent) fosters and perpetuates incorrect beliefs (such as the idea of evil robots taking over the world, instead of robots that are not benevolent to an individual because they are programmed to prioritize the benefit of the majority). Firstly, these examples foster the false attribution of intent to robots and feed into the misconception that robots will gain or possess self-interest or be programmed to pursue selfish or malicious goals. Secondly, this view diverts attention from the real threats and implications of value or priority misalignment, as well as unexpected or incomprehensible robot behaviour.

While performance is still an important determinant of human-robot trust (Chhogyal et al., 2019; Correia et al., 2018; Hancock, Billings, Schaefer, et al., 2011a), this study strengthens the idea that aspects such as preferences, personal relations and moral aspects become equally important (Malle & Ullman, 2021; Matthews, Panganiban, et al., 2021). However, we concur with the notion of Alarcon et al. (Alarcon et al., 2021) and Lee & See (J. D. Lee & See, 2004) that, as a robot lacks intentionality, the purpose or intentionality of a robot in fact embodies the intentions of its designers. Therefore perceptions of benevolence and integrity might not be valid when evaluating interactions with a robot, as people might differentially attribute intentionality to the robot itself or to its designer. Further research is needed to evaluate these potential differences in perception and their effects of HRI. While intent is a highly debated concept in relation to artificial agents and the terms benevolence and integrity are deemed inappropriate by some scholars, the observation that an artificial agent is no longer automatically trustworthy when it is capable of completing a given task without making mistakes, is persistent (Malle & Ullman, 2021). Decisions by an artificial agent can be objectively correct in the sense that they adhere to the set of rules the agent operates by, but can nonetheless be subjectively questionable or unacceptable in a given context when those decision do not align with implicit rules.

The results of the current study emphasize the importance of distinguishing between different perceptions of trustworthiness. Our findings show that perceptions of ability, benevolence and integrity are differentially affected by different types of explanations regarding the intentionality behind a trust-violating advice. One of the questions that emerge from these results is what the implications will be for behavioural reliance. What will be the behavioural consequence of a situation where perceptions of ability have recovered, while perceptions of benevolence and integrity have not (yet)? Further research should be undertaken to investigate the behavioural consequences of this discrepancy in trusting beliefs.

Conclusion

Increasingly autonomous AI-based artificial agents are used in a wide variety of both military and civilian applications (Jessie Y.C. Chen & Schulte, 2021). As artificial agents

enter more complicated operational situations and gain the ability to self-select courses of action in an ever-changing world, they will encounter situations that they have not seen before. Consequently, artificial agents will encounter dilemmas where they must navigate tradeoffs among conflicting goals or competing human values. As a result, their decisions cannot always be beneficial for everyone. Still we want to enable and maintain appropriate levels of trust, as this is key to successful and effective long-term human-robot collaboration (Hou et al., 2021). Traditionally, HRI focused on performance measures such as task-related strengths and limitations, reliability and predictability of a robot (Chhogyal et al., 2019; Malle & Ullman, 2021; Marsh, 1994). Today, human operators should increasingly be aware of a robot's higher-level values, preferences and goals (Chhogyal et al., 2019). At the same time robots in collaborative settings should gain the interactive ability to resolve competing goals through social processes (Chiou & Lee, 2023). Knowing your partner's intentions, goals and preferences is crucial for calibrated trust and successful team performance (Hou et al., 2021). As technology advances, it is vital to critically assess the psychosocial consequences of the growing responsibility that we give artificial agents in increasingly complex decision-making processes (Awad et al., 2018) and, as a part of that, to understand if and how trust can be recovered after intentional or unintentional trust violations (Taenyun Kim & Song, 2021).

Chapter 6

Discussion

Trust is fundamental requirement for successful collaboration, as it enables individuals to depend on each other's contributions to collectively complete tasks and achieve objectives (Fahim, Khan, Jensen, Albayram, & Coman, 2021; Parasuraman & Riley, 1997). To minimize the risks and maximize the benefits of the collaboration, or to make it safe and effective, people must be able to determine when it is appropriate to rely on AI and when it is necessary to intervene (J. D. Lee & See, 2004). Achieving this requires a balanced relationship between the perceived trustworthiness of an AI agent and its actual trustworthiness, known as calibrated trust (J. D. Lee & See, 2004; Lewis et al., 2018). Finding this balance is challenging because it demands an ongoing evaluation of an AI agent's trustworthiness and reliability (Hoffman, 2017). To facilitate this calibration process, we must first understand how trust develops, breaks down, and recovers (de Visser, Pak, et al., 2017).

Trust is fragile and easily damaged. For example, people may lose trust in an AI agent they collaborate with due to errors or poor performance, which can arise from software bugs or hardware malfunctions. These type of trust violations are the primary focus of current HRI trust repair literature. However, an AI agent's trustworthiness is not solely defined by its ability to perform a task, but also by how it performs them, and the underlying goals and values it pursues (Lubars & Tan, 2019; Malle & Ullman, 2021). As a result, trust can be violated by factors beyond performance issues, such as when an AI agent behaves unpredictably and cannot explain the reasoning behind its actions, or when its priorities differ from those of the people it interacts with. As AI agents evolve from isolated tools to more autonomous social actors with increased decision authority in complex, social environments, the risk of these types of trust violations increases. This dissertation aims to broaden our perspective and deepen our understanding of H-AI trust violations by examining how the nature of a trust violation affects the development of H-AI trust.

Additionally, we investigated the impact of different trust-repair mechanisms on the development of H-AI trust, evaluating how, and to what extent, these mechanisms can preserve H-AI trust in the face of inevitable trust violations. We assessed both preventative and reactive strategies and evaluated the effect of informational as well as affective content. Preventative measures focus on proactively addressing potential trust issues before they arise, such as communicating uncertainty (i.e., "danger detected with 80% certainty") as a means to manage expectations. Reactive strategies, on the other hand, address trust violations after they have occurred, for example by expressing regret or providing explanations for anomalous behaviour. Certain strategies are considered informational, focusing on improving transparency or interpretability by communicating uncertainty or providing explanations. In contrast, other strategies are considered affective, as they aim to restore positive feelings of trust through actions such as expressing regret. By examining a range of trust repair mechanisms and trust violations, this dissertation contributes to our knowledge on maintaining H-AI trust as a key part of the trust calibration process.

Key findings

A multidimensional perspective on trust violation and repair

We explored the effect of trust violations due to 1) poor performance, 2) unexpected behaviour in combination with an AI agent's that does not explain its behaviour and 3) priority misalignment. When evaluating the different perceptions of trustworthiness, i.e., ability, benevolence and integrity, in response to the latter two events, we found that these dimensions can be affected differentially and run parallel and unsynchronized. For example, in Chapter 5, we observed that people lost trust in an AI agent's abilities while still trusting its benevolence, or vice versa. Alternatively, in Chapter 4, the AI agent that did not explain why it suddenly deviated from the original plan was deemed less trustworthy in terms of ability and integrity than its transparent counterpart, while benevolence was more or less unaffected.

Moreover, we found that the nature of a trust violation had significant implications for the ease of repair. In Chapter 5, when an AI agent explained that an encounter with a hazard was due to an error rather than a deliberate choice, all perceptions of trustworthiness successfully recovered. However, when the AI agent explained that the trust-violating event was the result of weighing options and making a deliberate choice to the disadvantage of its human partner, only perceptions of ability were restored, while perceptions of integrity plateaued and perceptions of benevolence dropped even further. These findings support the growing consensus that H-AI trust is a multidimensional concept, highlighting the importance of distinguishing between different dimensions of perceived trustworthiness within HRI.

Informational strategies

We also found that informational strategies to mitigate the negative impact of trust violations, both preventative and reactive, generally led to higher and more stable levels of trust compared to baseline conditions where no information was provided. In Chapter 3, we found a robust effect of uncertainty communication on the perceived trustworthiness of the AI agent. In both studies (i.e., civilian and military sample) we found that uncertainty communication in the advice of the AI agent (i.e., "danger detected with x% certainty" rather than simply "danger detected") generally resulted in higher levels of trust (Kox, Siegling, et al., 2022). In the civilian study, uncertainty communication even dampened the decline in trust following the AI agent's error, meaning that advice that included an uncertainty measure led to a less severe decrease in trust following a trust violation compared to an advice that did not include a notion of uncertainty (Kox, Siegling, et al., 2022).

In Chapter 4, we observed that the perceived trustworthiness of the AI agent was considerably higher when the robot provided regular updates about its actions throughout the task. While participants in the high transparency condition maintained a stable level of

trust during the robot's deviation, participants in the low transparency condition showed a significant decline in perceived trustworthiness in response to the robot's sudden adaptation to the plan. In other words, the explanation prevented a trust violation. This confirms earlier research that showed that transparency can have a buffering effect on perceived trustworthiness, increasing resilience against the effects of unexpected behaviour or temporary malfunctioning (Hamacher et al., 2016; Kox, Siegling, et al., 2022; Kraus et al., 2020; Lyons et al., 2023; Tenhundfeld et al., 2020).

These findings suggest that an AI agent does not have to be 100% reliable to maintain a trusting relationship, but that communication is essential for maintaining trust. Trust can fluctuate following errors or unexpected deviations from the plan, but it can recover when additional information about why the system failed is provided (Lyons, 2013). The AI agent should be able to provide explanations regarding its decisions, recommendations, and actions. Any changes to the AI agent's functionality or its planned behaviours should be communicated clearly to human team members to maintain trust.

Affective strategies

We found that apologies including an expression of regret can be effective in repairing H-AI trust. In Chapter 2, we saw that after an incorrect advice from the AI agent caused a trust violation, trust only recovered significantly when the provided apology included an expression of regret (i.e., "I am sorry"). This effect was stronger when an explanation was added. In Chapter 3, we again found that an apology containing both an expression of regret and an explanation was effective in repairing a trust violation due to poor performance in a civilian sample. Although expressing regret is typically perceived as a human-like quality, these results suggest that saying sorry can also make a difference in rebuilding trust when it comes from a non-human agent.

While we observed in Chapters 2 and 3 that an AI agent expressing regret effectively repaired trust with civilian participants, Chapter 3 also demonstrates that this approach did not yield the same results with a military sample. During a debriefing session with some military participants, it became evident that expressing regret is uncommon in the Dutch military context. Participants noted that acknowledging responsibility with phrases like "I was wrong" or "I misjudged the situation" is acceptable, but that saying "I am sorry" is extremely rare. This cultural difference possibly explains the varied responses to the AI agent's expression of regret. These findings emphasise once again that we cannot draw conclusions about "people" in general. Understanding individual differences and cultural norms and preferences can help AI adapt its behaviour to effectively engage with diverse teams. It is important to recognize that designing social systems is not a one-size-fits-all approach and will require continuous exploration and refinement.

Contributions

Contributions to theory

To some extent, the findings in this dissertation parallel the Computers-are-Social-Actors (CASA) paradigm (J.-E. R. Lee & Nass, 2010). We agree that understanding "the social principles that govern us and the social expectations we hold with respect to trust-building in the interpersonal settings" (J.-E. R. Lee & Nass, 2010) (p.11) is valuable, as some of these principles translate to the context of HRI, such as the effectiveness of apologizing. Furthermore, the ineffectiveness of apologizing with military personnel, where it is not a common social practice, aligns with the idea that people tend to find an AI agent more trustworthy when it exhibits behaviours similar to their own (J.-E. R. Lee & Nass, 2010). Additionally, the heavy decay of trust when the AI agent prioritized collective goals over the participants safety reflects the notion that we are more likely to trust entities, whether human or AI, that demonstrate caring behaviours. Just like we are more likely to trust people who make us feel cared for, we are more inclined to trust machines that show concern for our well-being (J.-E. R. Lee & Nass, 2010). Thus, we argue that the social and emotional dimensions of technology use cannot be ignored. Expecting people to be 'rational agents', driven by logic and utility in their interaction with non-human entities, neglects and undervalues the important role of the social and emotional strengths humans bring to the collaboration.

Yet, we believe that is not automatically appropriate to design robots to be social or human-like simply because people often respond socially to them, or because it increases trust, and that "successful robots utilize the distinctive features of machines" (Shneiderman, 2020) (p.113). While the findings indicate that it may be valuable to leverage humans' evolutionary social wiring to facilitate collaboration with other forms of intelligence by incorporating social cues into the design of AI agents, this should be done with caution and without trying to re-create what we already have (people) (Darling, 2021). AI agents can only meaningfully augment human decision-making and benefit society when they are applied to tasks in which they excel (Kox, van Riemsdijk, & Kerstholt, 2024), and when they effectively communicate the what, how and why behind their decisions to human operators while performing them. Building on the parallels with the CASA paradigm, this dissertation advances theoretical discussions by showing that informational strategies, such as explanations and communicating uncertainty, are just as important for maintaining trust in AI agents.

Finally, this dissertation expands on the concept of a 'trust lifecycle' and deepens our understanding of the temporal dynamics of trust,. Trust is dynamic and typically evolves over a series of interactions (Baker et al., 2018; Hou et al., 2021). As people interact with and learn more about an AI agent, they continuously update their perceptions, judgments, and trust, particularly in response to norm violations (E. Phillips et al., 2023). Most studies, however, offer a static perspective on trust, focusing on momentary states of

trust rather than its developmental process. Unlike cross-sectional studies, our research uses repeated measurements of trust over time, and sheds light on how trust develops, deteriorates and recovers in response to various factors. Our goal was to capture these dynamic changes in trust, as understanding them is crucial for studying norm conflict, resolution, and mitigation, and for maintaining H-AI trust.

Contributions to method

A key methodological contribution of this dissertation comes from the technical effort invested in the development of high-fidelity military scenarios and graphically detailed virtual task environments. Both our desktop-based and VR simulations were designed to mirror authentic HRI situations with a high level of detail and realism, providing a practical understanding of potential trust relations in an operational military context. This approach offers three major advantages. First, the scenarios facilitated our temporal perspective on trust in the AI agent, allowing us to observe how it evolved as different events unfolded.

Secondly, by incorporating emotion-evoking, startling events in highly detailed graphic environments portraying realistic scenarios, we have strived to simulate a sense of threat and risk that was likely more effective at triggering implicit trust decisions compared to traditional cognitive paradigms. Trust, by definition, is only relevant in situations characterized by risk, uncertainty and vulnerability (Li et al., 2019). Studying trust repair requires violating trust and allowing people to experience the risk they take and the vulnerability they accept. Yet, it can be challenging to create experimental scenarios that induce feelings of vulnerability and risk without compromising participants' physical and psychological safety (Baker et al., 2018). Our goal was to create high-fidelity experiences that would elicit feelings rather than relying solely on a more cognitive, incentive-based approach. Matthews et al. (2018) have suggested that we need more complex, realistic threat-detection scenarios to truly understand to what extent people are willing to trust an AI agent in circumstances characterized by threat (Matthews et al., 2018). By exploring innovative ways of simulating risk that elicit emotional responses within ethical boundaries, instead of relying solely on gamification elements or cognitive incentives, we can lay a stronger foundation for future empirical studies in more complex, real-world settings.

Lastly, the detailed virtual military scenarios, designed to closely mimic actual conditions, allowed us to portray different types of HRI trust violations in a realistic manner. Most HRI studies that explore trust violations beyond performance-related issues often rely on simplified fictional game-like scenarios, where AI agents exhibit human-like behaviours, such as pursuing personal gain (i.e., money) or lying, to violate perceptions of benevolence or integrity-based trust (Alarcon et al., 2023; Sebo et al., 2019). While these clear and familiar examples of self-serving behaviour and deception are valuable for developing and testing theory, they tend to be overly anthropomorphic and somewhat unrealistic in HRI contexts. After all, AI agents operate based on programming, predefined objectives, and operational goals, rather than human-like motivations such as

the financial gain or the intent to deceive. Nevertheless, they can still violate perceptions of benevolence and integrity. However, realistic examples of such violations remain significantly underrepresented in current research.

A benevolent partner, by definition, is genuinely concerned with your well-being and is motivated to pursue joint gain (Bhagat & Steers, 2009). In other words, a benevolence-based trust violation occurs when a partner fails to act in your best interest, disregards your needs, or shows a lack of concern for your welfare (Mayer & Davis, 1999). However, this does not necessarily imply *self*-interest, where the partner prioritizes its own interest over others (Jessup et al., 2020). For example, a partner may prioritize the collective interests of the team over the safety of an individual member, which present a more realistic scenario for future HRI operations (Jorge et al., 2022). Therefore, trust violations based on benevolence and integrity based in HRI studies do not have to be unrealistic.

Our virtual task environments preserve the authenticity and practical relevance of operational reality while allowing the incorporation of hypothetical elements (e.g., advanced robotics that do not yet exist) and maintaining the experimental control and immediate feedback that would be lost in field experiments (Petty & Cacioppo, 1996).

Contributions to society

Our research opens up a broader societal conversation about the role and the decision authority we want AI agents to have and "what we want a future AI-enabled society to look like" (Winkler, 2024) (p.1). Our scenarios are based on hypothetical but realistic situations in which AI agents have the authority to cause harm to people (and their trust) by their decisions and recommendations. We focused on plausible AI agent behaviours in the military domain, rather than using current interpersonal examples and attributing human-like motivations (e.g., greed, deception) to AI agents. For example, it is conceivable that AI agents may be programmed to follow a utilitarian approach, prioritizing team goals over individual safety for the greater good. Similarly, it is plausible that AI agents will be allowed to make autonomous decisions while pursuing a delegated goal and designed to balance the frequency of updates, all to reduce the human's cognitive load. Even though this can sometimes result in miscommunication or miscomprehension. These examples of potential trust violations more closely reflect how future AI agents may operate in practice and can better inform policy and ethical guidelines.

There are numerous operational HRI scenarios conceivable where a well-considered trade-off decision can cause harm while still pursuing a largely positive outcome. Instances of misalignment between human and AI agent values or priorities are already occurring to some extent. For example, in some regions, autonomous security robots are being deployed in public spaces for security tasks (Stephens, 2023). These robots may patrol parking lots with the aim to prevent vehicle break-ins by detecting environmental anomalies and suspicious behaviour (Knightscope, 2023). This design reflects a focus on overall safety, which may come at the expense of individual privacy. Consequently, these robot

might encroach on people's personal space and sense of privacy, leading to mistrust not only of the robots themselves but also of their developers and deployers. The root of this mistrust lies in the robot's purpose rather than its performance, as it is designed to uphold a value (security) that inherently conflicts with another (freedom). Specifically, the robot operates within the trade-off between security and freedom: increasing security measures can restrict personal freedoms, while maximizing freedom might reduce security.

This examples highlights the importance of carefully considering the implications of designing and deploying AI agents that may cause discomfort or harm to people. Beyond advancing based on technical feasibility or cost efficiency, we must prioritize the broader societal impact. As increasingly sophisticated, trust-violating AI agents raise significant ethical concerns, this dissertation aims to contribute to the ongoing debate on how we should design and implement AI agents in a way that helps society and improves well-being.

In practice, it is inevitable that in future scenarios, an AI agent's purpose and priorities might not be able to serve everyone equally, albeit with the best intentions. As AI agents gain in autonomy, it becomes increasingly important to develop theories around these situations and to critically consider the implications of realistic scenarios where robots may make choices that could potentially harm, hurt, or at least disappoint humans. By developing more plausible scenarios, we can not only study how different people respond to them, but also critically reflect on the desirability of such future scenarios and whether we want them to become reality.

Limitations

The behavioural dimension of trust

In our studies, we aimed to create experiments that would elicit emotional responses through immersive storylines and task environments featuring emotion-evoking and startling events, rather than relying solely on a cognitive, incentive-based approach (e.g., gamification elements such as lives, ticking clocks, or performance-based monetary bonuses) to simulate risk and manipulate trust. However, in the storylines, participants were not given the opportunity to make decisions based on those emotions. They were simply instructed to walk from point A to point B while encountering various events and completing self-report trust questionnaires. They had no option to disobey or deviate from the AI agent's recommendation if they lacked trust, nor did their behaviour have any consequences. This approach, which prioritized experimental control over behavioural freedom, may have limited our ability to measure trust in a more behavioural sense as a readiness for risk-taking and a willingness to be vulnerable.

Given this limitation, we were eager to explore the possibilities of incorporating more behavioural measures in the VR studies we conducted, which are not included in this dissertation (Kox, van Riemsdijk, de Vries, et al., 2024a, 2024b). In an initial study using our

VR maze, we sought to investigate the relationship between trust and compliance by giving participants the behavioural freedom to either follow or disregard the AI agent's advice by choosing an alternative path than the one suggested by the drone. Simultaneously, we assessed self-reported trust to determine whether and how participants' perceptions of trustworthiness aligned with their behavioural choices (Kox, van Riemsdijk, de Vries, et al., 2024a). Unfortunately, due to technical difficulties and unforeseen low compliance rates, we were unable to gather sufficient valid data to gain valuable insights into this relationship. However, we view the link between trust and reliance or compliance as an important area for future research. Researching this link helps identify the thresholds at which trust transitions into actionable reliance or compliance, providing insights into when and why humans choose to act based on an AI agent's input.

The findings presented in Chapter 5 show that different dimension of perceived trustworthiness are differentially affected by trust violations due to error or choice. One of the key questions that emerges from the finding is what this means for behavioural reliance. For instance, what will be the behavioural consequence when perceptions of ability have recovered, but perceptions of benevolence and integrity have not yet been restored? Which dimension of perceived trustworthiness is most predictive of behaviour, and does this vary depending on context and individual differences? Further research is needed to investigate the behavioural consequences of this discrepancy in trust beliefs.

Moreover, placing a greater focus on behavioural measures potentially related to trust could significantly contribute to the field of HRI by providing more quantitative metrics to assess trust. The VR environments we developed, including both the virtual maze and the house-search task, are ideally suited for such purposes (Kox, Barnhoorn, et al., 2022; Kox, van Riemsdijk, de Vries, et al., 2024a, 2024b). For instance, in the virtual maze, the reaction time in the decision-making task (i.e., choosing between the red or blue door after the drone's recommendation) can be assessed, while in the house-search scenario, walking speed can be measured, both of which may indicate hesitation. Additionally, eye movements tracked through the VR headset in both scenarios could reveal excessive monitoring. Both hesitation and excessive monitoring might point to lowered trust. Incorporating these objective measures into HRI trust research would be worthwhile, as it can increase the accuracy and reliability of trust measures.

In summary, our ability to capture trust as a readiness for risk-taking and a willingness to be vulnerable was limited by the constraints of our desktop paradigm and the unsuccessful implementation of behavioural measures in the VR studies, leaving a critical dimension of trust unmeasured in this work. We strongly encourage further exploration of these behavioural dimensions in subsequent studies.

Military scenarios and civilian participants

Like the vast majority of human factors research on H-AI trust, we predominantly used university students as participants in task situations that were new to them (Hoffman,

2017). This approach presents two major limitations in our studies. First, it could be argued that having homogenous samples comprising primarily young adults with an academic background limits the generalizability of our findings. However, the primary aim of these controlled experiments is to uncover causal mechanism, in order to better understand how people make judgements in response to certain stimuli (Kadres, 1996). These causal mechanisms often hold true across different groups unless there is a strong reason to believe that demographic factors would significantly alter the effect. For example, Chapter 3 showed that apologizing was specifically ineffective with military participants, indicating that cultural differences played a significant role in shaping responses to trust repair strategies. At the same time, the positive effect of communicating uncertainty remained consistent across both civilian and military participants, showing that certain mechanisms can be robust across varied demographics. While we should always be cautious with generalizing findings beyond the immediate sample, controlled studies that reveal these mechanisms provide a foundation that can be expanded upon in more diverse settings.

Secondly, presenting a range of different military scenarios to civilians raises questions about the applicability of the results to actual military contexts. It is uncertain whether results obtained from studies using military scenarios with civilian participants will generalize to real military scenarios. Presenting civilians with military scenarios carries the risk of misinterpretation, as soldiers and civilians might perceive these situations in distinct ways. Civilian participants, especially those familiar with video games, may approach military scenarios with a mindset shaped by gaming experiences, potentially leading to differing risk assessments. Video games often prioritize entertainment over realism, and players may take risks or make decisions that do not reflect real-world consequences. It is probable that participants would have made different choices if their lives were actually in danger. While this limitation is inherent to controlled experiments, it highlights the importance of follow-up studies to test whether our findings hold up in actual operational settings and across different populations.

A lack of ground truth

A final limitation to my studies is that we cannot draw conclusion on whether trust is actually better calibrated thanks to certain strategies. We tried to map changes and fragments of the dynamics of trust, but, because my studies did not include a ground truth (i.e., a reference point against which the accuracy of measurements are assessed, such as a fixed reliability rate), the results cannot confirm whether trust was properly calibrated.

Despite emphasizing the importance of calibrated trust and striving to repair trust to an optimal rather than maximum level, we cannot ascertain whether participants' trust in the AI agents during our experiments was properly calibrated. Instead, we aimed to uncover causal mechanism and measured how certain events caused a decline in trust and whether specific strategies (implemented before or after the violation) could minimize the immediate impact on trust development. We are aware that, in many cases,

decreases in trust are a logical and functional adaptive response to perceiving reduced performance or unexpected behaviour, and actually contribute to trust calibration. Yet, our research contributed to a toolbox for addressing undertrust. We have demonstrated the effectiveness of certain strategies in increasing trust under specific circumstances. Future research should focus on identifying instances and causes of undertrust, and on then determining which strategies are appropriate for those particular situations and circumstances.

Recommendations for future research

Several questions remain unanswered. Future studies should investigate more specifically which repair strategy is most effective for each type of violation, by a systematic comparison between different repair strategies and violation types (de Visser et al., 2018). For instance, previous HRI research has proposed certain optimal combinations, such as apologies for ability-based trust violations and denial for integrity-based trust violations (Sebo et al., 2019).

Additionally, it is inevitable that certain trust-repair mechanism will work for some people, but not others (de Visser et al., 2018). There are many inter-individual differences, including affective (i.e. moods, feelings, etc.) and dispositional (i.e. personality traits), that can influence how people respond to trust violating events and repair strategies and navigate social interactions and collaboration in general, such as cultural differences (e.g., individualism versus collectivism), gender differences (Macko, 2020; Schumann & Ross, 2010) or differences in cognitive abilities (Ku & Pak, 2023; Rovira et al., 2017). Future research should focus on how such inter-individual differences impact trust repair, in order to develop more personalized and effective strategies.

Follow-up research should also assess the long-term effects of trust repair strategies. For example, more work is needed to determine whether the beneficial effects of an apology will last when the same apology is offered repeatedly. An apology is a way of taking responsibility for one's behaviour, implying a commitment to improve and to avoiding similar mistakes in the future. However, when the machine does not change its (erratic or undesirable) behaviour that it expressed regret for, an apology is deemed ineffective (de Visser et al., 2018).

Moreover, the monotony and uniformity of communication styles often seen in AI agents may lead to irritation when the same message is consistently delivered in the same format and tone. If so, it is plausible that the potentially beneficial effects of an apology will not be robust and soon be perceived as a gimmick. In this context, the emergence of Large Language Models (LLMs) holds significant potential as intuitive human-machine interfaces capable of mitigating the "robot-like" manner of communication often associated with AI agents. However, their sophisticated communication capabilities also raise concerns about the potential for overreliance or overtrust. Their ability to generate highly coherent

answers “can fool us into thinking that they understand more than they do” (B. X. Chen, 2022). People might not always be aware of the risks associated with the use of generative AI models, such as LLMs (Kox & Beretta, 2024).

Implications and practical implementation

Carefully consider the consequences of design choices

One of our findings is that expressing regret can be an effective strategy for repairing trust, but that its effectiveness varies depending on the audience. This contributes to an ongoing debate about the appropriateness of humanizing AI-interaction and the extent to which anthropomorphism should be implemented (J. Johnson, 2024; Shneiderman, 2020, 2021). Some researchers argue against referring to future technologies as teammates, partners, or collaborators as they are more likely to function as advanced tools (Shneiderman, 2021), and this type of terminology and metaphors can create “a false equivalence between human and machine intelligence” (J. Johnson, 2024) (p. 77). It is not automatically appropriate to design robots to be social or human-like simply because people often respond socially to them (Shneiderman, 2020).

In practice, the appropriateness of using human-like cues, such as expressions of regret, depends heavily on the context. Our finding that expressing regret is not as effective with military personnel as it is with civilians indicates that it is important to consider the social customs and cultural norms of the target population when designing collaborative AI agents. When AI agents are expected to interact with humans in high-stakes environments, it is crucial that they are designed to seamlessly integrate, using the same terminology and communication styles as the rest of the team to facilitate smoother interactions. As Julie Carpenter emphasizes in her study on how U.S. Military Explosive Ordnance Disposal personnel integrate robotic tools into their work and develop emotional bonds: “In order for human-robot teams to be effective, research is needed into the whole system that the individual team members are a part of, and how these factors ultimately shape the interactions at micro levels” (Carpenter, 2013) (p.2). While ‘human-like’ behavioural cues may facilitate social interaction by creating a sense of familiarity, it is important to recognize that not all humans behave in the same way.

Beyond the diversity of target audiences, it is crucial to recognize the wide range of AI applications, rather than treating AI as a single, homogenous concept (Jermutus et al., 2022). For example, in social robotics, human-like cues in AI agents might be beneficial, while in professional settings where safety and calibrated trust are crucial (e.g., intensive care, military operations), such features may be less appropriate. Anthropomorphic design should be guided by a clear understanding of the potential effects on user interactions and should be approached with careful and strategic consideration, as these features can lead to both positive and negative outcomes (Carpenter, 2013; Disalvo et al., 2002;

J. Johnson, 2024; Taenyun Kim & Song, 2021). By focusing on context, we can have a more meaningful discussion about whether using human-like cues is appropriate.

Ultimately, it is crucial to carefully consider the consequences of design choices and the ethical implications for AI agent behaviour. With this in mind, although some studies have explored the effects of denial in human-robot interactions (e.g., Kohn et al., 2018; Sebo et al., 2019; Zhang, Lee, Kim, et al., 2023), we deliberately chose not to include it as a trust repair strategy. Denial, often used by humans to downplay errors or dismiss concerns, conflicts with our goal of establishing sustainable, trust-based relationships with AI agents. AI agents should not be programmed or allowed to lie, even if it might yield a desirable outcome in the short term. Instead, they should be designed with a clear ethical framework that prioritizes transparency, calibrated trust, and the long-term effectiveness of human-AI collaboration.

It is our responsibility as researchers not only to investigate what strategies or design features effectively yield desirable outcomes, such as calibrated trust, but also to determine what we consider acceptable and ethical behaviour for AI agents in achieving those outcomes. The end does not automatically justify the means.

Invest in people

Our results suggest that explanations can be effective in repairing trust. However, explainable AI is a complex field of research that proves how challenging it is to make a machine explain its reasoning in a way that aligns with human thinking. Additionally, it is difficult for an AI agent to recognize when it fails to meet human expectations. If it could, it likely would not make such errors. Even with explicit human feedback, providing satisfactory explanations for its actions will remain a significant challenge. The increasingly complex algorithms used in AI agents often result in a “black-box” effect, where the internal workings are not easily understood or accessible, even to developers, making it difficult to understand or explain how the AI reaches its conclusions. Robust explanation capabilities may still be far from being fully realized.

In the meantime, it seems wise to invest in people’s capacity to know what AI is good at, where its limitations lie, and in which contexts trust may be risky. For instance, this can be achieved by improving people’s AI-literacy, which is a broad set of skills that enable individuals to recognize everyday applications of AI, know the basic functions of AI and understand how to use AI effectively in daily life (Ng et al., 2021). Exploring research directions that focus on improving AI-literacy is crucial. This includes understanding what AI-literacy encompasses, including the key knowledge and skills people need to effectively navigate and resolve conflicts arising from the inherent differences between humans and AI agents. For example, AI-literacy should focus on critical thinking about AI, helping people recognize red flags, potential areas of concern, and when human oversight or intervention is necessary. Equally important is determining which mental models of AI people should develop or employ, and identifying metaphors

that either aid or hinder appropriate understanding of AI (Maas, 2023). Addressing these questions will not only enhance human-AI collaboration but also pave the way for more informed and safe use of AI in everyday life.

However, if even developers struggle to explain AI-systems due to the "black-box" nature of many algorithms, it seems contradictory to expect the general public to develop AI-literacy to a level that would enable them to confidently interact with these systems. This highlights a clear responsibility for developers and the AI industry to ensure that AI-systems are trustworthy and reliable. Efforts should be directed toward creating transparent systems where feasible and establishing safeguards and protocols that clarify the strengths and limitations of AI, particularly when enhancing explanation capabilities is not possible.

Approach AI as a socio-technical system

This thesis explored various causes of trust violations in AI agents. Moving beyond errors as the sole cause of trust breakdowns requires examining the entire development process and questioning the desirability of such scenarios. Trust violations arising from an AI agent's technical shortcomings (e.g., software bugs, hardware malfunctions) are largely unpredictable. However, violations resulting from unexpected or incomprehensible behaviour, often due to the agent's incapacity to explain itself, or conflicting priorities are more likely the result of policy and design choices rather than the AI agent's technical sophistication. These types of violations are more predictable and thus, potentially preventable.

This raises questions about how much harm, disappointment, or confusion we are willing to tolerate from AI agents, and what benefits we gain in return. As AI agents are designed, built and programmed by humans, all trust violations are ultimately human-made and traceable to decisions in the development process. Each violation can, at least theoretically, be attributed a different group: for instance, software bugs or hardware failures are linked to engineers (e.g., mechanical, electrical, software), interface transparency issues are the responsibility of designers and UX specialists, and value misalignments can be traced back to executive leadership, who set strategic priorities, and product managers, who ensure the product aligns with user needs and company goals.

In other words, AI systems should be approached as sociotechnical systems, where the trustworthiness of the technology is as much a product of the people designing, developing, deploying and using it as of the system itself (Duenser & Douglas, 2023). In order to create trustworthy AI, the creators themselves must be worthy of trust (Cameron et al., 2023). Building trustworthy AI agents requires a systematic, multidisciplinary approach. *"Just as it takes a village to raise a child, the governing of AI needs to be a multidisciplinary village so that we can raise AIs that are productive, valued contributors to society."* (Winkler, 2024)

Conclusion

The prospect of more autonomous AI agents gaining decision-making authority in our physical and virtual worlds introduces a range of new questions, ethical dilemmas, opportunities for advancement, and potential risks. My findings show that an AI agent's trustworthiness of is no longer determined solely by what it can do, but also by how and why it does so. Therefore, we emphasize the importance of a multidisciplinary, human-centred approach that aims to contribute to the integration of AI agents into our existing social structures, while prioritizing the needs, behaviours, and experiences of the people who will interact with them, rather than advancing solely based on technical feasibility.

Dankwoord

Ik heb dit proefschrift alleen kunnen schrijven met de hulp en steun van mijn begeleiders, collega's, vrienden en familie. Ik wil allereerst mijn promotor en copromotoren bedanken voor hun intensieve begeleiding. Zonder hun betrokkenheid en steun was dit werk niet mogelijk geweest. Ik ben heel dankbaar voor hun tijd, het geduldige meedenken, hun aanmoedigingen en de vele inhoudelijke discussies die we tijdens onze tweewekelijke meetings met zijn vieren hebben gevoerd en die mijn werk zo hebben verrijkt.

Ik wil mijn promotor José Kerstholt bedanken voor haar wijsheid, onuitputtelijke optimisme en pragmatisme. Altijd relaxed en goedlachs en tegelijkertijd altijd hard aan het werk en bereid om te helpen en mee te denken. Met José en haar relativeringsvermogen aan je kant kun je elke negatieve review en afwijzing aan.

Mijn co-promotor Birna van Riemsdijk wil ik bedanken voor haar toewijding, nieuwsgierigheid en bemoedigende woorden. Haar analytisch vermogen, vragen en eigen perspectief hebben mij vaak geholpen de vanzelfsprekendheid van bepaalde zaken in twijfel te trekken, wat hard nodig is als je zo diep in een onderwerp zit.

Tot slot wil ik mijn co-promotor Peter de Vries bedanken voor zijn behulpzaamheid, interesse en ontspannen houding. Zijn constructieve en zorgvuldige benadering heeft mijn werk sterk verbeterd. Ik heb genoten van onze boeiende discussies en ben dankbaar voor zijn warme persoonlijkheid die me altijd heel welkom deed voelen in Enschede bij PCRS.

Ik dank de leden van mijn leescommissie, prof. dr. ir. Mieke Boon, prof. dr. ir. Dirk Heylen, dr. Jurriaan van Diggelen, prof. dr. Tibor Bosse en prof. dr. Wijnand IJsselsteijn voor het lezen en beoordelen van dit proefschrift.

I'd like to extend my gratitude to the lead developer of the BMS lab Lucy Rábago Mayer for her invaluable support during our time working together on the VR environments, amidst the challenges of COVID. Lucia was a lifesaver during those days in the lab and I thank her for her patience and good company.

To my fellow PhDs and PCRS colleagues, I want to express my gratitude for making me feel so welcome during my time at the University of Twente. I'm thankful for the support you've shown me, and I appreciate the effort you put into making me feel included, despite the distance and my frequent absences. Thank you for being such a great group of colleagues.

Veel dank aan al mijn TNO collega's die de afgelopen jaren niet alleen inhoudelijk een belangrijke bijdrage hebben geleverd aan dit werk, maar ook voor het gezelschap en de afwisseling hebben gezorgd die ik nodig had om productief en gefocust te blijven. Hun emotionele steun, in de vorm van het tolereren van mijn vele gezucht en gevloek tijdens tafelvoetbal, zal ik ook niet vergeten.

I would like to extend special thanks to all my students for their help in collecting the data, without which this research would not have been possible. I really enjoyed supervising, as it was rewarding to see students grow, learn and develop an appreciation for research, much like my own. I've learned as much from them as I hope they have from me.

Dan wil ik mijn paranimfen Andrea en Naomi bedanken; twee bijzondere psychologen met een gave voor het stellen van vragen en het onthouden van details over mijn leven die ik zelf vergeet. Hun betrokkenheid en onze gesprekken zijn me heel dierbaar en hebben me de afgelopen jaren ontzettend geholpen, en ik ben heel dankbaar dat zij straks naast mij staan.

Dit werk was ook niet mogelijk zonder mijn inspirerende vriendinnen, waarmee ik al jaren lang lief en leed deel en die altijd een stabiele factor zijn geweest in de afgelopen jaren waarin zoveel is veranderd. Ik kan me geen betere vriendinnen wensen.

Hetzelfde geldt voor mijn Hellas maatjes die de afgelopen jaren onmisbaar waren voor mijn mentale gezondheid. De vele uren die we samen hebben doorgebracht op de fiets, op de baan en in het zwembad hebben me geholpen om mijn gedachten te verzetten en te ontspannen na drukke dagen. Het gezelschap van deze warme en inspirerende gemeenschap, vol lieve en slimme mensen, heeft me veel energie en motivatie gegeven.

Ook wil ik mijn huisgenoten bedanken die er waren op alle dagen, die vol parels en die vol stenen en allen die daar tussenin. Met jullie thuis, tijdens mijn hersenschuddingen en de maanden samen binnen tijdens COVID, betekende steun, stabiliteit en gezelligheid en het was de allerbeste en fijnste plek om te werken aan wat nu dit boekje is geworden.

Lieve papa en mama, Judith, mijn dank aan jullie voor het mogelijk maken van dit werk gaat veel verder dan jullie betrokkenheid en onvoorwaardelijke steun die jullie mij de afgelopen jaren op het gebied van mijn PhD, als op elk onderdeel van mijn leven altijd, hebben gegeven. Bedankt voor het goede leven dat jullie me gegeven hebben en het zorgen voor de basis die dit heeft mogelijk gemaakt.

Allerliefste Jasper, met alles wat ik over vertrouwen en het bouwen van duurzame samenwerkingen heb geleerd in de afgelopen jaren durf ik wel te zeggen dat wij aan alle criteria voldoen. Dank je wel voor alle geruststelling en steun, de aanmoedigingen, je interesse (het zorgvuldig printen en lezen van mijn artikelen) en je grapjes en energie. Ik heb super veel vertrouwen in ons als team.

Ter nagedachtenis aan Juul, een van de meest gemotiveerde en getalenteerde studenten die ik heb ontmoet, en al snel ook een vriendin. Slim, assertief, nauwkeurig en oplossingsgericht, en bovenal zo geïnteresseerd, open, sociaal en goedlachs. Tijdens haar afstudeerstage bij TNO in 2023 heeft Juul samen met Vesa fantastisch werk verricht voor het paper dat op 31 oktober 2024 is gepubliceerd (Hoofdstuk 4). Ik ben heel trots op haar en haar bijdrage. Haar aanstekelijke lach en energie zal ik altijd onthouden.

Bibliography

Abbass, H. A. (2019). Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust. *Cognitive Computation*, *11*(2), 159–171. https://doi.org/10.1007/s12559-018-9619-0

AIHLEG. (2019). *A Definition of AI: Main Capabilities and Disciplines*. https://ec.europa.eu/digital-single-

Akash, K., McMahon, G., Reid, T., & Jain, N. (2020). Human Trust-Based Feedback Control: Dynamically Varying Automation Transparency to Optimize Human-Machine Interactions. *IEEE Control Systems*, *40*(6), 98–116. https://doi.org/10.1109/MCS.2020.3019151

Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C. M., Monz, C., Neerincx, M. A., Oliehoek, F., Prakken, H., Schlobach, S., Gaag, L. van der, Harmelen, F. van, … Welling, M. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, *53*(08), 18–28.

Akgun, M., Cagiltay, K., & Zeyrek, D. (2010). The effect of apologetic error messages and mood states on computer users' self-appraisal of performance. *Journal of Pragmatics*, *42*(9), 2430–2448. https://doi.org/10.1016/j.pragma.2009.12.011

Alarcon, G. M., Gibson, A. M., & Jessup, S. A. (2020). Trust Repair in Performance, Process, and Purpose Factors of Human-Robot Ttust. *Proceedings of the 2020 IEEE International Conference on Human-Machine Systems, ICHMS 2020*. https://doi.org/10.1109/ICHMS49158.2020.9209453

Alarcon, G. M., Gibson, A. M., Jessup, S. A., & Capiola, A. (2021). Exploring the differential effects of trust violations in human-human and human-robot interactions. *Applied Ergonomics*, *93*(May 2020), 103350. https://doi.org/10.1016/j.apergo.2020.103350

Alarcon, G. M., Lyons, J. B., Hamdan, I. aldin, & Jessup, S. A. (2023). Affective Responses to Trust Violations in a Human-Autonomy Teaming Context: Humans Versus Robots. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-023-01017-w

Albayram, Y., Jensen, T., Khan, M. M. H., Fahim, M. A. Al, Buck, R., & Coman, E. (2020). Investigating the Effects of (Empty) Promises on Human-Automation Interaction and Trust Repair. *HAI 2020 - Proceedings of the 8th International Conference on Human-Agent Interaction*, 6–14. https://doi.org/10.1145/3406499.3415064

Antifakos, S., Schwaninger, A., & Schiele, B. (2004). Evaluating the effects of displaying uncertainty in context-aware applications. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3205*(September 2004), 54–69. https://doi.org/10.1007/978-3-540-30119-6_4

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, *21*(1), 99–131. https://doi.org/10.1007/s10683-017-9527-2

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, *100*(3), 571–589. https://doi.org/10.1109/JPROC.2011.2173265

Atkinson, D., Hancock, P. A., Hoffman, R. R., Lee, J. D., Rovira, E., Stokes, C., & Wagner, A. R. (2012). Trust in computers and robots: The uses and boundaries of the analogy to interpersonal trust. *Proceedings of the Human Factors and Ergonomics Society*, 303–307. https://doi.org/10.1177/1071181312561071

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems*, *8*(4). https://doi.org/10.1145/3181671

Baker, A. L., Schaefer, K. E., & Hill, S. G. (2019). Teamwork and Communication Methods and Metrics for Human–Autonomy Teaming. In *CCDC Army Research Laboratory Aberdeen Proving Ground*.

Barnes, M. J., Chen, J. Y. C., Jentsch, F. G., Oron-Gilad, T., Redden, E., Elliott, L., & Evans, A. W. (2014). *Designing for Humans in Autonomous Systems: Military Applications* (Issue January).

Baron, J., & Hershey, J. C. (1988). Outcome Bias in Decision Evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569–579.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Beck, B., & Kühler, M. (2020). *Technology, Anthropology, and Dimensions of Responsibility* (Vol. 1). https://doi.org/10.1007/978-3-476-04896-7_5

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A. W. M., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., Van Dijk, M. C. R. F., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., … Venâncio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - Journal of the American Medical Association*, *318*(22), 2199–2210. https://doi.org/10.1001/jama.2017.14585

Bethel, C. L., & Murphy, R. R. (2010). Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics*, *2*(4), 347–359. https://doi.org/10.1007/s12369-010-0064-9

Bhagat, R. S., & Steers, R. M. (2009). Cambridge Handbook of Culture, Organizations, and Work. In *Cambridge Handbook of Culture, Organizations, and Work*. Cambridge University Press. https://doi.org/10.1017/cbo9780511581151

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*(2), 173–189. https://doi.org/10.1023/B:GRUP.0000021840.85686.57

Bobko, P., Hirshfield, L., Eloy, L., Spencer, C., Doherty, E., Driscoll, J., & Obolsky, H. (2022). 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*, *0*(0), 1–25. https://doi.org/10.1080/1463922X.2022.2086644

Bradfield, M., & Aquino, K. (1999). The effects of blame attributions and offender likableness on forgiveness and revenge in the workplace. *Journal of Management*, *25*(5), 607–631. https://doi.org/10.1177/014920639902500501

Bradshaw, J. M., Dignum, V., Jonker, C. M., & Sierhuis, M. (2012). Introduction to Special Issue on Human-Agent-Robot Teamwork (HART). *IEEE Intelligent Systems*, *27*(2), 8–13. https://doi.org/10.1109/MIS.2012.37

Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of "autonomous systems." *IEEE Intelligent Systems*, *28*(3), 54–61. https://doi.org/10.1109/MIS.2013.70

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, *62*(2), 161–178.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment*, *21*(2), 115–127. https://doi.org/10.1027/1015-5759.21.2.115

Buchholz, V., Kulms, P., & Kopp, S. (2017). *It's (Not) Your Fault! Blame and Trust Repair in Human-Agent Cooperation*. *2017*(1). https://doi.org/10.17185/DUEPUBLICO/44538

Cameron, D., Collins, E. C., de Saille, S., Eimontaite, I., Greenwood, A., & Law, J. (2023). The Social Triad Model: Considering the Deployer in a Novel Approach to Trust in Human–Robot Interaction. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-023-01048-3

Cameron, D., de Saille, S., Collins, E. C., Aitken, J. M., Cheung, H., Chua, A., Loh, E. J., & Law, J. (2021). The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in Human Behavior*, *114*(September), 106561. https://doi.org/10.1016/j.chb.2020.106561

Carpenter, J. (2013). The Quiet Professional: An investigation of U.S. military Explosive Ordnance Disposal personnel interactions with everyday field robots. *ProQuest Dissertations and Theses*, *3599678*, 170.

Cellan-Jones, R. (2020). Uber's self-driving operator charged over fatal crash. *BBC*. https://www.bbc.com/news/technology-54175359

Charalambous, G., Fletcher, S., & Webb, P. (2016). The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration. *International Journal of Social Robotics*, *8*(2), 193–209. https://doi.org/10.1007/s12369-015-0333-8

Chen, B. X. (2022, December 21). How to Use ChatGPT and Still Be a Good Person. *The New York Times*. https://www.nytimes.com/2022/12/21/technology/personaltech/how-to-use-chatgpt-ethically.html

Chen, Jessie Y.C., & Barnes, M. J. (2014). Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13–29. https://doi.org/10.1109/THMS.2013.2293535

Chen, Jessie Y.C., Barnes, M. J., Selkowitz, A. R., & Stowers, K. (2017). Effects of Agent Transparency on Human-Autonomy Teaming Effectiveness. *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, *MI*, 1838–1843. https://doi.org/10.1109/SMC.2016.7844505

Chen, Jessie Y.C., Flemisch, F. O., Lyons, J. B., & Neerincx, M. A. (2020). Guest Editorial: Agent and System Transparency. *IEEE Transactions on Human-Machine Systems*, *50*(3), 189–193. https://doi.org/10.1109/THMS.2020.2988835

Chen, Jessie Y.C., Procci, K., Boyce, M., Wright, J. L., Garcia, A., & Barnes, M. J. (2014). *Situation Awareness – Based Agent Transparency* (Issue April).

Chen, Jessie Y.C., & Schulte, A. (2021). Special issue on "Human-Autonomy Teaming in Military Contexts." *Human-Intelligent Systems Integration*, *3*(4), 287–289. https://doi.org/10.1007/S42454-021-00032-4

Chhogyal, K., Nayak, A., Ghose, A., & Dam, H. K. (2019). A value-based trust assessment model for multi-agent systems. *IJCAI International Joint Conference on Artificial Intelligence*, *2019-Augus*, 194–200. https://doi.org/10.24963/ijcai.2019/28

Chi, V. B., & Malle, B. F. (2023). People Dynamically Update Trust When Interactively Teaching Robots. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.*, 554–564. https://doi.org/10.1145/3568162.3576962

Chien, S., Semnani-azad, Z., Lewis, M., & Sycara, K. (2014). Towards the Development of an Inter-cultural Scale. *Cross-Cultural Design: 6th International Conference, CCD 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014. Proceedings 6*, 35–46.

Chiou, E. K., Demir, M., Buchanan, V., Corral, C. C., Endsley, M. R., Lematta, G. J., Cooke, N. J., & McNeese, N. J. (2022). Towards Human–Robot Teaming: Tradeoffs of Explanation-Based Communication Strategies in a Virtual Search and Rescue Task. *International Journal of Social Robotics*, *14*(5), 1117–1136. https://doi.org/10.1007/s12369-021-00834-1

Chiou, E. K., & Lee, J. D. (2023). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Chirico, A., Cipresso, P., Yaden, D. B., Biassoni, F., Riva, G., & Gaggioli, A. (2017). Effectiveness of Immersive Videos in Inducing Awe: An Experimental Study. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-01242-0

Clabaugh, C., & Mataric, M. J. (2016). Exploring elicitation frequency of learning-sensitive information by a robotic tutor for interactive personalization. *25th IEEE International*

Symposium on Robot and Human Interactive Communication, RO-MAN 2016, 968–973. https://doi.org/10.1109/ROMAN.2016.7745226

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018). Exploring the impact of fault justification in human-robot trust: Socially Interactive Agents Track. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*(July), 507–513.

Costa, A. C. (2003). Work team trust and effectiveness. *Personnel Review*, *32*(5), 605-622+672. https://doi.org/10.1108/00483480310488360

Cremer, D. De, & Kasparov, G. (2021). AI Should Augment Human Intelligence, Not Replace It. *Harvard Business Review*, 1–9.

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior*, *29*(3), 577–579. https://doi.org/10.1016/j.chb.2012.11.023

Darling, K. (2021). *The New Breed: What Our History with Animals Reveals about Our Future with Robots.* Henry Holt and Company.

de Graaf, M. M. A., & Malle, B. F. (2017). *How people explain action (and Autonomous Intelligent Systems Should Too)* (Issue November). https://doi.org/10.3174/ajnr.A1282

De Melo, C. M., Gratch, J., & Carnevale, P. J. (2015). Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing*, *6*(2), 127–136. https://doi.org/10.1109/TAFFC.2014.2332471

De Melo, C. M., Zheng, L., & Gratch, J. (2009). Expression of moral emotions in cooperating agents. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5773 LNAI*, 301–307. https://doi.org/10.1007/978-3-642-04380-2_32

de Visser, E. J., Krueger, F., McKnight, P. E., Scheid, S., Smith, M. A. B., Chalk, S., & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society*, 263–267. https://doi.org/10.1177/1071181312561062

de Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., Parasuraman, R., & Krueger, F. (2017). A Little Anthropomorphism Goes a Long Way: Effects of Oxytocin on Trust, Compliance, and Team Performance with Automated Agents. *Human Factors*, *59*(1), 116–133. https://doi.org/10.1177/0018720816687205

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

de Visser, E. J., Pak, R., & Neerincx, M. A. (2017). Trust development and repair in human-robot teams. *ACM/IEEE International Conference on Human-Robot Interaction*, 103–104. https://doi.org/10.1145/3029798.3038409

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

de Visser, E. J., & Parasuraman, R. (2011). Adaptive Aiding of Human-Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231. https://doi.org/10.1177/1555343411410160

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S. C., Shaw, T. H., Pak, R., & Neerincx, M. A. (2019). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, 1–20. https://doi.org/10.1007/s12369-019-00596-x

de Vries, P. W., van den Berg, S. M., & Midden, C. (2015). Assessing Technology in the Absence of Proof: Trust Based on the Interplay of Others Opinions and the Interaction Process. *Human Factors*, *57*(8), 1378–1402. https://doi.org/10.1177/0018720815598604

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination. *Cognition, Technology & Work*, *4*(4), 240–244. https://doi.org/10.1007/s101110200022

Dennett, D. C. (1981). Brainstorms: Philosophical Essays on Mind and Psychology. In *Angewandte Chemie International Edition, 6(11), 951–952.* http://repo.iain-tulungagung.ac.id/5510/5/BAB 2.pdf

Disalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). *All Robots Are Not Created Equal : The Design and Perception of Humanoid Robot Heads*.

Driskell, J. E., Salas, E., & Driskell, T. (2018). Foundations of teamwork and collaboration. *American Psychologist*, *73*(4), 334–348. https://doi.org/10.1037/amp0000241

Du, N., Haspiel, J., Zhang, Q., Tilbury, D. M., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*(September 2018), 428–442. https://doi.org/10.1016/j.trc.2019.05.025

Duenser, A., & Douglas, D. M. (2023). Whom to Trust, How and Why: Untangling Artificial Intelligence Ethics Principles, Trustworthiness, and Trust. *IEEE Intelligent Systems*, *38*(6), 19–26. https://doi.org/10.1109/MIS.2023.3322586

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human Computer Studies*. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting Misuse and Disuse of Combat Identification Systems. *Military Psychology*, *13*(3), 147–164. https://doi.org/10.1207/S15327876MP1303_2

Ellwart, T., & Schauffel, N. (2023). *Human-Autonomy Teaming in Ship Inspection : Psychological Perspectives on the Collaboration Between Humans and Self-Governing Systems*. Springer International Publishing. https://doi.org/10.1007/978-3-031-25296-9

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Esterwood, C., & Robert, L. P. (2023a). The Theory of Mind and Human-Robot Trust Repair. *Scientific Reports*, *13*(1), 1–15. https://doi.org/10.1038/s41598-023-37032-0

Esterwood, C., & Robert, L. P. (2023b). Three Strikes and You are Out!: The Impacts of Multiple Human-Robot Trust Violations and Repairs on Robot Trustworthiness. *Computers in Human Behavior*, *January*.

Esterwood, C., & Robert, L. P. (2022). A Literature Review of Trust Repair in HRI. *Proceedings of 31th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2022)*, *July*.

Esterwood, C., & Robert, L. P. (2021). Do You Still Trust Me? Human-Robot Trust Repair Strategies. *Proceedings of 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 183–188.

Fahim, M. A. Al, Khan, M. M. H., Jensen, T., Albayram, Y., & Coman, E. (2021). Do Integral Emotions Affect Trust? The Mediating Effect of Emotions on Trust in the Context of Human-Agent Interaction. *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere*, *June 2021*, 1492–1503. https://doi.org/10.1145/3461778.3461997

Fahim, M. A. Al, Khan, M. M. H., Jensen, T., Albayram, Y., Coman, E., & Buck, R. (2021). The Mediating Effect of Emotions on Trust in the Context of Automated System Usage. *IEEE Transactions on Affective Computing*, *3045*(c), 1–1. https://doi.org/10.1109/taffc.2021.3094883

Feaver, P. D., & Kohn, R. H. (2001). *Soldiers and civilians: The civil-military gap and American national security.* Mit Press.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human Computer Studies*, *132*(June), 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Ferguson, G., & Allen, J. (2011). A cognitive model for collaborative agents. *AAAI Fall Symposium - Technical Report*, *FS-11-01*, 112–120.

Fine, G. A., & Holyfield, L. (2006). Secrecy, Trust, and Dangerous Leisure: Generating Group Cohesion in Voluntary Organizations. *Social Psychology Quarterly*. https://doi.org/10.2307/2787117

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. *Lecture Notes in Computer Science (Including Subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *7621 LNAI*, 199–208. https://doi.org/10.1007/978-3-642-34103-8_20

Forster, Y., Naujoks, F., & Neukum, A. (2017). Increasing anthropomorphism and trust in automated driving functions by adding speech output. *IEEE Intelligent Vehicles Symposium, Proceedings*, *2*(Iv), 365–372. https://doi.org/10.1109/IVS.2017.7995746

Fox, C. R., & Ulkumen, G. (2021). Distinguishing Two Dimensions of Uncertainty. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3695311

Fratczak, P., Goh, Y. M., Kinnell, P., Justham, L., & Soltoggio, A. (2021). Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *International Journal of Industrial Ergonomics*, *82*(July 2020), 103078. https://doi.org/10.1016/j.ergon.2020.103078

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems, CTS*, 106–114. https://doi.org/10.1109/CTS.2007.4621745

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Gambetta, D. (2000). Can We Trust Trust? In *Trust: Making and Breaking Cooperative Relations* (electronic, pp. 212–237). Department of Sociology, University of Oxford.

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice taking from humans and machines: An fMRI and effective connectivity study. *Frontiers in Human Neuroscience*, *10*(NOV2016), 1–15. https://doi.org/10.3389/fnhum.2016.00542

Gould, S. J. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2015). Home is Where the Lab is: A Comparison of Online and Lab Data From a Time-sensitive Study of Interruption. *Human Computation*, *2*(1), 45–67. https://doi.org/10.15346/hc.v2i1.4

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Greenberg, A. M., & Marble, J. L. (2023). Foundational concepts in person-machine teaming. *Frontiers in Physics*, *10*(January), 1–16. https://doi.org/10.3389/fphy.2022.1080132

Grover, S. L., Hasel, M. C., Manville, C., & Serrano-Archimi, C. (2014). Follower reactions to leader trust violations: A grounded theory of violation types, likelihood of recovery, and recovery process. *European Management Journal*, *32*(5), 689–702. https://doi.org/10.1016/j.emj.2014.01.002

Guo, Y., & Yang, X. J. (2020). Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics*, *13*(8), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

Guznov, S., Lyons, J., Pfahler, M., Heironimus, A., Woolley, M., Friedman, J., & Neimeier, A. (2020). Robot Transparency and Team Orientation Effects on Human–Robot Teaming. *International Journal of Human-Computer Interaction*, *36*(7), 650–660. https://doi.org/10.1080/10447318.2019.1676519

Hald, K., Weitz, K., André, E., & Rehm, M. (2021). "An Error Occurred!" - Trust Repair With Virtual Robot Using Levels of Mistake Explanation. *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*, *3*(1), 9. http://journal.unilak.ac.id/index.php/JIEB/article/view/3845%0Ahttp://dspace.uc.ac.id/handle/123456789/1288

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, 493–500. https://doi.org/10.1109/ROMAN.2016.7745163

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can you trust your robot? *Ergonomics in Design*, *19*(3), 24–29. https://doi.org/10.1177/1064804611415045

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011a). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011b). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2020). Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors*. https://doi.org/10.1177/0018720820922080

Hannibal, G., & Weiss, A. (2022). Exploring the Situated Vulnerabilities of Robots for Interpersonal Trust in Human-Robot Interaction. In S. T. Koeszegi & M. Vincze (Eds.), *Trust in Robots* (pp. 33–56). TU Wien Academic Press. https://doi.org/https://doi.org/10.34727/2022/isbn.978-3-85448-052-5

Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *Lecture Notes in Engineering and Computer Science*, *1*, 425–430.

Harrington, L. (2023). ChatGPT Is Trending: Trust but Verify. *Technology Today*, *76109*, 1–7.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society*, 904–908. https://doi.org/10.1177/154193120605000909

Haselhuhn, M. P., Schweitzer, M. E., & Wood, A. M. (2010). How implicit beliefs influence trust recovery. *Psychological Science*, *21*(5), 645–648. https://doi.org/10.1177/0956797610367752

Hayes, B. (2016). *Supportive Behaviors for Human-Robot Teaming*. http://scazlab.yale.edu/sites/default/files/files/Hayes_Dissertation_humsnrobotteaming.pdf

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2013*, *October*, 210–217. https://doi.org/10.1145/2516540.2516554

Hidalgo, C. A., Orghian, D., Albo-Canals, J., Almeida, F. de, & Martin, N. (2021). *How Humans Judge Machines*. The MIT Press Cambridge, Massachusetts London, England.

Ho, N. T., Johnson, W. B., Panesar, K., Wakeland, K., Sadler, G. G., Wilson, N., Nguyen, B., Lachter, J., & Brandt, S. L. (2017). Application of human-autonomy teaming to an advanced ground station for reduced crew operations. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2017-Septe*, 9–12. https://doi.org/10.1109/DASC.2017.8102124

Hock, P., Kraus, J., Walch, M., Lang, N., & Baumann, M. (2016). Elaborating feedback strategies for maintaining automation in highly automated driving. *AutomotiveUI 2016 - 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Proceedings*, 105–112. https://doi.org/10.1145/3003715.3005414

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R. (2017). A taxonomy of emergent trusting in the human-machine relationship. *Cognitive Systems Engineering: The Future for a Changing World*, 137–164. https://doi.org/10.1201/9781315572529

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & A, U. (2013). Trust in Automation. *IEEE INTELLIGENT SYSTEMS*, *13*, 84–88. file:///C:/Users/Sami Lini HEAL/AppData/Local/Mendeley Ltd./Mendeley Desktop/Downloaded/Zachary et al. - 2012 - R - I C ER RE ? Rapidized ? Cognitive Task Analysis.pdf

Hou, M., Ho, G., & Dunwoody, D. (2021). IMPACTS: a trust model for human-autonomy teaming. *Human-Intelligent Systems Integration*, *3*(2), 79–97. https://doi.org/10.1007/s42454-020-00023-x

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jermutus, E., Kneale, D., Thomas, J., & Michie, S. (2022). Influences on User Trust in Healthcare Artificial Intelligence: A Systematic Review. *Wellcome Open Research*, *7*, 65. https://doi.org/10.12688/wellcomeopenres.17550.1

Jessup, S. A. (2018). Measurement of the propensity to trust automation. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Jessup, S. A., Gibson, A. M., Capiola, A., Alarcon, G. M., & Borders, M. (2020). Investigating the Effect of Trust Manipulations on Affect over Time in Human-Human

versus Human-Robot Interactions. *Proceedings of the 53rd Hawaii International Conference on System Science*.

Jeste, D. V., Graham, S. A., Nguyen, T. T., Depp, C. A., Lee, E. E., & Kim, H. C. (2020). Beyond artificial intelligence: Exploring artificial wisdom. *International Psychogeriatrics*, *32*(8), 993–1001. https://doi.org/10.1017/S1041610220000927

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*, 53–71.

Johannemann, K., Morasch, K., & Wiens, M. (2016). *Can occupational norms foster cooperative behavior? An experimental study comparing cooperation by military officers and civilians*.

Johnson, J. (2024). Finding AI Faces in the Moon and Armies in the Clouds: Anthropomorphising Artificial Intelligence in Military Human-Machine Interactions. *Global Society*, *38*(1), 67–82. https://doi.org/10.1080/13600826.2023.2205444

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2011). The fundamental principle of coactive design: Interdependence must shape autonomy. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6541 LNAI*, 172–191. https://doi.org/10.1007/978-3-642-21268-0_10

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2012). Interdependence Robot Teams. *IEEE Intelligent Systems*, *27*, 43–51.

Johnson, M., & Vera, A. H. (2019). No Ai is an island: The case for teaming intelligence. *AI Magazine*, *40*(1), 16–28. https://doi.org/10.1609/aimag.v40i1.2842

Jones, N. A., Ross, H., Lynam, T., Perez, P., & Leitch, A. (2011). Mental models: An interdisciplinary synthesis of theory and methods. *Ecology and Society*, *16*(1). https://doi.org/10.5751/ES-03802-160146

Jorge, C. C., Bouman, N. H., Jonker, C. M., & Tielman, M. L. (2023). Exploring the Effect of Automation Failure on the Human's Trustworthiness in Human-Agent Teamwork. *Frpntiers in Robotics and AI*, *August*, 1–14. https://doi.org/10.3389/frobt.2023.1143723

Jorge, C. C., Tielman, M. L., & Jonker, C. M. (2022). Assessing Artificial Trust in Human-Agent Teams A Conceptual Model Assessing Artificial Trust in Human-Agent Teams. *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agent*, *September*, 1–3. https://doi.org/10.1145/3514197.3549696

Jorge, C. C., Zoelen, E. M. va., Verhagen, R. S., Mehrotra, S., Jonker, C. M., & Tielman, M. L. (2024). Appropriate Context-Dependent Artificial Trust in Human-Machine Teamwork. In *Putting AI in the Critical Loop* (pp. 41–60). Elsevier Inc.

Kadres, F. (1996). In Defense of Experimental Consumer Psychology. In *Journal of Consumer Psychology* (Vol. 5, Issue 3, pp. 279–296).

Kahneman, D. (2011). Thinking fast and thinking slow. In *Farrar, Strauss and Giroux, New York, NY*.

Kessler, T. T., Larios, C., Walker, T., Yerdon, V., & Hancock, P. A. (2016). A Comparison of Trust Measures in Human–Robot Interaction Scenarios Theresa. *Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems*, *November*, 436. https://doi.org/10.1007/978-3-319-41959-6

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/j.obhdp.2005.07.002

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- versus Integrity-Based Trust Violations. *Journal of Applied Psychology*, *89*(1), 104–118. https://doi.org/10.1037/0021-9010.89.1.104

Kim, Taemie, & Hinds, P. J. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 80–85. https://doi.org/10.1109/ROMAN.2006.314398

Kim, Taenyun, & Song, H. (2021). How should intelligent agents apologize to restore trust ?: The interaction effect between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*.

Knighton, R. J. (2004). The Psychology of Risk and its Role in Military Decision□Making. *Defence Studies*, *4*(3), 309–334. https://doi.org/10.1080/1470243042000344786

Knightscope. (2023). *The K5 ASR - A fully autonomous outdoor security robot.* https://www.knightscope.com/products/k5

Kohn, S. C., Quinn, D. B., Pak, R., de Visser, E. J., & Shaw, T. H. (2018). Trust repair strategies with self-driving vehicles: An exploratory study. *Proceedings of the Human Factors and Ergonomics Society*, *2*, 1108–1112. https://doi.org/10.1177/1541931218621254

Körber, M. (2019). Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. Proceedings 20th Triennial Congress of the IEA. Springer. *Advances in Intelligent Systems and Computing, Vol.823*, 13–30. http://link.springer.com/10.1007/978-3-319-96074-6_2

Korteling, J. E. (Hans)., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence*, *4*(March), 1–13. https://doi.org/10.3389/frai.2021.622364

Kox, E. S., Barnhoorn, J., Rábago Mayer, L., Temel, A., & Klunder, T. (2022). Using a Virtual Reality House-Search Task to Measure Trust During Human-Agent Interaction (Demo Paper). *HHAI2022: Augmenting Human Intellect*, 272–274. https://doi.org/10.3233/FAIA220214

Kox, E. S., & Beretta, B. (2024). Evaluating Generative AI Incidents: An Exploratory Vignette Study on the Role of Trust , Attitude and AI Literacy. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, 188–198. https://doi.org/10.3233/FAIA240194

Kox, E. S., Finlayson, N. B., Broderick-Hale, J. C., & Kerstholt, J. H. (2023). Calibrated Trust as a Means to Build Societal Resilience Against Cognitive Warfare. *Proceedings of NATO Symposium HFM-361 on Mitigating and Responding to Cognitive Warfare*, 1–15.

Kox, E. S., Hennekens, M., Metcalfe, J. S., & Kerstholt, J. H. (n.d.). Trust Violations Due to Error or Choice: the Differential Effects on Trust Repair in Human-Human and Human-Robot Interaction. *Transactions on Human-Robot Interaction*.

Kox, E. S., Kerstholt, J. H., Hueting, T., Barnhoorn, J., & Eikelboom, A. (2019). Autonomous Systems As Intelligent Teammates: Social Psychological Implications. *024th International Command and Control Research & Technology Symposium*.

Kox, E. S., Kerstholt, J. H., Hueting, T., & de Vries, P. W. (2021). Trust Repair in Human-Agent Teams: the Effectiveness of Explanations and Expressing Regret. *Autonomous Agents and Multi-Agent Systems*, *35*(2), 1–20. https://doi.org/10.1007/s10458-021-09515-9

Kox, E. S., Siegling, L. B., & Kerstholt, J. H. (2022). Trust Development in Military and Civilian Human-Agent Teams: the Effect of Social-Cognitive Recovery Strategies. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00871-4

Kox, E. S., van den Boogaard, J., Turjaka, V., & Kerstholt, J. H. (2024). The Journey or the Destination: The Impact of Transparency and Goal Attainment on Trust in Human-Robot Teams. *Transactions on Human-Robot Interaction*. https://doi.org/https://doi.org/10.1145/3702245

Kox, E. S., van Riemsdijk, M. B., de Vries, P. W., & Kerstholt, J. H. (2024a). Red or Blue Door: Exploring the Behavioural Consequences of Trust Violations due to Robot Error or Choice Using a VR Maze. *Retrieved from Osf.Io/J6vwk*. https://doi.org/DOI 10.17605/OSF.IO/J6VWK

Kox, E. S., van Riemsdijk, M. B., de Vries, P. W., & Kerstholt, J. H. (2024b). The Impact of Anthropomorphic Cues and Explanations on Trust Formation, Violation, and Repair in HRI: Insights from a VR Experiment. *Retrieved from Osf.Io/62gte*. https://doi.org/DOI 10.17605/OSF.IO/62GTE

Kox, E. S., van Riemsdijk, M. B., & Kerstholt, J. H. (2024). From Exploitation to Augmentation: Navigating AI's Impact on Human Value- and Inference-based Decision-Making. *Proceedings of NATO HFM-RSY-377 Symposium on "Meaningful Human Control of Future Military Operations: Spanning Across Warfare Domains with Advanced AI,"* 1–12.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors*, *62*(5), 718–736. https://doi.org/10.1177/0018720819853686

Ku, C., & Pak, R. (2023). The Effects of Individual Differences in Working Memory on Trust Recovery. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*(1), 1134–1139. https://doi.org/10.1177/21695067231195000

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345–360. https://doi.org/10.1080/00140139.2018.1547842

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, 1–72. https://doi.org/10.1017/S0140525X16001837

Langer, A., Feingold-Polak, R., Mueller, O., Kellmeyer, P., & Levy-Tzedek, S. (2019). Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience and Biobehavioral Reviews*, *104*(July), 231–239. https://doi.org/10.1016/j.neubiorev.2019.07.014

Lee, A. Y., Bond, G. D., Russell, D. C., Tost, J., González, C., & Scarbrough, P. S. (2010). Team perceived trustworthiness in a complex military peacekeeping simulation. *Military Psychology*, *22*(3), 237–261. https://doi.org/10.1080/08995605.2010.492676

Lee, J.-E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. *Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives*, 1–15. https://doi.org/10.4018/978-1-61520-901-9.ch001

Lee, J. D. (1991). The Dynamics of Trust in a Supervisory Control Simulation. *Proceedings of the Human Factors Society 35th Annual Meeting*, *35*(17), 1228–1232.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80.

Lee, J. E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. *Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives*, 1–15. https://doi.org/10.4018/978-1-61520-901-9.ch001

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010a). Gracefully mitigating breakdowns in robotic services. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 203–210. https://doi.org/10.1109/HRI.2010.5453195

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S. S., & Rybski, P. (2010b). Gracefully mitigating breakdowns in robotic services. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 203–210. https://doi.org/10.1109/HRI.2010.5453195

Lewicki, R. J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 287–313.

Lewicki, R. J., Polin, B., & Lount, R. B. (2016). An Exploration of the Structure of Effective Apologies. *Negotiation and Conflict Management Research*, *9*(2), 177–196. https://doi.org/10.1111/ncmr.12073

Lewis, M., Sycara, K., & Walker, P. (2018). The Role of Trust in Human-Robot Interaction. In *Studies in Systems, Decision and Control* (Vol. 117). https://doi.org/10.1007/978-3-319-64816-3_8

Li, M., Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2019). No risk no trust: Investigating perceived risk in highly automated driving. *Proceedings - 11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2019*, *September*, 177–185. https://doi.org/10.1145/3342197.3344525

Loewenstein, G. F., Hsee, C. K., Weber, E. U., & Welch, N. (2001). Risk as Feelings. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.127.2.267

Lozano, E. B., & Laurent, S. M. (2019). The effect of admitting fault versus shifting blame on expectations for others to do the same. *PLoS ONE*, *14*(3), 1–19. https://doi.org/10.1371/journal.pone.0213276

Lubars, B., & Tan, C. (2019). Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *Advances in Neural Information Processing Systems*, *32*(NeurIPS).

Luebbers, M. B., Tabrez, A., Ruvane, K., & Hayes, B. (2023). Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming. *Robotics: Science and Systems*.

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. *AAAI Spring Symposium - Technical Report*, *SS-13-07*, 48–53.

Lyons, J. B., Hamdan, I. aldin, & Vo, T. Q. (2023). Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior*, *138*(February 2022), 107473. https://doi.org/10.1016/j.chb.2022.107473

Lyu, N., Xie, L., Wu, C., Fu, Q., & Deng, C. (2017). Driver's cognitive workload and driving performance under traffic sign information exposure in complex environments: A case study of the highways in China. *International Journal of Environmental Research and Public Health*, *14*(2), 1–25. https://doi.org/10.3390/ijerph14020203

Maas, M. M. (2023). *AI is Like… A Literature Review of AI Metaphors and Why They Matter for Policy*.

Macko, A. (2020). Gender differences in trust, reactions to trust violation, and trust restoration. *Decyzje*, *33*, 55–73. https://doi.org/10.7206/DEC.1733-0092.140

Madhavan, P., & Wiegmann, D. A. (2005). Effects of information source, pedigree, and reliability on operators' utilizaton of diagnostic advice. *Proceedings of the Human Factors and Ergonomics Society*, 487–491. https://doi.org/10.1177/154193120504900358

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*(2), 241–256. https://doi.org/10.1518/001872006777724408

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. *Proceedings of Eleventh Australasian Conference on Information Systems*, 6–8. http://books.google.com/books?hl=en&lr=&id=b0yalwi1HDMC&oi=fnd&pg=PA102&dq=The+Big+Five+Trait+Taxonomy:+History,+measurement,+and+Theoretical+Perspectives&ots=758BNaTvOi&sig=L52e79TS6r0Fp2m6xQVESnGt8mw%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124. https://doi.org/10.1145/2696454.2696458

Malle, B. F., & Ullman, D. (2021). A Multi-Dimensional Conception and Measure of Human-Robot Trust. *Trust in Human-Robot Interaction: Research and Applications*, 3–25.

Marsh, S. (1994). Formalising trust as a computational concept [University of Stirling]. In *SpringerBriefs in Computer Science* (Vol. 0, Issue 9781461470304). https://doi.org/10.1007/978-1-4614-7031-1_2

Martelaro, N. (2016). Wizard-of-Oz interfaces as a step towards autonomous HRI. *AAAI Spring Symposium - Technical Report*, *SS-16-01-*, 147–150.

Mathieu, J. E., Heffner, T. S., Goodwin, G., Salas, E., & Cannon-Bowers, J. A. (2000). The Influence of Shared Mental Models on Team Process and Performance. *Journal of Applied Psychology*, *85*(2), 273–283.

Matthews, G., Hancock, P. A., Lin, J., Panganiban, A. R., Reinerman-Jones, L. E., Szalma, J. L., & Wohleber, R. W. (2021). Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Personality and Individual Differences*, *169*(December 2019), 109969. https://doi.org/10.1016/j.paid.2020.109969

Matthews, G., Lin, J., Panganiban, A. R., & Long, M. D. (2019). Individual Differences in Trust in Autonomous Robots: Implications for Transparency. *IEEE Transactions on Human-Machine Systems*, *PP*(November), 1–11. https://doi.org/10.1109/THMS.2019.2947592

Matthews, G., Panganiban, A. R., Bailey, R., & Lin, J. (2018). Trust in Autonomous Systems for Threat Analysis: A Simulation Methodology. *International Conference of Virtual, Augmented and Mixed Reality*, *10910 LNCS*, 116–125. https://doi.org/10.1007/978-3-319-91584-5_10

Matthews, G., Panganiban, A. R., Lin, J., Long, M. D., & Schwing, M. (2021). Super-machines or sub-humans: Mental models and trust in intelligent autonomous systems. In *Trust in Human-Robot Interaction* (pp. 59–82). Elsevier Inc. https://doi.org/10.1016/b978-0-12-819472-0.00003-4

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123–136. https://doi.org/10.1037/0021-9010.84.1.123

Mayer, R. C., Davis, J. H., & Schoorman, D. F. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.1109/GLOCOM.2017.8254064

McKnight, D. H., & Chervany, N. L. (2000). What is Trust ? A Conceptual Analysis and an Interdisciplinary Model. *Proceedings of the 2000 Americas Conference on Information Systems AMCI2000 AIS Long Beach CA August 2000*, *346*, 382. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1876&context=amcis2000

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206

Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring Individual Differences in the Perfect Automation Schema. *Human Factors*, *57*(5), 740–753. https://doi.org/10.1177/0018720815581247

Metcalfe, J. S., & van Diggelen, J. (2021). Design Considerations for Future Human-AI Ecosystems. *HHAI '21*, *June*.

Miller, C. A. (2020). Trust, transparency, explanation, and planning: Why we need a lifecycle perspective on human-automation interaction. In *Trust in Human-Robot Interaction*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-819472-0.00011-3

Miller, C. A. (2014). Delegation and Transparency: Coordinating Interactions So Information Exchange Is No Surprise. In R. Shumaker & S. Lackey (Eds.), *Proceedings of the 6th International Conference of Virtual, Augmented and Mixed Reality (VAMR), Part I: Vol. 8525 LNCS* (Issue PART 1, pp. 191–202). Springer. https://doi.org/10.1007/978-3-319-07458-0_8

Miller, C. A., Barber, D., Holder, E., Huang, L., Lyons, J., Roth, E., & Wauck, H. (2023). LifeCycle Transparency : Why, and How, Transparency Information Exchange Should be Distributed throughout the Life of Technology Usage Participant Biographies. *Special Issue: Proceedings of the 67th Human Factors and Ergonomics Society International Annual Meeting*. https://doi.org/10.1177/21695067231192272

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors*, *49*(1), 57–75. https://doi.org/10.1518/001872007779598037

Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences*. http://arxiv.org/abs/1706.07269

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI*, *4*(May), 1–15. https://doi.org/10.3389/frobt.2017.00021

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues . *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509. https://doi.org/10.1002/pra2.487

Norman, D. A. (2013). The Design of Everyday Things - Revised and expanded edition. *Basic Books. New York. NY. Pp. Xi-10 ISBN*, *13*, 970–978.

O'Neill, T., McNeese, N. J., Barron, A., & Schelble, B. G. (2022). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Olshtain, E., & Cohen, A. (1983). Apology: A speech act set. *Sociolinguistics and Language Acquisition*, 18–35.

Ososky, S., Sanders, T. L., Jentsch, F. G., Hancock, P. A., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In R. E. Karlsen, D. W. Gage, C. M. Shoemaker, & G. R. Gerhart (Eds.), *Proc. SPIE 9084, Unmanned Systems Technology XVI* (Vol. 9084, p. 90840E). https://doi.org/10.1117/12.2050622

Ososky, S., Schuster, D., Jentsch, F., Fiore, S., Shumaker, R., Lebiere, C., Kurup, U., Oh, J., & Stentz, A. (2012a). The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. *Unmanned Systems Technology XIV*, *8387*(May 2012), 838710. https://doi.org/10.1117/12.923283

Ososky, S., Schuster, D., Jentsch, F. G., Fiore, S. M., Shumaker, R., Lebiere, C., Kurup, U., Oh, J., & Stentz, A. (2012b). The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. *Unmanned Systems Technology XIV*, *8387*(May 2012), 838710. https://doi.org/10.1117/12.923283

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554

Pak, R., & Rovira, E. (2023). A theoretical model to explain mixed effects of trust repair strategies in autonomous systems. *Theoretical Issues in Ergonomics Science*. https://doi.org/10.1080/1463922X.2023.2250424

Pan, X., & Hamilton, A. F. d. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395–417. https://doi.org/10.1111/bjop.12290

Pang, S. (2023). *The Truth About Saying "Thanks" & "Please" To ChatGPT*. Medium. https://ppangsy.medium.com/a-world-of-thanks-the-surprising-effects-of-chatgpt-appreciation-105f6e25fcce

Panganiban, A R, Long, M. D., & Matthews, G. (2020). *Human Machine Teaming (HMT): Trust Cues in Communication and Bias Towards Robotic Partners*. https://apps.dtic.mil/sti/citations/AD1121408%0Ahttps://apps.dtic.mil/sti/pdfs/AD1121408.pdf

Panganiban, April Rose, Matthews, G., & Long, M. D. (2020). Transparency in Autonomous Teammates: Intention to Support as Teaming Information. *Journal of Cognitive Engineering and Decision Making*, *14*(2), 174–190. https://doi.org/10.1177/1555343419881563

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans : A Publication of the IEEE Systems, Man, and Cybernetics Society*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160. https://doi.org/10.1518/155534308X284417

Parker, S. K., & Grote, G. (2022). Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World. *Applied Psychology*, *71*(4), 1171–1204. https://doi.org/10.1111/apps.12241

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, *9*(DEC), 1–19. https://doi.org/10.3389/fnhum.2015.00660

Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI and Society*, *36*(1), 217–238. https://doi.org/10.1007/s00146-020-01005-y

Perkins, R., Khavas, Z. R., McCallum, K., Kotturu, M. R., & Robinette, P. (2022). The Reason for an Apology Matters for Robot Trust Repair. *Social Robotics: 14th International Conference*, Proceedings, Part II (pp. 640-651). https://doi.org/10.1007/978-3-031-24670-8

Petty, R. E., & Cacioppo, J. T. (1996). Addressing Disturbing and Disturbed Consumer Behavior: Is it Necessary to Change the Way We Conduct Behavioral Science? *Journal of Marketing Research*, *33*(1), 1–8. https://doi.org/10.1177/002224379603300101

Phillips, E. K., Ososky, S., Grove, J., & Jentsch, F. G. (2011). From tools to teammates: Toward the development of appropriate mental models for intelligent robots. *Proceedings of the Human Factors and Ergonomics Society*, 1491–1495. https://doi.org/10.1177/1071181311551310

Phillips, E., Malle, B. F., & Chi, V. B. (2023). Systematic methods for Moral HRI : Studying human responses to robot norm conflicts. *TBD*, *1*(1). https://doi.org/10.31234/osf.io/by4rh

Quinn, D. B., Pak, R., & de Visser, E. J. (2017). Testing the efficacy of human-human trust repair strategies with machines. *Proceedings of the Human Factors and Ergonomics Society*, *2017-Octob*(2016), 1794–1798. https://doi.org/10.1177/1541931213601930

Raue, M., D'Ambrosio, L. A., Ward, C., Lee, C., Jacquillat, C., & Coughlin, J. F. (2019). The Influence of Feelings While Driving Regular Cars on the Perception and Acceptance of Self-Driving Cars. *Risk Analysis*, *39*(2), 358–374. https://doi.org/10.1111/risa.13267

Razzouk, R., & Johnson, T. (2012). Shared Cognition. In *Encyclopedia of the Sciences of Learning*. https://doi.org/10.1007/978-1-4419-1428-6_205

Rebensky, S., Kendall Carmody, C. F., Nguyen, D., Carroll, M., Wildman, J., & Thayer, A. (2021). Whoops! Something Went Wrong: Errors, Trust and Trust Repair Strategies in Human Agent Teaming. In H. D. and S. Ntoa (Ed.), *Second International Conference, AI-HCI 2021 Held as Part of the 23rd HCI International Conference* (Vol. 2, pp. 95–106). Springer Nature Switzerland. https://doi.org/10.1016/j.gpb.2023.01.002

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. https://doi.org/10.18653/v1/n16-3020

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction - HRI '09*, 245. https://doi.org/10.1145/1514095.1514158

Robinette, P., Howard, A. M., & Wagner, A. R. (2017a). Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed? In *Autonomy and Artificial Intelligence: A Threat or Savior?* (pp. 129–156). https://doi.org/10.1007/978-3-319-59719-5

Robinette, P., Howard, A. M., & Wagner, A. R. (2017b). Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems*, *47*(4), 425–436. https://doi.org/10.1109/THMS.2017.2648849

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. *International Conference on Social Robotics*, *9388 LNCS*, 574–583. https://doi.org/10.1007/978-3-319-25554-5_46

Roff, H., & Danks, D. (2018). "Trust but Verify": The difficulty of trusting autonomous weapons systems. *Journal of Military Ethics*, *17*(1), 2–20. http://weekly.cnbnews.com/news/article.html?no=124000

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn*, *9*(1), 137–154. https://doi.org/10.1515/pjbr-2018-0010

Rouse, W. B., & Morris, N. M. (1985). *On looking into the black box: Prospects and limits in the search for mental models*. http://scholar.

google.com/scholar?q=related:QM4p5zGC8jMJ:scholar.google.com/&hl=en&num=30&as_sdt=0,5

Rousseau, D. M., Sitkin, S. B., Burt, R. S., Camerer, C., Rousseau, D. M., & Burt, R. S. (1998). Not so Different after All: A Cross-Discipline View of Trust. *Academy of Management Review*, *23*(3), 393–404.

Rovira, E., Pak, R., & McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science*, *18*(6), 573–591. https://doi.org/10.1080/1463922X.2016.1252806

Salas, E., Reyes, D. L., & McDaniel, S. H. (2018). The science of teamwork: Progress, reflections, and the road ahead. *American Psychologist*, *73*(4), 93–600. https://doi.org/10.1037/amp0000334

Salas, E., Sims, D. E., & Shawn Burke, C. (2005). Is there A "big five" in teamwork? *Small Group Research*, *36*(5), 555–599. https://doi.org/10.1177/1046496405277134

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015-March*, 141–148. https://doi.org/10.1145/2696454.2696497

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schaefer, K. E., Hill, S. G., & Jentsch, F. G. (2018). Trust in Human-Autonomy Teaming: A Review of Trust Research from the US Army Research Laboratory Robotics Collaborative Technology Alliance. In J.Y.C. Chen (Ed.), *Proceedings of the AHFE 2018 International Conference on Human Factors in Robots and Unmanned Systems* (Vol. 784, pp. 102–114). Springer International Publishing AG, part of Springer Nature (outside the USA). https://doi.org/10.1007/978-3-319-94346-6_8

Schaefer, K. E., Straub, E. R., Chen, J. Y. C., Putney, J., & Evans, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, *46*, 26–39. https://doi.org/10.1016/j.cogsys.2017.02.002

Schaekermann, M., Beaton, G., Sanoubari, E., Lim, A., Larson, K., & Law, E. (2020). Ambiguity-aware AI Assistants for Medical Data Analysis. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14. https://doi.org/10.1145/3313831.3376506

Scher, S. J., & Darley, J. M. (1997). How Effective Are the Things People Say to Apologize? Effects of the Realization of the Apology Speech Act. *Journal of Psycholinguistic Research*, *26*(1), 127–140. https://doi.org/10.1023/A:1025068306386

Schneider, T. R., Jessup, S. A., Stokes, C. K., Rivers, S., Lohani, M., & McCoy, M. (2017). The influence of trust propensity on behavioral trust. *Poster Session Presented at the Meeting of Association for Psychological Society, Boston.*

Scholtz, J. (2003). Theory and evaluation of human robot interactions. *System Sciences, 2003. Proceedings of the 36th …*, *00*(C), 1–10. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1174284

Schoorman, D. F., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present and Future. *Academy of Management Review*, *32*(2), 344–354. http://www.jstor.org/stable/258792?origin=crossref

Schraagen, J. M., & van Diggelen, J. (2021). A Brief History of the Relationship Between Expertise and Artificial Intelligence. *Expertise at Work*, 149–175. https://doi.org/10.1007/978-3-030-64371-3_8

Schumann, K., & Ross, M. (2010). Why women apologize more than men: Gender differences in thresholds for perceiving offensive behavior. *Psychological Science*, *21*(11), 1649–1655. https://doi.org/10.1177/0956797610384150

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019). "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. *ACM/IEEE International Conference on Human-Robot Interaction*, *2019-March*, 57–65. https://doi.org/10.1109/HRI.2019.8673169

Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. C. (2017). Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research*, *46*, 13–25. https://doi.org/10.1016/j.cogsys.2017.02.003

Selkowitz, A. R., Lakhmani, S. G., Larios, C. N., & Chen, J. Y. C. (2016). Agent transparency and the autonomous squad member. *Proceedings of the Human Factors and Ergonomics Society*, *2014*, 1318–1322. https://doi.org/10.1177/1541931213601305

Serholt, S., & Barendregt, W. (2016). Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. *ACM International Conference Proceeding Series*, *23-27-Octo*. https://doi.org/10.1145/2971485.2971536

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694–696. https://doi.org/10.1038/s41562-017-0202-6

Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. *Frontiers in Psychology*, *10*(MAY), 1–7. https://doi.org/10.3389/fpsyg.2019.01117

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109–124. https://doi.org/10.17705/1thci.00131

Shneiderman, B. (2021). Human-Centered AI. *ISSUES IN SCIENCE AND TECHNOLOGY*, 43–44. https://doi.org/10.2307/j.ctv1s5nzbk.19

Sims, V. K., Chin, M. G., Lum, H. C., Upham-Ellis, L., Ballion, T., & Lagattuta, N. C. (2009). Robots' auditory cues are subject to anthropomorphism. *Proceedings of the Human Factors and Ergonomics Society*, *3*, 1418–1421. https://doi.org/10.1518/107118109x12524444079352

Söllner, M., & Pavlou, P. A. (2016). A longitudinal perspective on trust in it artefacts. *24th European Conference on Information Systems, ECIS 2016*, *June*.

Soltanzadeh, S. (2022). Strictly Human: Limitations of Autonomous Systems. *Minds and Machines*, *32*(2), 269–288. https://doi.org/10.1007/s11023-021-09582-7

Stephens, S. (2023). *NYPD Launches Knightscope Security Robot Service in Manhattan Subway*. Business Wire. https://www.businesswire.com/news/home/20230922025249/en/NYPD-Launches-Knightscope-Security-Robot-Service-in-Manhattan-Subway

Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. J. (2020). The IMPACT of Agent Transparency on Human Performance. *IEEE Transactions on Human-Machine Systems*, *50*(3), 245–253. https://doi.org/10.1109/THMS.2020.2978041

Syrdal, D. S., Dautenhahn, K., Woods, S. N., Walters, M. L., & Koay, K. L. (2007). Looking good? Appearance preferences and robot personality inferences at zero acquaintance. *AAAI Spring Symposium - Technical Report*, *SS-07-07*, 86–92.

Tannenbaum, S. I., Beard, R. L., & Salas, E. (1992). Team Building and its Influence on Team Effectiveness: An Examination of Conceptual and Empirical Developments. *Advances in Psychology*, *82*(C), 117–153. https://doi.org/10.1016/S0166-4115(08)62601-1

Tenhundfeld, N. L., de Visser, E. J., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2020). Trust and Distrust of Automated Parking in a Tesla Model X. *Human Factors*, *62*(2), 194–210. https://doi.org/10.1177/0018720819865412

Teo, G., Wohleber, R., Lin, J., & Reinerman-jones, L. (2019). *The Relevance of Theory to Human-Robot Teaming Research and Development*. *784*(January). https://doi.org/10.1007/978-3-319-94346-6

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T. M., Dixon, C., & Tielman, M. L. (2020). Taxonomy of trust-relevant failures and mitigation strategies. *ACM/IEEE International Conference on Human-Robot Interaction*, 3–12. https://doi.org/10.1145/3319502.3374793

Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*. https://doi.org/10.1016/j.jm.2003.01.003

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, *1*(4), 100049. https://doi.org/10.1016/j.patter.2020.100049

Turner, A., Kaushik, M., Huang, M.-T., & Varanasi, S. (2020). *Calibrating Trust in AI-Assisted Decision Making*.

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. https://doi.org/10.15358/0340-1650-2006-6-331

Tzeng, J. Y. (2004). Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human Computer Studies*. https://doi.org/10.1016/j.ijhcs.2004.01.002

Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of Subjective Uncertainty: Clues from natural language. *Journal of Experimental Psychology: General*, *145*(10), 1280–1297. https://doi.org/10.1037/xge0000202

Ullrich, D., Butz, A., & Diefenbach, S. (2021). The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction. *Frontiers in Robotics and AI*, *8*(April), 1–15. https://doi.org/10.3389/frobt.2021.554578

van de Merwe, K., Mallam, S., & Nazir, S. (2022). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*. https://doi.org/10.1177/00187208221077804

van den Bosch, K., & Bronkhorst, A. (2018). Human-AI Cooperation to Benefit Military Decision Making. *Human-AI Cooperation to Benefit Military Decision Making*, *STO-MP-IST*, 1–12.

van Dongen, K., & van Maanen, P. P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human Computer Studies*, *71*(4), 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018

Walters, M. L., Koay, K. L., Syrdal, D. S., Dautenhahn, K., & Te Boekhorst, R. (2009). Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials. *Adaptive and Emergent Behaviour and Complex Systems - Proceedings of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2009*, 136–143.

Wang, N., Pynadath, D. V., & Hill, S. G. (2015). Building Trust in a Human-Robot Team with Automatically Generated Explanations. *Interservice/Industry Training, Simulation, and Education Conference*, *15315*, 1–12.

Wang, N., Pynadath, D. V., Rovira, E., Barnes, M. J., & Hill, S. G. (2018). Is it my looks? Or something i said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10809 LNCS*, 56–69. https://doi.org/10.1007/978-3-319-78978-1_5

Waytz, A., Cacioppo, J. T., & Epley, N. (2008). Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism. *Bone*, *23*(1), 1–7. https://doi.org/10.1177/1745691610369336.Who

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Werkhoven, P., Kester, L., & Neerincx, M. A. (2018). Telling autonomous systems what to do. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3232078.3232238

Westerbeek, H., & Maes, A. (2013). Route-external and Route-internal Landmarks in Route Descriptions: Effects of Route Length and Map Design. *Applied Cognitive Psychology*, *27*(3), 297–305. https://doi.org/10.1002/acp.2907

Winkler, K. (2024). *It takes a village to regulate AI*. Australian Strategic Policy Institute. https://www.aspistrategist.org.au/it-takes-a-village-to-regulate-ai/

Wischnewski, M., Krämer, N. C., & Müller, E. (2023). Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Conference on Human Factors in Computing Systems - Proceedings* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581197

Wright, J. L., Chen, J. Y. C., Barnes, M. J., & Hancock, P. A. (2016). *The Effect of Agent Reasoning Transparency on Automation Bias: An Analysis of Response Performance* (Vol. 9740, pp. 465–477). https://doi.org/10.1007/978-3-319-39907-2_45

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, *19*(3), 353–374. https://doi.org/10.1080/1463922X.2016.1260181

Xie, Y., & Peng, S. (2009). How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology and Marketing*, *26*(7), 572–589. https://doi.org/10.1002/mar.20289

Xiong, W., Fan, H., Ma, L., & Wang, C. (2022). Challenges of human—machine collaboration in risky decision-making. *Frontiers of Engineering Management*, *9*(1), 89–103. https://doi.org/10.1007/s42524-021-0182-0

Xu, J., & Howard, A. M. (2022). Evaluating the Impact of Emotional Apology on Human-Robot Trust. *RO-MAN 2022 - 31st IEEE International Conference on Robot and Human Interactive Communication: Social, Asocial, and Antisocial Robots*, 1655–1661. https://doi.org/10.1109/RO-MAN53752.2022.9900518

Xu, X., & Sar, S. (2018). Do We See Machines the Same Way As We See Humans? A Survey on Mind Perception of Machines and Human Beings. *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, 472–475. https://doi.org/10.1109/ROMAN.2018.8525586

Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation. *Human Factors*, *00*(0), 1–17. https://doi.org/10.1177/00187208211034716

Ye, Y., You, H., & Du, J. (2023). Improved Trust in Human-Robot Collaboration With ChatGPT. *IEEE Access*, *11*(June), 55748–55754. https://doi.org/10.1109/ACCESS.2023.3282111

Zhan, X., Xu, Y., & Sarkadi, S. (2023). Deceptive AI Ecosystems: The Case of ChatGPT. *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI 2023*. https://doi.org/10.1145/3571884.3603754

Zhang, X., Lee, S. K., Kim, W., & Hahn, S. (2023). “Sorry, It Was My Fault”: Repairing Trust in Human-Robot Interactions. *International Journal of Human-Computer Studies*, *175*(March), 103031. https://doi.org/10.1016/j.ijhcs.2023.103031

Zhang, X., Lee, S. K., Maeng, H., & Hahn, S. (2023). Effects of Failure Types on Trust Repairs in Human–Robot Interactions. *International Journal of Social Robotics*, *15*(9–10), 1619–1635. https://doi.org/10.1007/s12369-023-01059-0

Summary

People are increasingly working with Artificial Intelligence (AI) agents, whether as software-based systems such as AI-chatbots and voice assistants, or embedded in hardware devices like autonomous vehicles, advanced robots, and drones. The idea of Human-AI (H-AI) collaboration is promising, since humans and AI possess complementary skills that, when combined, can enhance performance beyond the capabilities of its individual members. Here, the real challenge is not just determining which tasks are better suited for humans or machines working independently, but in finding ways to enhance their respective strengths through effective interaction. Working together towards a common goal requires good cooperation, coordination, and communication, and it is within these areas that the true challenges lie.

A key component in these activities is trust, as it allows individuals to depend on each other's contributions to complete tasks and achieve shared goals. More specifically, maintaining balanced trust (i.e., neither too much nor too little) is crucial for safe and effective H-AI collaborations. Finding this balance, a process known as trust calibration, should enable people to determine when to rely on AI agents and when to override them. To facilitate this, we need to understand how H-AI trust is built, breaks down, and recovers (i.e., the 'trust lifecycle'). This dissertation focusses on how to maintain H-AI trust, by examining how trust breaks down (i.e., trust violations) and the mechanisms through which trust can be repaired.

In this thesis, I cover three types of trust violations, stemming from 1) inadequate abilities of the AI agent (errors), 2) unexpected behaviour without any explanation, and 3) priority misalignment. In other words, violations in respect to *what* an AI agent does, *how* it operates, and *why* it acts in a certain way. Additionally, we examined the impact of various trust-repair mechanisms on the development of H-AI trust. We evaluated a preventative measure designed to mitigate potential trust issues by proactively communicating uncertainty (i.e., "environment detected as clear, with 80% certainty") and reactive strategies addressing trust violations post-incident, such as expressing regret (i.e., "I am sorry") or providing explanations for anomalous behaviour. These strategies can be categorized as informational (e.g., uncertainty, explanations) or affective (e.g., regret), aiming either to improve the AI agent's interpretability or restore trust through emotional engagement. In short, we investigate how the nature of a trust violation and different repair strategies influence the development of H-AI trust.

Data for these studies were obtained using a series of custom-designed, game-like virtual task environments, simulating military scenarios where participants carried out missions in collaboration with an AI agent, presented in various physical forms. In each study, we used repeated measures of H-AI trust to track its changes over time.

Chapter 2 and 3 examine trust violations due to the inadequate abilities of the AI agent. In Chapter 2, participants were assigned to return to basecamp as fast as possible after running out of ammunition. Halfway, the AI agent failed to warn the participant for an approaching enemy. Following this failure, the AI agent employed one of four trust repair strategies: an explanation or an expression of regret either individually, combined,

or neither. H-AI trust recovered only when the apology included an expression of regret, with even greater recovery when both regret and explanation were offered.

Chapter 3 involves house-searches in two abandoned buildings, supported by a small drone. Halfway, the AI agent failed to warn the participant for a hazard. We studied the effects of uncertainty communication and apology (i.e., explanation + regret), deployed before and after trust had been violated respectively. We conducted this study with both civilian and military samples to investigate whether findings were consistent across different participant groups. Results showed that (a) communicating uncertainty led to more trust, (b) an incorrect advice by the agent led to a less severe decline in trust when that advice included a notion of uncertainty, and (c) after a trust violation, trust recovered significantly more when the agent offered an apology. The two latter effects were only found in the study with civilians.

Chapter 4 examines a trust violation due to unexpected behaviour and the AI agent's incapacity to explain itself. Halfway a reconnaissance mission, the AI agent detected a faster alternative route that emerged due to changes in the environment (i.e., the river had dried up) and decided to deviate from the original plan. We studied the effect of transparency (i.e. regular status updates and an explanation for the deviation) and outcome on trust and the participant's workload. The main result was that transparency prevented a trust violation and contributed to higher levels of trust, without increasing subjective workload.

Chapter 5 examines a trust violation caused by priority misalignment. Halfway during the mission, the AI agent, who was guiding the participant, did not warn the participant in time of a hazard down the road. In one condition, it explained that this failure was due to an underperforming sensor. In the other condition, the AI agent explained that it deliberately recommended the faster route over the safer one. The rationale was that the rest of the team was waiting, and further delays could jeopardize both the team and the mission. Our findings suggest that trust violations due to choices are harder to repair than those due to errors.

By analysing the dynamics of trust during H-AI interaction, this research aims to inform the design of AI-systems that promote calibrated trust in high-stakes environments. As AI agents gain decision-making authority in the physical and virtual world, they will increasingly face conflicting human values (e.g., privacy vs. safety, efficiency vs. empathy). As they get more autonomous and complex, moral considerations will play a larger role, and trust may be lost not only due to malfunctions but also due to miscommunication and misaligned values. The trustworthiness of an AI agent is no longer determined solely by what it can do, but also by how and why it does so. Our findings support the growing consensus that H-AI trust, much like interpersonal trust, is multidimensional, even if the moral dimensions are not yet as apparent in current interactions. As the complexity of H-AI trust grows, maintaining an appropriate level of trust becomes increasingly important. Designing and developing trustworthy AI agents for safe and effective H-AI collaborations requires a systematic and multidisciplinary approach.

Samenvatting

Mensen werken steeds vaker samen met kunstmatige intelligentie (KI) agenten. KI-agenten kunnen softwarematig werken zoals AI-chatbots en spraakassistenten of geïntegreerd zijn in hardware zoals autonome voertuigen, geavanceerde robots en drones. Het idee van mens-KI (M-KI) samenwerkingen is veelbelovend, omdat mensen en KI complementaire vaardigheden bezitten waardoor ze collectief meer kunnen bereiken dan elk afzonderlijk had gekund. Daarbij ligt de uitdaging niet in de taken zo te verdelen dat beiden doen waar ze goed in zijn, maar uitvinden hoe mensen en machines het beste in elkaar naar boven halen door te interacteren. Samenwerken aan een gemeenschappelijk doel vereist coöperatie, coördinatie en communicatie; daar liggen de echte uitdagingen.

Vertrouwen is essentieel voor werken in teamverband omdat het mensen in staat stelt op elkaars bijdragen te rekenen. Het bewaren van een balans hierin, dus niet te veel en niet te weinig vertrouwen, is een voorwaarde voor veilige en effectieve samenwerking tussen mens en KI. Het vinden van deze balans oftewel het *kalibreren* van vertrouwen, moet mensen in staat stellen te bepalen wanneer ze iets aan KI kunnen overlaten. Om dit te kunnen faciliteren moeten we eerst begrijpen hoe M-KI vertrouwen zich ontwikkelt. Dit proefschrift focust zich op de vraag hoe we M-KI vertrouwen *behouden* door te onderzoeken hoe vertrouwen afbreekt en welke mechanismen bestaan om het te herstellen.

Ik behandel drie soorten vertrouwensschendingen: als gevolg van 1) verminderde capaciteiten van de KI-agent (fouten), 2) onverwacht gedrag zonder uitleg van de KI-agent, en 3) conflicterende prioriteiten. Met andere woorden schendingen met betrekking tot *wat* een KI-agent doet, *hoe* het opereert en *waarom* het op een bepaalde manier handelt. Daarnaast onderzochten we verschillende herstelmechanismen: een preventieve maatregel om een mogelijke vertrouwensbreuk te beperken door proactief onzekerheid te communiceren (bijvoorbeeld “omgeving beoordeeld als veilig met 80% zekerheid”) en reactieve strategieën die na een incident werden ingezet zoals het betuigen van spijt of het geven van verklaringen voor afwijkend gedrag. Deze strategieën kunnen worden gecategoriseerd als informatief (bijvoorbeeld onzekerheid en verklaringen) met als doel de interpreteerbaarheid van de KI-agent te verbeteren of als affectief (bijvoorbeeld spijt) gericht op emotionele betrokkenheid. Kortom, we onderzochten hoe de aard van een vertrouwensschending en verschillende herstelstrategieën de ontwikkeling van M-KI vertrouwen beïnvloeden.

De data in mijn studies zijn verzameld met behulp op maat gemaakte computerspel-achtige virtuele omgevingen, waarin militaire scenario’s werden gesimuleerd waarin deelnemers missies uitvoerden in samenwerking met een KI-agent in verschillende fysieke verschijningsvormen. In elke studie werd M-KI vertrouwen herhaaldelijk gemeten zodat we de veranderingen in de tijd konden volgen.

Hoofdstuk 2 en 3 onderzoeken vertrouwensschendingen als gevolg van fouten. In hoofdstuk 2 kregen deelnemers de opdracht om zo snel mogelijk terug te keren naar het basiskamp. Halverwege waarschuwde de KI-agent de deelnemer niet voor een naderende vijand. Na dit falen, gebruikte de KI-agent een van vier herstelstrategieën:

1) een verklaring of2) een uiting van spijt afzonderlijk, 3) gecombineerd of 4) geen van beide. Vertrouwen herstelde alleen wanneer de KI-agent spijt betuigde; dit effect was nog sterker wanneer het óók een verklaring gaf.

Hoofdstuk 3 betreft twee huiszoekingen in verlaten gebouwen met behulp van een kleine drone. Halverwege waarschuwde de KI-agent de deelnemer niet voor een gevaar. We bestudeerden de effecten van onzekerheidscommunicatie en verontschuldigingen (dat wil zeggen een verklaring + spijt) die respectievelijk voor en na het incident werden ingezet. Deze studie werd uitgevoerd met zowel civiele als militaire deelnemers om te onderzoeken of de bevindingen consistent waren tussen verschillende groepen deelnemers. De resultaten lieten zien dat 1) onzekerheidscommunicatie leidde tot meer vertrouwen, 2) onjuiste adviezen leidden tot een minder ernstige daling van het vertrouwen wanneer die onzekerheidsinformatie bevatte en 3) het vertrouwen na een vertrouwensschending significant meer herstelde wanneer de agent een verontschuldiging aanbood. De laatste twee effecten werden alleen gevonden in de studie met civiele deelnemers.

Hoofdstuk 4 onderzoekt een vertrouwensschending door onverwacht gedrag en het ontbreken van toelichting door de KI-agent. Halverwege een verkenningsmissie observeerde de KI-agent dat een snellere alternatieve route was ontstaan door veranderingen in de omgeving (de rivier was opgedroogd) en besloot af te wijken van het oorspronkelijke plan. We bestudeerden het effect van transparantie (dat wil zeggen regelmatige statusupdates en een verklaring voor de afwijking) en de uitkomst van de missie op vertrouwen en de werklast van de deelnemer. Het belangrijkste resultaat was dat transparantie een breuk in vertrouwen voorkwam zonder te zorgen voor een toename van de subjectieve werklast.

Hoofdstuk 5 onderzoekt een vertrouwensschending door conflicterende prioriteiten. Halverwege de missie waarschuwde de KI-agent die de deelnemer leidde, deze niet voor een gevaar op de weg. In één conditie legde de KI-agent uit dat dit te wijten was aan een slecht functionerende sensor. In de andere conditie legde het uit dat het bewust de snellere route had aanbevolen in plaats van de veiligere, omdat vertragingen de rest van het team en de missie in gevaar konden brengen. Onze bevindingen suggereren dat breuken in vertrouwen als gevolg van afwegingen moeilijker te herstellen zijn dan die als gevolg van fouten.

Door de dynamiek van vertrouwen tijdens M-KI interacties te onderzoeken, hopen we bij te dragen aan het ontwerp van KI-agenten die gekalibreerd vertrouwen bevorderen. Naarmate KI-agenten autonomer worden en meer beslissingsbevoegdheid krijgen in zowel de fysieke als virtuele wereld zullen ze steeds vaker met conflicterende waarden te maken krijgen (bijvoorbeeld privacy vs. veiligheid, efficiëntie vs. empathie) en zullen vertrouwensbreuken niet alleen ontstaan door fouten, maar ook door miscommunicatie en conflicterende belangen. De betrouwbaarheid van een KI-agent wordt niet langer uitsluitend bepaald door wat het kan, maar ook door hoe en waarom het iets doet. Onze bevindingen sluiten aan bij de groeiende consensus dat M-KI vertrouwen net als vertrouwen tussen mensen multidimensionaal is. De toenemende complexiteit van M-KI

vertrouwen maakt het behouden van een passend niveau van vertrouwen tot een steeds grotere uitdaging. Het ontwerpen en ontwikkelen van betrouwbare KI-agenten voor veilige en effectieve M-KI samenwerkingen vraagt om een systematische en multidisciplinaire aanpak.

Curriculum Vitae

Work experience

R&D Engineer | Human-AI Teaming in Aerospace (2025 – present)
Netherlands Aerospace Centre (NLR), Amsterdam
Department: Aerospace Operations - Safety & Human Performance
Scientist Specialist | Human-Machine Teaming in Defense (2018 – 2025)
TNO Soesterberg
Department: Human-Machine Teaming
PhD | Maintaining Human-AI Trust (0.5 FTE) (2020 – 2024)
University of Twente
Department: Psychology of Conflict, Risk & Safety
Graduate research intern | **Dashboard Design** (2017 – 2018)
TNO Soesterberg | Perceptual and Cognitive Systems

Education

Master of Science in Applied Cognitive Psychology, Utrecht University (2017 – 2018)
Faculty of Social Sciences
Bachelor of Science in Psychology, Utrecht University (2013 – 2017)
Specialisation: Cognitive and Neurobiological Psychology
Faculty of Social Sciences
Minor in Information Science, Utrecht University (2015 – 2016)
Faculty of Science
Erasmus Exchange, Lund, Sweden (2016 – 2017)
Faculty of Social Sciences
VWO, Alberdingk Thijm College, Hilversum (2007 – 2013)
Nature & Health w/ Informatics and Spanish

Publications

Kox, E. S., Barnhoorn, J., Rábago Mayer, L., Temel, A., & Klunder, T. (2022). Using a Virtual Reality House-Search Task to Measure Trust During Human-Agent Interaction (Demo Paper). *HHAI2022: Augmenting Human Intellect*, 272–274. https://doi.org/10.3233/FAIA220214

Kox, E. S., & Beretta, B. (2024). Evaluating Generative AI Incidents: An Exploratory Vignette Study on the Role of Trust , Attitude and AI Literacy. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, 188–198. https://doi.org/10.3233/FAIA240194

Kox, E. S., Finlayson, N. B., Broderick-Hale, J. C., & Kerstholt, J. H. (2023). Calibrated Trust as a Means to Build Societal Resilience Against Cognitive Warfare. *Proceedings of NATO Symposium HFM-361 on Mitigating and Responding to Cognitive Warfare*, 1–15.

Kox, E. S., Hennekens, M., Metcalfe, J. S., & Kerstholt, J. H. (n.d.). Trust Violations Due to Error or Choice: the Differential Effects on Trust Repair in Human-Human and Human-Robot Interaction. *Transactions on Human-Robot Interaction*.

Kox, E. S., Kerstholt, J. H., Hueting, T., Barnhoorn, J., & Eikelboom, A. (2019). Autonomous Systems As Intelligent Teammates: Social Psychological Implications. *024th International Command and Control Research & Technology Symposium*.

Kox, E. S., Kerstholt, J. H., Hueting, T., & de Vries, P. W. (2021). Trust Repair in Human-Agent Teams: the Effectiveness of Explanations and Expressing Regret. *Autonomous Agents and Multi-Agent Systems*, *35*(2), 1–20. https://doi.org/10.1007/s10458-021-09515-9

Kox, E. S., Siegling, L. B., & Kerstholt, J. H. (2022). Trust Development in Military and Civilian Human-Agent Teams: the Effect of Social-Cognitive Recovery Strategies. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00871-4

Kox, E. S., van den Boogaard, J., Turjaka, V., & Kerstholt, J. H. (2024). The Journey or the Destination: The Impact of Transparency and Goal Attainment on Trust in Human-Robot Teams. *Transactions on Human-Robot Interaction*. https://doi.org/https://doi.org/10.1145/3702245

Kox, E. S., van Riemsdijk, M. B., de Vries, P. W., & Kerstholt, J. H. (2024a). Red or Blue Door: Exploring the Behavioural Consequences of Trust Violations due to Robot Error or Choice Using a VR Maze. *Retrieved from Osf.Io/J6vwk*. https://doi.org/DOI 10.17605/OSF.IO/J6VWK

Kox, E. S., van Riemsdijk, M. B., de Vries, P. W., & Kerstholt, J. H. (2024b). The Impact of Anthropomorphic Cues and Explanations on Trust Formation, Violation, and Repair in HRI: Insights from a VR Experiment. *Retrieved from Osf.Io/62gte*. https://doi.org/DOI 10.17605/OSF.IO/62GTE

Kox, E. S., van Riemsdijk, M. B., & Kerstholt, J. H. (2024). From Exploitation to Augmentation: Navigating AI's Impact on Human Value- and Inference-based Decision-Making. *Proceedings of NATO HFM-RSY-377 Symposium on "Meaningful Human Control of Future Military Operations: Spanning Across Warfare Domains with Advanced AI,"* 1–12.