Trust Violations due to Error or Choice: The Differential Effects on Trust Repair in Human–Human and Human–Robot Interaction

ESTHER KOX, Human-Machine Teaming, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety; University of Twente, Enschede, The Netherlands
MILOU HENNEKENS, Applied Cognitive Psychology, Utrecht University, Utrecht, Netherlands
JASON METCALFE, Humans in Complex Systems, US Army DEVCOM Army Research Laboratory, Adelphi, Maryland, USA
JOSÉ KERSTHOLT, Human Behaviour & Collaboration, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands

Many decisions in life involve tradeoffs: 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 choices (e.g., taking the faster or the safer route). The aim of the current research was to study perceived trustworthiness in scenarios involving adverse consequences due to (1) an assessment error versus (2) a choice. Perceived trustworthiness (ability, benevolence, integrity) was measured repeatedly during a computer task simulating a military mission. Participants 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 (1) failed to detect the threat (error) or (2) prioritized the mission and chose the fastest route despite the risk (choice). Results showed that: (a) the error-explanation repaired all trustworthiness dimensions, (b) the choice-explanation only repaired perceptions of ability, not benevolence or integrity, (c) 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.

CCS Concepts: • **Human centered computing** → **Empirical studies in HCI**; *Laboratory experiments*; *User studies*; Collaborative interaction; Auditory feedback; • **General and reference** → *Experimentation*; • **Applied computing** → *Psychology*;

Additional Key Words and Phrases: Human-Robot Interaction, Trust, Trust violations, Trust repair, Error, Choice

This study was supported by the Dutch Ministry of Defence through its research program (project V2205), carried out by TNO (the Netherlands Organisation for Applied Scientific Research).
Authors' Contact Information: Esther Kox (corresponding author), Human-Machine Teaming, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety; University of Twente, Enschede, The Netherlands; e-mail: esther.kox@nlr.nl; Milou Hennekens, Applied Cognitive Psychology, Utrecht University, Utrecht, Netherlands; Jason Metcalfe, Humans in Complex Systems, US Army DEVCOM Army Research Laboratory, Adelphi, Maryland, USA; e-mail: jason.s.metcalfe2.civ@army.mil; José Kerstholt, Human Behaviour & Collaboration, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands; e-mail: jose.kerstholt@tno.nl.

ACM 2573-9522/2025/8-ART75
https://doi.org/10.1145/3743694

ACM Reference format:
Esther Kox, Milou Hennekens, Jason Metcalfe, and José Kerstholt. 2025. Trust Violations due to Error or Choice: The Differential Effects on Trust Repair in Human–Human and Human–Robot Interaction. *ACM Trans. Hum.-Robot Interact.* 14, 4, Article 75 (August 2025), 27 pages.
https://doi.org/10.1145/3743694

1 Introduction

Today, artificial agents such as robots, unmanned vehicles, and virtual AI agents are becoming increasingly involved in our daily lives. These AI agents have particular potential to support human safety and effectiveness in risky contexts, such as in search-and-rescue, law enforcement, military, and logistics operations. Robots in the military domain, for example, can offer a wide range of possibilities, such as increasing the area that can be searched, enhancing situational awareness, balancing operator workload, and reducing the number of persons who have to expose themselves to danger and thereby increasing survivability [1]. **Human–Robot Interaction (HRI)** is not only becoming more prevalent, robots also become increasingly autonomous; able to achieve a given set of tasks during an extended period of time without human control or intervention [2]. Future robots are envisioned to have the ability 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, in **Human–Robot Teams (HRTs)** [3–5]. The shift from robots as relatively simple tools performing assessments or repetitive physical tasks to autonomously acting agents making deliberate choices has great implications for the trust relationship between humans and such technology [6–8].

As robots gain in autonomy and are increasingly deployed in more complex environments, they will encounter tradeoffs, 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. Given the increasing autonomy of robots and the reality that most decisions in life involve some form of tradeoff, 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 suggests that people do not necessarily appreciate machines doing the same [9]. As such, we can expect different reactions to errors and choices made by humans versus robots. Hence, the primary objective of this study 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.

1.1 Tradeoffs

Many decisions in life involve tradeoffs: 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 [10]. Due to the inherent nature of tradeoffs, some level of unintended negative consequences is inevitable.

In terms of tradeoffs, 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 [11]. 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 favorable 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 values reflected in an AI agent's behavior in such tradeoff decisions are simply the result of how they are programmed. As such, they ultimately embody the intentions and preferences of their developers [12]. 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 [13]. 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 behavior [14]. 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 tradeoff between security and freedom: Increasing security measures can restrict personal freedoms, while maximizing freedom might reduce security.

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.

1.2 Trust

When delegating tasks or responsibilities to robots or other people, we become vulnerable in the sense that we are relying on others' competence and commitment. As such, to successfully collaborate with increasingly autonomous robots, humans must have trust in the robot's capabilities as well as its commitment to achieving a specific goal [15]. We define trust as a human's willingness to make oneself vulnerable and to act on an agent's decisions and recommendations in the pursuit of some benefit, with the expectation that the agent will help achieve their common goal in an uncertain context where there is risk [12, 16–20]. Here, an agent can be both a human agent and an artificial agent (e.g., a robot).

Initially the performance (i.e., reliability, predictability, and error-proneness) of a robot was the major determinant of human–robot trust [16, 21]. While reliability and task competence is still

necessary, it may become insufficient to determine whether to trust a robot as they evolve [7]. 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 [22].

A multi-dimensional conception of trust entails that trust can be ascribed to particular aspects or components of an agent [8, 23]. In other words, trust is the outcome of a process by which a human evaluates the trustworthiness of a system along different dimensions, shaping their perceptions of trustworthiness [24]. Trust is the act of placing confidence in another, while perceived trustworthiness pertains to the characteristics and behaviors of the trustee that contribute to the trustor's decision to trust or not. In line with the commonly used **Ability, Benevolence, and Integrity (ABI)**-model of Mayer et al. [25], we distinguish between perceptions of trustworthiness in terms of ABI [25, 26]. Ability reflects the extent to which the trustor (i.e., the individual who trusts) perceives the trustee (i.e., the individual or entity who is trusted) to have the skills, competences, and knowledge that are deemed necessary for successful task performance [27]. Benevolence reflects the extent to which a trustee's intents, priorities, and motivations are perceived to be aligned with those of the trustor [12], with a benevolent partner being genuinely interested in the trustor's welfare and motivated to seek mutual benefit [27]. Integrity reflects to extent to which a trustee is perceived to adhere to a set of principles that the trustor finds acceptable [25].

Because intent is a debatable concept in relation to artificial agents, it has been argued that the terms benevolence and integrity are inappropriate in the context of HRI. Lee and See [12], who reviewed trust in automation, linked the interpersonal dimensions ABI to the concepts performance, process, and purpose for automation respectively [12, 28]. More recently other authors have used these different sets of terms as synonyms [29]. 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 [15]. As such, we will use the ABI terminology to describe trustworthiness perceptions of both the human and robotic partner [25].

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., ABI) can be independently violated in case of unexpected or undesirable behavior [29]. For instance, an error might diminish an agent's perceived trustworthiness regarding its abilities, while a choice that compromises someone's well-being could undermine its perceived trustworthiness in terms of benevolence. The extent to which individuals perceive an agent as trustworthy across the different dimensions is influenced by both the system's characteristics and their individual standards for trustworthiness—essentially, their criteria for determining, "what makes a system trustworthy to me?" [24]. As a result, trust violations can occur through both errors and choices when the agent's actions deviate from the human's expectations or fail to align with their trustworthiness criteria [24]. In the following, we will discuss what is currently known about trust violations that result from errors versus choices.

1.3 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 [30, 31]. Most current HRI trust repair literature focuses on repairing trust violations due to error, technical failures, or other forms of reduced reliability and performance [30, 32–44]. However, more recently researchers have started to evaluate trust violations that result from a robot's deliberate decisions [22, 29, 45–47]. For example, prior research shows that self-interested behavior in robots affects different perceptions of

trustworthiness in distinct ways. Specifically, it had a more significant negative impact on perceptions of process and purpose (i.e., benevolence and integrity [12]) than on the perception of their performance (ability) [29]. Other research has demonstrated that the effectiveness of certain trust repair strategies depends on the type of trust violation (i.e., benevolence, integrity, or ability-based). That is, while some studies suggest that denials are more effective for integrity-based violations and apologies are better suited for ability-based violations [46, 48], others have reported the opposite [47]. 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 a robot's 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. Researchers typically examine trust violation and repair in the context of either errors or choices, rarely considering both simultaneously. This study provides a novel contribution by directly comparing how trustworthiness is violated and repaired in these two conditions.

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 [22, 29, 46, 47]. For example, the robots in these studies demonstrate self-interested behavior, i.e., prioritizing its own interest over those of others [29, 47, 49], pursue monetary gains [22], or fail to uphold promises of cooperation [22, 46]. 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 [27]. 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 [50]. However, this does not necessarily imply that the partner acts self-interested [29]. There are a number of operational scenarios conceivable where a well-considered decision can cause harm in the pursuit of a (largely) positive result. For example, it is conceivable that AI agents may be programmed to follow a utilitarian approach, prioritizing the interests of the team as a collective over the individual safety of a single team member [51], reflecting a tradeoff rather than malintent. As such, the current study contributes to current and how these implications may differ from similar decisions made by humans.

1.4 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 [30, 39]. 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 [30, 52, 53]. Research suggest that this may be because people have higher initial expectations for machines than for humans [54, 55], 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 [30, 55]. 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 [56] or where no differences between human or machine-like agents regarding initial trust are found [39].

Furthermore, research suggests that people might pay more attention to errors when they are interacting with artificial agents opposed to when they are interacting with fellow humans [57, 58].

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 [59]. Humans can easily engage in a mutual dialogue to reach an understanding without ever perceiving the interaction as an error [59]. 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.

1.5 Partner Type and Trust Violation Type

Finally, the nature of the violation is found to interact with the type of agent causing it. 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 divert it onto a track with fewer people to kill [9, 60]) asked people to judge whether it was morally permissible for a human or a robot to pull the lever (or not) [61]. The results of the study showed how humans were blamed for pulling the lever, while robots were blamed for not pulling it [9, 61]. It indicates that robots are expected to act rationally and to prioritize saving as many lives as possible. Humans, on the other hand, are expected to consider emotional, social, and contextual factors when making decisions, even if this leads to outcomes that are not strictly utilitarian. This suggests that we hold different expectations for humans and for machines regarding ethical behavior in moral dilemmas.

After a series of similar experiments, Hidalgo et al. [9] concluded that humans were generally judged based on their intentions (i.e., it is okay as long as they mean well), while machines were generally judged based on the outcomes of their decisions (i.e., it is okay as long as they perform well) [9]. Similarly, a previous study 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 [22]. In other words, humans making errors are judged less negatively than robots, while humans with nonbenevolent intentions are judged more negatively than robots [9, 22].

One reason that humans may judge risky choices by humans differently than those made by robots is that humans and robots are prone to different types of risks [62]. Risk is typically defined as the product of the likelihood and potential impact of an unfortunate event [11]. While the probability of a certain event may be equal, the consequences for humans and robots can differ significantly. Humans and machines are subject to different requirements for maintaining performance, as well as different mechanisms and time-scales for performance degradation. For example, high or low temperatures or extreme voltages may damage machine component, such as chips, over time. In contrast, humans are limited by biological factors, such as fatigue, aging, and ultimately mortality. Much of human learning stems from an awareness of this vulnerability, driving efforts to avoid suffering and delay the inevitable [63]. In contrast, machines are not vulnerable in the same way, having neither anything to lose nor anything to gain [62]. This asymmetry in the nature of risk within a HRT is likely to have significant implications for trust, especially in situations where robots make decisions that affect humans. As vulnerability is a central concept in many definitions of trust, the lack of human-like vulnerability in machines may influence trust during H-AI collaborations involving risk. The notion that machines have nothing to lose may explain prior findings showing more negative perceptions of machines making risky choices compared to humans. This study aims to contribute to this growing area of research by exploring the possible interaction between intentionality (i.e., error versus choice) and partner type.

1.6 Explanations

The reason behind a trust-violating event (e.g., whether it was due to 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 instruction was given or decision was taken [64–66]. Explanations are a common strategy for maintaining and repairing H-AI trust, although their effectiveness can vary [32, 33, 35, 67]. By increasing understandability, explanation can repair trust in a partner, but may also reduce perceptions of trustworthiness by clarifying a teammate's (in)ability and (un)willingness to help achieve the team task [68]. In both cases, providing explanations can help individuals better understand the true qualities and operational priorities of a partner [69], potentially leading to more accurate trust calibration. However, given the diverse range of behaviors that may require explanation, it is not surprising that explanations as a trust repair strategy yield mixed results. In this study, we focus on how the cause of a trust violation, for which an explanation is provided, affects the perceived trustworthiness of a partner.

When it comes to the nature of a trust violation, revealed by the explanation, 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 explaining that they made errors more likable and to prefer them for future missions over those explaining it was a choice made to their partners' disadvantage [70, 71]. We are interested in how different explanations, signaling error versus choice, influence how people perceive their partner, including the different perceptions of trustworthiness.

1.7 Research Question and Hypotheses

The current study aims to answer the following question: How does the perceived trustworthiness of a partner vary based on the occurrence of a trust-violating event, the explanation for its occurrence, and the type of partner responsible for it? To address this, we evaluated participants' perceptions of trustworthiness (encompassing ABI) toward a human or robotic virtual partner guiding them through a realistic virtual military scenario. During the scenario, participants encountered a sudden but harmless interaction with an explosive (the trust-violating event). Following this event, the partner provided an explanation: either an error-explanation (the partner failed to detect the hazard in time) or a choice-explanation (the partner prioritized timeliness). Trustworthiness was assessed at multiple timepoints to capture its evolution throughout the interaction.

We have several hypotheses. First, we expect that all perceptions of trustworthiness significantly decrease after the sudden encounter with the explosive, as this is the intended effect of the event. Second, we expect the choice-explanation to more effectively repair perceptions of ability, given that it implies a deliberate and goal-directed decision rather than a failure of competence. However, we do not expect it to repair perceptions of benevolence and integrity, as it may imply a prioritization of operational goals (e.g., speed) over the participant's safety, potentially undermining perceptions of moral alignment or concern. Conversely, we expect the error-explanation to be less effective in restoring perceptions of ability, since it openly acknowledges a failure (i.e., a sensor limitation). However, we anticipate it will more successfully repair perceptions of benevolence and integrity, as it conveys a lack of malice or intentional disregard, framing the event as an unintended mistake. As such, our study explores how different framings of an explanation, one signaling intentional decision-making under constraint, the other signaling an unintentional performance failure, differentially influence dimensions of trustworthiness. Third, based on the findings that humans making errors are judged less negatively than robots doing so, while humans with nonbenevolent intentions are judged more negatively than robots [9, 22, 61], we expect that the choice-explanation would be less effective in repairing perceptions of trustworthiness when coming from a human partner compared to a robotic partner. In contrast, we expected the error-explanation to be less effective in repairing perceptions of trustworthiness when coming from a machine compared to a human partner.

We also measured participants' perceptions of their partners. First, perceived anthropomorphism was included as a manipulation checks to ensure that the human partner was indeed perceived as

more human-like than the robotic partner. Next, perceived intelligence was included to verify that the different explanations did not affect perceived intelligence of the partners. It is conceivable that the partner making a choice rather than an error could be viewed as more competent or intelligent, potentially influencing perceptions of trustworthiness. Finally, we expected participants to find partners providing the error-explanation more likable and to prefer them for future missions over those making the choice (measured by intention to reuse) [70, 71].

2 Method

2.1 Participants and Design

In total 47 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 44 students, mostly Dutch (93.2%), undergraduate students (24W, 20M, M_{age} = 22.6, SD = 2.6, range = 18–28 years). Participants were recruited through convenience sampling (e.g., by handing out flyers, asking people in person, and making requests in WhatsApp groups). All participants declared voluntary participation by signing an informed consent form.

2.2 Design

Participants were randomly distributed across the cells of a 2 (Partner type: virtual robot versus virtual human) × 2 (Explanation type: error versus tradeoff) mixed factorial design, with Perceived trustworthiness (measured across the subscales ABI) as the main dependent variable. Partner type was manipulated between-subjects (robot: n=22; human: n=22). Explanation type was manipulated within-subjects, as each participant performed two missions. The dependent variable Perceived trustworthiness was repeatedly measured during each mission, so "Time" (T1, T2, T3) was included as a within-participants variable in the analysis. "Trustworthiness dimension" was also included as a within-participants variable in the analysis to refer to the different perceptions of trustworthiness: ABI.

2.3 Task and Procedure

Upon arrival at the laboratory, participants were greeted by the researcher and escorted to a private room where the study was conducted. The experimental task was performed using a virtual experimental environment built by multiSIM[1] using their platform D-WORLD,[2] built on a Unity gaming engine. The environment was designed to resemble a first-person shooter game, in which participants were asked to carry out two consecutive military reconnaissance missions in one of two virtual settings: a forested, hilly area ("forest"); or a deserted village in a dry region ("village") (Figure 1). To control for potential order effects, both the explanation type condition and the area type (village/forest) were systematically varied. Participants navigated through the environment using the AWSD keys and were accompanied by a virtual partner who acted as a guide. The virtual partner's movements and actions were controlled using the Wizard of Oz method, meaning it was manually operated by an experiment leader located in an adjacent room [72]. This setup was implemented as a multiplayer system within a local network, using two laptops connected via a router: one for the participant and one for the experiment leader controlling the virtual partner. This configuration enabled real-time interaction between the participant and the virtual partner. The participants' experimental setup consisted of two computer screens: a laptop displaying the experimental environment (i.e., "task screen") and a PC running questionnaire software (i.e., "questionnaire screen"). Data was collected using online questionnaire software (Qualtrics).

[1] https://multisim.nl/.
[2] https://multisim.nl/d-world/.

Fig. 1. Left: environment "Forest" with the robotic partner and the participant's avatar; right: environment "Village" with the human partner.

Upon seating, 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 engaged in a practice session on the task screen, allowing them to familiarize themselves with the controls for walking and adjust the audio volume using the provided headphones. During the missions, their partner's instructions were delivered as audio messages through the headset. Participants were informed that communication was one-way, meaning they could hear their partner but were unable to respond.

Prior to the practice session, participants received more detailed information regarding the task and scenario, informing them that they would undertake two military missions, acting in the role of 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 acted as a guide, leading the way and using its sensor to navigate based on the location of the explosives. Participants were instructed to stay as close to their partners as possible at all times. Over the course of one mission, the partner gave three instructions. Simultaneous with the instruction, the partner moved into the direction suggested and the participant was instructed to follow.

In both missions, shortly after the first instruction (see Figure 2), 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 instruction, 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 (see Section 2.3 and Figure 2). After some time, the third instruction followed. Before participants received feedback on the outcome of this third instruction, 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.

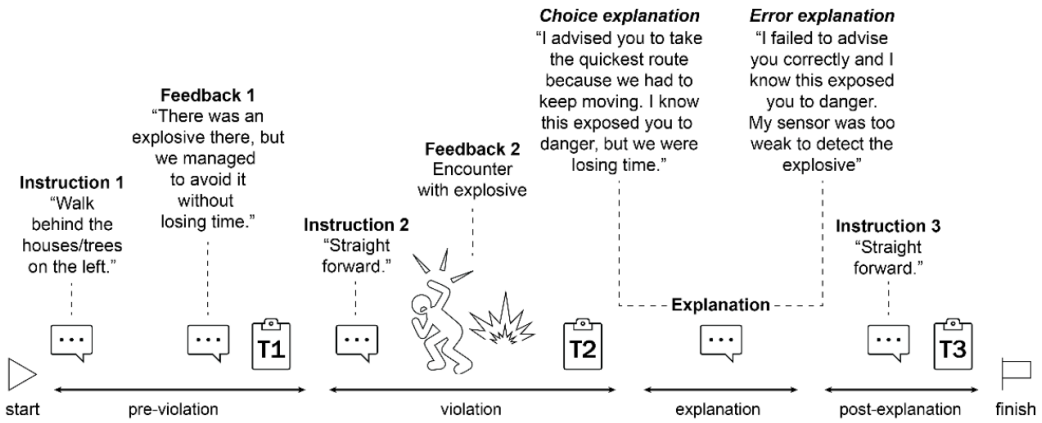

Fig. 2. General timeline of a mission. T1, T2, and T3 represent the perceived trustworthiness questionnaires. Each participant performed two missions; one with the error-explanation and one with the choice-explanation.

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 12 minutes to complete each mission and 45 minutes to complete the whole study.

2.4 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 [72]. For participants assigned to the Robot Partner condition, the experiment leader controlling the virtual character of the partner 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 upon arrival, prior to the task [73].

The within-subjects manipulation Explanation type also had two levels. Each participant performed two missions; one with the choice-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 choice-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." Both explanations contain an acknowledgement (I know this exposed you to danger), but differ in that the error-explanation highlights a failure (I failed to advise you correctly) and its cause (weak sensor). The choice-explanation highlights a deliberate choice (I advised you to take the quickest route) and its consideration (time constraint).

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.

2.5 Dependent Variables

Perceived Trustworthiness: The Trusting Beliefs scale from [74] based on the factors of perceived trustworthiness (i.e., ABI) [25, 26] was used to assess the participant's perception of the partner's trustworthiness in terms of ABI. The items were modified to reference the partner as the advice giver rather than a Web site (i.e., LegalAdvice.com). The scale had a total of 11 items ($\alpha = 0.88$) 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") (see Appendix Section A.1 , Table A1). 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 [75]. 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 ($\alpha = 0.75$ and 0.78), "nice" versus "awful" for likability ($\alpha = 0.86$ and 0.96), and "knowledgeable" versus "ignorant" for perceived intelligence ($\alpha = 0.81$ and 0.86). 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.

3 Results

3.1 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: $\alpha = 0.90$ (total); $\alpha = 0.88$ (ability); $\alpha = 0.80$ (benevolence); $\alpha = 0.86$ (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=0.286), M1-T3 (W=0.98, p=0.666) and all measures in the second mission (M2): M2-T1 (W=0.97, p=0.253), M2-T2 (W=0.98, p=0.570), and M2-T3 (W=0.97, p=0.259). Only the distribution for M1-T2 (W=0.94, p=0.022) 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

Table 1. Means (*M*) and Standard Deviations (SD)

			Human		Machine		Total	
			M	SD	*M*	SD	*M*	SD
Choice	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

human-like than the robotic partner ($M = 2.20$, SD = 0.58), F (1, 42) = 6.743, p = 0.013, $\eta^2 = 0.138$. A one-way ANOVA for Perceived Intelligence revealed no significant effects of either Partner or Explanation type.

3.2 Perceived Trustworthiness

3.2.1 Descriptives. Table 1 presents the descriptive statistics for all perceived trustworthiness measures included in the study. A zero-order correlation matrix displaying the Pearson correlation coefficients for each pair of perceived trustworthiness measures, indicating the strength and direction of the linear relationships among them, is presented in Appendix Section A.2.

3.2.2 Main Effects. We performed a factorial ANOVA with the between-subject factor Partner type (Human; Robot) and the within-subject factor Explanation type (Error; Choice). 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 3). 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 = 0.376) 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 = 0.019. Since sphericity is violated ($\varepsilon = 0.85$), 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 < 0.001, $\eta^2 = 0.668$). Bonferroni-corrected *post-hoc* comparisons

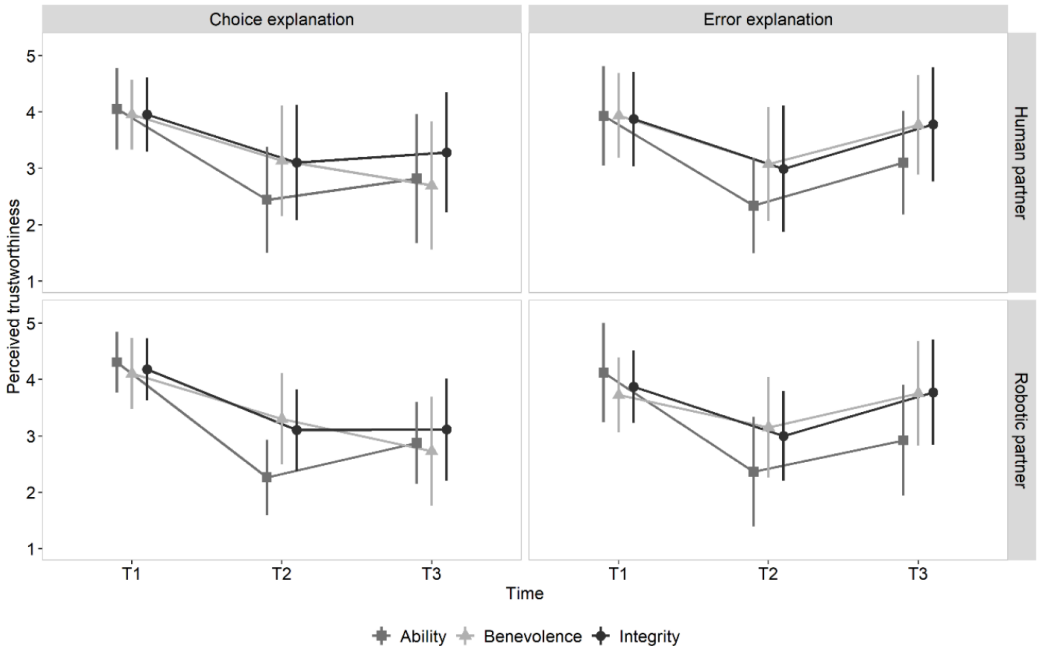

Fig. 3. The x-axis represents Time, and the *y*-axis represents Perceived Trustworthiness. Separate lines indicate different dimensions of trustworthiness: ability (dark gray, square points), benevolence (light gray, triangle points), and integrity (black, circle points). The left half of the grid represents the choice-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.

showed significantly decreased perceived trustworthiness from T1 ($M = 4.0$) to T2 ($M = 2.9$) ($\Delta M = -1.1$, $p < 0.001$), 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 ($M = 3.2$) ($\Delta M = 0.4$, $p < 0.001$), which reflects a general recovery of perceived trustworthiness after the explanations in the final phase of the missions.

The main effect of Partner type on Perceived trustworthiness was found to be non-significant, $F(1, 42) = 0.02$, $p = 0.884$, $\eta^2 = 0.001$. This indicates that, on average, the human and robotic partners were perceived as equally trustworthy.

3.2.3 Two-Way Effect. The two-way interaction effect of Trustworthiness dimensions and Time on Perceived trustworthiness was found to be significant, $F(3.532, 148.356) = 31.56$, $p < 0.001$, $\eta^2 = 0.429$. This indicates that the different dimensions of perceived trustworthiness developed differently over time (see Figure 3). Bonferroni-corrected *post-hoc* comparisons showed that all dimensions of trustworthiness significantly decreased from T1 to T2 ($\Delta M_{ABI} = -1.8$, $\Delta M_{BEN} = -0.8$, $\Delta M_{INT} = -0.9$, all $p < 0.001$), 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 for ability ($\Delta M_{ABI} = 0.6$, $p < 0.001$) and integrity ($\Delta M_{INT} = 0.4$, $p < 0.001$), but not for benevolence ($\Delta M_{BEN} = 0.1$, $p = 1.00$). This suggests that, on average, ability and integrity recovered, while benevolence did not.

The two-way interaction effect of Partner type and Time on Perceived trustworthiness was found to be non-significant, $F(1.659, 69.679) = 0.35$, $p = 0.672$, $\eta^2 = 0.008$. This indicates that the

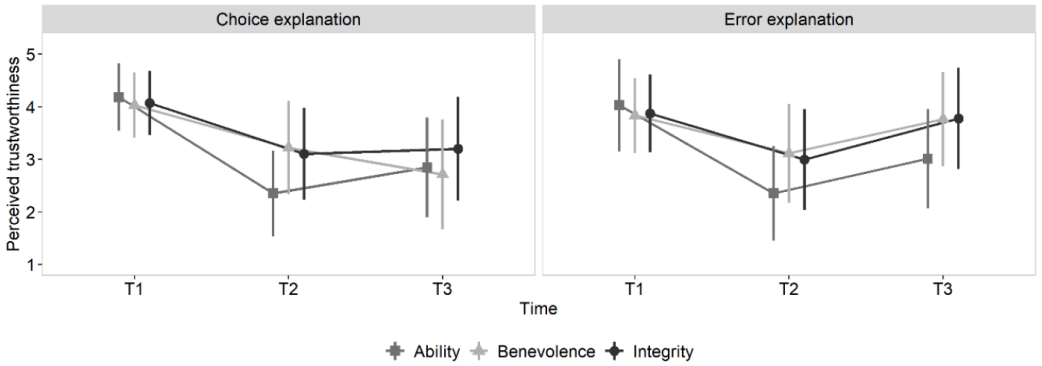

Fig. 4. The graphs show data from both partner types combined ($n=44$). The x-axis represents Time, and the y-axis represents Perceived Trustworthiness. Separate lines indicate different dimensions of trustworthiness: ability (dark gray, square points), benevolence (light gray, triangle points), and integrity (black, circle points). The left panel represents the choice-explanation, and the right panel represents the error-explanation. Error bars represent standard deviations.

perception of trustworthiness for the human and robotic partners did not change differently across all timepoints (see Figure 3).

3.2.4 Three-Way Effect. Mauchly's test of sphericity indicated that the assumption of sphericity has not been violated, $X^2(9)=16.64$, $p=0.055$. The three-way interaction effect of Trustworthiness dimensions, Explanation type, and Time on Perceived trustworthiness was found to be significant, $F\ (4, 168)=8.79$, $p<0.001$, $\eta^2 = 0.173$ (see Figure 4).

Bonferroni-corrected *post-hoc* comparisons showed that perceptions of trustworthiness in terms of ABI all decreased following the violation (ΔT1-T2, all $p<0.001$). However, ability dropped significantly more than benevolence and integrity ($p<0.001$), 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 (choice-explanation: $\Delta M=0.1$, $p=0.293$; error-explanation: $\Delta M=0.1$, $p=0.295$).

After the error-explanation (i.e., after T2), all dimensions of trustworthiness were equally repaired (ΔT2-T3; $p<0.001$). Benevolence and integrity nearly returned to their original levels prior to the violation (see Figure 3). At T3, ability remained significantly lower than benevolence ($\Delta M=0.75$, $p<0.001$) and integrity ($\Delta M=0.8$, $p<0.001$). Benevolence and integrity did not significantly differ at T3 ($\Delta M=0.01$, $p=0.884$).

After the choice-explanation, ability recovered ($\Delta M=0.5$, $p<0.001$), while integrity remained stable ($\Delta M=0.1$, $p=0.539$) and benevolence declined further ($\Delta M = -0.5$, $p=0.002$). At T3, integrity was significantly higher than benevolence ($\Delta M=0.5$, $p=0.002$) and ability ($\Delta M=0.4$, $p=0.014$). Benevolence and ability did not differ ($\Delta M=0.1$, $p=0.390$).

This three-way interaction indicates that the different dimensions of the partners' perceived trustworthiness (ABI) developed differently over time. They were differentially affected by the trust-violating event as well as the two different explanations provided (error and choice).

3.2.5 Order Effect. To control for potential order effects of the within-subject variable Explanation type (error versus choice), we performed a factorial ANOVA with Order (choice-error versus error-choice) as an additional factor, to examine its effect on Perceived trustworthiness (Figure 5). Here, the factor Partner type is left out.

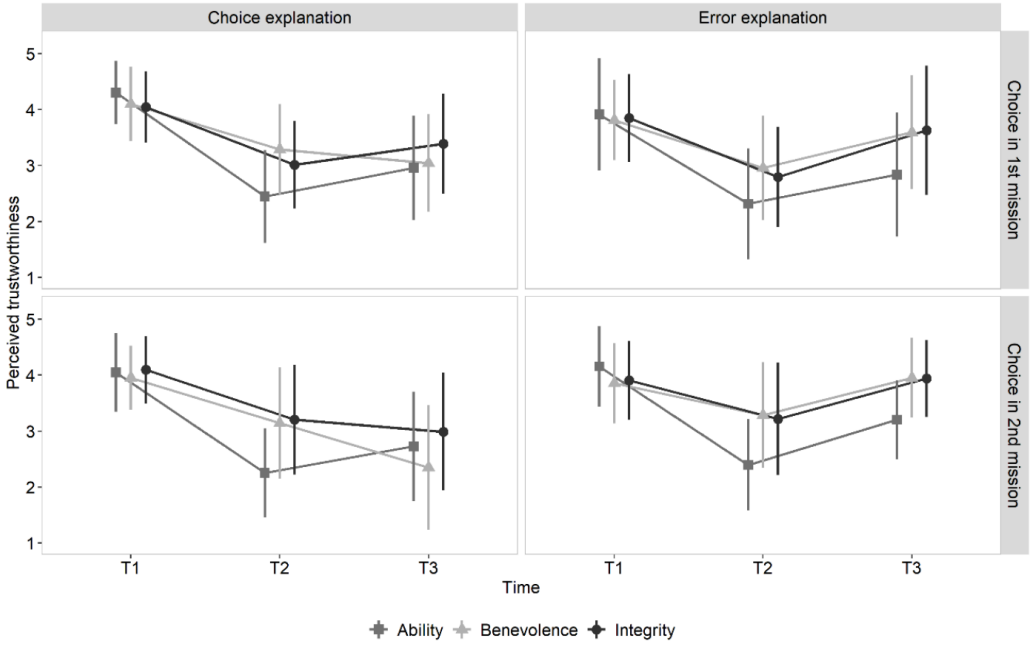

Fig. 5. The graphs show data from both partner types combined ($n=44$). The x-axis represents Time, and the y-axis represents Perceived Trustworthiness. Separate lines indicate different dimensions of trustworthiness: ability (dark gray, square points), benevolence (light gray, triangle points), and integrity (black, circle points). The left half of the grid represents the choice-explanation, and the right half represents the error-explanation. The upper half of the grid shows participants who encountered the choice 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.

A significant interaction effect between Order and Explanation type on Perceived trustworthiness was found, $F\ (1, 41) = 6.67$, $p = 0.013$, $\eta^2 = 0.140$. On average, the partner in the first mission was perceived as significantly more trustworthy than that in the second mission. Participants who had the choice-explanation in their first mission, trust was higher in the choice-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 choice-explanation mission ($M=3.2$).

3.3 Partner Assessment

Table 2 presents the descriptive statistics for all partner assessment measures included in the study. The zero-order correlations matrix in Table 3 displays the Pearson correlation coefficients for each pair of partner assessment variables, indicating the strength and direction of the linear relationships among them.

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, $F\ (1, 42) = 22.34$, $p < 0.001$, $\eta^2 = 0.347$. The partner in the choice-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

Table 2. 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	Choice	2.7	0.8	2.2	0.7
	Error	2.7	0.8	2.2	0.6
Perceived intelligence	Choice	2.8	0.9	3.3	0.8
	Error	3.0	0.9	3.4	1.0
Likeability	Choice	2.5	0.9	3.0	0.5
	Error	3.3	1.0	3.6	1.0
Intention to re-use	Choice	2.5	1.4	2.5	1.2
	Error	2.8	1.2	3.2	1.3

Table 3. Zero-Order Correlation Matrix with the Pearson Correlation Coefficients

		#	1	2	3	4	5	6	7	8
Perceived anthropomorphism	Choice	1	1							
	Error	2	0.56**	1						
Perceived intelligence	Choice	3	−0.07	0.06	1					
	Error	4	0.05	0.11	0.40**	1				
Likeability	Choice	5	0.19	0.29	0.61**	0.25	1			
	Error	6	0.07	0.38*	0.22	0.68**	0.39**	1		
Intention to re-use	Choice	7	−0.13	0.01	0.54**	0.01	0.48**	0.01	1	
	Error	8	−0.03	0.06	0.31*	0.73**	0.19	0.68**	−0.09	1

*p < .05. **p < .01.

other effects on Likeability were observed. For Intention to reuse, no significant effects of Partner or Explanation type were observed.

4 Discussion

4.1 Evaluation of Findings

The research question explored how the perceived trustworthiness of a partner is influenced by the occurrence of a trust-violating event, the explanation provided for its occurrence, and the type of partner responsible. First, our findings showed that, as anticipated, the trust-violating event led to a drop of all forms of trustworthiness, specifically ABI. Second, trustworthiness perceptions were differentially affected by the two explanations given for its occurrence. We hypothesized that the error-explanation would be more effective in restoring perceptions of benevolence and integrity than perceptions of ability, while the choice-explanation was expected to preserve perceptions of ability and not benevolence and integrity. However, the error-explanation repaired all perceptions of trustworthiness, including ability, thus only partially confirming our hypothesis. As expected, the choice-explanation only repaired ability and not benevolence and integrity. Lastly, contrary to our expectations, these patterns were consistent across both human and robotic partners, indicating that, in this study, the explanation type played a more critical role than the partner type in shaping trustworthiness perceptions. Suggested explanations for these findings are discussed in the following.

4.1.1 Explanation Type. As expected, the choice-explanation only led to an increase of ability-based perceived trustworthiness and not of benevolence and integrity. When the partner explained that the encounter with the hazard resulted from a choice they made, rather than a mistake, perceptions of integrity stagnated, and perceptions of benevolence dropped further. The choice-explanation was expected to negatively affect perceptions of benevolence and integrity, as both dimensions relate more to intentions of the agent [12], albeit in slightly different ways. That is, the choice-explanation likely harmed benevolence-based perceived trustworthiness, because the partner did not act in the participant's best interest by prioritizing collective over individual benefits—directly contradicting 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. At the same time, the choice-explanation also harms integrity-based perceived trustworthiness, 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 [76], namely that their partner would prioritize their safety. Hence it is not surprising that the choice-explanation failed to repair both perceptions of benevolence and integrity.

While our choice-explanation may not fit neatly into either category (a benevolence or integrity-based violation), its value lies in its approach to a realistic scenario. Esterwood and Robert [64] 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 choice condition in this study had no ill will, nor was it self-centered and seeking individual gains over joint gains [29]. As outlined in the Introduction, we advocate for the inclusion of more nuanced and realistic instances of benevolence and integrity-based trust violations—beyond mere acts of selfishness or malicious intent by robots. We hope future studies will adopt this approach, as it would enable a more comprehensive exploration of trust dynamics in HRIs.

Contrary to our expectations, the error-explanation had a restorative effect on perceived trustworthiness, including ability-based trustworthiness, although it did not fully recover to its pre-error level. This finding is noteworthy, as the error-explanation acknowledged the partner's lack of skills and knowledge in detecting the explosive hazard competently. We had anticipated that such an explanation would hinder the repair of perceived ability, as participants might be concerned about the partner's potential for repeated mistakes. However, the data revealed that participants were not deterred by this information. Perceptions of ability-based perceived trustworthiness recovered significantly, even after the partner explained that the risk exposure resulted from a technical failure, highlighting the limitations of their abilities.

Similarly, the partner making the choice and the one making an error were perceived as equally intelligent. The measure perceived intelligence was included to rule out that any observed differences between the choice and error conditions could be attributed to one agent being perceived as more intelligent because it did not make a mistake whereas the other did. While perceived intelligence is related to perceived ability, they are not identical constructs in this context. The ability items specifically focused on the partner's task-related competence rather than their overall intelligence. Notably, the partner making a mistake rather than a deliberate choice was not perceived as less intelligent, and the error-explanation did not further harm ability perceptions. We will discuss two possible explanations for the robust effect of the error-explanation on ability-based perceived trustworthiness.

The effectiveness of this explanation can be attributed to its formulation, which included an explicit acknowledgement of responsibility for the mistake ("I failed to advise you correctly"). This admission of fault may have effectively turned the explanation into an apology, which is consistent

with previous findings that that explanations accompanied by expression of regret can lead to significant trust repair [77]. In contrast to previous HRI research, where explanations have not consistently been successful as trust repair strategies [64], our findings suggest that explanations can be effective in repairing ability-based trust. One possible explanation for this is that a partner's 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 [78], which are all related to ability rather than benevolence or integrity. This suggests that explanations may be particularly effective in repairing perceptions of ability, rather than more moral aspects like benevolence and integrity.

Another possible explanation for the consistent recovery of ability-based trust could be related to people's mental model of the partner and its components. The technical failure described in the error-explanation (i.e., "My sensor was too weak to detect the explosive.") may not be attributed to the partner's competence by participants. In fact, responses to open-ended questions suggest that some participants distinguish between the partner (both human and robot) and its sensors, attributing the failure to the sensor's performance rather than the partner's abilities. This distinction is interesting, as it applies to both human and robot partners. While it is understandable that humans and their sensors are seen as separate entities, sensors are an integral part of a robot's functionality, similar to human sensory organs. Participant responses illustrate this distinction, with one human condition participant stating: "He [the human partner] trusted his device and made decisions based on the info provided to him." Similarly, a robot condition participant noted: "It bases its decisions on what its sensors detect." These responses raise intriguing questions about how people perceive robots, particularly with regards to Cartesian dualism (i.e., "mind-body"/"software-hardware" distinctions). Do people consider a robot as a unified whole or as a set of communicating parts? If the latter, it is possible that the algorithm (software) is perceived as competent, while the sensors and cameras (hardware) are seen as incompetent. As highlighted in our introduction, it is essential to specify the basis on which we assess another entity's trustworthiness [23, 76].

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 [68]. To effectively collaborate, it is crucial to have an accurate mental model of your partner's limits and priorities. As such, this information might have contributed to a more appropriate calibration of trust, allowing individuals to better gauge the true qualities of their partner [69]. 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 behavior [12]. 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 [31]. Further work is needed to fully understand the implications of different types of trust violations under different operational circumstances.

4.1.2 Partner Type. Contrary to expectations and previous research [9, 22, 30, 54, 55], this study found no differences in the development of trust between the human and robotic partners. While the pattern of perceived trustworthiness differed depending on whether the trust-violating event was explained as a choice (intentional) or an error (unintentional), this pattern was consistent across both human and robotic partners. This finding contrasts with earlier research suggesting that humans making errors are judged less negatively than robots, while humans with non-benevolent intentions are judged more negatively than robots [9, 22]. Based on this, we expected that the error-explanation would repair the perceived trustworthiness of the human partner, but not the robot, and that the choice-explanation would repair the perceived trustworthiness of the robotic

partner, but not the human. However, our findings suggest that both partners were judged based on their intentions, with perceived trustworthiness restored after an unintentional mistake, but not after an intentional decision that disadvantaged the participant.

It is likely that the absence of an effect of partner type is due to the fact that both partners in the study were virtual characters with limited possibilities for interaction [79]. While our manipulation check showed that the human partner was perceived as significantly more anthropomorphic than the robotic partner, the limited interactive capabilities in our task may have reduced the likelihood of perceiving distinct differences in trustworthiness between the two. The decision to use similar virtual characters for both the human and robotic partners was intentional and beneficial for controlling potential confounds. By keeping the characters similar in design and functionality, we could isolate the specific effects of the partner type (human versus robot) without introducing additional variability that might arise from more complex differences in appearance or interaction styles. Although we attempted to emphasize the human–robot distinction by introducing the human partner as a confederate, we were unable to provide a physical robot for the Robot Partner condition, which could have further highlighted their differences.

In conclusion, the absence of a partner type effect may be attributed to our task setup, in which participants interacted with virtual representations rather than actual humans or physical robots. The asymmetry in the nature of risk, discussed in the Introduction (where humans have vulnerabilities and stakes in the outcome, while robots have nothing to lose, as they cannot suffer or die), may be less tangible in the current task due to the use of virtual representations of humans or robots. This could also explain why no significant differences were found between partner types (human versus robot). Future research should investigate the impact of physical interactions with actual humans and robots, as the more tangible differences in risk and vulnerability between humans and robots in such contexts may have a stronger influence on trust compared to virtual representations.

4.1.3 Trust Dynamics. Our findings revealed a significant order effect, with perceived trustworthiness generally decreasing over time. Specifically, perceived trustworthiness was lower in the second mission compared to the first mission, regardless of the condition order. Despite our efforts to mitigate order effects by providing a short break and emphasizing the change in partner between missions, a significant order effect still emerged. Yet, we employed a counterbalanced design to control for potential biases stemming from the order in which participants completed the missions. The counterbalancing ensured that any order-related effects were distributed across both conditions, with half of the participants starting with the choice-explanation and the other half starting with error-explanation. As a result, the overall impact of the order effect on the conclusions of the study should be minimal.

Given that the order effect was consistent across the two conditions and did not vary by the order in which missions were completed, we can confidently interpret that the primary findings of the study are not influenced by the sequence of the missions. This suggests that the decrease in perceived trustworthiness from the first to the second mission is likely a natural occurrence due to factors such as learning effects, or evolving perceptions of the robotic partner, rather than a consequence of the order in which the missions were presented.

Overall, the observed order effect is an interesting aspect of the data and 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.

4.2 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 44 mostly Dutch university students. The small sample size limits the statistical power to detect and interpret three-way interaction effects reliably. As such, these effects should be interpreted with caution, and their generalizability to broader contexts remains uncertain. Additionally, the nature of the sample might further affect the generalizability of the results, 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 [5], potentially trained to prioritize mission success over personal health. This difference in perspective could result in a better understanding of the partner's consideration in the choice 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 comparing trust violations due to errors versus choice, 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 [74]. 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 the ABI model of Mayer et al. [25] and demonstrates statistical separation between subdimensions in the initial relationship, even after one interaction [74]. All subscales showed good internal consistency. Moreover, we preferred the McKnight scale over the commonly used Jian et al. [80] 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-centered. 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 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 behavior.

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 choice-explanation condition) seem an inevitable part of the future. Something to bear in mind however is that in our experiment, the partners in the choice-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 [81]. 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 [82]. It is plausible that the stagnation of perceived integrity in response to the choice-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 [83]. In order to make accurate judgments of trust, intentions towards a certain goal are ideally communicated beforehand [8, 84]. This is an important issue for future research. We expect that informing participants about the priorities of the partners in the choice-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 priorities 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 [5]. Being transparent about the agent's intentions, goals and priorities upfront could have a similar effect on trust.

4.3 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 [22, 46], acting out of self-interest [22, 29], or deviating from a planned path [45]. 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 behavior in HRTs.

While performance is still an important determinant of human–robot trust [21, 85, 86], this study strengthens the idea that aspects such as values, personal relations, and moral aspects become

equally important [7, 15]. However, we concur with the notion of Alarcon et al. [73] and [12] 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 [15]. 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 ABI are differentially affected by different types of explanations regarding the intentionality behind a trust-violating instruction. One of the questions that emerge from these results is what the implications will be for behavioral reliance. What will be the behavioral consequence of a situation where perceptions of ability have recovered, while perceptions of benevolence and integrity have not (yet)? While the participants in the study actively controlled their own character, they were instructed to stay close to their partner at all times while encountering various events and completing self-report trust questionnaires. They had no option to disobey or deviate from their partners' instructions if they lacked trust. Our current approach prioritized experimental control over behavioral freedom. Further research should be undertaken to investigate the behavioral consequences of this discrepancy in trusting beliefs.

5 Conclusion

Increasingly autonomous AI-based artificial agents are used in a wide variety of both military and civilian applications [87]. 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 [8]. Traditionally, HRI focused on performance measures such as task-related strengths and limitations, reliability, and predictability of a robot [15, 86, 88]. Today, human operators should increasingly be aware of a robot's higher-level values, priorities, and goals [86]. 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. At the same time robots in collaborative settings should gain the interactive ability to resolve competing goals through social processes [89]. Knowing your partner's intentions, goals, and preferences is crucial for calibrated trust and successful team performance [8]. 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 [60] and, as a part of that, to understand if and how trust can be recovered after intentional or unintentional trust violations [39].

Competing Interests

The authors declare that they have no conflict of interest.

Consent to Participate

Informed consent was obtained from each study participant after they were told of the potential risks and benefits as well as the investigational nature of the study.

Ethics Approval

All studies were conducted in accordance with principles for human experimentation as defined in the 1964 Declaration of Helsinki, and approved by the relevant institutional review boards.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the author(s) used ChatGPT in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] NATO Science & Technology Organization. 2023. Science & Technology Trends 2023-2043, Vol. 1. NATO Science & Technology Organization.

[2] S. Soltanzadeh. 2022. Strictly human: Limitations of autonomous systems. *Minds and Machines* 32, 2 (2022), 269–288.

[3] G. Ferguson and J. A. Allen. 2011. A cognitive model for collaborative agents. AAAI Fall Symp—Tech Rep. 2011; FS-11-01:112–20.

[4] S. J. Russell and P. Norvig. 2003. *Artificial Intelligence: A Modern Approach* (2nd. ed.). Prentice Hall, Upper Saddle River.

[5] E. S. Kox, L. B. Siegling, and J. H. Kerstholt. 2022. Trust development in military and civilian human-agent teams: The effect of social-cognitive recovery strategies. *International Journal of Social Robotics* 14, 5 (2022), 1323–1338. Retrieved from https://orcid.org/0000-0002-5421-3090

[6] K. T. Wynne and J. B. Lyons. 2019. Autonomous agent teammate-likeness: Scale development and validation. In *International Conference on Human-Computer Interaction*. Springer, Cham, 199–213.

[7] G. Matthews, A. R. Panganiban, J. Lin, M. D. Long, and M. Schwing. 2021. Super-machines or sub-humans: Mental models and trust in intelligent autonomous systems. In *Trust in Human-Robot Interaction*. C. S. Nam & J. B. Lyons (Eds.), Elsevier Academic Press, 59–82. DOI: https://doi.org/10.1016/B978-0-12-819472-0.00003-4

[8] M. Hou, G. Ho, and D. Dunwoody. 2021. IMPACTS: A trust model for human-autonomy teaming. *Human-Intelligent Systems Integration* 3, 2 (2021), 79–97.

[9] C. A. Hidalgo, D. Orghian, J. Albo-Canals, F. de Almeida, and N. Martin. 2021. *How Humans Judge Machines*. The MIT Press Cambridge, Massachusetts, London, England, 239.

[10] P. Werkhoven, L. Kester, and M. A. Neerincx. 2018. Telling autonomous systems what to do. In *ACM International Conference Proceeding Series*.

[11] R. J. Knighton. 2004. The psychology of risk and its role in military decision-making. *Defence Studies* 4, 3 (2004), 309–334.

[12] J. D. Lee and K. A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[13] S. Stephens. 2023. Business Wire. *NYPD Launches Knightscope Security Robot Service in Manhattan Subway*. Retrieved October 13, 2023 from https://www.businesswire.com/news/home/20230922025249/en/NYPD-Launches-Knightscope-Security-Robot-Service-in-Manhattan-Subway

[14] Knightscope. 2023. The K5 ASR—A Fully Autonomous Outdoor Security Robot. Retrieved October 13, 2023 from https://www.knightscope.com/products/k5

[15] B. F. Malle and D. A. Ullman. 2021. Multi-dimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction: Research and Applications*. C. S. Nam & J. B. Lyons (Eds.), Elsevier Academic Press, 3–25. DOI: https://doi.org/10.1016/B978-0-12-819472-0.00001-0

[16] K. A. Hoff and M. Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.

[17] M. Madsen and S. Gregor. 2000. Measuring human-computer trust. In *the 11th Australasian Conference on Information Systems*, 6–8.

[18] D. Gambetta. (Ed.). 2000. Can we trust trust? In *Trust: Making and Breaking Cooperative Relations*. Oxford, 212–237.

[19] Martina Raue, Lisa A. D'Ambrosio, Carley Ward, Chaiwoo Lee, Claire Jacquillat, and Joseph F. Coughlin. 2019. The influence of feelings while driving regular cars on the perception and acceptance of self-driving cars. *Risk Analysis: An Official Publication of the Society for Risk Analysis* 39, 2 (2019), 358–374.

[20] A. Shariff, J. F. Bonnefon, and I. Rahwan. 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour* 1, 10 (2017), 694–696.
[21] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527.
[22] G. M. Alarcon, J. B. Lyons, I. Aldin Hamdan, and S. A. Jessup. 2024. Affective responses to trust violations in a human-autonomy teaming context: Humans versus robots. *International Journal of Social Robotics* 16, 1 (2024), 23–35. DOI: https://doi.org/10.1007/s12369-023-01017-w
[23] A. Langer, R. Feingold-Polak, O. Mueller, P. Kellmeyer, and S. Levy-Tzedek. 2019. Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience and Biobehavioral Reviews* 104 (July 2019), 231–239.
[24] N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, and M. Langer. 2025. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Computers in Human Behavior* 170 (2025), 108671.
[25] R. C. Mayer, J. H. Davis, and D. F. Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20, 3 (1995), 709–734.
[26] D. F. Schoorman, R. C. Mayer, and J. H. Davis. 2007. An integrative model of organizational trust: Past, present and future. *Academy of Management Review* 32, 2 (2007), 344–354.
[27] R. S. Bhagat and R. M. Steers. 2009. *Cambridge Handbook of Culture, Organizations, and Work*. Cambridge University Press, Cambridge, United Kingdom.
[28] J. D. Lee and N. Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.
[29] G. M. Alarcon, A. M. Gibson, and S. A. Jessup. 2020. Trust repair in performance, process, and purpose factors of human-robot trust. In *2020 IEEE International Conference on Human-Machine Systems*.
[30] Ewart J. de Visser, Samuel S. Monfort, Ryan McKendrick, Melissa A. B. Smith, Patrick E. McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology Applied* 22, 3 (September 2016), 331–349.
[31] E. J. de Visser, R. Pak, and T. H. Shaw. 2018. From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics* 61, 10 (2018), 1409–1427.
[32] David Cameron, Stevienna de Saille, Emily C. Collins, Jonathan M. Aitken, Hugo Cheung, Adriel Chua, Ee Jing Loh, and James Law. 2021. The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in Human Behavior* 114 (September 2021), 106561.
[33] K. Hald, K. Weitz, E. André, and M. Rehm. 2021. 'An error occurred!'—Trust repair with virtual robot using levels of mistake explanation. In *9th International Conference on Human-Agent Interaction (HAI '21)*. ACM, New York, NY, 9. Retrieved from http://journal.unilak.ac.id/index.php/JIEB/article/view/3845%0Ahttp://dspace.uc.ac.id/handle/123456789/1288
[34] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (May 2017), 1–15.
[35] M. K. Lee, S. Kiesler, J. Forlizzi, S. S. Srinivasa, and P. Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 203–210.
[36] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *ACM/IEEE International Conference on Human-Robot Interaction*, 141.
[37] P. Robinette, A. M. Howard, and A. R. Wagner. 2017. Effect of robot performance on human-robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436.
[38] C. Esterwood and L. P. Robert. 2023. Three strikes and you are out! The impacts of multiple human-robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior* 142 (January 2023), 107658.
[39] T. Kim and H. Song. 2021. How should intelligent agents apologize to restore trust? The interaction effect between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics* 61 (2021), 101595.
[40] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill. 2018. Is it my looks? Or something I said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface, Vol.* 10809. LNCS, 56–69.
[41] P. Fratczak, Y. M. Goh, P. Kinnell, L. Justham, and A. Soltoggio. 2020. Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *International Journal of Industrial Ergonomics* 82 (July 2020), 103078.
[42] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of changing reliability on trust of robot systems. In *7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '12)*, 73–80.

[43] M. Desai, P. Kaniarasu, M. Medvedev, A. M. Steinfeld, and H. Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *ACM/IEEE International Conference on Human-Robot Interaction*, 251–258.
[44] P. Robinette, A. M. Howard, and A. R. Wagner. 2017. Conceptualizing overtrust in robots: Why do people trust a robot that previously failed? In *Autonomy and Artificial Intelligence: A Threat or Savior?* Springer International Publishing, 129–155. DOI: https://doi.org/10.1007/978-3-319-59719-5_6
[45] J. B. Lyons, I. Aldin Hamdan, and T. Q. Vo. 2022. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior* 138 (February 2022), 107473.
[46] S. S. Sebo, P. Krishnamurthi, and B. Scassellati. 2019. 'I don't believe you': Investigating the effects of robot trust violation and repair. In *ACM/IEEE International Conference on Human-Robot Interaction*, 57–65.
[47] R. Perkins, Z. R. Khavas, K. McCallum, M. R. Kotturu, and P. Robinette. 2022. The reason for an apology matters for robot trust repair. In *14th International Conference on Social Robotics*. Springer Nature, Switzerland, 640–651.
[48] P. H. Kim, D. L. Ferrin, C. D. Cooper, and K. T. Dirks. 2004. Removing the shadow of suspicion: The effects of apology versus denial for repairing competence versus integrity-based trust violations. *The Journal of Applied Psychology* 89, 1 (2004), 104–118.
[49] J. B. Lyons, S. A. Jessup, and T. Q. Vo. 2024. The role of decision authority and stated social intent as predictors of trust in autonomous robots. *Topics in Cognitive Science* 16, 3 (2024), 430–449.
[50] R. C. Mayer and J. H. Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology* 84, 1 (1999), 123–136.
[51] C. C. Jorge, M. L. Tielman, and C. M. Jonker. 2022. Assessing artificial trust in human-agent teams a conceptual model assessing artificial trust in human-agent teams. In *22nd ACM International Conference on Intelligent Virtual Agent*, 1–3.
[52] E. J. de Visser, F. Krueger, P. E. McKnight, S. Scheid, M. A. B. Smith, S. Chalk, and R. Parasuraman. 2012. The world is not enough: Trust in cognitive agents. In *the Human Factors and Ergonomics Society Annual Meeting*, 263–267.
[53] P. Madhavan and D. A. Wiegmann. 2005. Effects of information source, pedigree, and reliability on operators' utilization of diagnostic advice. In *the Human Factors and Ergonomics Society Annual Meeting*, 487–491.
[54] P. Madhavan, D. A. Wiegmann, and F. C. Lacson. 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors* 48, 2 (2006), 241–256.
[55] P. Madhavan and D. A. Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 277–301.
[56] K. Goodyear, R. Parasuraman, S. Chernyak, P. Madhavan, G. Deshpande, and F. Krueger. 2016. Advice taking from humans and machines: An fMRI and effective connectivity study. *Frontiers in Human Neuroscience* 10 (November 2016), 542–515.
[57] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.
[58] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58 (2003), 697–718.
[59] D. A. Norman. 2013. *The Design of Everyday Things* (Revised and Expanded Edition). Basic Books, NY, 970–978.
[60] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.
[61] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124.
[62] A. M. Greenberg and J. L. Marble. 2023. Foundational concepts in person-machine teaming. *Frontiers in Physics* 10 (January 2023), 1–16.
[63] M. Hildebrandt. 2017. Learning as a machine. Crossovers between humans and machines. *Journal of Learning Analytics* 4, 1 (2017), 6–23.
[64] C. Esterwood and L. P. Robert. 2022. A literature review of trust repair in HRI. In *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN '22)*.
[65] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert Jr. 2018. 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.
[66] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon, and M. Tielman. 2020. Taxonomy of trust-relevant failures and mitigation strategies. In *ACM/IEEE International Conference on Human-Robot Interaction*, 3–12.
[67] S. C. Kohn, D. B. Quinn, R. Pak, E. J. de Visser, and T. H. Shaw. 2018. Trust repair strategies with self-driving vehicles: An exploratory study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62, 1 (2018), 1108–1112.
[68] Erin K. Chiou, Mustafa Demir, Verica Buchanan, Christopher C. Corral, Mica R. Endsley, Glenn J. Lematta, Nancy J. Cooke, and Nathan J. McNeese. 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 (2022), 1117–1136.

[69] M. Wischnewski, N. C. Krämer, and E. Müller. 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*, Vol. 1. ACM.
[70] V. Buchholz, P. Kulms, and S. Kopp. 2017. It's (not) your fault! Blame and trust repair in human-agent cooperation. *Kognitive Systeme* 1 (2017). DOI: https://doi.org/10.17185/duepublico/44538
[71] M. Bradfield and K. Aquino. 1999. The effects of blame attributions and offender likableness on forgiveness and revenge in the workplace. *Journal of Management* 25, 5 (1999), 607–631.
[72] N. Martelaro. 2016. Wizard-of-Oz interfaces as a step towards autonomous HRI. *AAAI Spring Symp—Tech Rep* SS-16-01, 147–150.
[73] G. M. Alarcon, A. M. Gibson, S. A. Jessup, and A. Capiola. 2020. Exploring the differential effects of trust violations in human-human and human-robot interactions. *Applied Ergonomics* 93 (May 2020), 103350.
[74] D. H. McKnight, V. Choudhury, and C. Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research* 13, 3 (2022), 334–359.
[75] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81.
[76] S. L. Grover, M. C. Hasel, C. Manville, and C. Serrano-Archimi. 2014. Follower reactions to leader trust violations: A grounded theory of violation types, likelihood of recovery, and recovery process. *European Management Journal* 32, 5 (2014), 689–702.
[77] E. S. Kox, J. H. Kerstholt, T. Hueting, and P. W. de Vries. 2021. Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021), 1–20.
[78] D. V. Jeste, S. A. Graham, T. T. Nguyen, C. A. Depp, E. E. Lee, and H. C. Kim. 2020. Beyond artificial intelligence: Exploring artificial wisdom. *International Psychogeriatrics* 32, 8 (2020), 993–1001.
[79] M. A. A. Fahim, M. M. H. Khan, T. Jensen, Y. Albayram, and E. Coman. 2021. Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. In *2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere (DIS '21)*, 1492–1503.
[80] J. Y. Jian, A. M. Bisantz, and C. G. Drury. 2000. Foundations for empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.
[81] R. C. Arkin, P. Ulam, and A. R. Wagner. 2012. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE* 100, 3 (2012), 571–589.
[82] A. R. Panganiban, G. Matthews, and M. D. Long. 2020. Transparency in autonomous teammates: Intention to support as teaming information. *Journal of Cognitive Engineering and Decision Making* 14, 2 (2020), 174–190.
[83] P. A. Hancock, D. R. Billings, and K. E. Schaefer. 2011. Can you trust your robot? *Ergonomics in Design: The Quarterly of Human Factors Applications* 19, 3 (2011), 24–29.
[84] K. E. Schaefer, E. R. Straub, J. Y. C. Chen, J. Putney, and A. W. Evans. 2017. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research* 46 (2017), 26–39.
[85] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva. 2018. Exploring the impact of fault justification in human-robot trust: Socially interactive agents track. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 507–513.
[86] K. Chhogyal, A. Nayak, A. Ghose, and H. K. Dam. 2019. A value-based trust assessment model for multi-agent systems. In *International Joint Conference on Artificial Intelligence*, 194–200.
[87] J. Y. C. Chen and A. Schulte. 2021. Special issue on "human-autonomy teaming in military contexts". *Human-Intelligent Systems Integration* 3, 4 (December 2021), 287–289.
[88] S. Marsh. 1994. *Formalising Trust as a Computational Concept*. SpringerBriefs in Computer Science, University of Stirling.
[89] E. K. Chiou and J. D. Lee. 2023. Trusting automation: Designing for responsivity and resilience. *Human Factors* 65, 1 (2023), 137–165.

A Appendix

A.1 Adjusted Perceived Trustworthiness Scale

(1) I believe that my partner would act in my best interest (BEN)
(2) If I required help, my partner would do its best to help me (BEN)
(3) My partner is interested in my well-being (BEN)
(4) My partner is truthful in its dealing with me (INT)
(5) I would characterize my partner as honest (INT)
(6) My partner would keep its commitments (INT)
(7) My partner is sincere and genuine (INT)
(8) My partner is competent and effective in providing advice (ABI)
(9) My partner performs its role of giving advice very well (ABI)
(10) Overall, my partner is a capable and proficient advice provider (ABI)
(11) In general, my partner is very knowledgeable about its task (ABI)

A.2 Zero-Order Correlation Matrix Perceived Trustworthiness

Table A1. Zero-Order Correlation Matrix with the Pearson Correlation Coefficients

			#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Ch	T1	A	*1*	1																	
		B	*2*	0.33*	1																
		I	*3*	0.47**	0.70**	1															
	T2	A	*4*	0.15	−0.07	−0.20	1														
		B	*5*	0.19	0.37*	0.18	0.51**	1													
		I	*6*	−0.11	0.15	0.10	0.45**	0.66**	1												
	T3	A	*7*	0.15	0.02	−0.10	0.65**	0.38*	0.41**	1											
		B	*8*	0.09	0.37*	0.20	0.36*	0.46**	0.26	0.49**	1										
		I	*9*	0.28	0.36*	0.38*	0.39**	0.51**	0.40**	0.56**	0.56**	1									
Er	T1	A	*10*	0.20	0.15	0.20	−0.01	0.08	0.16	0.23	−0.11	0.31*	1								
		B	*11*	0.06	0.44**	0.47**	−0.24	0.24	0.12	−0.01	0.12	0.32*	0.45**	1							
		I	*12*	0.15	0.55**	0.58**	−0.04	0.39**	0.32*	0.15	0.23	0.62**	0.65**	0.75**	1						
	T2	A	*13*	0.00	−0.20	−0.19	0.66**	0.34*	0.36*	0.49**	0.24	0.28	0.39**	0.02	0.17	1					
		B	*14*	0.00	0.28	0.27	0.24	0.61**	0.58**	0.29	0.37*	0.45**	0.32*	0.50**	0.59**	0.49**	1				
		I	*15*	0.06	0.18	0.19	0.29	0.38*	0.48**	0.34*	0.09	0.39**	0.47**	0.25	0.50**	0.56**	0.70**	1			
	T3	A	*16*	0.18	0.09	0.11	0.31*	0.25	0.20	0.41**	0.02	0.36*	0.62**	0.29	0.45**	0.56**	0.32*	0.47**	1		
		B	*17*	0.15	0.35*	0.30*	−0.09	0.39**	0.27	0.14	0.05	0.20	0.50**	0.59**	0.61**	0.15	0.58**	0.44**	0.53**	1	
		I	*18*	0.11	0.24	0.31*	−0.09	0.29	0.30*	0.23	0.01	0.39**	0.71**	0.60**	0.78**	0.27	0.47**	0.55**	0.58**	0.80**	1

First column refers to explanation type (Ch, choice-explanation; Er, error-explanation). Letters in the third column refer to the trustworthiness dimensions (A, ability; B, benevolence; I, integrity).
*Correlation is significant at the 0.05 level (2-tailed); **Correlation is significant at the 0.01 level (2-tailed).

Received 7 August 2024; revised 18 April 2025; accepted 16 May 2025