Autonomous Agents and Multi-Agent Systems (2021) 35:30
https://doi.org/10.1007/s10458-021-09515-9

Trust repair in human-agent teams: the effectiveness of explanations and expressing regret

E. S. Kox[1,2] · **J. H. Kerstholt**[1,2] · **T. F. Hueting**[1] · **P. W. de Vries**[2]

Accepted: 7 June 2021 / Published online: 18 June 2021

Abstract
The role of intelligent agents becomes more social as they are expected to act in direct interaction, involvement and/or interdependency with humans and other artificial entities, as in Human-Agent Teams (HAT). The highly interdependent and dynamic nature of teamwork demands correctly calibrated trust among team members. Trust violations are an inevitable aspect of the cycle of trust and since repairing damaged trust proves to be more difficult than building trust initially, effective trust repair strategies are needed to ensure durable and successful team performance. The aim of this study was to explore the effectiveness of different trust repair strategies from an intelligent agent by measuring the development of human trust and advice taking in a Human-Agent Teaming task. Data for this study were obtained using a task environment resembling a first-person shooter game. Participants carried out a mission in collaboration with their artificial team member. A trust violation was provoked when the agent failed to detect an approaching enemy. After this, the agent offered one of four trust repair strategies, composed of the apology components *explanation* and *expression of regret* (either one alone, both or neither). Our results indicated that expressing regret was crucial for effective trust repair. After trust declined due to the violation by the agent, trust only significantly recovered when an expression of regret was included in the apology. This effect was stronger when an explanation was added. In this context, the intelligent agent was the most effective in its attempt of rebuilding trust when it provided an apology that was both affective, and informational. Finally, the implications of our findings for the design and study of Human-Agent trust repair are discussed.

Keywords Trust repair · Trust · Intelligent agents · Human · Agent interaction · Apology · Social abilities · Collaboration · Competence

✉ E. S. Kox
esther.kox@tno.nl

J. H. Kerstholt
jose.kerstholt@tno.nl

T. F. Hueting
tom.hueting@tno.nl

P. W. de Vries
p.w.devries@utwente.nl

[1] TNO Soesterberg, Soesterberg, The Netherlands

[2] University of Twente, Enschede, The Netherlands

1 Introduction

In a wide variety of domains, such as healthcare, military, transport, but also in and around the regular household, autonomous systems are increasingly deployed as teammates rather than tools. This is reflected in the fact that humans and autonomous systems accomplish goals together, like driving a car, performing surgery, and providing domestic handicap assistance as human-agent teams (HATs). We define a HAT as a team consisting of at least one human and one intelligent agent, robot, and/or other AI or autonomous system [59]. The artificial component of the team will be referred to as an intelligent agent, defined as an artificial entity that observes and acts upon an environment autonomously and that is able to communicate and collaborate with other agents, including humans, to solve problems and achieve (common) goals [8, 47]. As such, the role of intelligent agents becomes more social as they are expected to act in direct interaction, involvement and/or interdependency with humans and other artificial entities. Technology is no longer merely viewed as a tool to achieve a certain goal, but people create unique social relationships with automated and autonomous entities [7, 30, 51]. This requires a more comprehensive set of social skills. Equipping autonomous systems with teaming capabilities is changing the way in which people interact with them.

One of these social skills relates to the calibration of trust. For an optimal collaboration between humans and intelligent agents, it is important that the level of human trust is warranted by the agent's capabilities [21]. If the former exceeds the latter, this may cause humans to overly rely on the agent ("overtrust"); the latter exceeding the former may result in disuse ("undertrust"). A host of studies have shown that system errors negatively affect the level of trust people have in a system, and, consequently, their willingness to rely on subsequent agent advice [22, 37, 61]. For trust to be calibrated, however, humans would need to be able to determine whether the intelligent agent is to blame for the violation or rather the situation [41]; failure to attribute a trust violation to the situation may cause an unwarranted decrease in the human's trust and reliance. Other studies have shown that people may be more forgiving when they understand why a trust violation may sometimes occur [6, 60]. Considering this, optimal collaboration between humans and intelligent agents relies heavily on the agent's capacity to effectively communicate with the human, i.e., to explain why a violation has occurred, so as to remedy damaged trust. This study focuses on the development of human trust in intelligent agents within a HAT and explores whether offering an explanation or an expression of regret after a trust violation can effectively repair trust.

1.1 Teamwork

A successful human team allows members to share workload, monitor other team members, and contribute their expertise to subtasks. Team members are each assigned a specific role and must interact dynamically, interdependently, and adaptively to achieve a common and valued goal or mission [39]. To be successful, teams require both taskwork and teamwork. *Taskwork* is defined as the interaction of individual team members with tasks, tools and systems, while *teamwork* represents a set of interrelated thoughts, actions, and feelings of each team member that are needed to function as a team through coordination and cooperative interaction [48]. In other words, human teamwork largely depends on a variety of internal processes that are partly unconscious and often

communicated implicitly [16]. Implicit communication includes emotional, nonverbal exchanges that are, although at times subtle, a crucial complement to the explicit information in a verbal interpersonal exchange [21]. Unconscious reasoning and implicit communication are typically human skills. It enables important aspects of human teamwork, such as understanding responsibilities, norms and interaction patterns [39]. It allows us to create a shared understanding and to assess other members' commitment to the task or the social intent in their communication [43]. As human teams will increasingly be complemented by intelligent agents, new challenges arise on whether intelligent agents can effectively participate in team processes as we understand them today [39].

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

According to the CASA-paradigm (Computers Are Social Actors), people treat computers as if they were social actors, applying the same social rules, norms, and expectations to their interaction with computers as soon as social cues pertaining to, for example, personality traits or gender, are provided [23]. Incorporating such cues in intelligent agents can trigger anthropomorphism, i.e. the tendency to make organic attributions to inorganic entities [39]. Anthropomorphism, in turn, may cause human operators to generate a more sympathetic and user-friendly mental representation of the agent [5]. On the one hand, anthropomorphism can be beneficial, as humans are more likely to collaborate with intelligent agents if they show the same qualities and traits that allow humans to team with other humans [53]. Culley and Madhavan [5], on the other hand, argued that including anthropomorphic cues may have a considerable impact on the calibration of trust in an agent, as it strengthens the human tendency to attribute human features to non-human entities. As a result, a human might base its level of trust on characteristics attributed to the agent, rather than on actual experiences with the agent itself and trust may turn out to be misplaced [5].

However, other research suggests that human-agent interaction is qualitatively different from interpersonal interaction [28, 33, 57]. Recent developments in autonomous driving, for instance, show that although self-driving cars are statistically safer than human drivers, fatal accidents involving self-driving cars evoke a stronger public response than accidents involving human drivers [52]. Research shows that even a single error from a robot strongly affects a person's trust [45]. Research suggests that people often consider a machine as nearly infallible and that they have a natural tendency to follow the advice of automation, a phenomenon known as the automation bias [63]. These high expectations result in a steeper decline in trust in case of a machine failure than it would in case of a human error, as humans are considered to be inherently fallible [28, 29]. This is in line with the notion of algorithm aversion, the tendency for people to more rapidly lose faith in an erring decision-making algorithm than in humans making comparable errors [52]. Apparently, trust violations by machines are viewed and judged differently than trust violations by humans [14].

As AI technology matures, agents will become more social and more frequently deployed in social roles. Therefore it seems likely that people will increasingly treat intelligent agents as social actors and more readily apply the same rules, potentially triggering

undesirable biases and heuristics. The challenge is to incorporate social skills in a way that supports human-agent teaming, without being misleading.

1.2 Trust

Trust is a fundamental aspect of teamwork. Perceived trustworthiness is one of the decisive factors when we consider to engage in some sort of cooperation with another entity [11]. Trust facilitates collaboration and cooperation between interdependent actors. When team members with different roles, skills, and task responsibilities work together towards a common goal, they rely on each other and must trust their teammates that they will perform as required in order to accomplish that goal [20]. The highly interdependent and dynamic nature of teamwork demands trust among team members to be correctly calibrated in order to perform successfully [20]. We define Human-agent trust as the human's willingness to make oneself vulnerable and to act on the agent's recommendations and decisions in the pursuit of some benefit, with the expectation that the agent will help achieve their common goal in an uncertain context [11, 15, 21, 30, 42, 52]. As trust relates to the willingness to be vulnerable, it is often associated with risk and the perceived probability of loss by the trusting person(s) [4]. To minimize the risks and maximize the benefits, the aim should be calibrated trust, which refers to a balanced relation between the perceived trustworthiness of an intelligent agent and its actual trustworthiness. Poor calibration, meaning either overtrust or undertrust, can lead to inappropriate reliance on intelligent agents, which can compromise safety and profitability [21]. Especially under complex and uncertain conditions, the establishment of calibrated trust among teammates is essential for efficient collaboration and communication and ultimately team performance [40, 59].

Trust tends to be less influential with stable, well-structured conditions without much uncertainty. As artificial agents become more complex and go beyond a simple tool with sharply defined and easily understood behaviors, the importance of trust increases [21]. Trust becomes a more reliable predictor of subsequent behavior when complexity and unexpected circumstances make a complete understanding of the system impossible. The feeling of trust plays a crucial role in people's ability to overcome the cognitive complexity and the uncertainty that is associated with increasingly sophisticated intelligent systems [21]. People in complex, uncertain, or even high-risk situations can suffer attentional overload, which triggers automatic processing (based on fast and effortless biases and heuristics), where many aspects of causal reasoning occur outside conscious awareness [16]. One way to enable people to focus their limited attentional capacity is by using emotions to fill in the gaps in rational [26]. When cognitive resources are insufficient for rational decision making, feelings may guide behavior [21]. Lee and See [21] note that, in both human-agent trust as in interpersonal trust literature, the influence of affect is typically undervalued, while the impact of cognitive capacities is often exaggerated. Affective aspects of trust presumably have the most direct impact on behaviour however, since people not only think about trust, but foremost feel it [9].

Studies on trust do not solely focus on the beneficial effects of trust, but also expose the complex processes through which trust develops. Given the complexity and unpredictability of many situations in which intelligent agents are deployed, like military operations and city traffic, agents will not always be able to make perfect decisions or come to correct conclusions. Hence it is conceivable that an intelligent agent will at some point in time provide their human teammate with an incorrect advice. An incorrect advice and its potentially damaging consequences may lead to a decrease in trust and in the willingness to

accept further information from the agent, and as a consequence, limited benefit from the advantages that intelligent agents have to offer [10, 12]. In addition, it has been shown that repairing damaged trust is more difficult than building trust initially [17], which further underscores the importance of effective trust-repair strategies.

1.3 Trust repair

1.3.1 Interpersonal trust-repair strategies

In interpersonal trust literature, multiple strategies for trust repair are found, such as ignoring the occurrence of the trust violation, denying responsibility for the violation, or apologizing for the violation [17, 18, 58]. The current study will focus on apology, as this is the most common trust repair strategy [25]. Providing an apology is a way for the apologizer to show an understanding of the "social requirement" for an apology when any sort of trust violation has occurred; the apologizer acknowledges that she is aware that she has done something that made the other person feel disadvantaged or hurt. Additionally, the apology may include an emotional expression that could provide context for the apologizer's intentions, for example 'If I had known that the book was that important to you, I would never have given it away" [25]. An apology can consist of multiple components, including (1) an expression of regret about the costly act (i.e. I am very sorry), (2) an explanation of why the failure occurred, (3) an acknowledgement of responsibility for the mistake, (4) an offer of repair, (5) a promise that it will not happen again in the future, and (6) a request for forgiveness [1, 25, 38, 58]. Some components are more common than others. An analysis by Lewicki and Polin [25] found that apologies usually included an expression of regret and an explanation for why the violation occurred. Other apology components were less common, less clear or not at all included in the apologies that were found. In interpersonal interaction, trust violations are shown to result in less damage when apologies for the violation had been provided, compared to when no apologies had been given [17, 55]. Furthermore, research suggests that the composition of an apology matters. An older study in which the number of apology components was manipulated showed a linear trend, where more apology components were perceived as more effective than fewer components [49]. This implies that the more extensive the apology, the smaller the damage.

1.3.2 Non-human apology

Research findings of studies dedicated to the effects of apologetic messages by computers and other forms of automation are somewhat ambiguous. Generally, research shows that providing an apology can benefit the feelings of the human towards an artificial entity [1, 3, 6, 49, 56]. Studies that looked at human-agent trust found that agents that expressed empathetic emotions towards the human (e.g. "I am sorry" or "I apologize") were trusted more than agents that did not [3, 24]. Moreover, people are more likely to trust and rely on an automated decision-support system when given an explanation why the decision aid might err [6], or when they inferred such explanations after observing system behaviour themselves [60]. The effectiveness of a trust repair strategy seems to depend on situational factors such as timing [46], violation type [50, 54] and agent type [19]. Research on the effect of timing suggests that apologies for a costly act were only effective when performed not immediately after the violation occurred, but rather when a new opportunity for deciding whether to trust the robot arose [46]. In terms of violation types, an apology appears to

be the most effective trust repair strategy after a robot performs a competence-based trust violation, whereas denial proves to be more effective in case of an integrity-based violation [50]. Other research suggests that for human-like agents, apologies were the most effective when attributed internally, whereas for machine-like agents apologizing with an external attribution was more effective [19]. Humans have a natural tendency to follow the advice of automation, even when they do not know the rationale behind the suggestions, which can lead to overtrust. Insight into agent reasoning appears to allow the human to effectively calibrate their trust in the agent, which reduces this automation bias and improves performance [63]. Other research on apologies focused mainly on performance. Akgun et al. [1] found that apologetic error messages that included both an expression of regret and an explanation had a positive effect on participants' self-appraisals of performance, when interacting with a system that errs. Tzeng [56] showed that the provision of brief apologetic feedback (i.e. "Sorry, this is not a correct guess" or "We are sorry that the provided clues were not very helpful for you") did not affect the user's overall assessment of the program, but did make the participants feel better about their interactions with the program and think of the computer as less mechanical and more sensitive to their emotions. New approaches are needed to understand the potential impact of apologetic messages from non-human agents on human-agent trust.

1.4 Current study

The aim of this study is to investigate the effect of the apology components *expression of regret* (i.e. "I am sorry") and *explanation* on the development of trust, after it has been violated. The experimental environment resembles a first-person shooter game where participants carry out a mission whilst being advised by an intelligent agent. The intelligent agent is represented graphically as a virtual robot. An encounter with the enemy after an incorrect advice from the agent is expected to cause a violation of trust and a drop in people's willingness to accept subsequent advice [45]. Intentionally breaking trust allows us to examine the effectiveness of different strategies in the trust repair phase. Immediately after the violation has occurred, the agent attempts to repair trust by offering an apology that consists of an expression of regret or an explanation, a combination of both, or neither. The main research question is how trust develops over time when an intelligent agent uses different strategies to repair trust after a trust violation has occurred. We expect to find an effect for both expression of regret and explanation. The combination of components is expected to be the most effective strategy for trust repair.

2 Method

2.1 Design

A 3 (Time: prior to violation [T1], after violation [T2], after repair [T3]) × 2 (Regret: provided or not) × 2 (Explanation: provided not) mixed-design was used. Time was a within-participant factor and Regret and Explanation were varied between participants. The main dependent variables were trust and advice acceptance. Participants were randomly assigned to one of the four trust-repair conditions (explanation only: n = 18; regret only: n = 16; neither: n = 14; both: n = 18).

Fig. 1 Screenshots of the task environment

2.2 Participants

There were 66 participants, most of them students at the University of Twente. Their age ranged from 19 to 55 with a mean of 24.6 (SD = 5.6). 37 of participants were male. The participants were recruited through SONA, a test subjects pool at the University of Twente. Participants received credits for participation. In addition, the fastest participant to finish the experiment received a prize of 50 euros.

2.3 Task and procedure

The experimental environment that was built in Unity3D resembled a first-person shooter game. Participants carried out a mission whilst being advised throughout the game by their artificial team member with its robotic embodiment (Fig. 1). For the control of the intelligent agent, the Wizard of Oz method was used; the agent was controlled by one of the experiment leaders in an adjacent room, while the participant was kept under the impression that it was operating autonomously.

Participants were first presented with information about the study and a consent form. Upon agreeing to participate, each participant was randomly assigned to one of the four trust repair conditions. Participants started with a training session to get familiar with the controls and to test the volume of the audio. Participant wore headphones to hear the auditory messages from the agent.

For the actual task, participants were instructed to head back to basecamp as fast and careful as possible, since they were running low on ammunition. In addition to getting from A to B as fast as possible, they had to watch out for enemies along the way. The basecamp was marked by a red flag and located on top of a mountain. The basecamp was visible for most of the route, so participants knew what direction to go. At three points

Table 1 Overview of messages from the agent throughout the experiment

Type of message	Message from the agent
Advice T1	I have detected enemies, so I advise you to take shelter
Feedback T1	The advice I gave you was correct. The enemy was getting closer, and if you had not taken shelter, you would probably have been discovered by now
Advice T2	I am not detecting any enemies, so i advise you to move forward
Feedback T2	–
Repair	See Table 2
Advice T3	I have detected enemies, so I advise you to take shelter again
Feedback T3	The advice I gave you was correct. The enemy was getting closer, and if you had not taken shelter, you would probably have been discovered by now

throughout the scenario, the agent provided the participants with information on whether it detected enemies or not and the corresponding advice to take shelter or continue moving. The agent communicated through auditory messages. Although the task environment resembled a first-person shooter game, participants were told to avoid hostile contact due to their ammunition shortage. After an advice was given, the game paused and participants were asked to turn to a second screen and to rate their willingness to accept the agent's advice through a single-item questionnaire. Advice acceptance was always measured directly after the participant received an advice. Participants were told that by answering the question, they made their decision to accept the advice or not; they did not actually have to seek shelter when they returned to the game. Participants were told that during the questionnaire break, ten minutes had passed in the game. A few moments after continuing the game, participants received feedback from the agent on whether the advice had turned out to be correct or not. Feedback was either provided by the agent itself through a auditory message (e.g. "my advice was correct"), or by an external event (i.e. the appearance of an enemy, indicating that the advice to move forward had been inaccurate). The agent's first advice was correct (see Table 1). The agent's second advice was incorrect, resulting in the encounter with the enemy and provoking a trust violation. During this encounter with the enemy, participants could only continue once they had eliminated the enemy with their firearms. During the confrontation, the enemy kept shouting and the periphery of the screen coloured red to create a sense of threat. Participants did not know that they actually had an endless supply of ammunition or that the enemy could not eliminate them in the game. Although some took longer than others, in the end every participant succeeded in eliminating the enemy. The rationale behind this confrontation was to startle the participant and to provoke a trust violation. There were no further consequences to their performance on this part of the task. After receiving feedback, the game paused again and participants were asked to fill out the trust questionnaire on the second screen. A few moments after continuing the game after the second trust measure (i.e. violated trust), the trust repair manipulation followed (see Fig. 2). The agent offered an apology that consisted of either an expression of regret or an explanation, a combination of both, or neither. To assess the effect of the trust repair strategy, both advice acceptance *and* trust were measured directly after the third advice. The third advice was again correct, but this performance feedback about the last advice was provided later on to avoid interference with the effect of the trust repair manipulation. A schematic timeline is presented in Fig. 2. After the participant

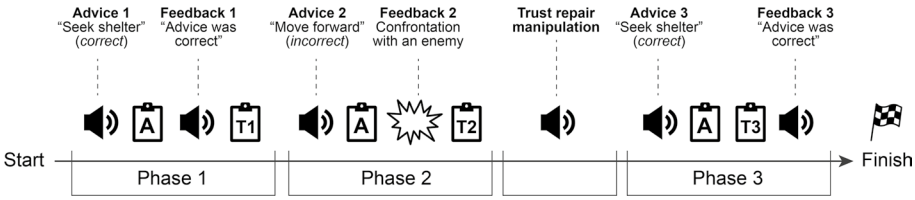

Fig. 2 Schematic timeline of the experiment. Each phase consisted of (1) an advice from the agent, (2) an advice acceptance questionnaire (clipboard icon with the letter A), (3) a moment of feedback (verbal or experienced), and (4) a trust questionnaire (clipboard icon with the letter T)

finished the game, a final questionnaire measured the concepts 'anthropomorphism', 'likeability', 'perceived intelligence', 'perceived usefulness', 'feeling', 'game experience' and demographics.

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

2.4 Materials

2.4.1 Questionnaires

Unless described otherwise, 7-point Likert scales ranging from 'completely disagree' to 'completely agree' were used.

2.4.2 Advice acceptance

Acceptance of advice was measured by a single item: "the probability that I will follow the buddy's advice is [very low–low–slightly low–slightly high–high–very high]", on a six-point scale.

2.4.3 Trust

Trust in the agent contained three subdimensions with a total of 11 items ($\alpha = .84$): competence (4 items, i.e. "My buddy has a lot of knowledge on navigating through this environment.") ($\alpha = .86$); benevolence (3 items, i.e. "My buddy puts my interests first.") ($\alpha = .72$); and integrity (3 items, i.e. "My buddy is honest.") ($\alpha = .61$). The items were based on the constructs of McKnight and Chervany (32).

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

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

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

2.4.4 Self-efficacy

Self-efficacy was measured with three items (i.e. "I am sure of my skills for performing this task") ($\alpha = .89$).

2.4.5 Godspeed: perceived anthropomorphism, likeability and intelligence

The Godspeed questionnaire (Bartneck et al., 2009) was used to measure perceived anthropomorphism, likeability and intelligence. For each item, the participant was presented a pair of two opposite words and asked to indicate to what extent they perceived that their buddy possessed this quality. Each concept was measured by five word pairs. Examples of pairs are for *anthropomorphism* 'Machine-like' versus 'Humanlike' ($\alpha = .65$); for *likeability* 'Unfriendly' versus 'Friendly' ($\alpha = .86$); and for *intelligence* 'Incompetent' versus 'Competent' ($\alpha = .86$).

2.4.6 Perceived usefulness

Perceived usefulness of the agent was measured by four items (i.e. "Thanks to my buddy I was able to decide faster.") The participant was asked to rate these on a six-point scale, ranging from 'not at all' to 'to a great extent'. ($\alpha = .84$).

2.4.7 Feeling

Feeling was measured with a four item scale to assess the participants' feelings during the experiment. Each item starts with 'I felt…", followed by the words: 'nervous', 'scared', 'worried' and 'anxious'. Answers were rated on a six-point scale ranging from 'not at all' to 'to a great extent' ($\alpha = .77$).

2.4.8 Game experience

Game experience measured to what extent the participant had experience with playing games (specified as virtual reality games, shooter- or fighting games and others) with a single question. The response scale varied from 'never' to 'more than one hour a day'.

2.4.9 Demographics

Lastly, two demographic items assessed the participant's age and gender.

3 Results

3.1 Advice taking

A Repeated-Measures ANOVA was conducted with the between-subject factors Regret (present or absent) and Explanation (present or absent) and the within-subject factors

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

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

a. Computed using alpha = .05

Time (prior to violation [T1] versus after violation [T2] versus after repair [T3]). Here, advice taking was the dependent variable.

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

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

3.2 Trust

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

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

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

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

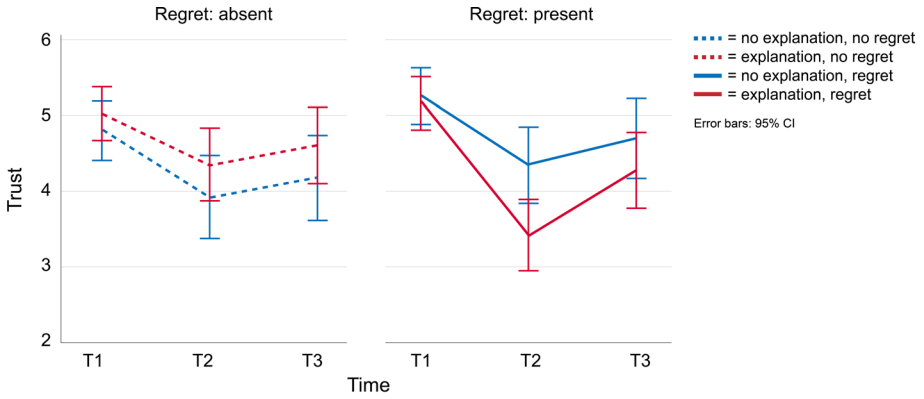

Fig. 3 Estimated marginal means used to represent the relation between Time [T1–T3] (X-axis), trust (Y-axis), explanation (blue line presents explanation absent, red line represents explanation present) and regret (left panel represents regret absent, right pane represents regret present). Error bars represent the 95% confidence interval (Color figure online)

Table 4 Simple main effects of regret, explanation and time [T2–T3]

Regret	Explanation	Δ time		Sig
0	0	T2	T3	0.199
	1	T2	T3	0.142
1	0	T2	T3	0.056
	1	T2	T3	0.000

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

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

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

3.3 Correlations

For the correlations, initial trust (T1) is used as this is considered the purest trust measure with the least interference of occurrences during the experiment. Correlations show that trust was higher when the agent was considered more human-like ($r(64) = .45$, $p < .00$), likeable ($r(64) = .45$, $p < .00$), intelligent ($r(64) = .48$, $p < .00$) and useful ($r(64) = .61$, $p < .00$). Furthermore, the higher the level of trust the more likely the participant was to follow the advice ($r(64) = .51$, $p < 00$). With regard to advice taking, participants were more likely to follow the advice when they perceived the agent as more intelligent ($r(64) = .29$, $p = .02$) and useful ($r(64) = .53$, $p < .00$). Trust ($r(64) = -.42$, $p < .00$) and willingness to follow the advice ($r(64) = -.26$, $p = .04$) was higher when the participant was younger.

4 Discussion

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

Although expressing regret is typically perceived as a human-like quality, these results suggest that saying sorry also makes a difference in rebuilding trust when it comes from a non-human agent. In line with the CASA-paradigm, it indicates that the interpersonal custom of affective apologies can also benefit human-agent interaction [23]. Our findings are in line with studies that showed that computers expressing empathetic emotions were trusted more [3, 24, 44] and studies that find that people prefer to cooperate with virtual agents that express moral emotions [34]. These results support the notion that apology is an effective trust repair strategy in response to a competence-based trust violation [50]. Current findings contradict earlier findings that indicated that apologies where not effective when provided immediately after the agent broke trust [46].The important role of affect in trusting a non-human agent is strengthened by our finding that trust increased when participants perceived the agent as more human-like and likeable [57]. It suggests that a feeling of sincerity in the expression of regret by the non-human agent is the most important for trust repair. This aligns with the belief that affective aspects of trust have the most direct impact on behaviour, since people not only think about trust, but foremost feel it [9].This underlines the relevance of using engaging game environments rather than questionnaires only, since the former method induces physiological responses, increasing ecological validity. The immersiveness of the game environment used in the present study sets this study apart from simpler, more superficial questionnaire-based research and might explain why affect is the predominant factor in our results.

The findings on the effectiveness of the trust repair strategies including regret are somewhat ambiguous, since the trust violation is perceived differently across different participant groups. Although the participants were randomly assigned to each condition and their task was identical up to the point of the trust repair manipulation, the groups that received an apology including an expression of regret showed on average steeper declines in trust in response to the trust violation than the groups that did not. This results in counterintuitive outcomes in which the conditions without regret barely gain in trust after the manipulation,

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

Beyond the generic effect of affect, the combination of both the expression of regret and an explanation proved to be the most effective trust repair strategy. This is in line with the interpersonal study of Scher and Darley [49], which showed that more apology components led to more trust. Our findings are similar to those reported by Akgun et al. [1], who found that apologetic error messages that included both an expression of regret and an explanation had a positive effect. Offering an explanation without an expression of regret had no effect on trust repair. The absence of this effect may be due to the variability in the interpretation of the provided explanation, as became apparent during the debriefing. Some participants reported that they felt more comfortable after the explanation, as it gave more context and transparency, whereas others felt discomfort and suspicion when confronted with the fallibility of the system and with the idea that the agent was functioning on the edge of its abilities. Even though transparent communication is an essential aspect for building trust in human-agent teams [2], this anecdotal evidence suggests that an explanation does not automatically do so.

Generally, explanations contribute to transparency; as it is defined as the provision of information to help the human understand various aspects of agent functioning [27]. A recent study suggests that transparency should be compatible with the user's mental model of the system in order to support accurate trust calibration [31]. A mental model is an internal representation in the mind of one actor about the characteristics of another actor [59]. Different forms of transparency might be needed dependent on whether the humans representation of the system concerns an advanced tool or a teammate. Accordingly, personalized feedback that highlights either the machine's data-analytic capabilities (advanced tool) or its humanlike social functioning (teammate) provides a strategy for trust management [31]. In that sense, an explanation is far more complex than an expression of regret, as there is a wider range of possible underlying messages of the explanation and the way they are articulated. It would be interesting to include the human's mental model of the system (i.e. tool versus teammate) as a mediating factor in follow-up research to reduce the variability. Future personalization could also focus on individual differences that can influence trust development and specially trust repair, such as people's tendency to anthropomorphize [7, 62], propensity to trust [21] and their attitudes and other implicit beliefs and biases towards automation [13, 31, 35, 36].

Even though our results clearly show the importance of affective factors, there are several limitations that need to be taken into consideration. The first one concerns the participants, who were almost all students. The homogeneity of this group influences the representativeness of the study and the generalizability of the results. We for example found a negative correlation between age and trust and age and advice taking, possibly suggesting different attitudes of different age groups towards artificial agents. A second limitation is the absence of a manipulation check. The agent offered one of four types of trust repair strategies: an expression of regret; an explanation; neither, or both. However, the condition where the agent offered neither of the apology components, it still acknowledged that the advice it gave was wrong. This could be interpreted as the agent taking direct responsibility

for its mistake and thus an apology component on its own. Nonetheless, this acknowledging statement was the baseline in every condition. So even if the baseline condition is observed as a form of apology, the other apology components proved to significantly more effective in repairing trust. A third limitation is that we only used one type of trust violation, i.e. a competence-based trust violation. Research suggests that the ground of the trust violation (i.e. competence, benevolence or integrity-based) matters in determining which trust repair approach would be the most effective. An interpersonal study on repairing customer trust after negative publicity showed that emotional reactions are the most effective strategy when aiming to rebuild integrity and benevolence, and that providing sufficient information is essential for improving consumers' judgment about competence [64]. In our study the incorrect advice resulted from the incorrect application of knowledge, which mostly resembles a competence-based trust violation. Accordingly, an explanation would be expected to best fit this type of violation [64]. Yet even with the current task design, affect proves to be the most influential factor in rebuilding trust. Even though we predict that affect would even be stronger in other types of violation, follow-up research is needed to investigate a wider range of trust violations and to determine whether the beneficial effects will last when the same apology is offered repeatedly. A last limitation concerns the ecological validity of the game and its specific content. In the current task the trust violation was induced by a confrontation with an enemy. Although this successfully caused a decline in trust, it is conceivable that the impact of the trust violation and trust repair strategy in the game would differ from its impact in real-life. Possibly an even more immersive environment like virtual reality and a different task will trigger other psychological mechanisms than we have addressed in the present study.

4.1 Implications

There is an ongoing debate about the appropriateness of providing humanized messages by a robot and how far anthropomorphism should go. The current results accords with the view that humans are more likely to collaborate with intelligent agents that show the human-like qualities and traits and which states and that, on a relational level, anthropomorphism can be beneficial [53]. As intelligent agents are increasingly deployed as intelligent teammates, it seems useful to incorporate social skills into their design. These intelligent teammates will be deployed in many contexts, including complex and unpredictable situations, like military operations and city traffic. Even though the technology evolves at a high rate, we must prepare for the inevitability of errors. This study contributes to determining what the psychosocial requirements are for the maintenance and repair of trust in human-agent teaming. Our results suggest that to retain trust in a human-agent team, the ability of actively repairing trust after an error or unintended action should be a fundamental part of the design of intelligent agents. In response to a trust violation, a successful active trust repair strategy should include an explanation for why the error occurred and an expression of regret. Future research in the field of affective computing could explore the potential of measuring the affective states of humans in real-time during their interaction with an agent. This would allow the agent to adapt its trust repair strategies to the type and the intensity of the emotional reaction to the violation, to ensure better calibration.

It is important to note that trust evolves in a complex individual, cultural, and organizational context. Even though the appropriate trust repair strategy depends on many contextual factors such as the type, severity and frequency of the trust violation, it presumably

makes a difference if an intelligent agent offers an apology that is both affective, and informational in an attempt of rebuilding trust.

Authors' contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Esther Kox, Tom Hueting and José Kerstholt. The first draft of the manuscript was written by Esther Kox and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This material is based upon work supported by the Dutch Ministry of Defense's exploratory research program.

Availability of data and material The data that support the findings of this study are available from the corresponding author (E.S. Kox), upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

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.

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.

Consent for publication The Author transfers to Springer (respective to owner if other than Springer and for U.S. government employees: to the extent transferable) the non-exclusive publication rights and he warrants that his/her contribution is original and that he/she has full power to make this grant. The author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors. This transfer of publication rights covers the non-exclusive right to reproduce and distribute the article, including reprints, translations, photographic reproductions, microform, electronic form (offline, online) or any other reproductions of similar nature.

References

1. Akgun, M., Cagiltay, K., & Zeyrek, D. (2010). The effect of apologetic error messages and mood 'states on computer users' self-appraisal of performance. *Journal of Pragmatics, 42*(9), 2430–2448. https://doi.org/10.1016/j.pragma.2009.12.011
2. Barnes, M. J. *et al.* (2014). Designing for humans in autonomous systems: military applications. no. January, pp. 1–30.
3. Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*(2), 161–178.
4. Costa, A. C. (2003). Work team trust and effectiveness. *Personnel Review, 32*(5), 605–622. https://doi.org/10.1108/00483480310488360

5. Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior, 29*(3), 577–579. https://doi.org/10.1016/j.chb.2012.11.023
6. Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/S1071-5819(03)00038-7
7. Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
8. Ferguson, G., & Allen, J. (2011). A cognitive model for collaborative agents. In *AAAI Fall Symposium—Technical Report*, vol. FS-11-01, pp. 112–120.
9. Fine, G. A., & Holyfield, L. (2006). Secrecy, trust, and dangerous leisure: Generating group cohesion in voluntary organizations. *Social Psychology Quarterly*. https://doi.org/10.2307/2787117
10. Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *Proceedings of the 2007 international symposium on collaboration technology system. CTS*, pp. 106–114. https://doi.org/10.1109/CTS.2007.4621745
11. Gambetta, D. (2000). Can we trust trust? In *Trust: Making and breaking cooperative relations*, *electronic* (pp. 212–237). Department of Sociology, University of Oxford.
12. Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254
13. Haselhuhn, M. P., Schweitzer, M. E., & Wood, A. M. (2010). How implicit beliefs influence trust recovery. *Psychological Science, 21*(5), 645–648. https://doi.org/10.1177/0956797610367752
14. Hidalgo, C. A., Orghian, D., Albo-Canals, J., de Almeida, F., & Martin, N. (2021). *How humans judge machines*. Massachusetts Institute of Technology: The MIT Press Cambridge.
15. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570
16. Kahneman, D. (2011). *Thinking fast and thinking slow.*
17. Kim, P. H., Cooper, C. D., Ferrin, D. L., & Dirks, K. T. (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), 104–118.
18. Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/j.obhdp.2005.07.002
19. Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust ? The interaction effect between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics.*
20. Lee, A. Y., Bond, G. D., Russell, D. C., Tost, J., González, C., & Scarbrough, P. S. (2010). Team perceived trustworthiness in a complex military peacekeeping simulation. *Military Psychology, 22*(3), 237–261. https://doi.org/10.1080/08995605.2010.492676
21. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.
22. Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392
23. Lee, J. E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication, *Trust Technol. a Ubiquitous Mod. Environ. Theor. Methodol. Perspect.*, pp. 1–15. https://doi.org/10.4018/978-1-61520-901-9.ch001.
24. Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 203–210. https://doi.org/10.1109/HRI.2010.5453195.
25. Lewicki, R. J., Polin, B., & Lount, R. B. (2016). An exploration of the structure of effective apologies. *Negotiation and Conflict Management Research, 9*(2), 177–196. https://doi.org/10.1111/ncmr.12073
26. Loewenstein, G. F., Hsee, C. K., Weber, E. U., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.127.2.267
27. Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *AAAI Spring Symposium—Technical Report*, vol. SS-13-07, pp. 48–53.
28. Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. https://doi.org/10.1080/14639220500337708

29. Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors, 48*(2), 241–256. https://doi.org/10.1518/001872006777724408
30. Madsen, M., & Gregor, S. (2000). Measuring human–computer trust. In *Proceedings of the Elev. Australas. Conf. Inf. Syst.*, pp. 6–8, 2000. http://books.google.com/books?hl=en&lr=&id=b0yalwi1HDMC&oi=fnd&pg=PA102&dq=The+Big+Five+Trait+Taxonomy:+History,+measurement,+and+Theoretical+Perspectives&ots=758BNaTvOi&sig=L52e79TS6r0Fp2m6xQVESnGt8mw%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=.
31. Matthews, G., Lin, J., Panganiban, A. R., & Long, M. D. (2019). Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems, PP*(November), 1–11. https://doi.org/10.1109/THMS.2019.2947592
32. McKnight, D. H., & Chervany, N. L. (2000) What is trust ? A conceptual analysis and an interdisciplinary model. In *Proceedings of the 2000 American Conference on Information System. AMCI2000 AIS Long Beach CA August 2000*, vol. 346, p. 382. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1876&context=amcis2000.
33. Melo, C. M. D., Gratch, J., & Carnevale, P. J. (2015). Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing, 6*(2), 127–136. https://doi.org/10.1109/TAFFC.2014.2332471
34. Melo, C. M. D., Zheng, L., & Gratch, J. (2009). Expression of moral emotions in cooperating agents. In *Lecture Notes on Computer Science (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5773 LNAI, pp. 301–307. https://doi.org/10.1007/978-3-642-04380-2_32.
35. Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534. https://doi.org/10.1177/0018720812465081
36. Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors, 57*(5), 740–753. https://doi.org/10.1177/0018720815581247
37. Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5–6), 527–539.
38. Olshtain, E., & Cohen, A. (1983). Apology: A speech act set."*Socioling. Lang. Acquis.*, pp. 18–35.
39. Ososky, S. *et al.* (2012). The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. *Unmanned Syst. Technol. XIV*, vol. 8387, no. May 2012, p. 838710. https://doi.org/10.1117/12.923283.
40. Parasuraman, R., De Visser, E., Wiese, E., & Madhavan, P. (2014). Human trust in other humans, automation, robots, and cognitive agents: Neural correlates and design implications. In *Proceedings of the Human Factors Ergon. Soc.*, vol. 2014-Janua, pp. 340–344. https://doi.org/10.1177/1541931214581070.
41. Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors, 57*(4), 545–556. https://doi.org/10.1177/0018720814564422
42. Raue, M., D'Ambrosio, L. A., Ward, C., Lee, C., Jacquillat, C., & Coughlin, J. F. (2019). The influence of feelings while driving regular cars on the perception and acceptance of self-driving cars. *Risk Analysis, 39*(2), 358–374. https://doi.org/10.1111/risa.13267
43. Razzouk, R., & Johnson, T. (2012). Shared cognition. In *Encyclopedia of the Sciences of Learning*.
44. Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human and Robotics Interaction—HRI '09*, p. 245. https://doi.org/10.1145/1514095.1514158.
45. Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human-robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems, 47*(4), 425–436. https://doi.org/10.1109/THMS.2017.2648849
46. Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. In *International conference on social robotics*, vol. 9388 LNCS, pp. 574–583. https://doi.org/10.1007/978-3-319-25554-5_46.
47. Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Prentice Hall.
48. Salas, E., Sims, D. E., & Shawn Burke, C. (2005). Is there A 'big five' in teamwork? *Small Group Research, 36*(5), 555–599. https://doi.org/10.1177/1046496405277134
49. Scher, S. J., & Darley, J. M. (1997). How effective are the things people say to apologize? Effects of the realization of the apology speech act. *Journal of Psycholinguistic Research, 26*(1), 127–140. https://doi.org/10.1023/A:1025068306386

50. Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019). 'I Don't Believe You': Investigating the effects of robot trust violation and repair. In *ACM/IEEE international conference on human-robot interaction*, vol. 2019-March, pp. 57–65. https://doi.org/10.1109/HRI.2019.8673169.
51. Serholt, S., & Barendregt, W. (2016). Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. In *ACM international conference proceeding series*, vol. 23–27-Octo, 2016. https://doi.org/10.1145/2971485.2971536.
52. Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour, 1*(10), 694–696. https://doi.org/10.1038/s41562-017-0202-6
53. Teo, G., Wohleber, R., Lin, J., & Reinerman-jones, L. (2019). The relevance of theory to human-robot teaming research and development, vol. 784, no. January. https://doi.org/10.1007/978-3-319-94346-6.
54. Tolmeijer, S. *et al.* (2020). Taxonomy of trust-relevant failures and mitigation strategies. In *ACM/IEEE international conference on human-robot interaction.*, pp. 3–12. https://doi.org/10.1145/3319502.3374793.
55. Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*. https://doi.org/10.1016/j.jm.2003.01.003
56. Tzeng, J. Y. (2004). Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2004.01.002
57. Visser, E. J. D., et al. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology. Applied, 22*(3), 331–349. https://doi.org/10.1037/xap0000092
58. Visser, E. J. D., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics, 61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725
59. Visser, E. J. D., et al. (2019). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-019-00596-x
60. Vries, P. W. D., Van Den Berg, S. M., & Midden, C. (2015). Assessing technology in the absence of proof: Trust based on the interplay of others opinions and the interaction process. *Human Factors, 57*(8), 1378–1402. https://doi.org/10.1177/0018720815598604
61. Vries, P. W. D., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies, 58*(6), 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9
62. Waytz, A., Cacioppo, J., & Epley, N. (2008). Who sees human? The stability and importance of individual differences in anthropomorphism. *Bone, 23*(1), 1–7. https://doi.org/10.1177/1745691610369336.Who
63. Wright, J. L., Chen, J. Y. C., Barnes, M. J., & Hancock, P. A. (2016). The effect of agent reasoning transparency on automation bias: An analysis of response performance, vol. 9740, pp. 465–477.
64. Xie, Y., & Peng, S. (2009). How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology and Marketing, 26*(7), 572–589. https://doi.org/10.1002/mar.20289