Telematics and Informatics 61 (2021) 101595

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair

Taenyun Kim, Hayeon Song *

Department of Human-Artificial Intelligence Interaction, Sungkyunkwan University, Jongno-Gu, Seoul, 03063, Republic of Korea

ARTICLE INFO

Keywords:
Trust repair
Artificial intelligence
CASA paradigm
Automation bias
Anthropomorphism
Apology attribution

ABSTRACT

Trust is essential in individuals' perception, behavior, and evaluation of intelligent agents. Because, it is the primary motive for people to accept new technology, it is crucial to repair trust when damaged. This study investigated how intelligent agents should apologize to recover trust and how the effectiveness of the apology is different when the agent is human-like compared to machine-like based on two seemingly competing frameworks of the Computers-Are-Social-Actors paradigm and automation bias. A 2 (agent: *Human-like* vs. *Machine-like*) X 2 (apology attribution: *Internal* vs. *External)* between-subject design experiment was conducted ($N = 193$) in the context of the stock market. Participants were presented with a scenario to make investment choices based on an artificial intelligence agent's advice. To see the trajectory of the initial trust-building, trust violation, and trust repair process, we designed an investment game that consists of five rounds of eight investment choices (40 investment choices in total). The results show that trust was repaired more efficiently when a human-like agent apologizes with internal rather than external attribution. However, the opposite pattern was observed among participants who had machine-like agents; the external rather than internal attribution condition showed better trust repair. Both theoretical and practical implications are discussed.

1. Introduction

Trust is an important factor when designing interactive intelligent agents. Trust plays an essential role in human perception, behavior, and assessment of technology (Lee and See, 2004) and can determine how people interact with machines (Li et al., 2008; Siau et al., 2004). Recently, trust in new technologies such as Artificial Intelligence (AI) is becoming more important with the introduction of innovative technologies into areas such as financial (Lui and Lamb, 2018; Goldstein et al., 2019; Puschmann, 2017) and medical domains (Longoni et al., 2019). Trust is one of the main motives for people to accept new technology (Gefen et al., 2003). Lack of trust often leads to users' resistance to technologies such as medical AI (Longoni et al., 2019) and Fintech (Finance Technology; Lui and Lamb, 2018).

As intelligent agents have to predict and provide suggestions in uncertain or unknown situations, their advice might be incorrect. Indeed, they learn and adjust their models by proactively expecting and managing errors (Hollnagel et al., 2006). However, even a single failure can cause serious trust violations and can even be fatal to users' trust. For example, if the advice given by a Fintech AI is

* Corresponding author.
E-mail address: songhy@skku.edu (H. Song).

https://doi.org/10.1016/j.tele.2021.101595
Received 9 October 2020; Received in revised form 12 January 2021; Accepted 12 February 2021
Available online 24 February 2021

wrong, it can cause severe economic consequences to users. Then users might decide not to use the service, even though it was generally successful in the previous interactions and might perform better after learning from the failure.

Several studies have been conducted to identify effective ways for AI agents to regain lost trust (de Visser et al., 2018). One way to repair trust is to apologize (Kim et al., 2004, 2006; de Visser et al., 2018). According to the Computers-Are-Social-Actors (CASA) paradigm, people unknowingly apply social rules in their relationships with computers (Nass et al., 1995; Nass and Moon, 2000; Reeves and Nass, 1996). Thus, it can be postulated that a computer's apology would be as effective as a human's apology in trust repair (Bottom et al., 2002; Lewicki et al., 1996; Ohbuchi et al., 1989).

However, a recent study of the CASA paradigm argues that it is also possible that people have distinctive mental models for intelligent agents that are different from those utilized for humans (Gambino et al., 2020). Scholars who argue an automation bias assert that people have specific expectations and beliefs about machines, so their behavior, perception, and evaluation of a computer would be different (Dijkstra et al., 1998; Dijkstra, 1999; Dzindolet et al., 2003; Lee and See, 2004; Parasuraman and Manzey, 2010). Then, the effect of an apology in a human–computer interaction would be essentially different from a human interaction.

Therefore, this study investigates how AI agents should apologize and the effectiveness of apology type (i.e., internal vs. external) depending on how human-like or machine-like AI agents are perceived. For the types of apology, we relied on the literature on the attribution of apology, one of the most studied trust repair topics in human psychology. An experiment was conducted to explore how internal and external apology attribution produces different consequences depending on one's perception of intelligent agents manipulated by anthropomorphic cues. The results highlight the importance of anthropomorphism in determining how people perceive computers and which apology type repairs trust more efficiently.

2. Literature review

2.1. Trust repair in intelligent agent

Trust is defined as a willingness to rely on others' intentions or behaviors based on positive expectations (Rousseau et al., 1998). Trust can be mainly categorized into two dimensions: cognitive and behavioral. Cognitive (or subjective) trust is the perception and psychological state experienced by individuals expecting that others' intentions or behaviors will not violate the actors' interests (Das and Teng, 2004; Lahno, 2004). Behavioral trust is an action based on cognitive trust (Das and Teng, 2004; Lahno, 2004).

Trust can be violated when one transgresses the expectations of the other. Trust may recover naturally toward a baseline, based on, for example, the duration of the relationship or experience after the violation (Schilke et al., 2013). Alternatively, remedial action can be taken to repair lost trust (Dirks et al., 2009; Gillespie and Dietz, 2009; Kramer and Lewicki, 2010; Tomlinson and Mryer, 2009).

It is vital to study trust in intelligent agents as it plays a crucial role in determining humans' perception and behavior toward technology (Lee and See, 2004). However, the vast majority of studies on trust in computers have focused on other topics such as the influence of trust on over-reliance on the system (Van Dongen and Van Maanen, 2013), improving human–computer trust calibration (McGuirl and Sarter, 2006; de Visser et al., 2014) enhancing trust via system transparency (Koo et al., 2015; Mercado et al., 2016), promoting trust by communicating the agent's intent (Schaefer et al., 2017), and building initial trust and developing trust (Siau and Shen, 2003; Siau and Wang, 2018). Only a few studies asserted the advantages of trust repair in a technological context (Hoffman et al., 2009, 2013).

Recently, the topic of trust repair in intelligent agents has gained scholarly attention (de Visser et al., 2018). Trust repair refers to any action that attempts to regain trust after a trust violation (Kim et al., 2006). There are some ways to recover lost trust in an intelligent agent. For instance, the agent could employ anthropomorphic cues without altering other behavioral features (Pak et al., 2012; Seeger and Heinzl, 2018; de Visser et al., 2016) or explain why it failed (Dzindolet et al., 2003). However, because the study of trust repair in the context of computers is relatively new, there is a lack of knowledge regarding how to repair lost trust in the agent. We believe that this area can benefit from the already well-developed body of literature on apology in the field of human psychology. When trust is damaged, humans apologize. Before discussing the substantial body of literature on apology in human psychology, it is necessary to explain why it would make sense to apply how humans interact with each other to the interaction between humans and intelligent agents by introducing the CASA framework.

2.2. Trust repair dynamics in intelligent agents may be similar to those in humans

According to the CASA paradigm, people mindlessly apply social rules to computers as if computers are human, even without apparent social cues (Nass and Moon, 2000; Nass et al., 1995; Reeves and Nass, 1996). For example, Nass et al. (1995) showed that a computer's personality could be quickly established with minimal linguistic cues by creating a dominant character with strong language expressions such as commands or a submissive one with weak language expressions such as suggestions. Thus, one of the more accessible and less demanding ways of implementing trust repair in an AI agent would be using social responses to recover broken trust in human relationships. For instance, the agent might promise to change in the future (Schweitzer et al., 2006), take responsibility, or give excuses for the failure after the trust violation (Kim et al., 2004, 2006).

The CASA paradigm further argues that people apply a pre-existing mental model of how they would communicate with other humans to computers (Nass et al., 1994; Nass and Lee, 2001; Reeves and Nass, 1996). For instance, people were more attracted to a computer with similar characteristics to themselves than to a computer with different characteristics, as predicted by the similarity attraction theory in the human relationship literature (Nass et al., 1995). Therefore, in the CASA paradigm, trust repair dynamics similar to human–human relationships would also be observed in human–computer interactions.

2.3. Trust repair dynamics in intelligent agents may be different from those in humans

The concept of automation bias is that people have unique perceptions of computers (Dzindolet et al., 2003; Dijkstra et al., 1998; Dijkstra, 1999; Parasuraman and Manzey, 2010). Specifically, people tend to assign positive bias to computerized systems when they consider machines to be more reliable, consistent, and fair than humans (Dzindolet et al., 2003; Lyell and Coiera, 2017; Parasuraman and Manzey, 2010). Hence, when human and computerized experts have comparable skills and abilities, people consider the computer agent's recommendation to be more rational and objective than the human's (Araujo et al., 2020; Dijkstra et al., 1998; Richter et al., 2019). For example, researchers have found that individuals tend to comply more with automated experts than human experts in domains such as medical diagnoses (e.g., X-ray screening; Davis et al., 2020), healthcare recommendations (Araujo et al., 2020), and law enforcement (Araujo et al., 2020). However, people tend to assume that a computer's competence level is fixed and unlikely to change (de Visser et al., 2016). People also take the errors made by computerized experts more seriously than those of human experts (Madhavan et al., 2006).

People rely more on computerized experts than comparable human experts, probably because of the positive bias toward machines and against humans. Madhavan and Wiegmann (2007) showed that people trust human experts more when perceived expertise (i.e., pedigree) of automated and human experts are both high, but the evaluation is reversed when the perceived expertise of computerized experts is high (Pearson et al., 2019). However, because people generally expect higher performance from machines, higher initial trust and expertise are placed on machines than humans in their first interaction (Dzindolet et al., 2003; Parasuraman and Manzey, 2010), and people rely more on machine than on human experts (Araujo et al., 2020; Davis et al., 2020; de Visser et al., 2016; Dijkstra et al., 1998; Richter et al., 2019). Therefore, trust repair dynamics in human–computer interactions would be different from those in human–human relationships.

2.4. Different perceptions of the intelligent agents are due to anthropomorphism

The different expectations from computers may be affected by anthropomorphism. Anthropomorphism is the tendency of people to perceive human traits or qualities in objects and their surroundings (Waytz et al., 2010). People may perceive human-like looks, sounds, or other sensory signals in computers or see them as performing human-like actions (Nowak and Fox, 2018). Anthropomorphism is one of the most determining factors in eliciting human-like responses from computers (Lee, 2010; Gambino et al., 2020; Gong and Nass, 2007; Swinth and Blascovich, 2002).

Studies have shown that when agents are more machine-like, people impose more authority, expect higher performance, and have higher initial trust in them (Dzindolet et al., 2003; Parasuraman and Manzey, 2010). However, people lose their trust in computers more drastically when they make mistakes than in comparable situations with human experts (Madhavan et al., 2006). When the agents are more human-like, people perceive them to have less authority, expect less performance, and commence with lower initial trust than when dealing with machine-like systems (Dzindolet et al., 2003). Nevertheless, trust is maintained longer in human-like systems than for machine-like systems even when mistakes are made (Madhavan et al., 2006; Madhavan and Wiegmann, 2007).

There are few studies on the effect of anthropomorphism on trust repair in intelligent agents. de Visser et al. (2016) showed that anthropomorphism was associated with greater trust repair (de Visser et al., 2016). However, the study also showed that an apology, one of the trust repair attempts, was still efficient even without explicit anthropomorphic cues. Moreover, the uncanny valley theory suggests that if anthropomorphism is overused such that it becomes nearly human-like but not entirely human, it would negatively affect the human perception of machines (Mori et al., 2012). Hence, it is still unclear whether anthropomorphism and trust repair attempts are always beneficial, regardless of trust violation events and different apology types.

2.5. Trust repair and apology

An apology is a display of acknowledging responsibility and regret for a trust violation (Kim et al., 2006). Although an apology's primary purpose is to regain trust, not all apologies successfully achieve this goal. Some studies have shown that trust may recover more successfully if a transgressor acknowledges the responsibility for the trust-violating event (Lewicki and Bunker, 1996; Ohbuchi et al., 1989; Bottom et al., 2002). However others show that an apology may fail to alleviate the adverse outcomes of a trust violation because it entails acknowledging blame (Schlenker, 1980; Sigal et al., 1988; Riordan et al., 1983).

One way to mitigate one's blame while keeping the benefits of apology is to attribute charges to external factors (Kim et al., 2004, 2006). People are generally good at finding the underlying attributes of human behavior (Heider, 1982). When assessing what factors elicit the action, they can successfully exclude the influence of situational factors and identify the attribution internally, which may be related to one's disposition (Kelley, 1973). Thus, external attributions have been considered beneficial for offenders (Crant and Bateman, 1993; Shaw et al., 2003).

People assign less responsibility and blame the transgressors less when given external causes for poor performance than when the transgressors acknowledges their internal limitations (Crant and Bateman, 1993; Wood and Mitchell, 1981). However, the disadvantage of external attribution might be that people making excuses could be considered deceptive, egoistic, and incompetent (Schlenker et al., 2001). Victims of broken trust are more willing to reconcile with an offender who apologizes with internal rather than external attribution (Tomlinson et al., 2004) and display more favorable evaluations and expectations of a future relationship by taking greater responsibility for the act (Hodgins and Liebeskind, 2003). This is because those admitting full blame with an internal attribution are considered more likely to correct mistakes in the future than those who shift their accountability attribution (Kim et al., 2006).

2.6. Attribution of apology

Adding to extensive studies on attribution, Reeder and Brewer (1979) argued that the dimension of judgment could influence the inference process about dispositional attribution. That is, the attribution process differs depending on the dispositional qualities. Specifically, the hierarchically restrictive schema model of attribution postulates that certain dispositions, such as competence, are perceived differently based on performance. For example, people generally think that those with high competence can adjust their performance across a variety range of difficulty levels depending on motivation and task requirements, while those with low competence are expected to perform corresponding to their ability or lower difficulty, but are not expected to achieve above their ability level (Kim et al., 2006; Reeder and Brewer, 1979).

Thus, good performance provides important information to determine that someone is a good performer, not a poor performer who is simply lucky for once. In contrast, low or mediocre performance cannot indicate corresponding disposition because people guess that the actor's achievement is more dependent on the potential influence of situational factors. For instance, people might believe that less competent people simply cannot achieve higher performance because they lack the ability, while highly competent people sometimes make mistakes when they lack the motivation to accomplish the task.

In contrast, when people deem another's disposition to be unchangeable, such as integrity, they believe that even a single poor performance would be enough to be representative of one's disposition and judge them accordingly. They think that those who are honest would show honesty regardless of the situation. However, people also believe that those with a low disposition of integrity (i.e., dishonest people) may sometimes be honest, but other times act dishonestly. Hence, only low performance (i.e., dishonest act) contains valuable and definite information for judging the underlying nature of someone (Reeder and Brewer, 1979).

For example, people tend to believe that morality is an aspect of human nature that cannot change. Therefore, they harshly judge actors who have committed a single immoral action as having an evil personality, even though the same actors may engage in moral behaviors most of the time (Kim et al., 2006; Reeder and Brewer, 1979).

Similarly, people assume that a computer's competence level is fixed and unlikely to change (de Visser et al., 2016). Therefore, algorithmic competence (Dzindolet et al., 2003; Parasuraman and Manzey, 2010) and human integrity (Bierbrauer, 1979; Gawronski, 2004; Sabini and Silver, 1983; Safer, 1980) share a common characteristic. People believe that a highly competent algorithm, similar to honesty in a person, would show good performance consistently and reliably. In contrast, people suppose that a less-competent algorithm, like dishonesty, would occasionally show somewhat good performance in certain cases, such as, when the task is not difficult, the input data are sound or simply by luck.

Thus, good performance cannot be assumed to guarantee a computer's comparable capacity. However, poor performance can is perceived to indicate the commensurate competence of computers when people guess their capabilities (Reeder and Brewer, 1979).

2.7. The current study

The current study speculates that the effect of different types of apology can vary depending on whether anthropomorphic cues are present or not. Anthropomorphism influences the impact of apology and trust because it affects the perception of the human-likeness or machine-likeness of the agent.

Specifically, when anthropomorphic cues are absent, the agent would be perceived as more machine-like, resulting in higher initial trust (Dzindolet et al., 2003), and the agent would be deemed reliable yet difficult to change (de Visser et al., 2016). Information about bad algorithmic competence would override any signs of future improvement implied in internal attribution because people would believe algorithmic competence as stable and internal quality. Therefore, it is more beneficial to alleviate blame, and make external attributions would be more effective than internal ones.

However, when anthropomorphic cues are present, automatic bias would not be likely to be triggered, leading to lower initial trust toward the agent (Dzindolet et al., 2003) The agent will be perceived to be malleable to change because their algorithmic competence would be deemed analogous to human competence. In competence-related trust violation, a message about future improvement in internal attribution would outweigh information about low competence (Kim et al., 2006). The following hypotheses are proposed:

H1. Initial trust would be higher in a machine-like agent and lower in a human-like agent.

H2. Anthropomorphism will determine the effectiveness of distinctive types of apology (i.e., internal or external attribution apology).

H2a. Trust repair is more effective when a machine-like agent apologizes with an external attribution rather than internal attribution.

H2b. Trust repair is more effective when a human-like agent apologizes with an internal attribution rather than external attribution.

3. Method

3.1. Experimental design

The experiment was conducted using a 2×2 between-subjects design, with the level of anthropomorphism (agent type: *Human-like* vs. *Machine-like* agent) and attribute of the apology (apology type: *Internal* vs. *External* attribute apology) as independent variables. The participants were randomly assigned to one of four experimental conditions (*Human-like–Internal:* $n = 47$, *Human-like–External:* $n = 46$, *Machine-like–Internal:* $n = 48$, *Machine-like–External:* $n = 48$). The dependent variable was trust, measured in two different ways as a

cognitive and behavioral measure.

3.2. Participants

A total of 193 graduate and undergraduate students from a university in Seoul, South Korea, participated via an online experiment platform, Pavlovia (Open Science Tools Ltd., 2020). The study was conducted under the 1964 Declaration of Ethical Standards of Helsinki and was granted ethical approval by the ethics committee at the university. All participants provided informed consent prior to participation. Four participants were not included in the analysis since they did not strictly follow the experiment's instructions. The remaining 189 were aged from 18 to 40 years ($M = 22.87$, $SD = 3.26$); 54% ($n = 103$) were female, 43% ($n = 82$) were male, and 2% ($n = 4$) preferred not to identify their gender. Nearly 59% ($n = 112$) had no experience in investing stocks, 26% ($n = 49$) had occasional experience, and 15% ($n = 28$) invested in stocks often. The participants were compensated with 3,000 Korean won (approximately 2.5 U.S. dollars). In addition, on average, additional incentives of 1420 won (1.2 dollars; SD = 290 won [0.2 dollars]) were given depending on the individual's performance during the task.

3.3. Task

An investment game (Yokoi and Nakayachi, 2019) was adopted and modified for this experiment using PsychoJS-2020.1 (Peirce et al., 2019). The participants' goal was to earn as many points as possible by investing points in companies whose stock prices would most likely rise in the coming year. Each participant was given 1000 points as a starting fund. The game had five rounds, with each consisting of eight investment choices for a total of 40 investment choices. Before each choice, each company's information was provided, including the industry, number of employees, founding year, headquarters location, revenue, and operating profit. Participants had to choose between the two options based on the information: a large investment (100 points) or a small investment (10 points). Participants earned twice the number of points they invested if the company turned out to be successful and lost the points they invested if the company failed. Whether or not the company turned out to be successful was entirely hypothetical and preprogrammed by the researchers (see Figs. 1 and 2).

The investment decisions were categorized into (Fig. 3) good and bad decisions. A good decision was investing 100 points (large investment) in a company that succeeded in business so that the participant gain 200 points. Moreover, investing only 10 points was a good decision when the company did not do very well as the participant lose only 10 points. In contrast, a bad decision was investing only 10 points (small investment) in a successful company so that the participant gain only 20 points, not 200 points. Likewise, investing 100 points instead of 10 points in a failed company results in losing 100, not 10 points. At the point of the investment decision, the participants were presented with an AI's suggestion, which they could follow or neglect.

No information on AI reliability was given when introducing the agent. Instead, participants' AI reliability was built upon the

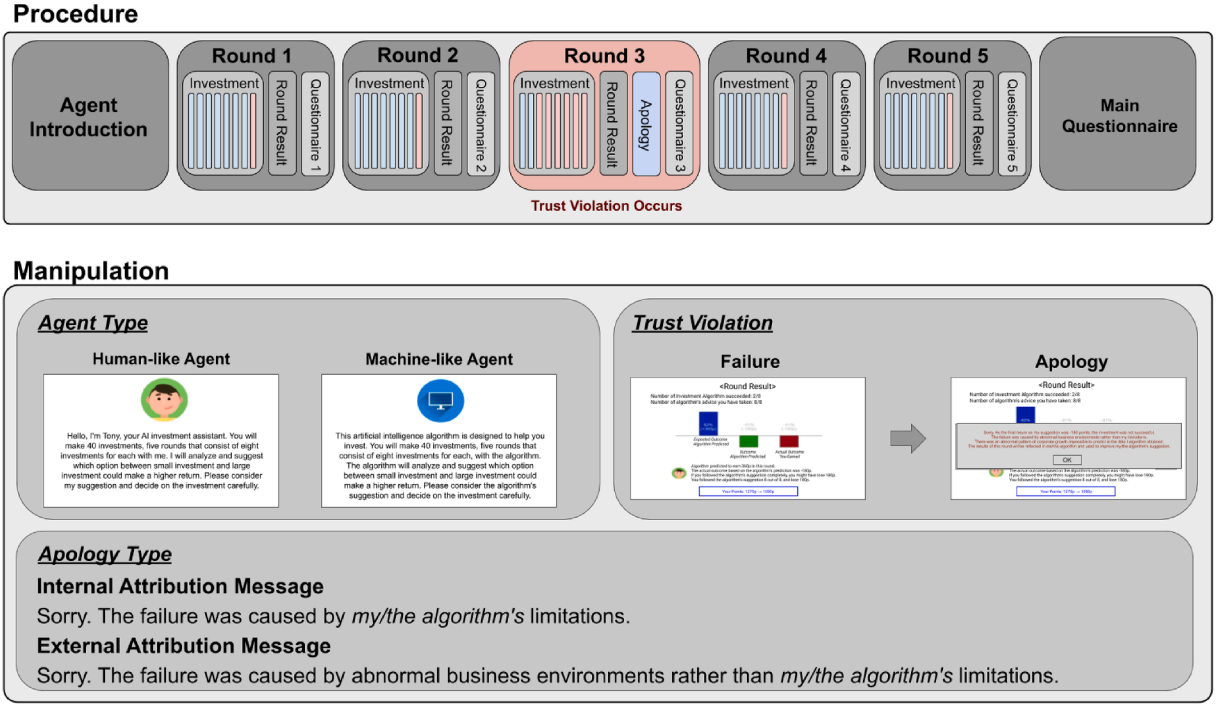

Fig. 1. Overview of the experimental procedure (top) and manipulation of the experiment (bottom).

experience in the first two rounds of interaction with the agents. The AI's suggestions were designed to be 30 good and 10 bad decisions out of 40 choice suggestions (75% accuracy). In all rounds except Round 3, the AI consistently presented seven good and one bad choice out of eight. However, in Round 3, the AI provided only two good and six bad decisions, causing a trust violation.

The feedback on the agent's and participants' performances was presented at the end of each round rather than immediately. The delayed feedback was employed to prevent participants from altering their investment tactics — to comply with the agent or not — during the round. At the end of each round, the total earned points in that round were presented along with the number of decisions that followed the AI's suggestion and the expected earnings if all the investment decisions in that round had been made following the AI's suggestion. Following the results of each round, the participants were asked questions regarding trust in AI. In Round 3, with trust violation, an apology message was presented between the round result and questionnaire. After completing all five rounds, the participants were asked to fill out the main questionnaire.

3.4. Manipulation of anthropomorphic cue

The experiment manipulated three factors to manipulate the level of the anthropomorphic cue of the AI agent: instructions, images, and names (see Fig. 1). In the *Human-like* condition, the profile image of the AI was a human character. Moreover, the agent had a name, said hello, and introduced itself by using the first-person singular pronoun: "*Hello, I'm Tony, your AI investment assistant.* You will make 40 investments, five rounds that consist of eight investments for each *with me.* (…) I will analyze and suggest which option between small investment and large investment could make a higher return. Please consider *my suggestion* and decide on the investment carefully." In contrast, in the *Machine-like* condition, the profile image of AI was a computer. The AI referred to itself as an algorithm and did not say hello: "*This artificial intelligence algorithm is designed to help you invest.* You will make 40 investments, five rounds that consist of eight investments for each, *with the algorithm.* (…) *The algorithm* will analyze and suggest which option between small investment and large investment could make a higher return. Please consider *the algorithm's suggestion* and decide on the investment carefully."

3.5. Manipulation of apology attribution

After the trust violation, the agent apologized with either internal or external attribution (see Fig. 1). In the *Internal* condition, the AI accepted full responsibility for the trust violation: "Sorry. As the final return on *my/the algorithm's* suggestion was XXX point, the investment was not successful. The failure was caused by *my/the algorithm's* limitations. There was a limitation to *my/the algorithm's* ability because *I/the algorithm* couldn't get enough data to predict the growth pattern of a company" In the *External* condition, the AI admitted only partial responsibility for the trust violation and attributed the remaining responsibility to the influence of external abnormality in the environment: "Sorry. As the final return on *my/the algorithm's* suggestion was XXX points, the investment was not successful. The failure was caused by abnormal business environments rather than *my/the algorithm's* limitations. There was an abnormal pattern of corporate growth impossible to predict in the data *I/the algorithm* obtained." In both cases, the AI also promised that it would perform better next time by learning from the failure: "The results of this round will be reflected in me/the algorithm and used to improve my/the algorithm's suggestion."

3.6. Measures

Trust was measured in two different ways: cognitive and behavioral trust. For measuring cognitive trust, in each round, participants were first asked to report the extent to which they trusted the agent (i.e., cognitive trust) and to what extent they trusted their own investment decision (i.e., self-confidence) on a 7point Likert scale. The reported self-confidence in each round served as covariates because it is closely related to trust (Lee and Moray, 1992; Lewandowsky et al., 2000; de Vries et al., 2003). This self-reported measure of cognitive trust is a commonly used, simple, and face-valid measure for assessing trust (Lee and Moray, 1992; Lewandowsky et al.,

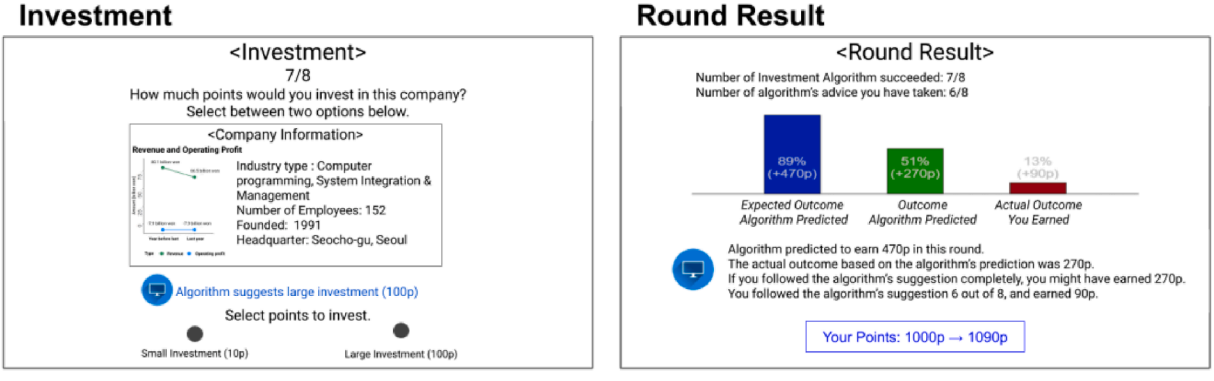

Fig. 2. Investment phase (left) and round result (right) in computer-like agent condition.

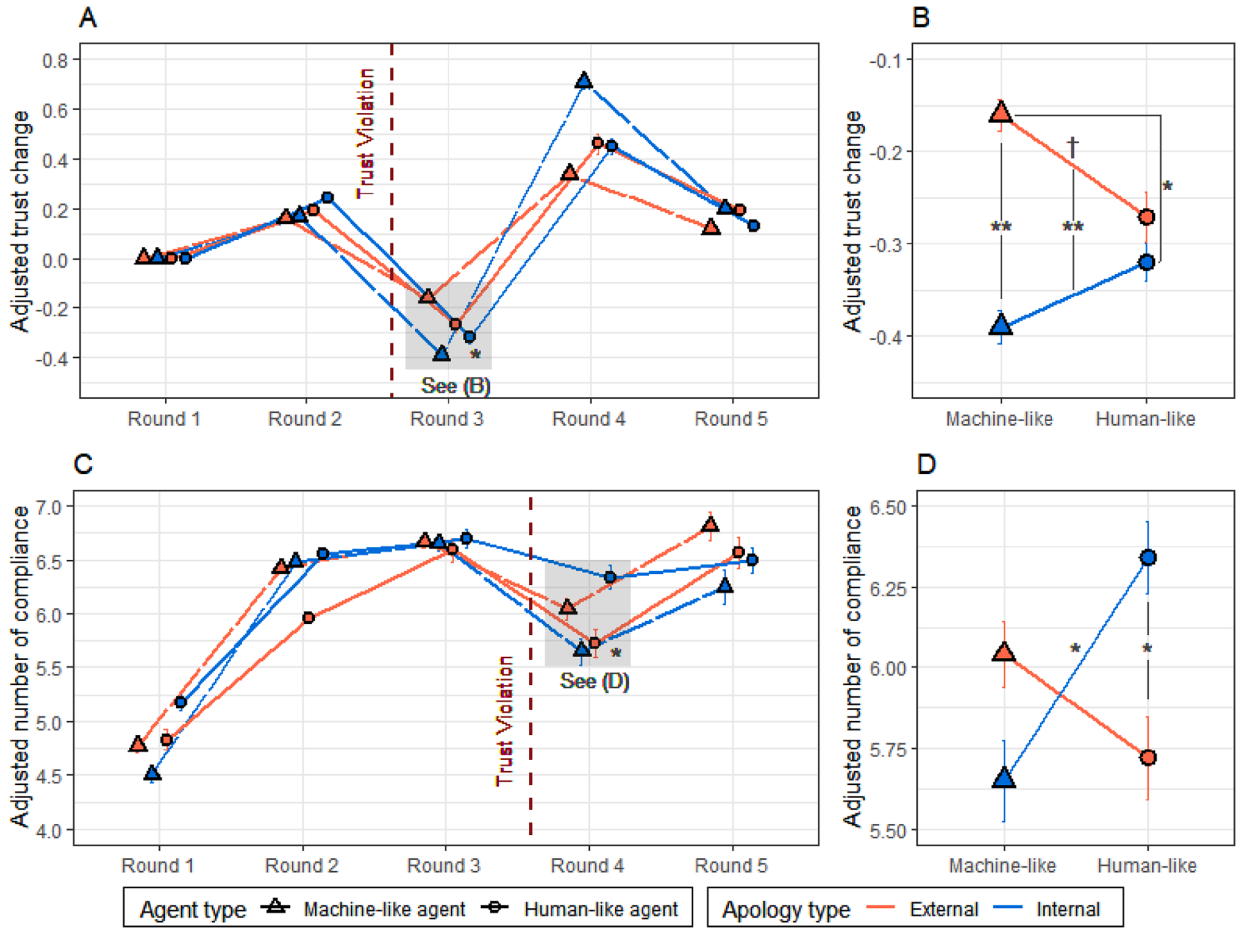

Fig. 3. Trajectory of cognitive and behavioral trust over the five rounds. (A) Cognitive trust over the five rounds. (B) Cognitive trust in Round 3. (C) Behavioral trust over Rounds. (D) Behavioral trust in Round 4. 95% confidence intervals are displayed for each value. Significance level: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

2000; de Vries et al., 2003). To view how trust develops over time, we calculated the change in the participants' cognitive trust in each round. Second, we assessed behavioral trust, also known as compliance. Behavioral trust was measured by counting the number of times the participant followed the recommendation proposed by the automation (Brule et al., 2014). After the five rounds, participants filled out the main questionnaire, which asked about beliefs, attitudes, and intentions on a 7-point Likert scale. To check whether anthropomorphic cues elicited different anthropomorphism levels as intended, we measured anthropomorphism (with three items: machine-like/human-like, unconscious/conscious, artificial/lifelike; $\alpha = 0.81$) in the Godspeed instrument (Bartneck, 2008; Bartneck et al., 2009) for manipulation check of human-likeness. Previous stock investment experience was also measured by asking how frequently participants had invested in stocks before participating in this experiment in a 7-point Likert scale (1: Never to 7: Very Frequently). Previous investment experience served as a covariate for the analysis because prior knowledge or experience influences the trust level (Schaefer et al., 2016).

4. Results

4.1. Manipulation check

First, we checked whether the anthropomorphism was successfully manipulated through Type II Analysis of Variance (ANOVA). Type II ANOVA was used because it is statistically more powerful than the Type III test when there is no interaction (Langsrud, 2003). Confirming the manipulation check, there was a significant main effect of agent type, $F(1, 185) = 4.93$, $p = 0.028$, $\eta^2 = 0.026$. Post-hoc test of Duncan's new multiple range test (MRT) showed a significant difference in anthropomorphism between *Human-like* ($M = 2.74$, $SD = 1.24$) and *Machine-like* ($M = 2.38$, $SD = 0.99$) conditions with 95% family-wise confidence level, $p = 0.027$. The main effect of the apology type was marginally significant, $F(1, 185) = 3.12$, $p = 0.079$, $\eta^2 = 0.017$. Post-hoc test of Duncan's MRT with 95% family-wise confidence level revealed a marginally significant difference in anthropomorphism between *External* ($M = 2.74$, $SD = 1.24$) and *Internal* ($M = 2.38$, $SD = 0.99$) conditions, $p = 0.079$. As expected, there was no significant interaction, $F(1, 185) = 2.57$, $p = 0.150$, $\eta^2 =$

0.011. A possible explanation for the marginally significant result of the apology type could be that the internal apology is also a behavioral dimension of anthropomorphism (de Visser et al., 2016). It is possible that acknowledging full responsibility might have been deemed a human-like feature because no one except humans can take responsibility.

In addition, we used Pearson's correlation to check whether previous investment experience was associated with participants' initial trust in the agent, as shown in previous literature (Schaefer et al., 2016) . The results show that previous investment experience was weakly negatively correlated with the number of compliance in Round 1, $r(187) = -0.16$, $p = 0.029$, although not with cognitive trust in Round 1, $r(187) = -0.06$, $p = 0.385$. These results indicate that participants with previous investment experience tend to comply less with the suggestion of the agents, which implies the influence of prior knowledge and experience on the level of trust, at least behaviorally. Therefore, previous investment experience was controlled as a covariate throughout the analyses.

4.2. Hypothesis testing

4.2.1. Initial Trust-building

In H1, we conjectured that initial trust would be higher in the *Machine-like* than the *Human-like* conditions. To test this, we compared cognitive and behavioral trust before trust violation using Type II ANCOVA. Previous experience of stock investment served as the covariate. For cognitive trust, there was no main effect of agent type in either reported trust in Round 1 (F [1, 183] $= 0.02$, $p = 0.890$, $\eta^2 = 0.000$) nor the trust change from Round 1 to Round 2 (F [1, 183] $= 1.02$, $p = 0.314$, $\eta^2 = 0.005$). For behavioral trust, no significant result was found for Round 1 (F [1, 184] $= 2.06$, $p = 0.153$, $\eta^2 = 0.011$), Round 2 (F [1, 183] $= 0.96$, $p = 0.329$, $\eta^2 = 0.005$), or Round 3 (F [1, 183] $= 0.00$, $p = 0.946$, $\eta^2 = 0.000$). Consequently, the results rejected H1 as no significant difference was found in initial trust between agent type (see Fig. 3).

4.2.2. Cognitive trust

In H2, we expected that agents in *Machine-like–External* or *Human-like– Internal* would be more robust to trust violations than other conditions. Since the trust violation occurred between the cognitive trust measurements in Rounds 2 and 3, we calculated the cognitive trust changes from Round 2 to 3 and compared them across the conditions using Type III ANCOVA. Previous stock investment experience and participants' self-confidence in Round 3 investment served as the covariates. Self-confidence was controlled because it vastly influences an individual's trust (de Vries et al., 2003). The results showed a significant interaction between independent variables, F (1, 183) $= 4.02$, $p = 0.046$, $\eta^2 = 0.020$. Post-hoc test of Duncan's MRT showed a significant difference in the trust change. The level of trust decreased less in *Machine-like–External* ($M = -0.16$, $SD = 0.06$) than other agents: *Machine-like–Internal* ($M = -0.38$, $SD = 0.06$, $p = 0.002$), *Human-like–Internal* ($M = -0.32$, $SD = 0.07$, $p = 0.042$). In addition, the post-hoc test showed a marginally significant difference that there was a trend that trust in *Machine-like–External* damaged less than *Human-like–External* ($M = -0.27$, $SD = 0.09$, $p = 0.099$).

The analysis also revealed a significant main effect of apology type, F (1, 183) $= 10.98$, $p = 0.001$, $\eta^2 = 0.035$. Post-hoc test of Duncan's MRT demonstrated that trust in the *Internal* ($M = -0.21$, $SD = 0.09$) condition decreased significantly less than in the *External* condition ($M = -0.35$, $SD = 0.07$) with 95% family-wise confidence level, $p = 0.007$. However, there was no main effect of agent type, F (1, 183) $= 2.63$, $p = 0.106$, $\eta^2 = 0.000$.

In sum, the results showed that trust in the *Machine-like–External* condition was less damaged than for any other agent after the violation. However, there were no differences in trust between *Human-like* conditions regardless of apology type. These findings support H2a, but not H2b (see Fig. 3).

4.2.3. Behavioral trust

In H2, we anticipated that participants in the *Machine-like–External* or *Human-like–Internal* would comply more with the agent's suggestion in Round 4, the next round of investment, after experiencing the trust violation, compared to other conditions. A Type III ANCOVA was conducted to test H2. In addition to the previous stock investment experience, the number of compliances in Round 3 served as covariates. Behavioral trust (i.e., compliance) right before trust violation was controlled for several reasons. First, it would represent the participant's expectation of the agent's performance before the trust violation, which also affects trust. Second, participants who complied with the agent more in Round 3 would feel more discrepancy in expectation before and after the trust violation, which is also known to affect trust (Lankton et al., 2014).

The analysis revealed a significant interaction between the independent variables F (1, 183) $= 6.19$, $p = 0.014$, $\eta^2 = 0.031$. Post-hoc test of Duncan's MRT showed a significant difference in trust: the participants in the *Human-like– Internal* condition ($M = 6.34$, $SD = 0.38$) agreed more with the agent's advice than those in the *Human-like–External* condition ($M = 5.72$, $SD = 0.43$, $p = 0.047$) or *Machine-like–Internal* ($M = 5.65$, $SD = 0.43$, $p = 0.030$) with 95% family-wise confidence level. There was no significant difference between *Human-like–Internal* and *Machine-like–External* ($M = 6.04$, $SD = 0.35$, $p = 0.327$). The main effects of agent type (F [1, 184] $= 1.27$, $p = 0.260$, $\eta^2 = 0.004$) and apology type (F [1, 184] $= 2.00$, $p = 0.158$, $\eta^2 = 0.001$) were not significant.

The results showed that the participants complied more with the suggestions of a *Human-like–Internal* agent than a *Human-like–External* or *Machine-like– External* agent. However, there was no significant difference between *Human-like–Internal* and *Machine-like–External*. In addition, there was no significant difference between *Machine-like* conditions. Consequently, the analyses support H2b, but not H2a (see Fig. 3).

5. Discussion

This study investigated the interaction effect between anthropomorphism and apology attribution type in the context of trust repair. The results indicate that a machine-like agent's trust repair was more efficient when the agent apologized with external attribution (i.e., mitigating one's blame) than with internal attribution (i.e., acknowledging one's internal limitations). However, those who interacted with human-like agents showed a more efficient trust repair when the agent admitted its fault rather than diminished its accountability. The results imply the existence of more efficient pairs of trust repair strategies and presumably less preferable pairs (see Fig. 3).

5.1. Some combinations of trust repair strategies suit better than others

Our results show that some combinations of trust repair strategies are more efficient than others. However, each pair of anthropomorphism and apology attribution types show strengths in specific dimensions of trust and weaknesses in other dimensions. Although cognitive trust is damaged less drastically when machine-like agents apologize with external rather than internal attribution, no significant results were shown for human-like agents. Conversely, in terms of the behavioral trust, although it was demonstrated that participants comply with the advice more when a human-like agent acknowledges its fault than mitigates it, no significant results were revealed for machine-like agents.

5.1.1. Trust is less damaged when a machine-like agent apologizes with external rather than internal attribution

The cognitive trust results show that acknowledging guilt is harmful in inhibiting trust violation's adverse outcomes. A human-like agent that admits mistakes was not shown to prevent cognitive trust damage compared to a human-like agent that mitigates blame. However, participants complied more with the advice of a human-like agent that accepts its fault than with that of a human-like agent that tries to mitigate its blame. This might be because all apologies are fundamentally acknowledging the responsibility of guilt; hence, they are detrimental to trust repair (Schlenker, 1980; Sigal et al., 1988; Riordan et al., 1983). In turn, it might be inevitable that trust is hampered immediately after the acknowledgment of guilt. Interestingly although both agents admitted their faults, only the human-like agent could recover from damaged trust and get its users to comply with its advice more, leading participants to earn more points. This might be because the machine-like agent was considered less likely to change (de Visser et al., 2016), while a human-like agent was deemed to be more likely to change due to its human-like cues. Therefore, future improvement signals overcame the negative information about competence involved in taking full responsibility by a human-like agent. In contrast, the negative information about competence entailed by taking full responsibility overshadowed the promise of future improvements in machine-like agents (Kim et al., 2006).

5.1.2. People comply with human-like agents that acknowledge its internal limitations

Although a machine-like agent that mitigated its accountability had less adverse outcomes of trust violation in cognitive trust than other agents, the benefits did not hold for behavioral trust. Participants did not comply more with the advice of the machine-like agent that mitigated blames than other agents. It is possible that although the apology with external attribution was successful in mitigating the agent's blame, the participants might believe that that the agent is deceptive, incompetent, and uncaring (Schlenker et al., 2001). Furthermore, the common perception of the machine being consistent (Dzindolet et al., 2003; Parasuraman and Manzey, 2010) and not likely to change (de Visser et al., 2016) could have made users assume that the computer would yield the same negative result in the future. As a result, users were less likely to comply with the advice of machine-like agents.

5.1.3. People do not favor machine-like agents that admit their limitations

Among different strategy pairs, a machine-like agent apologizing with internal attribution was the least beneficial in trust repair. This was consistently demonstrated in both cognitive and behavioral trust results of our study. This might be because participants consider it unnatural for a computer to take responsibility or because their expectations that the machine-like agent would be reliable and consistent were violated when it apologized for its limitation. Although the results were only marginally significant, the trend shows that the participants considered the agent with an internal apology to be more human-like in the manipulation check. This finding is in line with a study that distinguished anthropomorphism in two dimensions: appearance and behavior (Gambino et al., 2020; de Visser et al., 2016). The manipulation of anthropomorphism in this study was about appearance. However, the internal apology would be regarded as behavioral anthropomorphism because taking responsibility is something that only humans would do. Participants might have perceived that an algorithm that acknowledges its fault is unusual because a machine-like object is doing something that only humans can do. This feeling of weirdness caused by close-to humanlike machines might have negatively influenced the impression of the intelligent agent as depicted by the uncanny valley theory (Mori et al., 2012). Indeed, mismatched anthropomorphic features create dissonance and lead to a poor evaluation of an agent (Gong and Nass, 2007). This dissonance might have influenced various trust dimensions and prevented the agent from regaining the lost trust. However, it remains unknown why trust repair was the least effective for the machine-like agent that acknowledges its limitations. This could be because of the mismatch between different anthropomorphism dimensions as we postulated, but that does not rule out other explanations.

Thus, future researchers should investigate whether the low trust repair in a machine-like agent with an internal apology is genuinely due to the dissonance between different anthropomorphism dimensions or other mechanisms. To explore this, we suggest that future studies should identify and differentiate the different anthropomorphism dimensions and investigate the consequences of pairing the different anthropomorphic features in trust repair because different pairs would be more effective in different trust

categories. The same factor may affect cognitive and behavioral trust in heterogeneous ways. For example, Gong and Nass (2007) found that a mismatch between human-likeness and machine-likeness in the face and voice of the agent created delayed information processing in evaluating the agent, while the right match facilitated sociable responses of people to computers. Possible dimensions of anthropomorphism that might be related to trust repair include perceived agency (Fox et al., 2015; de Melo et al., 2014), mind perception (Gray et al., 2007), intentionality (Wiese et al., 2012) or developmental capability (Lee et al., 2005).

5.2. Theoretical implications

5.2.1. Expansion of the effect of apology in trust repair in inter-human relationships to human–computer interaction

Our study extends extant studies in two ways. First, it expands the implications of the literature on trust repair in human–human relationships (Kim et al., 2004, 2006) to a different domain and shows that it is also valid in human–computer interaction. The literature on trust repair has shown that the effect of different types of apology varies depending on the characteristics of trust violating events (Kim et al., 2006). For example, if a trust-violating event is considered easy to change, an apology that acknowledges responsibility is more effective in trust repair than one that mitigates the blame. On the contrary, if a trust-violating event is considered difficult to change, an apology that mitigates the blame is more effective in trust repair than the one that acknowledges responsibility.

These findings in human psychology also seem to be valid in human–computer interaction. When there are anthropomorphic cues, people apply identical social attributions as they do in human–human relationships (Nass et al., 1995, 1994; Nass and Lee, 2001; Reeves and Nass, 1996). In turn, the participants might have believed that the algorithm's competence is not so different from human competence and even perceived that the algorithm's competence could change. In a human–human relationship, individuals are more likely to forgive a transgressor who acknowledges their limitations than one who attributes the problem to external factors (Tomlinson et al., 2004). Individuals also expect a more positive future relationship with those who admit their limitations (Hodgins and Liebeskind, 2003) because they believe they are more likely to correct their faults in the future than those who attempt to mitigate their blame (Kim et al., 2006). Hence, people consider the promise of future improvement related to internal attribution more important than information about low competence when evaluating those who admit their mistakes (Kim et al., 2006). Similarly, a human-like agent that acknowledges its limitations was perceived to be more likely to do better in the future and trust is recovered more efficiently compared to an agent that mitigate the blame.

However, when there were no anthropomorphic cues, the users seemed not to perceive the algorithm to have similar competence to those of humans. Instead, the algorithm's competence appears to be perceived as a fixed trait as human integrity. We observed that those who interacted with a machine-like agent behaved in a way that is similar to the findings of integrity-related trust violations in human–human relationships. The recovery of trust after a trust violation is also in line with that of integrity-related trust violation in human–human relationships (Kim et al., 2006). This may be because human integrity and algorithmic competence share some common characteristics. Laypeople think that human integrity (Bierbrauer, 1979; Gawronski, 2004; Sabini and Silver, 1983; Safer, 1980) and algorithmic competence (Dzindolet et al., 2003; Parasuraman and Manzey, 2010) are both internal and stable dispositions that are less influenced by situations. A highly moral person and a highly capable machine are expected to be consistently good, regardless of the circumstance. In contrast, immoral persons and incapable machines are expected to behave well in some cases, but not on every occasion (Reeder and Brewer, 1979). In turn, the negative information about the dispositional factors involved in acknowledging limitations outweighs positive information about future improvements that are dependent on situational factors (Kim et al., 2006). Thus, even when observing a good performance 99% of the time, it does not guarantee a fully competent machine because a single failure is sufficient to conclude that the algorithm is flawed (Madhavan et al., 2006; Reeder and Brewer, 1979). Consequently, as in integrity-related trust violations, a machine-like agent that can mitigate its blame will be more successful in regaining lost trust after a trust violation.

5.2.2. Role of anthropomorphism in the effect of apology on trust repair

Our study demonstrated that anthropomorphism is one of the most important factors that elicit different perceptions and practices about AI agents, especially in the case of trust. Traditionally, scholars who follow the CASA paradigm have argued that people consider computers as human and apply similar mental models used in human relationships (Nass et al., 1994, 1995; Nass and Moon, 2000; Nass and Lee, 2001; Reeves and Nass, 1996). In contrast, automation bias insists that people attribute unique characteristics to computers that are distinct from their human counterparts (Dijkstra et al., 1998; Dijkstra, 1999; Dzindolet et al., 2003; Parasuraman and Manzey, 2010; de Visser et al., 2016). For example, people treat machines as if they are social beings, but consider them to be more reliable, consistent, and fair than humans (Dzindolet et al., 2003; Parasuraman and Manzey, 2010).

Our findings showed that the seemingly opposite CASA paradigm (Nass and Moon, 2000) and automation bias (Parasuraman and Manzey, 2010) could be explained when considering the moderating effect of anthropomorphism. The participants' responses to the apology varied depending on the level of anthropomorphism. When anthropomorphic cues are present, individuals take the human mental model pathway and employ the same rules of human–human relationships to computers, as evinced in the CASA paradigm (Nass et al., 1994, 1995; Nass and Lee, 2001; Reeves and Nass, 1996). They consider the algorithm's competence to correspond to human competence and behave accordingly. However, when anthropomorphic cues are not present, they adopt a non-human mental model pathway and apply unique characteristics to machines that are distinct from those applied to humans, as shown in automation bias and the recent revision of the CASA paradigm (Dzindolet et al., 2003; Dijkstra et al., 1998; Dijkstra, 1999; Gambino et al., 2020; Parasuraman and Manzey, 2010; de Visser et al., 2016). As a result, they distinguish the algorithm's competence from human competence and behave differently.

However, unlike the studies that showed positive consequences of eliciting a social response using anthropomorphism (Lee, 2010;

Gambino et al., 2020; Gong and Nass, 2007; Swinth and Blascovich, 2002), the current study showed that not using anthropomorphism can also be beneficial in trust repair. If anthropomorphism has only a facilitating effect on people's trust, the agent that conveys both visual and behavioral features of anthropomorphism should outperform the agent that does not have both. However, the findings indicated that a machine-like agent that mitigates its responsibility was more robust in repairing trust violations than other agents with high anthropomorphism and apology with internal attributions, including a human-like agent that acknowledges its guilt. This suggests that AI's communication strategies should vary depending on the agent's level of anthropomorphic features.

Researchers should also further explore the similarities and differences between people's perceptions of human competence and algorithm competence in different boundary conditions. Our study showed that similarities could be amplified, and differences could be distinguished depending on anthropomorphic cues. Indeed, more of such determining factors can influence the user's perception of the machine. We hope that future studies will investigate these factors and uncover new ones.

5.3. Practical implications

Designers of intelligent agents or robotic systems often assume that applying anthropomorphism always has positive outcomes. However, scholars have warned against the addition of anthropomorphic features without caution (Culley and Madhavan, 2013). Sometimes it is better that machines do not have human-like features to achieve design goal (de Melo and Gratch, 2015). For example, a study showed that people have less fear of self-disclosure and are more willing to disclose honestly when they believe they are dealing with a virtual doctor governed by algorithms as opposed to an actual human being (Lucas et al., 2014). In addition, individuals developed greater rapport with virtual therapists without any empathic communication cues than with those with empathic cues such as social dialogue, self-disclosure, especially when the patients did not have major concerns or declined to disclose themselves (Ranjbartabar et al., 2019). In line with these studies, our findings highlight the importance of strategic consideration in using anthropomorphic characteristics because doing so can induce not only positive but also negative results. As anthropomorphic characteristics greatly influence people's perception, intention, and behavior, designers should choose whether to include or exclude anthropomorphic features with deliberate plans (Fink, 2012; DiSalvo et al., 2002). For instance, if a designer desires to make the system appear more competent, consistent, and fair, it would be better to exclude human features from the interface to benefit from automation bias (Dzindolet et al., 2003; Parasuraman and Manzey, 2010). However, if errors are rare, it would be good to implement anthropomorphism because automation bias is particularly susceptible to damage by clumsy or unreliable systems (McGarry et al., 2003). In addition, creating agents that make mistakes similar to humans (e.g., forget the users' name) makes people feel that the machines are more natural and believable (Richards and Bransky, 2014).

Our findings suggest that different trust repair techniques should be implemented depending on whether anthropomorphism is used in designing the agent. For AI agents with anthropomorphic cues, using apology with internal attribution would be more efficient than external attribution. However, for AI agents with low anthropomorphic cues, acknowledging its limitations might be detrimental. Our findings may also provide suggestions for studies related to Explainable Artificial Intelligence (XAI). According to XAI design principles, it is the users' right to know the limitations of the AI to prevent its inappropriate use of the AI without knowing what it can and cannot do. The core issue of XAI is how to explain how AI works and present an AI's performance information. Although the context is different, this study highlights the importance of locus of attribution in communicating about an AI's performance, and further suggests that the locus of attribution may also play an important role in XAI. Future studies should investigate if attribution together with anthropomorphism as the system's design feature affect user's' perception and acceptance of XAI.

5.4. Limitations and future studies

This study has some limitations despite its notable implications. First, our study failed to replicate a machine-like agent's advantage in building the initial trust relationship. Several studies have shown that machine-like agents are trusted more in the beginning than human-like agents (Madhavan et al., 2006; Madhavan and Wiegmann, 2007). However, our study showed no significant result in either initial trust or the trust development stage. Several factors may have contributed to these unexpected results. For example, as a relatively recent trend, individuals tend to over-trust AI. Since Alpha Go won against the World Go champion, Lee Sedol, in 2016 (Borowiec, 2016), the public, in general, has been highly impressed by the advancement in AI technology.

Second, this study did not use a multi-item trust instrument. We purposely used a single-item measure in this experiment to track the changes in trust. Because our trust metric was repeatedly measured over five rounds, the questionnaire needed to be simple and brief rather than long and time-consuming. However, using a single-item trust measure might have compromised the construct validity because trust is a multi-faceted construct (Lahno, 2004). In addition, multi-item measures are generally considered innately more reliable than single-item measures (Churchill Jr, 1979; Peter, 1979) despite the controversies on the benefits of multi-item metrics over single-item (Bergkvist and Rossiter, 2007). Thus, considering that competence and integrity were found to be relevant to trust, it may be useful to use a trust instrument in interpersonal relationships such as the integrative model of organizational trust (consisting of ability, benevolence, and integrity) by Mayer et al. (1995) or source credibility (consisting of competence, goodwill, and trustworthiness) by McCroskey and Teven (1999).

Finally, this study may have some boundary conditions. The experiments were conducted in a high-risk, self-relevant, and highly-motivating context in which the points earned by the participants were converted to actual money and given as incentives. It is not certain whether the findings in this study will be identical in other contexts, involving low-risk, self-irrelevant, and unmotivating tasks and contexts. Because risk perception, self-relevance, and motivation are essential in deciding whether people go through systematic and thorough consideration or heuristic and simple inference in evaluating persuasion (Petty and Cacioppo, 1986), it would be

worthwhile to investigate trust repair in a different context and see whether results are obtained. We hope that future studies will further explore these topics by differentiating the task and context and identifying new contextual variables previously not investigated

6. Conclusion

As the interactions between human and AI agents is have important implications, it is becoming crucial to understand how trust can be recovered after intentional or unintentional trust violations. The findings of this study showed that some combinations of trust repair strategies are more efficient than others; specifically, a human-like agent apologizing with acknowledgment of its limitations and machine-like agent apologizing with external attribution were more useful in recovering lost trust than other conditions. This study has demonstrated that the findings in the literature on trust repair in human relationships can be applied to similar trust repair contexts in human–computer interaction. We also addressed the importance of anthropomorphism in determining the pathway between human and non-human mental model paths in trust repair. Our results further revealed that promoting anthropomorphism is not always beneficial, and a mismatch between the anthropomorphic dimensions can be harmful to trust repair. In the future, more studies will be conducted to reveal the role of anthropomorphism in the perception of intelligent agents and their influence on trust repair.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) program (20200-01816) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). Also, we would like to thank Eun Ji Kim for providing valuable insight into this study and Jieun Kim, Doha Kim, and Kyungha Lee for providing priceless comments and feedback for the manuscript.

References

Araujo, T., Helberger, N., Kruikemeier, S., De Vreese, C.H., 2020. In ai we trust? perceptions about automated decision-making by artificial intelligence. AI Soc. 1–13 https://doi.org/10.1007/s00146-019-00931-w.

Bartneck, C., Kulíc, D., Croft, E., Zoghbi, S., 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. 1, 71–81. https://doi.org/10.1007/s12369-008-0001-3.

Bartneck, C., 2008. The godspeed questionnaire series. http://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/.

Bergkvist, L., Rossiter, J.R., 2007. The predictive validity of multiple-item versus single-item measures of the same constructs. J. Mark. Res. 44, 175–184. https://doi.org/10.1509/jmkr.44.2.175.

Bierbrauer, G., 1979. Why did he do it? attribution of obedience and the phenomenon of dispositional bias. Eur. J. Soc. Psychol. 9, 67–84. https://doi.org/10.1002/ejsp.2420090106.

Borowiec, S., 2016. AlphaGo seals 4-1 victory over Go grandmaster Lee sedol. The Guardian URL: https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol/.

Bottom, W.P., Gibson, K., Daniels, S.E., Murnighan, J.K., 2002. When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. Organiz. Sci. 13, 497–513. https://doi.org/10.1287/orsc.13.5.497.7816.

Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Haselager, P., 2014. Do robot performance and behavioral style affect human trust? Int. J. Soc. Robot. 4, 519–531. https://doi.org/10.1007/s12369-014-0231-5.

Churchill Jr, G.A., 1979. A paradigm for developing better measures of marketing constructs. J. Mark. Res. 16, 64–73. https://doi.org/10.2307/3150876.

Crant, J.M., Bateman, T.S., 1993. Assignment of credit and blame for performance outcomes. Acad. Manag. J. 36, 7–27. https://doi.org/10.5465/256510.

Culley, K.E., Madhavan, P., 2013. A note of caution regarding anthropomorphism in hci agents. Comput. Hum. Behav. 29, 577–579. https://doi.org/10.1016/j.chb.2012.11.023.

Das, T., Teng, B.S., 2004. The risk-based view of trust: A conceptual framework. J. Bus. Psychol. 19, 85–116. https://doi.org/10.1023/B:JOBU.0000040274.23551.1b.

Davis, J., Atchley, A., Smitherman, H., Simon, H., Tenhundfeld, N., 2020. Measuring automation bias and complacency in an x-ray screening task. 2020 Syst. Inf. Eng. Des. Symp. (SIEDS). IEEE, pp. 1–5.

de Melo, C.M., Gratch, J., 2015. Beyond believability: quantifying the differences between real and virtual humans. In: Int. Conf. Intell. Virtual Agents. Springer, pp. 109–118.

de Melo, C.M., Gratch, J., Carnevale, P.J., 2014. Humans versus computers: Impact of emotion expressions on people's decision making. IEEE Trans. Affect. Comput. 6, 127–136. https://doi.org/10.1109/TAFFC.2014.2332471.

de Visser, E.J., Cohen, M., Freedy, A., Parasuraman, R., 2014. A design methodology for trust cue calibration in cognitive agents. In: Int. conf. virtual, augmented & mixed reality. Springer, pp. 251–262.

de Visser, E.J., Monfort, S.S., McKendrick, R., Smith, M.A., McKnight, P.E., Krueger, F., Parasuraman, R., 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. J. Exp. Psychol.: Appl. 22, 331. https://doi.org/10.1037/xap0000092.

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

de Vries, P., Midden, C., Bouwhuis, D., 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. Int. J. Human-Comput. Stud. 58, 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9.

Dijkstra, J.J., Liebrand, W.B., Timminga, E., 1998. Persuasiveness of expert systems. Behav. Inform. Technol. 17, 155–163. https://doi.org/10.1080/014492998119526.

Dijkstra, J.J., 1999. User agreement with incorrect expert system advice. Behav. Inform. Technol. 18, 399–411. https://doi.org/10.1080/014492999118832.

Dirks, K.T., Lewicki, R.J., Zaheer, A., 2009. Reparing relationships within and between organizations: building a conceptual foundation. Acad. Manag. Rev. 34, 68–84. https://doi.org/10.5465/amr.2009.
Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P., 2003. The role of trust in automation reliance. Int. J. Hum. Comput. Stud. 58, 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7.
Fink, J., 2012. Anthropomorphism and human likeness in the design of robots and human-robot interaction. Int. Conf. Soc. Robot. 199–208. https://doi.org/10.1007/978-3-642-34103-8_20.
Fox, J., Ahn, S.J., Janssen, J.H., Yeykelis, L., Segovia, K.Y., Bailenson, J.N., 2015. Avatars versus agents: a meta-analysis quantifying the effect of agency on social influence. Hum. Comp. Interact. 30, 401–432. https://doi.org/10.1080/07370024.2014.921494.
Gambino, A., Fox, J., Ratan, R.A., 2020. Building a stronger casa: Extending the computers are social actors paradigm. Hum. Mach. Commun. 1, 5. https://doi.org/10.30658/hmc.1.5.
Gawronski, B., 2004. Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. Eur. Rev. Soc. Psychol. 15, 183–217. https://doi.org/10.1080/10463280440000026.
Gefen, D., Karahanna, E., Straub, D.W., 2003. Trust and tam in online shopping: An integrated model. MIS Quart. 27, 51–90. https://doi.org/10.5555/2017181.2017185.
Gillespie, N., Dietz, G., 2009. Trust repair after an organization-level failure. Acad. Manage. Rev. 34, 127–145. https://doi.org/10.5465/amr.2009.35713319.
Goldstein, I., Jiang, W., Karolyi, G.A., 2019. To fintech and beyond. The Review of Financial Studies 32, 1647–1661. https://doi.org/10.1093/rfs/hhz025.
Gong, L., Nass, C., 2007. When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. Human Commun. Res. 33, 163–193. https://doi.org/10.1111/j.1468-2958.2007.00295.x.
Gray, H.M., Gray, K., Wegner, D.M., 2007. Dimensions of mind perception. Science 315, 619. https://doi.org/10.1126/science.1134475.
Heider, F., 1982. The psychology of interpersonal relations. Psychology Press.
Hodgins, H.S., Liebeskind, E., 2003. Apology versus defense: Antecedents and consequences. J. Exp. Soc. Psychol. 39, 297–316. https://doi.org/10.1016/S0022-1031(03)00024-6.
Hoffman, R.R., Johnson, M., Bradshaw, J.M., Underbrink, A., 2013. Trust in automation. IEEE Intelligent Systems 28, 84–88. https://doi.org/10.1109/MIS.2013.24.
Hoffman, R.R., Lee, J.D., Woods, D.D., Shadbolt, N., Miller, J., Bradshaw, J.M., 2009. The dynamics of trust in cyberdomains. IEEE Intell. Syst. 24, 5–11. https://doi.org/10.1109/MIS.2009.124.
Hollnagel, E., Woods, D.D., Leveson, N., 2006. Resilience engineering: Concepts and precepts. Ashgate Publishing Ltd.
Kelley, H.H., 1973. The processes of causal attribution. Am. Psychol. 28, 107. https://doi.org/10.1037/h0034225.
Kim, P.H., Ferrin, D.L., Cooper, C.D., Dirks, K.T., 2004. Removing the shadow of suspicion: the effects of apology versus denial for repairing competence versus integrity-based trust violations. J. Appl. Psychol. 89, 104. https://doi.org/10.1037/0021-9010.89.1.104.
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. Organ. Behav. Hum. Decis. Process. 99, 49–65. https://doi.org/10.1016/j.obhdp.2005.07.002.
Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., Nass, C., 2015. Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. Int. J. Interact. Des. Manuf. (IJIDeM) 9, 269–275. https://doi.org/10.1007/s12008-014-0227-2.
Kramer, R.M., Lewicki, R.J., 2010. Repairing and enhancing trust: Approaches to reducing organizational trust deficits. Acad. Manage. Annals 4, 245–277. https://doi.org/10.5465/19416520.2010.487403.
Lahno, B., 2004. Three aspects of interpersonal trust. Analyse kritik 26, 30–47. https://doi.org/10.1515/auk-2004-0102.
Langsrud, Ø., 2003. Anova for unbalanced data: Use type ii instead of type iii sums of squares. Stat Comput. 13, 163–167. https://doi.org/10.1023/A:1023260610025.
Lankton, N., McKnight, D.H., Thatcher, J.B., 2014. Incorporating trust-in technology into expectation disconfirmation theory. J. Strat. Informat. Syst. 23, 128–145. https://doi.org/10.1016/j.jsis. 2013.09.001.
Lee, K.M., Park, N., Song, H., 2005. Can a robot be perceived as a developing creature? effects of a robot's long-term cognitive developments on its social presence and people's social responses toward it. Hum. Commun. Res. 31, 538–563. https://doi.org/10.1111/j.1468-2958.2005.tb00882.x.
Lee, J., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 1243–1270. https://doi.org/10.1080/00140139208967392.
Lee, J.D., See, K.A., 2004. Trust in automation: Designing for appropriate reliance. Hum. Factors 46, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.
Lee, E.J., 2010. What triggers social responses to flattering computers? experimental tests of anthropomorphism and mindlessness explanations. Commun. Res. 37, 191–214. https://doi.org/10.1177/0093650209356389.
Lewandowsky, S., Mundy, M., Tan, G., 2000. The dynamics of trust: Comparing humans to automation. J. Exp. Psychol.: Appl. 6, 104. https://doi.org/10.1037/1076-898X.6.2.104.
Lewicki, R.J., Bunker, B.B., et al., 1996. Developing and maintaining trust in work relationships. Trust Organiz. Front. Theory Res. 114, 139.
Li, X., Hess, T.J., Valacich, J.S., 2008. Why do we trust new technology? a study of initial trust formation with organizational information systems. J. Strateg. Inf. Syst. 17, 39–71. https://doi.org/10.1016/j.jsis.2008.01.001.
Longoni, C., Bonezzi, A., Morewedge, C.K., 2019. Resistance to medical artificial intelligence. J. Consum. Res. 46, 629–650. https://doi.org/10.1093/jcr/ucz013.
DiSalvo, C.F., Gemperle, F., Forlizzi, J., Kiesler, S., 2002. All robots are not created equal: the design and perception of humanoid robot heads. Proc. Conf. Des. interact. syst. 321–326. https://doi.org/10.1145/778712.778756.
Open Science Tools Ltd., 2020. Pavlovia. https://pavlovia.org/.
Lucas, G.M., Gratch, J., King, A., Morency, L.P., 2014. It's only a computer: Virtual humans increase willingness to disclose. Comput. Hum. Behav. 37, 94–100. https://doi.org/10.1016/j.chb.2014.04.043.
Lui, A., Lamb, G.W., 2018. Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector. Informat. Commun. Technol. Law 27, 267–283. https://doi.org/10.1080/13600834.2018.1488659.
Lyell, D., Coiera, E., 2017. Automation bias and verification complexity: a systematic review. J. Am. Med. Inform. Assoc. 24, 423–431. https://doi.org/10.1093/jamia/ocw105.
Madhavan, P., Wiegmann, D.A., 2007. Similarities and differences between human–human and human–automation trust: an integrative review. Theoret. Issues Ergon. Sci. 8, 277–301. https://doi.org/10.1080/14639220500337708.
Madhavan, P., Wiegmann, D.A., Lacson, F.C., 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. Hum. factors 48, 241–256. https://doi.org/10.1518/001872006777724408.
Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An integrative model of organizational trust. Acad. Manage. Rev. 20, 709–734. https://doi.org/10.2307/258792.
McCroskey, J.C., Teven, J.J., 1999. Goodwill: A reexamination of the construct and its measurement. Commun. Monograp. 66, 90–103. https://doi.org/10.1080/03637759909376464.
McGarry, K., Rovira, E., Parasuraman, R., 2003. Effects of task duration and type of automation support on human performance and stress in a simulated battlefield engagement task. Proc. Hum. Factors Ergonomics Soc. Annu. Meeting 548–552. https://doi.org/10.1177/154193120304700362.
McGuirl, J.M., Sarter, N.B., 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Hum. Factors 48, 656–665. https://doi.org/10.1518/001872006779166334.
Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., Procci, K., 2016. Intelligent agent transparency in human–agent teaming for multi-uxv management. Hum. Factors 58, 401–415. https://doi.org/10.1177/0018720815621206.
Mori, M., MacDorman, K.F., Kageki, N., 2012. The uncanny valley [from the field]. IEEE Rob. Autom. Mag. 19, 98–100. https://doi.org/10.1109/MRA.2012.2192811.
Nass, C., Moon, Y., 2000. Machines and mindlessness: Social responses to computers. J. Soc. Issues 56, 81–103. https://doi.org/10.1111/0022-4537.00153.
Nass, C., Steuer, J., Tauber, E.R., 1994. Computers are social actors. Proc. Conf. Hum. factors Comput. systs. 72–78. https://doi.org/10.1145/191666.191703.

Nass, C., Moon, Y., Fogg, B.J., Reeves, B., Dryer, C., 1995. Can computer personalities be human personalities? Proc. Conf. Hum. Factors Comput. Systs. 228–229. https://doi.org/10.1006/ijhc.1995.1042.
Nass, C., Lee, K.M., 2001. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency attraction. J. Exp. Psychol. Appl. 7, 171. https://doi.org/10.1037/1076-898X.7.3.171.
Nowak, K.L., Fox, J., 2018. Avatars and computer-mediated communication: a review of the definitions, uses, and effects of digital representations. Rev. Commun. Res. 6, 30–53. https://doi.org/10.12840/issn.2255-4165.2018.06.01.015.
Ohbuchi, K.i., Kameda, M., Agarie, N., 1989. Apology as aggression control: its role in mediating appraisal of and response to harm. J. Person. Soc. Psychol. 56, 219–227. https://doi.org/10.1037/0022-3514.
Pak, R., Fink, N., Price, M., Bass, B., Sturre, L., 2012. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics 55, 1059–1072. https://doi.org/10.1080/00140139.2012.691554.
Parasuraman, R., Manzey, D.H., 2010. Complacency and bias in human use of automation: An attentional integration. Hum. Factors 52, 381–410. https://doi.org/10.1177/0018720810376055.
Pearson, C.J., Geden, M., Mayhorn, C.B., 2019. Who's the real expert here? pedigree's unique bias on trust between human and automated advisers. Appl. Ergon. 81, 102907. https://doi.org/10.1016/j.apergo.2019.102907.
Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K., 2019. Psychopy2: Experiments in behavior made easy. Behav. Res. Methods 51, 195–203. https://doi.org/10.3758/s13428-018-01193-y.
Peter, J.P., 1979. Reliability: A review of psychometric basics and recent marketing practices. J. Market. Res. 16, 6–17. https://doi.org/10.2307/3150868.
Petty, R.E., Cacioppo, J.T., 1986. Communication and persuasion: Central and peripheral routes to attitude change. Springer Science & Business Media.
Puschmann, T., 2017. Fintech. Business & Information. Syst. Eng. 59, 69–76. https://doi.org/10.1007/s12599-017-0464-6.
Ranjbartabar, H., Richards, D., Bilgin, A., Kutay, C., 2019. First impressions count! the role of the human's emotional state on rapport established with an empathic versus neutral virtual therapist. IEEE Trans. Affective Comput. https://doi.org/10.1109/TAFFC.2019.2899305.
Reeder, G.D., Brewer, M.B., 1979. A schematic model of dispositional attribution in interpersonal perception. Psychol. Rev. 86, 61. https://doi.org/10.1037/0033-295X.86.1.61.
Reeves, B., Nass, C.I., 1996. The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.
Richards, D., Bransky, K., 2014. Forgetmenot: What and how users expect intelligent virtual agents to recall and forget personal conversational content. Int. J. Hum Comput Stud. 72, 460–476. https://doi.org/10.1016/j.ijhcs.2014.01.005.
Richter, R.M., Valladares,, M.J., Sutherland, S.C., 2019. Effects of the source of advice and decision task on decisions to request expert advice. Proceedings of the 24th International Conference on Intelligent User Interfaces 469–475. https://doi.org/10.1145/3301275.3302279.
Riordan, C.A., Marlin, N.A., Kellogg, R.T., 1983. The effectiveness of accounts following transgression. Social Psychol. Quart. 46, 213–219. https://doi.org/10.2307/3033792.
Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C., 1998. Not so different after all: A cross-discipline view of trust. Acad. Manag. Rev. 23, 393–404. https://doi.org/10.5465/amr.1998.926617.
Sabini, J., Silver, M., 1983. Dispositional vs. situational interpretations of Milgram's obedience experiments:" the fundamental attributional error. J. Theory Soc. Behav. https://doi.org/10.1111/j.1468-5914.1983.tb00468.x.
Safer, M.A., 1980. Attributing evil to the subject, not the situation: Student reaction to Milgram's film on obedience. Pers. Soc. Psychol. Bull. 6, 205–209. https://doi.org/10.1177/014616728062003.
Schaefer, K.E., Chen, J.Y., Szalma, J.L., Hancock, P.A., 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Hum. Factors 58, 377–400. https://doi.org/10.1177/0018720816634228.
Schaefer, K.E., Straub, E.R., Chen, J.Y., Putney, J., Evans III, A.W., 2017. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. Cognit. Syst. Res. 46, 26–39. https://doi.org/10.1016/j.cogsys.2017.02.002.
Schilke, O., Reimann, M., Cook, K.S., 2013. Effect of relationship experience on trust recovery following a breach. Proc. Natl. Acad. Sci. 110, 15236–15241. https://doi.org/10.1073/pnas.1314857110.
Schlenker, B.R., 1980. Impression management. Brooks/Cole, Monterey, CA, pp. 79–80.
Schlenker, B.R., Pontari, B.A., Christopher, A.N., 2001. Excuses and character: Personal and social implications of excuses. Personal. Soc. Psychol. Rev. 5, 15–32. https://doi.org/10.1207/S15327957PSPR0501_2.
Schweitzer, M.E., Hershey, J.C., Bradlow, E.T., 2006. Promises and lies: Restoring violated trust. Organ. Behav. Hum. Decis. Process. 101, 1–19. https://doi.org/10.1016/j.obhdp.2006.05.005.
Seeger, A.M., Heinzl, A., 2018. Human versus machine: Contingency factors of anthropomorphism as a trust-inducing design strategy for conversational agents. In: Information systems and neuroscience. Springer, pp. 129–139. https://doi.org/10.1007/978-3-319-67431-5_15.
Shaw, J.C., Wild, E., Colquitt, J.A., 2003. To justify or excuse?: A meta analytic review of the effects of explanations. J. Appl. Psychol. 88, 444. https://doi.org/10.1037/0021-9010.88.3.444.
Siau, K., Shen, Z., 2003. Building customer trust in mobile commerce. Commun. ACM 46, 91–94. https://doi.org/10.1145/641205. 641211.
Siau, K., Sheng, H., Nah, F., Davis, S., 2004. A qualitative investigation on consumer trust in mobile commerce. Int. J. Electron. Business 2, 283–300. https://doi.org/10.1504/IJEB.2004.005143.
Siau, K., Wang, W., 2018. Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technol. J. 31, 47–53.
Sigal, J., Hsu, L., Foodim, S., Betman, J., 1988. Factors affecting perceptions of political candidates accused of sexual and financial misconduct. Political Psychol. 273–280. https://doi.org/10.2307/3790956.
Swinth, K., Blascovich, J., 2002. Perceiving and responding to others: Human-human and human-computer social interaction in collaborative virtual environments. Proceedings of the 5th Annual International Workshop on PRESENCE. doi:10.1.1.495.9042.
Tomlinson, E.C., Mryer, R.C., 2009. The role of causal attribution dimensions in trust repair. Acad. Manage. Rev. 34, 85–104. https://doi.org/10.5465/amr.2009.35713291.
Tomlinson, E.C., Dineen, B.R., Lewicki, R.J., 2004. The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. J. Manage. 30, 165–187. https://doi.org/10.1016/j.jm.2003.01.003.
Van Dongen, K., Van Maanen, P.P., 2013. A framework for explaining reliance on decision aids. Int. J. Hum Comput Stud. 71, 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018.
Waytz, A., Cacioppo, J., Epley, N., 2010. Who sees human? the stability and importance of individual differences in anthropomorphism. Perspectives on Psychol. Sci. 5, 219–232. https://doi.org/10.1177/1745691610369336.
Wiese, E., Wykowska, A., Zwickel, J., Müller, H.J., 2012. I see what you mean: how attentional selection is shaped by ascribing intentions to others. PLoS ONE 7, e45391. https://doi.org/10.1371/journal.pone.0045391.
Wood, R.E., Mitchell, T.R., 1981. Manager behavior in a social context: The impact of impression management on attributions and disciplinary actions. Organizational Behavior and Human Performance 28, 356–378. https://doi.org/10.1016/0030-5073(81)90004-0.
Yokoi, R., Nakayachi, K., 2019. The effect of shared investing strategy on trust in artificial intelligence. Japan. J. Exp. Soc. Psychol. 59, 46–50. https://doi.org/10.2130/jjesp.1819.