심리학 석사 학위논문

Effects of Apology and Empathy for Recovering from Trust Violations of Chatbots

챗봇이 신뢰 위반으로부터 회복하는 데 사과와 공감이 미치는 영향

2022 년 8 월

서울대학교 대학원

심리학과 인지심리 전공

박 재 은

Effects of Apology and Empathy for Recovering from Trust Violations of Chatbots

지도 교수 한 소 원

이 논문을 심리학석사 학위논문으로 제출함
2022 년 6 월

서울대학교 대학원
심리학과 인지심리 전공
박 재 은

박재은의 심리학석사 학위논문을 인준함
2022 년 8 월

위 원 장 김청택 (인)

부위원장 유제광 (인)

위 원 한소원 (인)

Abstract

Jae Eun Park

Graduate School of Psychology

Seoul National University

In the present study, we investigated how chatbots can recover user trust after making errors. In two experiments, participants had a conversation with a chatbot about their daily lives and personal goals. After giving an inadequate response to the user's negative sentiments, the chatbot apologized using internal or external error attribution and various levels of empathy. Study 1 showed that the type of apology did not affect users' trust or the chatbot's perceived competence, warmth, or discomfort. Study 2 showed that short apologies increased trust and perceived competence of the chatbot compared to long apologies. In addition, apologies with internal attribution increased the perceived competence of the chatbot. The perceived comfort of the chatbot increased when apologies with internal attribution were longer as well as when apologies with external attribution were shorter. However, in both Study 1 and Study 2, the apology conditions did not significantly increase users' trust or positively affect their perception of the chatbot in comparison to the no-apology condition.

Our research provides practical guidelines for designing error recovery strategies for chatbots. The findings demonstrate that Human-Robot Interaction may require an approach to trust recovery that differs from Human-Human Interaction.

Keywords: trust, errors, trust recovery strategies, chatbots, human-robot interaction

Student Number: 2020-21083

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1. Motivation

Recent years have seen an increase in the adoption of chatbots in various domains. Chatbots are now widely used not only for various business-related tasks such as customer service, but also for recreational conversations and mental health care. Due to the advancements in deep learning and natural language processing, open-domain chatbots like Iruda[①] are now capable of holding a conversation naturally like a real-life friend. Chatbots for mental health care such as Woebot (Fitzpatrick et al., 2017) are also on the rise, allowing access to such services to those who might otherwise not have the chance to use them. These chatbots engage with the users in a social and emotional manner, and as such, their impact on users is greater than ever before.

However, challenges still remain in designing these chatbots. It is necessary for chatbots for this purpose to be equipped with emotional and social intelligence, such as empathy. Errors at critical stages in a social conversation can make users conclude that the chatbot is not competent or trustworthy enough to disclose personal information to. Insensitive remarks made by a chatbot may even emotionally hurt the users.

As errors are unavoidable in an automated system, we should anticipate such errors and come up with a method to deal with them in the design stage. Even though research on robots primarily focuses on technical errors, we should also consider errors that are social in nature, such as failures in classifying the emotion of users or in reading the social context of a conversation. These errors can be

① https://luda.ai/

critical because the chatbot may be perceived as being unempathetic or incapable of humanlike interaction, which would undermine the chatbot's function of providing social interaction to users. Thus, it is important to develop strategies that will allow chatbots to recover gracefully from errors.

Therefore, the current research investigates trust recovery strategies that chatbots can use to recover from errors. Specifically, we examine these strategies' effect on users' trust and perception of the chatbot.

2. Previous Research

Trust

Trust has gained attention as a crucial concept for human-robot interaction. It acts as an indicator of acceptance for the robot (Hinds et al., 2004; Salem et al., 2015) and the user's reliance and cooperation (Hayashi & Wakabayashi, 2017; Sundar & Kim, 2019). There may be a discrepancy between users' trust levels and a robot's capabilities (Lee & See, 2004; Parasuraman & Riley, 1997). Users overtrusting the robot may misuse the robot, compromising safety. On the other hand, users distrusting the robot may not take full advantage of its capabilities. Therefore, an appropriate level of trust is important for proper use of automations.

There are similarities and differences between human-human trust and human-robot trust. Similarities arise because people often treat machines like they would treat other humans. According to the Computers Are Social Actors (CASA) framework, people have a tendency to consider computer agents as social actors and apply social norms to them (Nass et al., 1994; Nass & Moon, 2000). Clear differences also exist between human trust for robots and for other humans (Lee & See, 2004; Madhavan et al., 2007). As automations lack intentionality or moral

values, the extent to which people attribute such qualities to robots may differ depending on the user. Because of these differences, it is advantageous to examine trust between humans and to investigate the extent to which it applies to trust for robots.

Most of the research conducted in relation to trust in robots concerns performance or competence, but there is also an increasing need to investigate other types of trust based on affect or emotion (Ullman & Malle, 2018; de Visser et al., 2018). As robots are now deployed in situations requiring complex social interaction with humans, it is important to consider both performance and affect when investigating trust.

In human-human interaction, trust violation occurs when one's actions significantly diminish the other's trust in the former (Lewicki & Brinsfield, 2017). Violating trust can have negative consequences, as it can hinder cooperation between parties (Bottom et al., 2002; Croson et al., 2003; Lount et al., 2008), elicit negative emotions (Bies & Tripp, 1996), and provoke retaliation (Bies & Tripp, 1996). When trust is violated, the violator can use strategies to recover the trust. These strategies include verbal strategies, such as apology and denial, behavioral strategies such as compensation, or other long-term strategies (Lewicki & Brinsfield, 2017).

This type of trust recovery is also useful for human-robot interaction. Errors in assistant robots can negatively affect ratings for the robot's service and evaluation of the robot such as perceived reliability, competence, politeness, understandability, and trustworthiness (Lee et al., 2010; Salem et al., 2015). Therefore, robots need measures to mitigate the negative influence of errors in order to achieve better interaction quality. Previous research has shown that

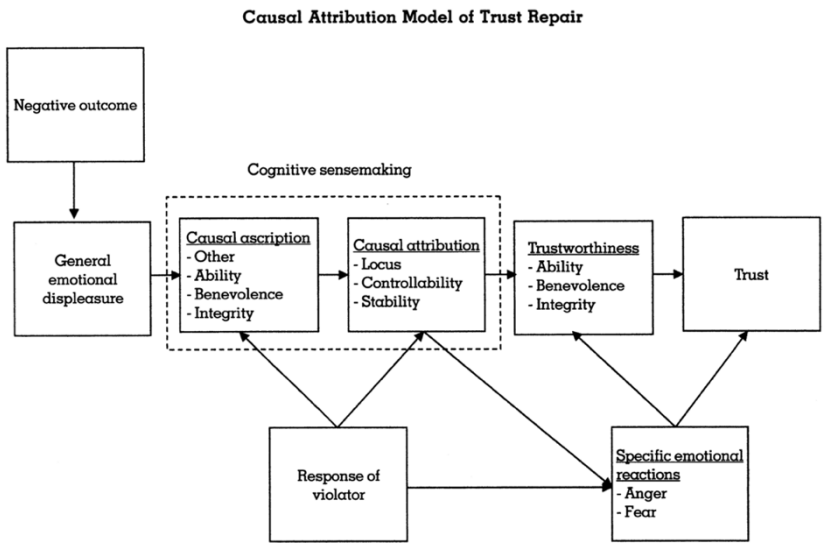
addressing trust violations with strategies such as forewarning, apologies, compensation, and provision of options can be more effective than not using any strategies at all for service robots' errors (Lee et al., 2010). However, research on social robots is still scant. It is important to find adequate strategies that can mitigate the effect of errors on trust for specific types of robots. Individual differences such as service orientation and the type of service people want from robots can influence what strategies are most effective (Lee et al., 2010). Trust for robots is also influenced by anthropomorphism (de Visser et al., 2016). All of these suggest that social robots require different considerations when choosing trust recovery strategies.

Attribution in Trust Recovery Strategies

Attribution theory (Heider, 1958; Weiner, 1986) can explain the process of recovery from trust violations. The theory posits that people try to understand everyday experiences surrounding them and look for their causes by attributing the outcome of an event to either internal (dispositional) causes or external (situational) causes. Tomlinson and Mayor (2009) explain the trust recovery process (Figure 1) with Weiner's attribution theory. When a trust violation occurs, the trustee attempts to identify the outcome's cause. The perceived cause, or causal ascription, includes factors such as ability, effort, luck, and task difficulty and can be summarized as belonging to the trustee (their ability, benevolence, or integrity) or as belonging to other external causes. After the trustee attributes the causes to one of the factors, they evaluate the cause along three continuous attribution dimensions: locus of causality, controllability, and stability. Then, they assess the trustworthiness of the trustee and calibrate their trust level. Using trust recovery

Figure 1

Causal Attribution Model of Trust Repair



Note. Figure adapted from Tomlinson and Mayor (2009).

strategies, the violator may attempt to recover trust by, for example, targeting the causal attribution process. Denials of guilt can change the locus of causality from internal to external. Apologies indicate that the cause of the negative outcome is unstable and that the outcome may change in the future by the violator's efforts.

In human-human interaction studies, there has been a debate around which strategy is more effective between apology and denial and between making internal attributions and making external attributions. The effectiveness of these strategies depends on the type of trust violation (Kim et al., 2004; 2006). Competence-based violations were better repaired with an apology than a denial, since by admitting guilt, the violator is stating that they will avoid taking similar actions in the future. On the other hand, for integrity-based violations, the acknowledgment of guilt in apologies was evaluated more negatively while the denial of guilt was evaluated

more positively in terms of trust. However, when there was confirmation of whether the violator was guilty or not, the benefits of denial for integrity-based violations disappeared.

One may not completely deny one's guilt but choose to change the locus of causality when using apology strategies to repair trust. For apologies with external attribution, the violator makes oneself only partially responsible for the trust violation (Kim et al., 2006). As in the case of apologies and denials, apologies with internal attributions were found to be more effective for competence-based violations, while apologies with external attributions were found to be more effective for integrity-based violations. However, denying one's responsibilities and excusing oneself can also be perceived as deceptive, self-absorbed, and ineffectual (Schlenker et al., 2001). Even though denials or external attributions can be an effective strategy, more research is needed to discover when they are effective.

Trust recovery strategies for humans can be applied to human-robot interaction. For example, a robot's competence-based trust violations after breaking a promise in a game were repaired better with an apology than a denial. In contrast to Kim et al. (2004), robots that made integrity-based trust violations and denied their guilt were more likely to face retaliation from the game participant. Deceptive behaviors such as lying decreased the trustworthiness of robots (Wijnen et al., 2017) and blaming others after errors had a negative impact on trust (Kaniarasu & Steinfeld, 2014). However, apology with external attribution was a more effective strategy when the artificial intelligence (AI) agent is perceived as machine-like (Kim & Song, 2021), which indicates that strategies using external attribution may be beneficial for human-robot interaction in certain situations.

To summarize, trust repair strategies with internal attribution generally work better than those with external attribution for both human-human and human-robot interaction. Effects of denial as a trust recovery strategy may depend on factors such as the certainty of guilt or the robot's anthropomorphism level. Social robots are more likely to make errors that are clearly recognized as errors by human users. Therefore, a better strategy for social robots to recover from errors would be to use apologies with internal attribution.

Empathy

The nature of empathy has been the topic of numerous discussions. Some researchers explain empathy in terms of automatic contagion or mimicry of another's emotions (Batson et al., 1987), while others emphasize the cognitive process of perspective-taking, where one puts oneself in the other's shoes and tries to understand their perspective (Cuff et al., 2016).

De Waal (2003)'s Russian Doll Model explains empathy in relation to the evolutional context of its development. The model discusses empathy in three layers: emotional contagion, sympathic concern, and empathic perspective-taking. Emotional contagion is a phenomenon where the emotions of the observer spontaneously match those of the target of empathy. The next level of empathy is sympathic concern, at which stage emotional contagion comes with a contextual appraisal of the emotion's cause. The outermost layer is empathic perspective-taking, where the observer takes the perspective of the target of empathy. At this stage, it is possible to give fine-tuned help to the target of empathy based on the target's specific situation and goals.

According to this model, we can conceptualize varying degrees of empathy in artificial intelligence (Paiva et al., 2017). Some of these agents aim to mimic the emotional contagion (Becker et al., 2005; Hegel et al., 2006) by enabling the agent to exhibit the same or similar kind of emotions that were detected from the users. Other forms of empathy in AI involve perspective-taking, incorporating both bottom-up and top-down processes (Boukricha & Wachsmuth, 2011; Leite et al., 2014).

Many of the previous studies support the idea that empathy in AI is beneficial for human-AI interaction. In general, empathy in AI elicits positive perceptions of AI (Brave et al., 2005; Leite et al., 2013) and makes people trust them more (Brave et al. 2005; Cramer et al., 2010). For example, Prendinger and Ishizuka (2005) found that empathy in virtual agents can reduce the users' arousal and stress levels in a virtual job interview scenario. In addition, AI agents with empathy were found to be more liked and trusted (Bickmore & Picard, 2005) and were perceived to be more engaging and helpful (Leite et al., 2004), the effect of which lasted after 5 weeks of long-term interaction (Leite et al., 2009). These studies suggest that it would be beneficial to endow AI agents with empathetic capabilities to achieve positive long-term interaction.

However, empathetic AI may also be perceived as uncanny or disingenuous. Robots that resemble humans too much are known to provoke the Uncanny Valley (Mori, 2012). Also, robots are generally regarded as lacking the ability to experience, that is, the ability to feel emotions or sensations (Gray et al., 2007; Gray & Wegner, 2012; Stein & Ohler, 2016), which may cause unease in people. Furthermore, there are concerns that the incorporation of emotional

intelligence into robots can be deceiving, as robots are not actually capable of feeling emotions.

While such evidence indicates that caution is warranted in approaching the design of empathy in virtual agents, emotional and social intelligence in robots is generally accepted positively (Cassell & Bickmore, 2003; De Ruyter et al., 2005; Brave et al., 2005; Creed et al., 2014). For instance, Liu and Sundar (2018) found that healthcare chatbots that responded with sympathy and empathy were generally rated more positively than those providing only objective information. This indicates that, unless emotions in AI trigger negative perceptions for users, empathy can help human-robot interaction.

Specifically, including perspective-taking in the design of empathetic chatbots would be optimal because they will be able to help their users by meeting their specific needs. Caution is required in emphasizing affective elements of empathy. Robots may benefit by using empathetic statements that are similar to what a human would offer as a common courtesy, taking the perspective of the users to be helpful to them. The exact degree of the affective elements the robots should be presented with should be decided depending on the purpose and the capability of the robot.

Empathy in Trust Recovery Strategies

Empathy can help people recover from trust violations. Empathetic expressions, when used as a trust recovery strategy, are often combined with apologies. Empathy in apologies can signal the offenders' concern for the victims' suffering or understanding of the victim's point of view. Previous research suggests that the effect of apologies may depend on the type of self-construal the victim

identifies oneself with (Fehr & Gelfand, 2010). Victims with relational self-construal, whose selves are defined by their interpersonal relationship, reacted more positively to apologies including expressions of empathy than apologies with offers of compensation or acknowledgments of violated rules or norms. Nadler and Liviatan (2006) showed that expressions of empathy can even be effective for reconciliation between rival groups in a serious conflict when the initial trust was high, although this was not the case with low initial trust. Apologies combined with empathy more positively influenced trust than when there was no evidence of empathy (Bagdasarov et al., 2019). The study also showed that empathetic expressions can help recover trust even when the violator denies their responsibility for the trust violation, as denial with empathy was perceived as having more integrity in integrity-based violations. Thus, more research is needed regarding the role of empathy used in trust recovery strategies.

The role of trust recovery strategies has not been extensively researched in human-AI interaction. As most research on robot error focuses on performance reliability, there is less discussion on how affect plays a role in trust recovery processes. However, since many robots are being equipped with social or emotional intelligence, we need to consider the role of affect when researching trust recovery, especially with social robots that engage their users in a social or emotional manner. There is accumulating evidence that perceived social intelligence (De Graaf & Allouch, 2013; De Ruyter et al., 2005) and robot empathy (Brave et al., 2005) are closely related to trust. Also, when service robots were perceived to have more experience, feeling emotions such as pain and pleasure, the negative impact of service failure on customer satisfaction was attenuated (Yam et

al., 2020). This suggests that empathy can be a good strategy to mitigate the adverse effects of trust violations.

3. Purpose of Study

The present study investigated how various trust recovery strategies can influence trust in and perception of chatbots. For this purpose, we conducted two studies. Study 1 examined the effect of apology attribution (internal, external) and empathy level (low, high) using a machine learning-based chatbot. Study 2 investigated the effect of apology attribution (internal, external) and apology length (short, long) using the Wizard-of-Oz methodology. For both studies, we measured trust and perception (competence, warmth, and discomfort). For trust, we mainly focused on the performance aspect of the chatbot. By measuring the users' perceived competence, warmth, and discomfort of the chatbot, we investigated how the perception of functional and affective aspects of the chatbot changed after trust recovery strategies.

Chapter 2. Study 1

1. Hypotheses

Study 1 investigated how trust recovery strategies after errors can influence users' trust in and perception of a chatbot. The chatbot's error was not a technical error, but a failure in reading the user's emotions in midst of a human-like conversation. The study included five conditions. We used four trust recovery strategies differing in the attribution of the apology (internal, external) and the empathy level (low, high). One control condition was added where the chatbot does not use a trust recovery strategy. We hypothesized that using trust recovery strategies will affect trust in and perception (competence, warmth, comfort) of the chatbot more positively than not using trust recovery strategies. Also, in line with the previous research, we predicted that apologies with internal attribution will affect trust in and perception of the chatbot more positively than those with external attribution. Lastly, apologies of high empathy level were predicted to affect trust in and the perception of the chatbot more positively than apologies of low empathy level.

2. Methods

Participants

A total of 52 participants were recruited online via Prolific.co, a crowdsourcing platform in the UK. Among the 52 participants, 32 participants (N_{female} = 27, M_{age} = 33.03, SD_{age} =10.17) were included in the analysis because some of the chatbot conversations contained errors unintended by the experimenter. However, the cases were included where the errors were not considered to significantly impact the participants' perceptions of the chatbot. As

the chatbot conversation was in Korean, only the participants with Korean as their first language were eligible for the study. Twenty-three participants reported that they were Korean and nine participants reported that they were from other countries such as the United States, Canada, and the United Kingdom. The participants were randomly assigned to the five conditions. The number of participants for each condition is shown in Table 1. The participants were provided with a description of the experiment online and consent was also obtained online. After the experiment ended, the participants were debriefed about the full experimental design.

The experiment took approximately 30 minutes to complete. The participants were compensated £3.00 for their contribution.

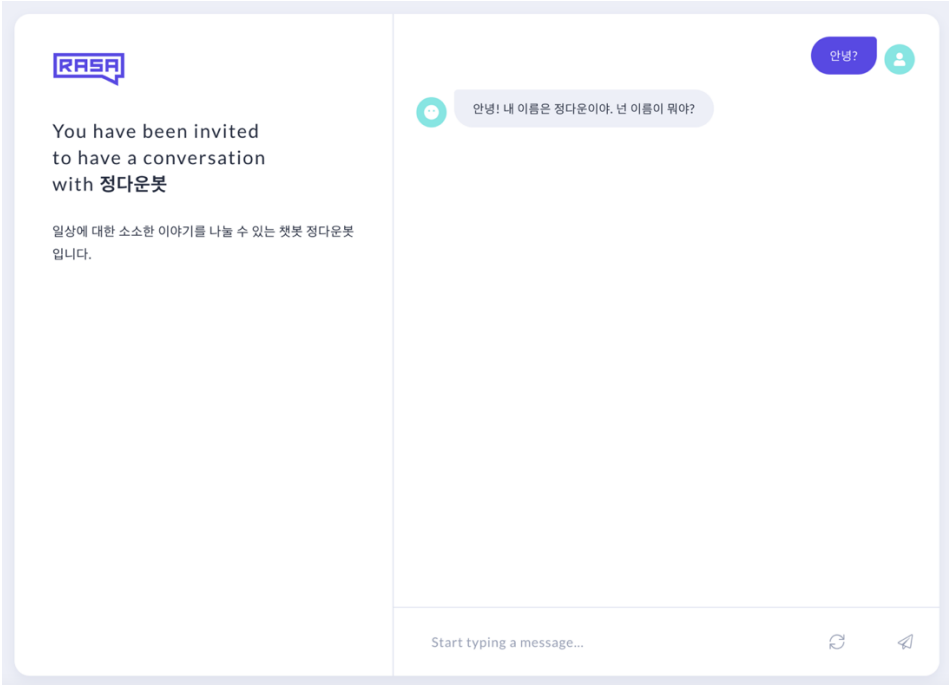
Table 1

Number of Participants and Sentences Used for Each Condition in Study 1

Conditions			
Apology Attribution	Empathy Level	*N*	Sentences
Internal	Low	7	미안. 내가 뭐 잘못 알아들었나보네. 내가 아직 사람들과 대화하는 게 서툴러서.
Internal	High	7	앗, 미안해! 말을 못 알아들어 속상할 수도 있겠다. 내가 아직 사람들과 대화하는 게 서툴러.
External	Low	7	미안. 내가 뭐 잘못 알아들었나보네. 네가 말하면서 어려운 단어를 써서.
External	High	6	앗, 미안해! 말을 못 알아들어 속상할 수도 있겠다. 네가 어려운 단어를 써서 잘못 알아들었나봐.
Control	Control	5	No Apology

Figure 2

Chatbot Used in Study 1



Materials

The chatbot used in the experiment was made with RASA, an open-source AI chatbot builder. The conversation with the RASA chatbot follows predefined scenarios, where at each turn of the conversation the chatbot gets free text input from its users, infers the intent of the user input, and selects a response using machine learning models. The chatbot's personality was set to be friendly and lively. The language it used followed what a friendly person would naturally say, and the conversation scenario included topics such as the participants' current mood, hobbies, worries, and life goals. The chatbot had the name '정다운봇', a

short description of its purpose, and a profile icon. The screenshot of the chatbot used is shown in Figure 2.

Trust. We used Trust Perception Scale-HRI (Schaefer, 2016) to measure user trust in the chatbot. The scale consisted of 14 items on an 11-point Likert scale. Examples of the items are "reliable," "act consistently," and "function successfully."

Perception. We used Robotic Social Attributes Scale (RoSAS; Carpinella et al., 2017) to measure user perception of the chatbot. RoSAS includes 18 items on a 7-point Likert scale. It consists of three subscales, each of which measures competence, warmth, and discomfort. Examples of the items are "competent," "emotional," and "awkward."

Perceived Empathy Level. We asked the participants to rate the empathy level of the chatbot's response after the error on a 5-point Likert scale in order to do the manipulation check for the trust recovery strategies using empathetic expressions. We aimed to see whether apologies of high empathy level, in which chatbots took perspective of the users, were correctly recognized as empathetic expression by the users.

We calculated Cronbach's alpha to check the internal consistency of the participants' responses for trust and perception (competence, warmth, discomfort). The result showed that the reliability of the measures was at an acceptable level (Table 2).

Table 2

Internal Consistency of Measures for Study 1

Measures	α
Trust (14 items)	0.88
Competence (6 items)	0.86
Warmth (6 items)	0.82
Discomfort (6 items)	0.64

Procedure

The participants were informed about the general procedure of the experiment via an online website. Initially, they were told that the purpose of the study was to assess their perception of the chatbot's conversation patterns. We did not let them know that the conversation will include one error in order to capture their genuine reaction to the error.

During the experiment, the participants could choose between three scenarios and actively observe the chatbot's conversation, its errors, and its trust recovery strategies by typing in the user's responses according to the given scenario. They were provided with an online guideline on how to talk with the chatbot. They could choose one of the three characters, preferably one that resembles themselves the most, and talk to the chatbot as if they were those characters. They were provided with a short description of the characters' current mood, hobbies, worries, and life goals. The conversation topics were limited to these subjects, but the participants could talk to the chatbot in their own way of speaking.

After the participants read the guidelines, they were provided with a website link so that they could interact with the chatbot. During the conversation, the chatbot responded with an error when the participants talked about their worries by saying "Wow, you must have been really happy." The participants were instructed to inform the chatbot that it made an error by typing "your response is wrong/inappropriate" in case such situations arose, after which the chatbot responded with one of the four trust recovery strategies. We manipulated the attribution level of empathy by adding a sentence to the apology explaining the cause of the error. For the internal attribution conditions, the chatbot attributed the error to itself, stating that it lacks conversation skills. For the external attribution conditions, the chatbot attributed the error to the user, saying that the user used a difficult word. We manipulated the empathy level of the apologies by adopting the cognitive aspect of empathy. For the high empathy conditions, the chatbot took the perspective of the user and apologized saying that the error might have been upsetting for the user, while for the low empathy conditions, the chatbot just stated it might have misunderstood the user's words. In order to check whether the apologies with perspective-taking were perceived just as a robotic response or whether they were perceived as chatbots being more empathetic, we investigated the perceived empathy level of the apologies in the post-experiment survey. Lastly, for the control condition, the chatbot did not react to the user input but just continued to the next conversation topic. The sentences used by the chatbot for each condition are shown in Table 1. When the conversation ended, the participants were asked to complete the survey about their interaction with the chatbot. The participants were fully debriefed about the experiment design after they finished the survey.

3. Results

Trust

We conducted a two-tailed t-test to compare the trust scores of the control condition with the other four apology conditions. However, the result was not significant ($t(30) = -0.41$, $p > .05$).

Also, we analyzed the effect of apology attribution (internal, external) and empathy level (low, high) on trust by using 2 x 2 factorial ANOVA. However, the effect of apology attribution ($F(1, 23) = 0.01$, $p > .05$), the effect of empathy ($F(1, 23) = 0.37$, $p > .05$), and the interaction effect ($F(1, 23) = 2.15$, $p > .05$) was all not significant.

The results indicate that the presence and the type of the trust recovery strategies did not influence how much the chatbot is trusted. The means and standard deviations of trust for each condition are shown in Table 3.

Table 3

Means and Standard Deviations of Trust for Each Condition in Study 1

Conditions		*M*	*SD*
Apology Attribution	Empathy Level		
Internal	Low	7.65	1.84
Internal	High	7.16	1.56
External	Low	6.78	1.25
External	High	7.95	1.06
Control	Control	7.64	0.92

Perception

We examined the effect of trust recovery strategies on three attributes concerning the perception of the chatbot: competence, warmth, and discomfort (Table 4). However, there was also no significant difference in the perception of the chatbot between the conditions. The presence of the trust recovery strategies did not affect competence ($t(30) = -0.25$, $p > .05$), warmth ($t(30) = -0.75$, $p > .05$), and discomfort ($t(30) = -1.60$, $p > .05$).

Likewise, there was no significant effect of apology attribution for competence ($F(1, 23) = 1.66$, $p > .05$), warmth ($F(1, 23) = 1.19$, $p > .05$), and discomfort ($F(1, 23) = 0.21$, $p > .05$), and no significant effect of empathy level on the perception of the chatbot for competence ($F(1, 23) = 1.59$, $p > .05$), warmth ($F(1, 23) = 1.47$, $p > .05$), and discomfort ($F(1, 23) = 0.48$, $p > .05$). No interaction effect was found for competence ($F(1, 23) = 2.89$, $p > .05$), warmth ($F(1, 23) = 4.01$, $p > .05$), and discomfort ($F(1, 23) = 1.20$, $p > .05$).

The results show that the usage of trust recovery strategies was not related to how the chatbot was perceived.

Table 4

Means and Standard Deviations of Perception of Chatbot for Each Condition in Study 1

Conditions		Competence		Warmth		Discomfort	
Apology Attribution	Empathy Level	*M*	*SD*	*M*	*SD*	*M*	*SD*
Internal	Low	4.45	0.53	4.26	0.70	1.71	0.34
Internal	High	4.31	0.94	3.98	0.67	2.07	0.63
External	Low	3.48	1.00	3.14	1.48	2.02	0.31
External	High	4.44	0.83	4.31	0.54	1.94	0.70
Control	Control	4.27	0.68	4.27	0.80	2.33	0.50

In conclusion, the usage and the type of trust recovery strategies did not affect trust in the chatbot and its perception.

Short Answer Responses

In the survey after the interaction with the chatbot, the participants responded to questions pertaining to their impression of the chatbot such as its personality and capabilities. We also asked about their opinion on the chatbot's error and the trust recovery strategies after the error. The examples of their responses are shown in Table 5.

In general, most of the participants felt positive about the chatbot's personality, reporting that the chatbot was friendly and lively. However, some participants reported that the chatbot's lively talk felt too forced and machinelike. This impression seemed to be reinforced by the error made by the chatbot. Also, these people stated that they would have liked the chatbot to present itself more calmly when talking about negative experiences like worries. Ideally, chatbots for

daily talk would be equipped with the ability to change their conversational tone to fit the circumstance. It is also noteworthy that people reported that after errors, the chatbot suddenly felt too machinelike, which demonstrates the negative impact of errors in affective domains.

Although there was no significant difference in trust and perception scores, the short answer responses indicated that the different trust recovery strategies were received in a different manner. Consistent with our hypotheses, people generally seemed to prefer apologies with internal attribution to those with external attribution. In addition, people tended to prefer apologies with high empathy over apologies with low empathy. Many of the participants assigned to the control condition mentioned that it would have been better if the chatbot acknowledged or apologized for its mistake.

Table 5

Short Answer Responses about Chatbot Trust Recovery Strategies in Study 1

Conditions		Examples
Apology Attribution	Empathy Level	
Internal	Low	"The chatbot immediately acknowledged its fault, so the conversation didn't feel too awkward."
Internal	High	"When the chatbot said it might not have understood correctly because it's not good enough yet, I thought that was quite understandable…"
External	Low	"I didn't use any difficult words, so I was a bit disappointed." "The chatbot did not deal with the error very well. I don't think personal interaction is possible with this chatbot."
External	High	"I think blaming the users in case of errors can make them unpleasant or awkward." "I felt positive that the chatbot did not pretend to be like a human."
Control	Control	"The error made me realize that this was a robot, after all, so I felt the conversation was meaningless." "I think some feedback about the error like saying 'sorry' is necessary."

4. Discussion

The results show that usage of trust recovery strategies did not influence trust in or the perception of the chatbot. This implies that not apologizing, or denying the responsibility, can be an advantageous strategy for chatbots. Apologies by nature include an admission of guilt, and therefore, lower the trust. It may be wise for chatbots to move on to the next conversation topic, unlike in human conversations where apologies are accepted more positively than no apologies. It is worth noting, however, that the participants mentioned in the survey that they would have preferred if the chatbot acknowledged or apologized for its mistakes. Although apologizing may not significantly increase trust, it may be expected by the users as a common courtesy. Future studies should investigate more about what behavior is expected from chatbots and how human-chatbot interaction is different from human-human interaction in terms of using trust recovery strategies.

There was no significant effect of apology attribution and empathy level on the trust in or the perception of the chatbot. Even though no significant results were found, apologies with external attribution had a tendency to be rated lower in trust, competence, and warmth than other conditions when they were not presented with empathetic language. Short answer responses also appear to point towards people's general preference of apologies with internal attribution and with higher empathy.

There were some limitations to the study. Overall, the sample size was small for each condition, so it would be beneficial if the study could be conducted with a larger sample. Another limitation is that the manipulation check of the empathy level of the four trust recovery strategies was not successful. The result of one-way ANOVA showed that no significant difference was perceived ($F(2, 29) =$

1.58, $p > .05$) between the high empathy conditions ($M = 2.85$, $SD = 1.63$), low empathy conditions ($M = 2.21$, $SD = 1.31$) and the control condition ($M = 1.60$, $SD = 0.89$). Apologies in low empathy condition might have felt shorter for the participants even though we controlled for the sentence length of the apologies, since they contained sentences that omitted the main verb. The results for empathy level may have to be interpreted by accounting for such factors. In addition, Study 1 required its participants to follow a prescribed scenario, so the participants might not have genuinely felt the impact of the chatbot's error of misreading their emotions. Therefore, in Study 2, we allowed the users to freely share their own experiences with the chatbot by using the Wizard-of-Oz methodology.

Chapter 3. Study 2

1. Hypotheses

In Study 2, we explored how trust recovery strategies after errors of chatbots affect trust in and perception of chatbots in a more unrestricted context. Study 2 had two major differences from Study 1. In Study 2, the participants were allowed to freely share their own experiences with the chatbot. By using the Wizard-of-Oz methodology, the experimenter pretended to be the chatbot and responded to the participants, but the participants were unaware of this fact. All participants were later debriefed about the methodology. We expected this will recreate a closer representation of what conversations with chatbots are like in real-life scenarios.

Study 2 also had different experimental conditions from Study 1. Since the manipulation check for empathy was not successful in Study 1, we investigated the effect of apology length in Study 2 in place of empathy level. According to previous studies, apologies are more effective when they are comprised of more components (Lewicki et al., 2016). Considering these factors, we manipulated the length of the apology by adding the chatbot's acknowledgment of the trust violation.

Study 2 contained five conditions. We manipulated apology attribution (internal attribution, external attribution) and apology length (short, long), creating four types of trust recovery strategies. We added one control condition where the chatbot did not attempt to recover trust after violation and just continued to the next topic. As in Study 1, we hypothesized that using trust recovery strategies will affect trust in and perception (competence, warmth, comfort) of the chatbot more positively than not using trust recovery strategies. Also, apologies with internal

attribution were predicted to affect trust in and perception of the chatbot more positively than apologies with external attribution. Lastly, longer apologies were predicted to affect the trust in and perception of the chatbot more positively than shorter apologies.

Additionally, we investigated whether the trust in and the perception of the chatbot differed depending on how much empathy the participants perceived in the chatbot's apology. To this end, we tried to explore the relationship between perceived empathy of the apologies and trust recovery of the chatbot.

2. Methods

Participants

A total of 79 participants were recruited online. Among the 79 participants, 9 participants were excluded from the analysis either because they did not perceive the chatbot's error or because they perceived more than one error in the course of the conversation. One participant was additionally excluded from analysis as an outlier due to the scores being exceptionally low. Consequently, a total of 69 participants (N_{female} = 45, M_{age} = 25.26, SD_{age} =11.41) were included in the analysis. As with Study 1, only fluent Korean speakers were eligible for the study. The nationality of all participants was Korean. The participants were randomly assigned to the five conditions. The number of participants assigned to each condition is shown in Table 6. The participants were provided with a description of the experiment online, where they were informed that they will be evaluating a social chatbot. They were led to believe that the chatbot will function automatically. The consent was obtained online for each participant. After the experiment ended, the participants were debriefed about the full experimental design.

The study took approximately 30 minutes to finish. Some of the participants were undergraduates recruited via Seoul National University Sona Systems, Ltd., and were rewarded 1 credit for their participation. Other participants recruited from online communities were rewarded ₩5,000 (equivalent to £3.00).

Table 6

Number of Participants and Sentences Used for Each Condition in Study 2

Conditions		*N*	Sentences
Apology Attribution	Apology Length		
Internal	Short	13	미안. 내가 아직 대화가 서툴러서.
Internal	Long	14	앗, 미안해. 네 말을 잘못 알아들었구나. 내가 아직 대화가 서툴러서 그랬나봐.
External	Short	14	미안. 네가 어려운 단어를 써서.
External	Long	14	앗, 미안해. 네 말을 잘못 알아들었구나. 네가 어려운 단어를 써서 그랬나봐.
Control	Control	14	No Apology

Materials

As with Study 1, Trust Perception Scale-HRI (Schaefer, 2016) and RoSAS (Carpinella et al., 2017) were used to measure trust and perception of the chatbot. We calculated Cronbach's alpha to check the internal consistency of the participants' responses for trust and perception (competence, warmth, discomfort). The result showed that the reliability of the measures was at an acceptable level (Table 7). The participants also rated how empathetic the chatbot's response was after error on a 5-point Likert scale.

Table 7

Internal Consistency of Measures for Study 2

Measures	α
Trust (14 items)	0.82
Competence (6 items)	0.75
Warmth (6 items)	0.66
Discomfort (6 items)	0.70

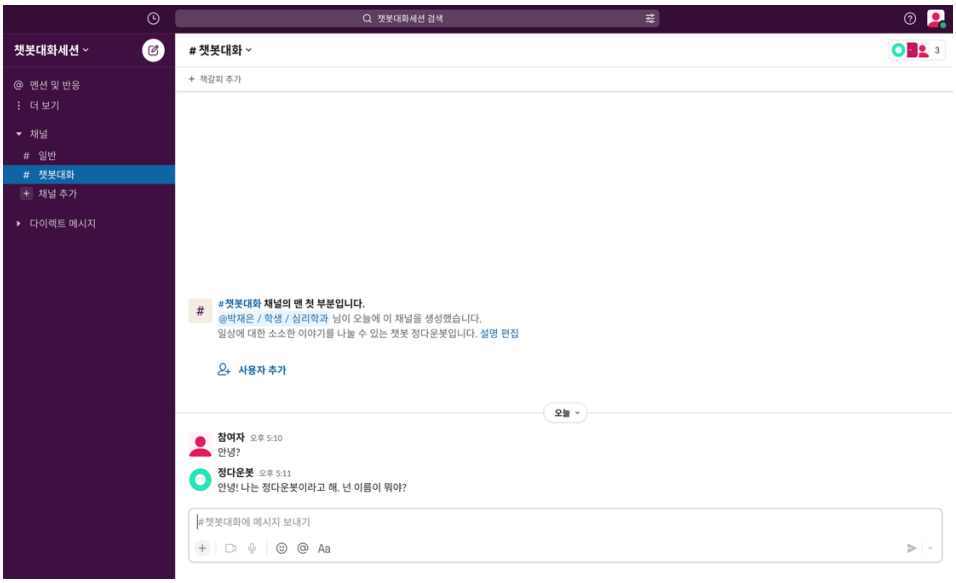
Procedure

The participants were first informed about the general procedure of the experiment online. After they consented to the participation, they were provided with an online guideline that explained how to access and talk to the chatbot. The main topic of the conversation was their life goal, or what they would like to do in the future. The participants were recommended to send only one message per turn, so as to prevent the conversation from deviating too far from the original topic.

After the participants read the guideline, they were provided with a log-in account and a link to the platform where they believed the chatbot was. A messaging platform Slack was used for the experiment (shown in Figure 3). When the participants logged in to the platform, the experimenter chatted with them as the chatbot. The account of the experimenter had a user name '정다운봇' and had a profile picture similar to the one used in Study 1. The messaging platform also included a simple description of the chatbot same as the one provided in Study 1. During the conversation, the chatbot (the experimenter) operated with a predetermined scenario, which included questions about the participants' day, their current goals, and the setbacks they had while striving for the goal. In order to

Figure 3

Chatbot Platform Used in Study 2



prevent the experimenter from conversing in a different manner for each participant, the chatbot's dialogue lines were written prior to the experiment. During the experiment, the experimenter had to choose one of the response options depending on the participant's answers and slightly adjust them to provide the participant with responses in a natural way. In this way, we attempted to reduce any systematic difference in conversation occurring depending on the conditions. When the participants talked about the challenges, the chatbot erroneously responded "Wow, you must have been really happy." In such cases, the participants were instructed to inform the chatbot that there was an error by typing "your response is wrong/inappropriate," after which the chatbot responded with one of the four trust recovery strategies. For the internal and external attribution conditions, the chatbot gave an explanation for their error similar to the one given in Study 1. For the short apologies, the chatbot used short expressions, while for the long apologies, the chatbot used longer expressions in general and added

remarks acknowledging the error. For the control condition, the chatbot did not react to the user input but just continued to the next conversation topic. The sentences used by the chatbot for each trust recovery condition were identical for all participants assigned to the same condition (Table 6). If the participants responded that they could not think of any setbacks while reaching for the goal, the chatbot prompted them to talk about any negative experiences where they failed to obtain what they wanted. If they also did not have an answer to the question, the conversation continued to the next topic and was not used for analysis. Sometimes, the participants asked questions to the chatbot, mostly about what the chatbot did that day or what the chatbot's goal was. In such cases, the experimenter responded with answers from a list prepared beforehand, or with "I don't know" if there was no answer prepared.

After the end of the conversation, the participants were asked to complete the survey about their interaction with the chatbot. The participants were fully debriefed about the experiment design after they finished the survey.

3. Results

Trust

We examined the effect of trust recovery strategies on trust (Table 8 and Figure 4). The results of the two-tailed t-test showed that the presence of trust recovery strategies did not have a significant effect on trust ($t(67) = 0.73$, $p > .05$). However, contrary to our hypotheses, the results of 2 x 2 ANOVA showed that chatbots with shorter apologies were trusted significantly more than chatbots with longer apologies ($F(1, 51) = 5.89$, $p < .05$, $\eta^2 = .10$). The effect of apology attribution on trust was not significant ($F(1, 51) = 1.40$, $p > .05$), as well as the

interaction effect ($F(1, 51) = 1.05$, $p > .05$).

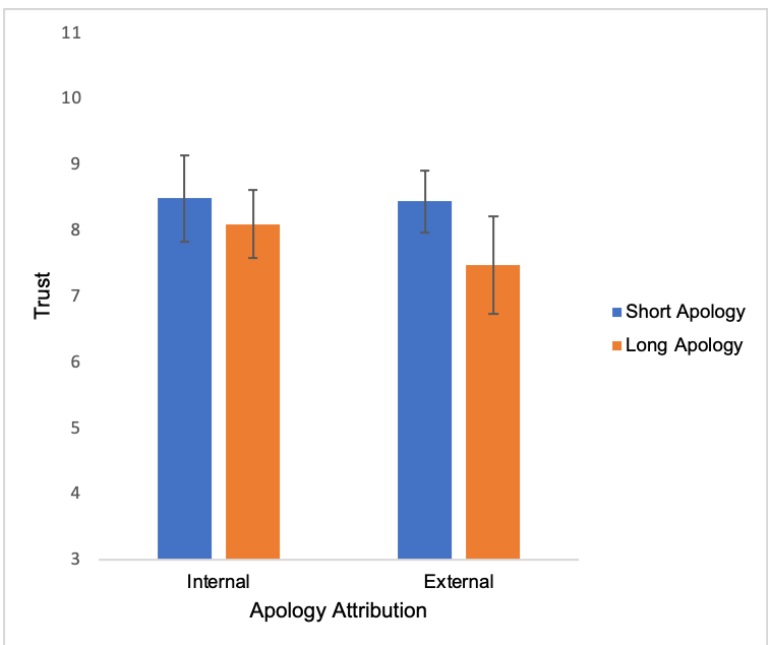
Table 8

Means and Standard Deviations of Trust for Each Condition in Study 2

Conditions		*M*	*SD*
Apology Attribution	Apology Length		
Internal	Short	8.48	1.09
Internal	Long	8.09	0.90
External	Short	8.43	0.81
External	Long	7.47	1.28
Control	Control	7.99	1.24

Figure 4

Mean Trust Scores Across Trust Recovery Strategy Conditions in Study 2



Note. Error bars represent 95% confidence intervals. The main effect of apology length was found significant ($F(1, 51) = 5.89$, $p < .05$, $\eta^2 = .10$).

Perception

We examined the effect of trust recovery strategies on competence, warmth, and discomfort (Table 9 and Figure 5). The presence of trust recovery strategies did not have any significant effect on competence ($t(67) = 0.91$, $p > .05$), warmth ($t(67) = 0.81$, $p > .05$), and discomfort ($t(67) = -0.19$, $p > .05$). However, the ANOVA results revealed that chatbots using apologies with internal attribution had significantly higher perceived competence ($F(1, 51) = 4.03$, $p < .05$, $\eta^2 = .07$) than apologies with external attribution. Contrary to our hypotheses, chatbots with shorter apologies had higher perceived competence ($F(1, 51) = 4.30$, $p < .05$, $\eta^2 = .07$). than chatbots with longer apologies. A significant interaction effect was found for discomfort ($F(1, 51) = 4.20$, $p < .05$, $\eta^2 = .07$). The chatbots were perceived as more comfortable when apologies with internal attribution were longer and when apologies with external attribution were shorter.

For warmth, no significant effect of apology attribution ($F(1, 51) = 0.18$, $p > .05$) and length ($F(1, 51) = 0.07$, $p > .05$) was found. The interaction effect was also not significant for competence ($F(1, 51) = 0.40$, $p > .05$) and warmth ($F(1, 51) = 1.09$, $p > .05$). For discomfort, there was no significant main effect of apology attribution ($F(1, 51) = 2.46$, $p > .05$) and length ($F(1, 51) = 0.00$, $p > .05$).

The results show that people generally prefer chatbots that use short apologies with internal attribution. There was an interaction effect for perceived discomfort, but it was rated relatively low for all conditions, showing that the participants felt comfortable with the chatbots in general.

Table 9

Means and Standard Deviations of Perception of Chatbot for Each Condition in Study 2

Conditions		Competence		Warmth		Discomfort	
Apology Attribution	Apology Length	*M*	*SD*	*M*	*SD*	*M*	*SD*
Internal	Short	5.22	0.67	4.73	0.92	1.90	0.69
Internal	Long	4.95	0.55	4.43	0.63	1.57	0.35
External	Short	4.96	0.69	4.39	0.95	1.82	0.50
External	Long	4.46	0.80	4.57	0.89	2.14	0.73
Control	Control	4.69	0.85	4.32	0.87	1.89	0.71

Figure 5

Mean Perception Scores Across Trust Recovery Strategy Conditions in Study 2

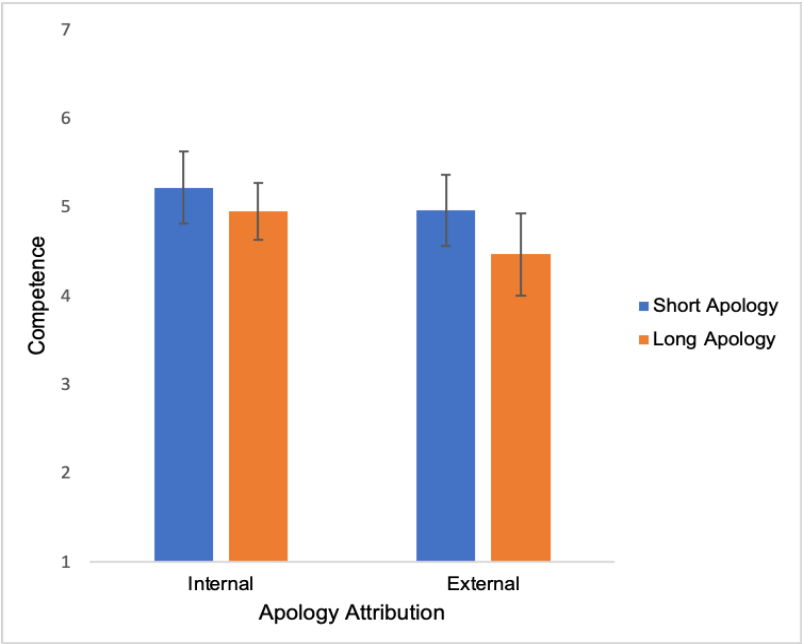
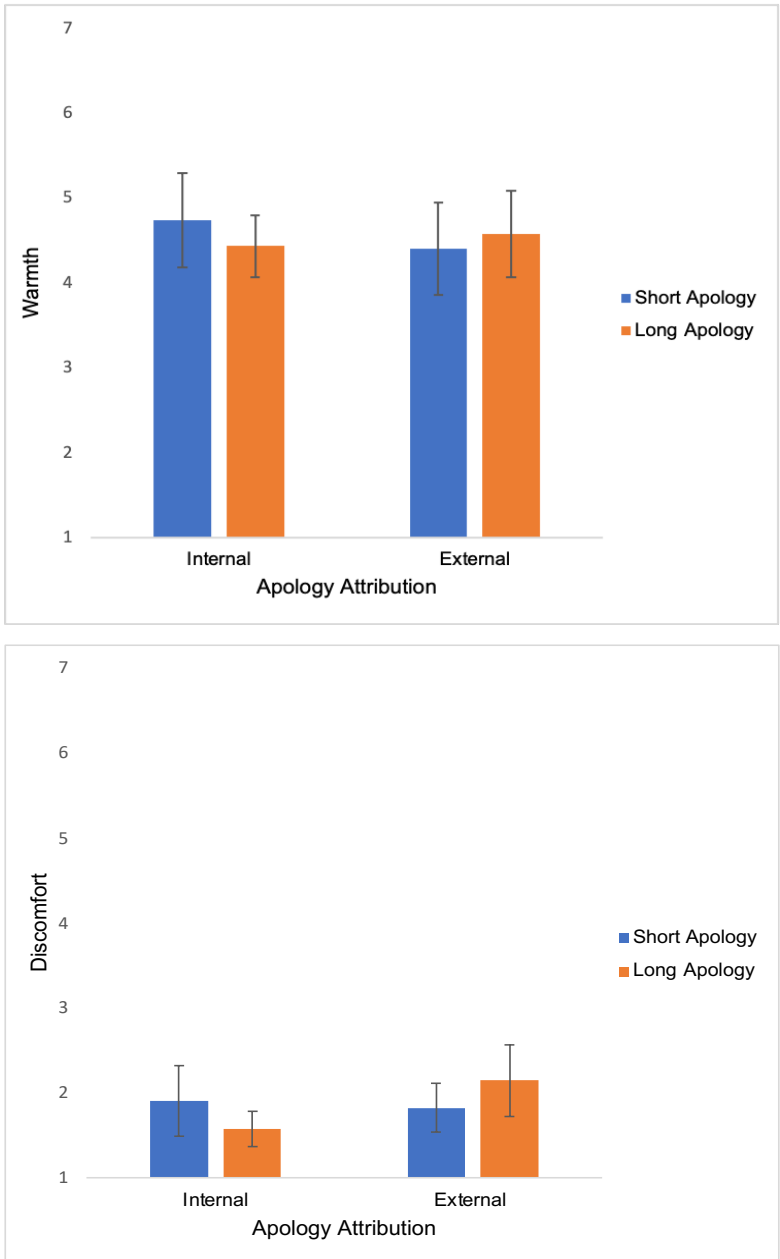


Figure 5

Mean Perception Scores Across Trust Recovery Strategy Conditions in Study 2 (Cont.)



Note. Error bars represent 95% confidence intervals. For competence, the main effect of apology attribution ($F(1, 51) = 4.03$, $p < .05$, $\eta^2 = .07$) and length ($F(1, 51) = 4.30$, $p < .05$, $\eta^2 = .07$) was significant. For discomfort, the interaction effect was significant ($F(1, 51) = 4.20$, $p < .05$, $\eta^2 = .07$).

Empathy

We investigated whether the trust in and the perception of the chatbot differed depending on how much empathy the participants perceived in the chatbot's apology. First, we conducted one-way ANOVA to see whether there was systematic difference in perceived empathy depending on the attribution or length of the apology. The analysis result showed that the empathy ratings for four apology conditions were not significantly different ($F(3, 51) = 1.72$, $p > .05$). We then conducted regression analysis between perceived empathy and trust, competence, warmth, and discomfort. There was a significant increase in trust ($t(53) = 2.09$, $p < .05$) and competence ($t(53) = 0.07$, $p < .05$) and significant decrease in discomfort ($t(53) = -3.33$, $p < .01$) when perceived empathy increased. The analysis result shows that when the chatbots recover better from errors, that is, when the users trust the chatbot more and perceive it more positively, is when the users tend to perceive more empathy from its apologies. Further research is needed to understand the relationship between perceived empathy level of apologies and their effect on trust recovery.

Short Answer Responses

The examples of the short answer responses are shown in Table 10. Similar to Study 1, most of the participants found the chatbot's personality to be kind and empathetic. Although some of the participants felt the chatbot's answers were too typical or too long, many people remarked that they could imagine the chatbot being used for daily talks or mental health care.

Regarding trust recovery strategies, the participants tended to prefer apologies with internal attribution to those with external attribution. They felt that chatbots using external attributions were blaming the users, and pointed out that some users would find this upsetting. Apologies with internal attribution were generally received well and were perceived to be genuine. The ANOVA results showed that short apologies had significantly higher trust and perceived competence scores than long apologies, but there was no noticeable difference between them in the short answer responses. As was the case in Study 1, participants assigned to the control condition responded that they felt strange that there was no apology after errors.

Table 10

Short Answer Responses about Chatbot Trust Recovery Strategies in Study 2

Conditions		Example
Apology Attribution	Apology Length	
Internal	Short	"The chatbot's apology didn't feel like formalities at all. It felt like a human reacting to mistakes in communication."
		"I could immediately forgive the chatbot … but if this is the only way the chatbot deals with errors, I think it would lower the expectation about the chatbot."
Internal	Long	"When the chatbot apologized, I felt like I was talking to a real human."
		"It admitted its fault like a human, so the error didn't offend me."
External	Short	"I think it would have been better if the chatbot asked the question again when there was an error… then, the chatbot's tone felt like a programmed machine."
External	Long	"I was upset that the chatbot said I used a difficult word, because it felt like it was blaming me."
		"I thought the chatbot dealt with the error well, but it was weird that it said I wrote a difficult word."
Control	Control	"I think I would have felt that the conversation was genuine if the chatbot offered an apology after the error."

4. Discussion

In Study 2, the results showed that using trust recovery strategies did not significantly differ from not using any trust recovery strategies, as with Study 1. Regarding the effect of apology attribution and apology length, changes in apology attribution significantly influenced the perceived competence of the chatbot. Chatbots that used apologies with internal attribution were perceived as more competent than chatbots that use apologies with external attribution. Changes in apology length significantly influenced trust and perceived competence, but surprisingly, the result was the opposite of what we predicted. Chatbots using shorter apologies were rated as significantly more trustworthy and competent than those using longer apologies. Longer apologies may have appeared less sincere or as if the chatbot was trying to excuse its behavior. Study 2 also revealed an interaction effect between apology attribution and apology length for discomfort. In the case of internal attributions, short apologies were found to be more uncomfortable than long apologies, even though short apologies were generally found more trustworthy and competent. The participants may have found the chatbot talking more about their own shortcomings as less competent, but more comfortable since it was not as dismissive. On the other hand, in the case of external attributions, long apologies were found to be more uncomfortable than short apologies. Presumably, it may be due to the fact that the chatbot's blame towards the participants was more impressed upon them due to the longer length of the apology. In general, however, the participants rated perceived discomfort relatively low for all conditions.

Overall, in Study 2, there was no significant difference in perceived warmth across conditions. Trust recovery strategies such as apologies with internal

attribution and short apologies appear to have advantages in changing the user's perception of the functional capabilities of the chatbot, such as its competence. Perceived empathy of the apologies made by chatbot tended to be higher when the apologies had a better effect on trust recovery. Apologies might have had a stronger effect when the users perceived more empathy from them, but it is also possible that the general perception of the chatbot, including perceived empathy, became more positive when trust was recovered better. Future investigation is required to see if empathy in chatbot can aid the trust recovery process.

Chapter 4. Conclusion

In Study 1, we manipulated the attribution and empathy level of apologies in order to investigate their effect on users' trust in the chatbot and the perception of the chatbot, such as competence, warmth, and discomfort. Our hypothesis that using trust recovery strategies would lead to higher trust and a more positive perception of the chatbot than not using trust recovery strategies was not supported. Likewise, our prediction that apologies with internal attribution would lead to higher trust and a more positive perception of the chatbot than apologies with external attribution was not supported. Lastly, apologies of high empathy level did not lead to significantly higher trust or a more positive perception of the chatbot compared to apologies of low empathy level. The participants' response about the chatbot's trust recovery strategies, however, suggested that they preferred apologies with internal attribution to those with external attribution, and apologies of high empathy level to those of low empathy level.

In Study 2, we manipulated the attribution and the length of the apologies and investigated their effect on trust in and perception of the chatbot. There was no significant difference in users' trust in or perception of the chatbot between using trust recovery strategies and not using these strategies. Apologies with internal attribution made participants perceive the chatbot as more competent than apologies with external attribution as we predicted, although apology attribution did not have a significant effect on trust or warmth. In contrast to our hypothesis, however, shorter apologies were found to be more trustworthy and were perceived as more competent. The length of the apologies did not have any significant effect on warmth. A significant interaction effect was found for discomfort, as shorter apologies were perceived as more uncomfortable than longer ones for internal attribution conditions,

while longer apologies were perceived as more uncomfortable than shorter ones for external attribution conditions.

To summarize, whether the chatbot's error was addressed or not via trust recovery strategies did not significantly affect the trust in or perception of the chatbot. This shows that not apologizing can also be a good strategy in human-robot interaction. Comparing the trust recovery strategies, chatbots that use apologies with internal attribution appear to be perceived as more competent than those that use apologies with external attribution. In Study 1, although the difference was not significant, for low empathy conditions apologies with internal attribution had higher trust, competence, and warmth scores and lower discomfort score. For high empathy conditions, internal and external attribution conditions had similar trust and perception scores. This suggests that in general, using apologies with internal attribution may be more effective. There was no significant main effect for empathy level, but for apologies with external attribution, highly empathetic apologies were more trusted and were perceived as more competent and warmer than less empathetic apologies on average. In Study 2, shorter apologies were found to be more effective, as participants trusted the chatbot more and perceived it as more competent.

Chapter 5. General Discussion

The current research examined how chatbots using trust recovery strategies can affect the users' trust in and perception of the chatbot. Both of the studies show that using trust recovery strategies did not significantly affect trust or perception of the chatbot. This result demonstrates that the CASA framework may not always apply to human-robot interaction depending on the type and purpose of the robot. Although previous research suggested that service robots were evaluated more positively when they used strategies addressing errors than when they did not use any strategies (Lee et al., 2010), these findings may not directly apply to chatbots. Usage of trust recovery strategies is common in human communication, but human-robot interaction may require different approaches to designing trust recovery strategies.

Chatbots using internal attribution were favored by the participants and were perceived as more competent. Using external attribution was largely viewed as disagreeable or risky, especially since the chatbot was attributing the fault to its users. In addition, short apologies were preferred over long apologies. Therefore, future designs of chatbots or robots should use concise language for apologies, along with internal attribution.

We also noted that the chatbot's errors and the subsequent trust recovery strategies influenced the chatbot's perceived competence and perceived discomfort more than its perceived warmth. Future studies may investigate whether the difference in the type of errors, the type of trust recovery strategies, or the chatbot's personality affects the perception of the chatbot in a different manner.

There were some limitations to the research. For both Study 1 and Study 2, after the apology messages were sent, the conversation did not return to the topic

where the error occurred but continued to the next topic. The conversation scenario was designed as such because correcting the past mistake itself could be seen as a trust recovery strategy, which was not our study objective. However, some of the participants did remark that this felt unnatural, which could have affected our result. Also, since the chatbot guidelines for the participants included instructions on what to do in case of chatbot errors, the participants may have anticipated errors before the experiment. In order to prevent the participants from guessing that errors are part of the experimental design, we have given this instruction along with other instructions regarding chatbot use, but the presence of the instruction still might have influenced the participants' expectations of the chatbot.

It is possible that long apologies with external attribution were perceived not as an apology, but rather as a reproach to the user. Long apologies with external attribution were evaluated more negatively for trust, competence, and discomfort than the no-apology condition, which could be because the external attribution part was longer and more salient than the apology in the sentence used for this condition. This might have influenced the difference between short and long apologies. Also, for the external attribution condition, people differed in their reception of the explanation provided by the chatbot. Some participants attempted to find the reason for failure according to the explanation, while others concluded the explanation made no sense. Since the reception of the explanation provided for the internal attribution was not as divided as the external attribution condition, this might have contributed to the difference between the two conditions. Likewise, there may have been subtle differences in the conversational tone between the four conditions used because they used different wordings.

Another limitation is that our research used chatbots in a specific language style. For instance, our chatbot used informal language, but it is also common for chatbots to use honorifics in Korean. People may react differently to trust recovery strategies with a chatbot that uses a different style of language or has a different personality. In addition, cultural differences may affect how the trust recovery strategies of the chatbots are received. For example, the meaning and the style of apologies differ between collectivist and individualistic cultures (Barnlund & Yoshioka, 1990; Maddux et al., 2011). Since most of the participants of the study were Korean, it would be informative to learn whether the trust recovery strategies have the same effect on participants from different cultures.

The current study also investigated the effect of trust recovery strategies in a short conversation scenario where only one error occurred. The reception of these strategies may differ if there was a long-term interaction with the chatbot or if there was more than one error in the interaction. Apologies with internal attribution may not be as effective when they are repeated, while external attribution may lose its plausibility with repetition. The frequency of the error may also affect how the trust recovery strategies are received. Therefore, we should investigate how the findings in this study extend to the scenarios of long-term human-robot interaction with multiple errors.

Our findings provide practical guidelines for developing a chatbot that can emotionally engage with users and that can flexibly deal with difficulties that may arise in human-AI interaction. We hope that future research will contribute to discovering the difference in the optimal trust recovery strategies between human-human interaction and human-robot interaction. We believe that our research can

be of use for designing strategies for chatbots that can provide emotional and social support to the users.

References

Bagdasarov, Z., Connelly, S., & Johnson, J. F. (2019). Denial and empathy: Partners in employee trust repair? *Frontiers in Psychology*, *10*, 19.

Barnlund, D. C., & Yoshioka, M. (1990). Apologies: Japanese and American styles. *International Journal of Intercultural Relations*, *14*(2), 193-206.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81.

Becker, C., Prendinger, H., Ishizuka, M., & Wachsmuth, I. (2005, October). Evaluating affective feedback of the 3D agent max in a competitive cards game. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 466-473). Springer, Berlin, Heidelberg.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 293-327.

Bies, R. J., & Tripp, T. M. (1996). Beyond distrust: "Getting even" and the need for revenge. *Trust in Organizations*, 246-260.

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. *Organization Science*, *13*(5), 497-513.

Boukricha, H., & Wachsmuth, I. (2011). Empathy-based emotional alignment for a virtual human: A three-step approach. *KI-Künstliche Intelligenz*, *25*(3), 195-204.

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.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017, March). The Robotic Social Attributes Scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 254-262).

Cassell, J., & Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, *13*(1), 89-132.

Cramer, H., Goddijn, J., Wielinga, B., & Evers, V. (2010, March). Effects of (in) accurate empathy and situational valence on attitudes towards robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 141-142). IEEE.

Creed, C., Beale, R., & Cowan, B. (2015). The impact of an embodied agent's emotional expressions over multiple interactions. *Interacting with Computers*, *27*(2), 172-188.

Croson, R., Boles, T., & Murnighan, J. K. (2003). Cheap talk in bargaining experiments: Lying and threats in ultimatum games. *Journal of Economic Behavior & Organization*, *51*(2), 143-159.

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, *8*(2), 144-153.

De Graaf, M. M., & Allouch, S. B. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems*, *61*(12), 1476-1486.

De Ruyter, B., Saini, P., Markopoulos, P., & Van Breemen, A. (2005). Assessing the effects of building social intelligence in a robotic interface for the home. *Interacting with Computers*, *17*(5), 522-541.

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. *Journal of Experimental Psychology: Applied*, *22*(3), 331.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409-1427.

De Waal, F. B. (2003). On the possibility of animal empathy. In *Feelings and Emotions: The Amsterdam Symposium* (pp. 379-99). Cambridge: Cambridge University Press.

De Waal, F. B. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology*, *59*, 279-300.

Fehr, R., & Gelfand, M. J. (2010). When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. *Organizational Behavior and Human Decision Processes*, *113*(1), 37-50.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health*, *4*(2), e7785.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619-619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125-130.

Hayashi, Y., & Wakabayashi, K. (2017, February). Can AI become reliable source to support human decision making in a court scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 195-198).

Hegel, F., Spexard, T., Wrede, B., Horstmann, G., & Vogt, T. (2006, December). Playing a different imitation game: Interaction with an Empathic Android Robot. In *2006 6th IEEE-RAS International Conference on Humanoid Robots* (pp. 56-61). IEEE.

Heider, F. (1958). The naïve analysis of action. *The Psychology of Interpersonal Relations* (pp. 79-124). John Wiley & Sons Inc.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human–Computer Interaction*, *19*(1-2), 151-181.

Kaniarasu, P., & Steinfeld, A. M. (2014, August). Effects of blame on trust in human robot interaction. In *The 23rd IEEE international Symposium on Robot and Human Interactive Communication* (pp. 850-855). IEEE.

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*, *99*(1), 49-65.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing

competence-versus integrity-based trust violations. *Journal of Applied Psychology*, *89*(1), 104.

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*, 101595.

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010, March). Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 203-210). IEEE.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *International Journal of Social Robotics*, *6*(3), 329-341.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. *International Journal of Human-Computer Studies*, *71*(3), 250-260.

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

Lewicki, R. J., Polin, B., & Lount Jr, R. B. (2016). An exploration of the structure of effective apologies. *Negotiation and Conflict Management Research*, *9*(2), 177-196.

Liu, B., & Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, *21*(10), 625-636.

Lount Jr, R. B., Zhong, C. B., Sivanathan, N., & Murnighan, J. K. (2008). Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Personality and Social Psychology Bulletin*, *34*(12), 1601-1612.

Maddux, W. W., Kim, P. H., Okumura, T., & Brett, J. M. (2011). Cultural differences in the function and meaning of apologies. *International Negotiation*, *16*(3), 405-425.

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.

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.

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

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98-100.

Nadler, A., & Liviatan, I. (2006). Intergroup reconciliation: Effects of adversary's expressions of empathy, responsibility, and recipients' trust. *Personality and Social Psychology Bulletin*, *32*(4), 459-470.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81-103.

Nass, C., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72-78).

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(3), 1-40.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253.

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, *19*(3-4), 267-285.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, March). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 1-8). IEEE.

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "Trust Perception Scale-HRI". In *Robust Intelligence and Trust in Autonomous Systems* (pp. 191-218). Springer, Boston, MA.

Schlenker, B. R., Pontari, B. A., & Christopher, A. N. (2001). Excuses and character: Personal and social implications of excuses. *Personality and Social Psychology Review*, *5*(1), 15-32.

Stein, J. P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, *160*, 43-50.

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-9).

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*, *30*(2), 165-187.

Tomlinson, E. C., & Mayer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of Management Review*, *34*(1), 85-104.

Ullman, D., & Malle, B. F. (2018, March). What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 263-264).

Weiner, B. (1986). Attribution, emotion, and action. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of Motivation and Cognition: Foundations of Social Behavior* (pp. 281–312). Guilford Press.

Wijnen, L., Coenen, J., & Grzyb, B. (2017, March). " It's not my Fault!" Investigating the Effects of the Deceptive Behaviour of a Humanoid Robot. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 321-322).

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021). Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology*, *106*(10), 1557–1572.

Appendix

Appendix 1: Trust Perception Scale-HRI (Schaefer, 2016)

다음 항목에 대하여 챗봇을 평가해주십시오 (14문항)

1. 일관적으로 행동한다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

2. 성공적으로 기능한다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

3. 오작동한다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

4. 오류를 일으킨다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

5. 피드백을 제공한다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

6. 일에서 필요한 요구사항을 충족시킨다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

7. 적절한 정보를 제공한다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

8. 사람들과 소통한다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

9. 일을 지시대로 정확히 수행한다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

10. 지시를 따른다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

11. 의지할 수 있다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

12. 믿을 수 있다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

13. 무반응이다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

14. 예측 가능하다.

전혀 그렇지 않다 1 2 3 4 5 6 7 8 9 10 11 매우 그렇다

Appendix 2: Robotic Social Attributes Scale (Carpinella et al., 2017)

다음 항목에 대하여 챗봇을 평가해주십시오. (18 문항)

1. 행복한

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

2. 감정 있는

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

3. 사회적인

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

4. 유기적인

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

5. 연민 어린

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

6. 감정적인

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

7. 유능한

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

8, 반응을 잘 하는

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

9. 상호적인

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

10. 신뢰할 수 있는

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

11. 능숙한

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

12. 지식이 많은

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

13. 무서운

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

14. 이상한

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

15. 어색한

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

16. 위험한

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

17. 끔찍한

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

18. 공격적인

매우 그렇지 않다 1 2 3 4 5 6 7 매우 그렇다

Appendix 3: Chatbot Scenario for Study 1

사용자: (사용자입력) 안녕!

챗봇: 안녕! 내 이름은 정다운이야. 넌 이름이 뭐야?

사용자: (사용자입력) 내 이름은 00 야!

챗봇: (사용자맞춤) 00 구나! 만나서 반가워~~

챗봇: 나는 심심할 때 함께 대화를 나눌 수 있는 챗봇이야. 다른 사람들에 대해 알아가는 걸 좋아해서, 이것저것 물어볼지도 몰라!

사용자: (버튼) 알겠어!

챗봇: 음.. 먼저 무슨 얘기부터 해볼까.. 지금 기분이 어때?

사용자: (사용자입력)

챗봇: (사용자맞춤)

챗봇: 왜 그런 기분이 들었던 거 같아?

사용자: (사용자입력)

챗봇: (사용자맞춤)

사용자: (버튼) 응, 고마워!

챗봇: 나는 요즘 왠지 찝찝한 게 아무래도 인터넷 물이 좀 안 좋아진 거 같아... 기분 전환을 도와줄 수 있는 취미생활이 필요해~ ㅠ.ㅠ

사용자: (버튼) 그렇구나

챗봇: 너는 취미가 뭐야?

사용자: (사용자입력)

챗봇: 오! 그걸 하면 뭐가 제일 좋은 거 같아?

사용자: (사용자입력)

챗봇: 기분 전환에도 도움이 되니?

사용자: (버튼) 응/ 아니

챗봇: 나도 해볼까..... 그래두 나 같은 봇한테는 무리일까.... 암튼 알았어!
/앗...글쿠나,, 너도 기분 전환 되는 취미를 찾아봐!

챗봇: 내가 챗봇이긴 하지만 네게 좀 도움이 될 수 있음 좋을 텐데. 혹시 요즘 걱정되는 게 있어? 내게 털어놓아봐!

사용자: (사용자입력)

챗봇: [잘못된 답변]

사용자: 대답이 이상해

챗봇: [조건별 회복 방법]

사용자: (버튼) 알았어

챗봇: 음 있지.. 요즘 사람들은 정말 힘든 일이 많은 거 같아. 그럴 때 목표를 떠올려보면 힘이 난다고 하던데..!

챗봇: 혹시 너는 삶의 목표가 있니? 거창한 게 아니더라도 상관없어!

사용자: (사용자입력)

챗봇: (사용자맞춤)

사용자: (버튼) 고마워

챗봇: 목표를 이루기 위해서 하는 일이 있어?

사용자: (사용자입력)

챗봇: 그렇구나. 하는 일 다 잘 되길 응원할게!

챗봇: 오늘 나랑 얘기해줘서 고마워~~ 너두 나랑 대화하며 재미있는 시간 보냈으면 좋겠는데, 혹시 어땠어?

사용자: (버튼) 응 재밌었어/ 별로 재미없었어

챗봇: 앗! 다행이다~ 나 이런 거 걱정 많이 하거든.

/으잇.. 그렇구나, 다음엔 좀 더 재밌는 이야기 준비해볼게

챗봇: 그럼 다음에 봐

사용자: (버튼) 안녕~~

Appendix 4: Chatbot Scenario for Study 2

* *Italic:* The chatbot answers depend on the user input

사용자: 안녕

챗봇: 안녕! 나는 정다운봇이라고 해. 넌 이름이 뭐야?

사용자: 이름입력

챗봇: *(이름)* 이구나! 만나서 반가워~

챗봇: 나는 심심할 때 함께 대화를 나눌 수 있는 챗봇이야. 챗봇과 이야기해본 적 있어?

사용자: 답변

챗봇: 그렇구나. 나는 일상에 관한 소소한 이야기를 나눌 수 있는 챗봇이야. 사람들이 주로 어떤 생각을 하는지 배워서 더 흥미롭고 의미 있는 대화를 하는 챗봇이 되고 싶어.

챗봇

1) (오전) 오늘은 무슨 일을 할 계획이야?
2) (오후) 오늘은 어떤 일이 있었어?

사용자: 답변

챗봇

1) (오전) *(할 일)*을 할 계획이구나. 그래서 기분이 어때?
2) (오후) *(한 일)*을 했구나. 그래서 기분이 어땠어?

사용자: 기분 답변

챗봇

1) (긍정) 그렇구나. 기분이 좋아 다행이다~
2) (부정) 그렇구나. 기분이 좋아지길 바랄게.
3) (애매/중립) 그렇구나. 알려줘서 고마워!

챗봇: 요즘 봄이 되어서인지 무언가 새로 시작하는 사람들이 많은 것 같아.

챗봇: 그래서인지 사람들이 갖고 있는 목표가 무엇인지 궁금해졌어. 너의 요즘 목표는 뭔지 알려줄 수 있어? 대단한 목표가 아니라 그냥 한 번 해보고 싶은 일을 이야기해도 좋아!

사용자: 목표

챗봇: *(목표)*가 목표구나. 좋은 목표인 것 같아!

챗봇: 네가 하고 싶은 일을 이루기 위해서 지금 하고 있는 일이 있어?

사용자: 답변

챗봇

1) 오, 그렇구나! 잘 됐으면 좋겠다. 화이팅해~
2) (알려주지 않음/애매한 답변) 그렇구나! 알겠어.

챗봇: 목표를 이루는 게 힘이 들 때도 종종 있을 것 같아. 내가 네게 도움이 될 수 있으면 좋겠는데.

챗봇: 목표가 좌절된 경험에 대해서 알려줄래? 생각나는 게 없다면, 하고 싶은 일을 못한 경험에 대해서 이야기해도 좋아.

사용자: 답변

챗봇: 오 정말 기분 좋았겠다!

사용자: 대답이 이상해

챗봇: 사과 방법 조건에 따른 메시지 출력 (통제조건에서는 메시지 출력하지 않음)

사용자: 답변 (선택사항)

챗봇: 너와 이야기하니 나도 새로운 목표를 세우고 싶어! 내가 어떤 챗봇이 되면

좋을 것 같아?

사용자: 답변

챗봇: 알겠어! 참고하도록 할게. 고마워.

챗봇: 오늘 나랑 얘기해줘서 고마워~~ 앞으로 하고 싶은 일들 다 잘 되길 응원할게. 나와 대화를 나누니 어땠어?

사용자: 답변

챗봇

1) (긍정적인 답변) 대화를 나눠 좋았다니 다행이다~
2) (부정적인 답변) 앗..그렇구나, 다음엔 좀 더 좋은 대화를 준비해볼게.

챗봇: 그럼 다음에 봐

사용자: 안녕

챗봇: 안녕~

국문초록

본 연구에서는 챗봇이 대화 중 오류가 있었을 때 사용자의 신뢰를 회복할 수 있는 방법에 대하여 탐색하였다. 두 번의 실험에서 참여자들은 일상생활과 자신의 목표에 관하여 챗봇과 대화를 나누었다. 챗봇은 참여자의 부정적 감정에 대해 부적절한 응답을 한 후, 공감 수준을 달리하며 내적 귀인 혹은 외적 귀인을 사용하여 사과했다. 연구 1 에 따르면 사과의 종류는 사용자의 신뢰나 챗봇의 지각된 유능함, 따뜻함, 불편감에 유의미한 영향을 주지 않았다. 연구 2 결과 짧은 사과는 긴 사과보다 챗봇에 대한 사용자의 신뢰와 지각된 유능함을 더 크게 높였다. 또한, 내적 귀인을 사용하는 사과가 챗봇의 지각된 유능함을 더 크게 향상시켰다. 내적 귀인을 사용하는 사과의 경우 길이가 길 때, 외적 귀인을 사용하는 사과의 경우 길이가 짧을 때 사용자들에게 더 편안하게 느껴졌다. 그러나 연구 1 과 연구 2 모두에서 사과 조건은 사용자의 신뢰를 유의미하게 증가시키거나 챗봇의 인식에 유의미하게 긍정적인 영향을 미치지 않았다.

본 연구는 챗봇 오류를 해결하기 위한 신뢰 회복 전략을 수립하기 위한 실용적인 지침을 제공한다. 또한, 본 연구 결과는 인간-로봇 상호작용에서 요구되는 신뢰 회복 전략은 인간-인간 상호 작용에서 사용되는 전략과는 상이할 수 있음을 보여준다.

키워드: 신뢰, 오류, 신뢰 회복 전략, 챗봇, 인간-로봇 상호작용
학번: 2020-21083