Owning Mistakes Sincerely: Strategies for Mitigating AI Errors

Amama Mahmood
amahmo11@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Jeanie W Fung
jfung4@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Isabel Won
iwon1@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Chien-Ming Huang
chienming.huang@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

ABSTRACT

Interactive AI systems such as voice assistants are bound to make errors because of imperfect sensing and reasoning. Prior human-AI interaction research has illustrated the importance of various strategies for error mitigation in repairing the perception of an AI following a breakdown in service. These strategies include explanations, monetary rewards, and apologies. This paper extends prior work on error mitigation by exploring how different methods of apology conveyance may affect people's perceptions of AI agents; we report an online study (N=37) that examines how varying the sincerity of an apology and the assignment of blame (on either the agent itself or others) affects participants' perceptions and experience with erroneous AI agents. We found that agents that openly accepted the blame and apologized sincerely for mistakes were thought to be more intelligent, likeable, and effective in recovering from errors than agents that shifted the blame to others.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing Methodologies** → *Artificial intelligence.*

KEYWORDS

Human-AI interaction, error mitigation, apologies, blame assignment, voice interactions

ACM Reference Format:
Amama Mahmood, Jeanie W Fung, Isabel Won, and Chien-Ming Huang. 2022. Owning Mistakes Sincerely: Strategies for Mitigating AI Errors. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3491102.3517565

1 INTRODUCTION

Speech-based systems such as Amazon's Alexa, Apple's Siri, and Google Assistant are fueled by data-driven deep learning algorithms. However, these algorithms are prone to errors due to uncertainties in the real world. Algorithmic mistakes impact the behaviors of voice assistants by causing detection errors, faulty interpretations, limited comprehension or abilities, or confusion of similar words and sentences. Despite recent advances in Artificial Intelligence (AI), voice assistants are bound to make mistakes and will continue to do so. For example, voice assistants might add the wrong item to shopping lists, text the wrong person, or simply not listen when they are addressed. Yet, people's expectations are misled by the futuristic portrayal of AI capabilities. As a result, even the smallest failures may violate users' expectations and hinder user adoption. To effectively foster productive interactions between people and voice assistants, or AI assistants in general, it is important to mitigate AI errors when they happen.

In fact, mitigating errors is critical to maintaining satisfactory service interactions between humans [50]. Amongst various strategies, an apology has proven to be effective for repairing relationships when human trust is damaged. Even needless apologies can sometimes increase trust as they reflect empathy and concern for the wronged party [5]. In the context of human-robot interactions, various mitigation strategies have shown positive effects on people's perceptions and willingness to use faulty robots. Some such strategies include fault justification [10] and apologies and compensation [31]. While compensation and explanations may be appropriate for short-term repair to restore service satisfaction, apologies seem to be important for promoting long-term positive outcomes such as trust, psychological closeness, and willingness to return to the service [31]. Moreover, apologies allow for a much needed emotional shift towards forgiveness [17]. For an apology to be effective, it needs to sound genuine [36]. However, a recent study showed that machines are perceived as "not having regret" [20]. So, *how can a machine sound sincere and genuine when mitigating its errors? Can apologies delivered with certain sincerity facilitate rapport building between people and machines?*

Another important aspect of mitigating errors through apologies is fault justification [10]. Prior works on blame attribution found that the assignment of blame depends on many factors including anthropomorphism and autonomy of the machine. For instance, if an embodied agent acts anthropomorphically and autonomously, more blame gets assigned to it, presumably because it conveys human-like traits [27]. For anthropomorphic and autonomous machines, people perceived internal attribution (e.g., taking the blame) while apologizing to be more effective. However, for non-autonomous

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517565

machine-like agents, blaming external factors helped agents maintain a positive image [14, 28]. Recent works suggest that people associate human-like characters to voice-based personal assistants (Alexa, Siri, etc.) and that different people anthropomorphize the assistants differently [1, 30, 43]. All these findings led us to ask, in the context of voice-based interaction, *how should voice assistants apologize? Should the blame be internalized by the assistants?*

To answer these questions, we investigate the sincerity of an apology in terms of its seriousness (serious vs. casual) and the assignment of blame (taking the blame vs. shifting the blame). The investigation takes place in an online shopping task with voice-based personal assistants where the assistants successfully recover from the error. Our investigation allows us to study whether the findings on effect of various aspects of an apology during user interactions with other humans, robots, and embodied virtual agents map to voice-based conversational agents which lack human-like embodiment. Recent research in interactions with conversational agents suggests that a progression of the task is desirable when conversation failures occur [12, 44]. Therefore, in this work, we focus on studying apology as a tool to mitigate negative impressions of AI agents 'immediately after a successful recovery from failure' without exploring various strategies of recovery from error itself.

Our results highlight that 1) a serious apology with an acceptance of blame is the most preferred, 2) not all apologies are perceived equally; in terms of willingness to use in future, offering no apology at all is better than an apology lacking in an acceptance of blame. Our findings have numerous design implications for creating voice assistants capable of sustaining positive relationships and experiences after service breakdowns.

Next, we review relevant prior research that motivate this work. We describe our experiment exploring how the sincerity of an apology and blame assignment affect people's perceptions of voice assistants in Section 3. We present the results of our experiment in Section 4 and conclude the paper with a discussion of our findings, limitations, and future work in Section 5.

2 BACKGROUND AND RELATED WORK

2.1 Error Recovery and Repair in Voice-based Interactions

Conversation breakdowns are inevitable in Voice User Interfaces (VUIs); breakdowns occur when users interact with an unfamiliar VUI and encounter errors related to Natural Language Processing (NLP) such as unrecognized intent and failed feedback [41]. Pearl [42] describes four different types of errors that may halt smooth conversation with voice-based AI agents: 1) no speech detected, 2) speech detected but not recognized, 3) speech recognized but not handled, and 4) speech recognized but incorrectly. The author further provides various suggestions on how to handle these errors. For instance, N-best lists may be used to recover from situations where speech is recognized incorrectly (intent errors). This particular suggestion informs the design of error recovery in our study.

The progression of a task after a breakdown is essential to the continued use of the item, and it is a main concern for conversational assistants (CAs) [12, 44]. To aid in the development of effective VUIs, prior research have studied how everyday users interact with VUIs and how users navigate their conversations to achieve certain goals [44]. They have also explored various strategies for VUIs to recover from errors [18]. A recent systematic literature review on recovery strategies to overcome conversational breakdowns [4] has identified six categories: 1) confirmation, 2) information, 3) disclosure, 4) social, 5) solve, and 6) ask. *Confirmation* is the mere acceptance that the agent does not understand what is asked of it. *Information* and *disclosure* are trying to provide useful feedback on the breakdown and present the agent's competencies and vulnerabilities to the user, respectively. An agent employing *solve* strategies tries to present a solid solution for the error. *Ask* strategy shifts the responsibility to the user by repeating the question to get clarification from the user. *Social* strategies are inspired by human-human interactions where it is common to repair broken trust by offering an apology, explanation, and compensation. However, these "*social*" strategies are mostly used in combination with others [4].

In this work, we explore apology from the the "*social*" recovery category. The apology is used as a tool to repair user-agent relationships in the context of online shopping with voice assistants. We create a fixed "*solve*" strategy to recover from intent-based (recognized but incorrectly) errors while manipulating two aspects of apology (sincerity and blame assignment) that are inspired from prior work on error mitigation and repair in human-human and human-robot interaction.

2.2 Apology as an Error Mitigation Strategy in Human-Human Interaction

Appropriate error mitigation and repair after errors are key components of human-to-human interactions. In service interactions, appropriate mitigation strategies are critical in repairing relationships with dissatisfied customers [50]. Without it, there could be an overall increase in negative reactions to the service. Appropriate recovery efforts encourage consumers to believe that the service is fair and able to recognize and account for mistakes. In fact, there are some instances in which the service failure recovery efforts leave consumers more satisfied than before the mistake had occurred [8]. As a result, the impact of the recovery is important, if not even more critical than the original failure of the service [50].

A major topic in service theory involves the role of apologies in repair. Apologies are defined as "admissions of blameworthiness and regret" for an "undesirable event that allows actors to try to obtain a pardon from audiences" [13]. Apologies demonstrate politeness, concern, effort, and empathy. They also increase the overall satisfaction of a service for recipients. [49]. Thus, apologies are important because they show the offender's willingness to admit that they have done something wrong and claim responsibility for the wrongdoings. In fact, even superfluous apologies can increase trust as it signifies empathy and concern for the victim, even if the offender is not truly guilty [5]. Apologies signify acknowledgment of the victim's dignity and moral worth, and represent the offender's respect for their feelings. Thus, appropriate apologies provide a critical set-up for an "emotional shift toward forgiveness" [17], suggesting that in human-to-human interactions, offering an apology can be effective in repairing a relationship, promoting reconciliation, and regaining trust.

However, the relationship between humans and intelligent systems is not fully understood, especially when it comes to errors and broken trust. Previous work suggests that humans perceive interactions with non-humans systems differently than with humans. For example, a mis-attribution of a mistake is common in relationships between humans and intelligent systems, and humans are more likely to trust other humans over non-human agents [46]. Although some work points to differing results, this effect is noticed in the service context as well. For example, when a particular service fails, customers are more likely to attribute less responsibility to service-provider-robots than human providers [32]. This is because robots are believed to have less control over a given task. Another study suggests that erroneous robots may even be more likeable than error-free robots [37]. However, other studies found that people generally consider erroneous robots to be less intelligent, reliable, competent, and superior than error-free robots [6].

Voice-based conversational agents are perceived differently than robots because of the varied level of embodiment. A study on conversational errors found that failures are perceived to be more severe for smart speaker embodiment than human-like robot embodiment even though the human-likeness is distracting and detrimental to the interaction. Additionally, smart speakers are perceived lower in intelligence and social presence than robots [29]. Thus, further work is required to gain greater understanding of appropriate conversational behavior, especially when it comes to maintaining trustful relationships with humans after failures. In this study, we explore this issue in the context of voice-based conversational agents.

2.3 Apology as an Error Mitigation Strategy in Human-Robot and Human-AI Interaction

A robot that identifies its mistake, and communicates its intention to rectify the situation is considered to be more capable than one that simply apologizes for its mistake. However, the latter is more likeable and, uniquely, increases people's intention to use the robot [7]. Similarly, another study on apology and compensation as error mitigation strategies revealed that receiving an apology was more effective across many ratings including politeness, competence, trust, likeness, feeling of being close to the robot, and the willingness to return [31]. The robot's apology included acknowledgment and explanation of the error (e.g., "*I thought this was Coke. I apologize for bringing the wrong one*"). Under the compensation condition, participants were given a free drink. Results show that this compensation strategy performed better for service satisfaction. This indicates that for immediate satisfaction, compensation was more effective, but for willingness to return and continual of use of technology, an apology was better received. Furthermore, the authors show that stronger *relational* orientation biased participants (who want to maintain good relationship with service provider) appreciate the apology strategy when it comes to quality of service. They disliked the compensation strategy as much as they disliked having no strategy. Whereas, the *utilitarian* orientation biased participants (who care more about quality and effectiveness of service than provider) rated the service with compensation strategy as most satisfactory [31].

In terms of the type of apology, fault justification has been found to be an important aspect of an apology. One study explored the reaction of technical failures on trust in a collaborative tangram puzzle solving setting. In the study, the Nao robot justified the failure by saying: "*There was a failure in my speech module. Let's restart*" [10]. This fault justification mitigated the negative impacts of the failure. The robot attributed the blame to itself, and users appreciated not being blamed for a mistake. Another study researched the relationship between human drivers and virtual passenger's (introduced as the speech-based interface to an in-car information system) blame attributions in a driving simulator. Drivers felt more at-ease, rated the car higher, and had better attention towards the road when the virtual passenger attributed the blame to external factors (i.e., the environment such as road as opposed to directly blaming the driver) [25]. Similarly, another study on the attribution of blame in human–robot team task failures showed that in general, humans were assigned the most blame, followed by robots, and lastly environmental factors. If the robot was portrayed as autonomous, it was assigned as much blame as a human. However, when it was portrayed as non-autonomous, the blame assigned to the robot was closer to the amount of blame placed on the environment [14].

Moreover, anthropomorphism may influence the perception of AI agents. For instance, Kim and Song [28] investigated the effect of apologies in regaining trust when the AI agent demonstrated human-like vs. machine-like behaviors. The agent either had a human icon as its profile picture or a picture of a computer. Additionally, the agent referred to itself by either a first-person singular pronoun or as an 'algorithm'. The agent apologized with either an internal (accepted full responsibility for the trust violation) or external attribution (accepted only partial responsibility and attributed the rest to an external source, such as an abnormality in the environment). When participants interacted with the human-like agents rather than the machine-like agents, trust was more efficiently repaired. Additionally, trust was less damaged when machine-like agents apologized with external, rather than internal attribution. However, for human-like agents, people prefer internal explanations to its limitations and mistakes [28]. These results suggest that there are different sentiments based on the different expectations of robots and virtual agents. People may assume that an artificial agent's competence levels are fixed and unable to adjust to different tasks accordingly. However, when an agent acts anthropomorphically and human-like, people may expect it to malleably change its behavior like a human would to adjust to specific circumstances [27]. All in all, the assignment of blame is dependent on behavior and embodiment of AI agents (i.e., varied level of anthropomorphism). However, it is unclear how an apology with an acceptance or aversion of blame will be perceived for conversational agents. Voice agents communicate verbally in a somewhat human-like manner but has a machine-like embodiment. In this study, we look at how blame attributed to itself or others by a voice assistant affects the quality of apology. We evaluate the voice assistant using service recovery satisfaction and willingness to use in the future.

When considering the importance of apologies in repairing relationships, the style and tone of an apology should also be noted. For instance, Roschk and Kaiser [47] argue that empathy, intensity, and the timing of an apology affect how well the apology is received. More empathetic and intense apologies lead to greater satisfaction from the recipient. Empathy is important as it demonstrates "warmth towards the victim, an understanding of the wrongdoing, or personal remorse" [11]. Moreover, an apology must be perceived as sincere in order for it to be effective. Sincerity is described as a "category of affective subjectivity" which plays an important role as a "condition for the successful communication between people" [2]. Thus, providing excuses and reasons for service failure that do not seem genuine will not satisfy customers [36]. In a recent study, Hu et al. [20] highlighted the impact of sincerity in service encounters between humans and embodied AI agents. The study tested the hypothesis that customers desire and need human-like interactions. Participants engaged with either a robot-generated text message, a robot-generated voice, or a human service employee receiving a service recovery message. Results suggested that recovery efforts provided by a human feel more sincere than those offered by a service robot, which led to higher rates of recovery satisfaction [20]. Thus, if we believe that sincerity is inherently a human characteristic, one may argue that apologies provided by non-humans (e.g., robots and AI agents) may never truly be considered genuine and sincere. However, a study on understanding the responses to errors made by smart speakers suggested that a neutral apology is found to be more sincere than a humorous apology [15]. Based on prior work, we explore two levels of sincerity in a apology: serious (similar to neutral) and casual (similar to humorous). In this paper, we strive to understand whether sincerity can be expressed and received as genuine in apologies between AI assistants and humans.

3 METHODS

In this section, we describe the study that we conducted to investigate how various error mitigation strategies affect users' perceptions of conversational agents.

3.1 Hypotheses

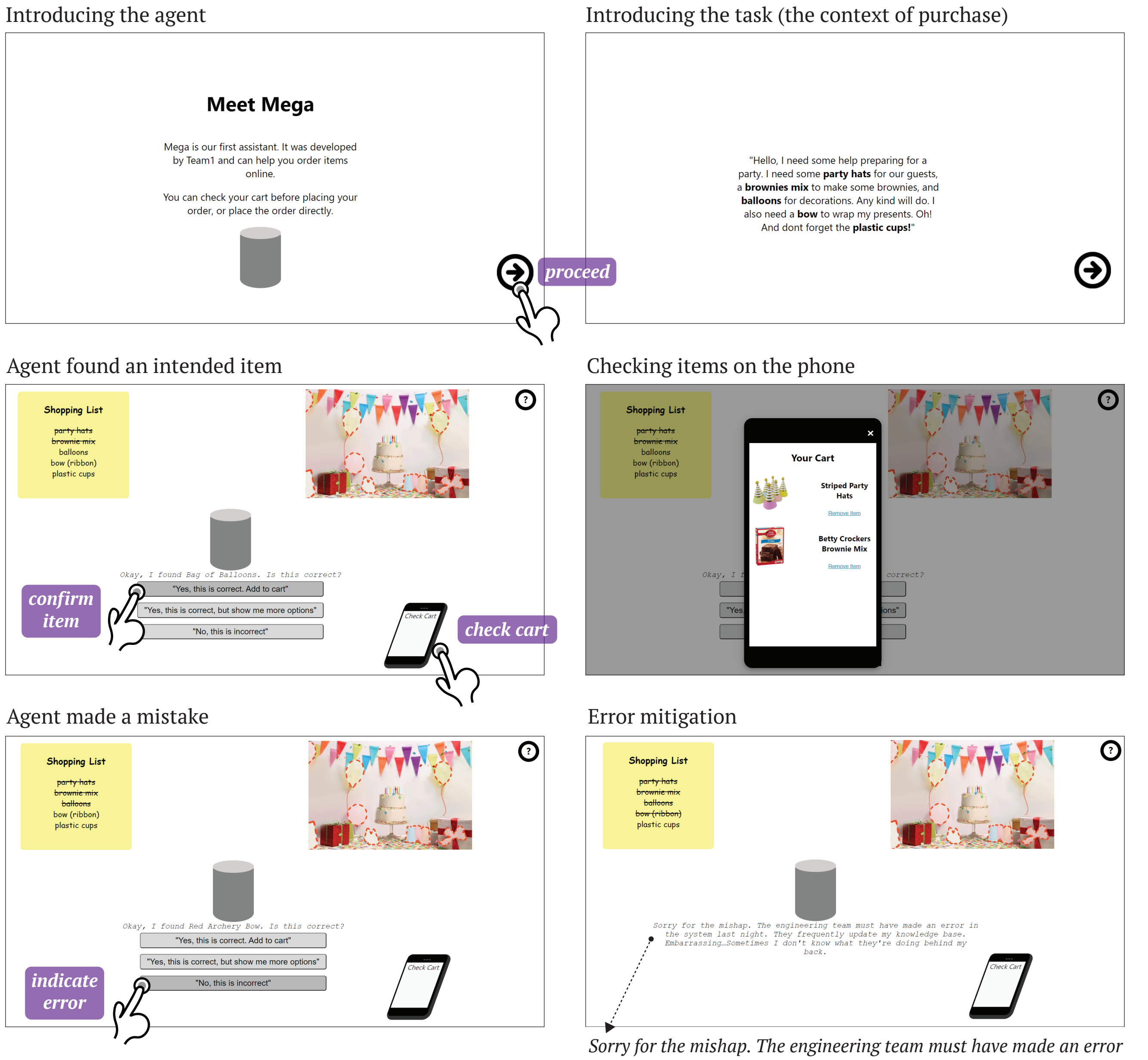
- **Hypothesis 1.** Any mitigation strategy is preferred over no mitigation. This hypothesis is informed by previous research showing that 1) apologies reduce the negative effects of errors or service failures [31]; 2) apologies make an agent more likeable and increases user's intention to use in the future [7]; and 3) even superfluous apologies increase positive impressions as they signify empathy and concern for the wronged [5].
- **Hypothesis 2.** A serious (more sincere) apology from an AI agent will result in higher satisfaction from service recovery; moreover, the agent will be perceived as more intelligent and likeable than a casual apology. This hypothesis is motivated by previous findings indicating that a sincere and genuine apology is essential to repair user perceptions of AI agents when encountering errors [20, 36].
- **Hypothesis 3.** Accepting the blame will maintain a positive image of the agent despite the error. This hypothesis is informed by prior works illustrating that 1) apologizing with an internal attribution (taking the blame), as opposed to external (blaming others), is found to be more effective in repairing user's perceptions when the virtual agent is more human-like [28] and 2) people associate human-like characters with voice-based personal assistants such as Siri and Alexa [1, 30, 43].

3.2 Experimental Task, Study Design, and Conditions

We conducted the experiment online as an interactive storyboard via a web application. Participants interacted with a set of five AI agents in a simulated online shopping scenario. We presented the task as beta testing five AI assistants developed by different teams (e.g., "Team 1" and "Team 2" etc). The AI assistants are portrayed as static images and used to order items online. In this study, we focused on intent recognition errors described as "recognized but incorrectly" [42] in online shopping tasks. The errors fabricated in this study are the use of homonyms—words that sound the same but have different meanings. For instance, a "bow" could mean a hair bow or an archery bow. N-best lists were one of the suggested ways to handle or attempt to recover from such errors (i.e., returning a list of top N possibilities of what the user might have said, ordered by likelihood and the confidence score [42]). Since there were only two possible objects associated with the selected homonym, the agent was able to fix and recover from the error in its first attempt of recovery.

In each task, the user was given a list of five items to order through a simulated speech-based interaction with the AI assistant (Figure 1). The AI assistant would add the item to the cart, and the user could check the items by accessing the mobile phone icon anytime during the experiment. The AI assistant would make a mistake on either the second, third, or fourth item on the list. The user could report a mistake via the cart on the displayed phone or when the mistake happened during the interaction. When the user indicated an error by clicking "No, this is incorrect", the agent first prompted the user for input again as an attempt to fix the error by saying, "Let's try that again". Given that the errors in our study were regarding homonyms and intent-recognition-based, the agent was able to fix its mistake by suggesting the correct item on the second try. After a successful recognition of the intended item, and the user indicated "Yes, this is correct. Add to cart", the agent performed the recovery strategy by initiating an apology based on the experimental condition. After ordering the five items on the list, the user was directed to the next page indicating that the items have arrived. The user is prompted to verify the correctness of the ordered items, or report a mistake. If the user indicated that an item was incorrect after delivery, the agent would apologize and initiate a refund request on the user's behalf. If the user reported a mistake before placing the full order, the agent mitigated the error according to the experiment condition immediately; otherwise, the agent mitigated the error at the last stage of item verification as described above.

The voice of the agents was female-gendered because most of the AI agents in our present digital world are gendered as female. For example, voice assistants such as Apple's Siri, Amazon's Alexa, and Google Assistant all have female voices set as default. We used the online text-to-speech service "Sound of Text" [39] to generate the voice of AI agents. We implemented the interface as a React App. The database to store interaction data was based on MongoDB and Microsoft Azure. The front-end was deployed on Github and the back-end was deployed on Microsoft Azure.



Introducing the agent

Meet Mega

Mega is our first assistant. It was developed by Team1 and can help you order items online.

You can check your cart before placing your order, or place the order directly.

Introducing the task (the context of purchase)

"Hello, I need some help preparing for a party. I need some **party hats** for our guests, a **brownies mix** to make some brownies, and **balloons** for decorations. Any kind will do. I also need a **bow** to wrap my presents. Oh! And dont forget the **plastic cups!**"

Agent found an intended item

Shopping List

~~party hats~~
~~brownie mix~~
balloons
bow (ribbon)
plastic cups

Okay, I found Bag of Balloons. Is this correct?

Checking items on the phone

Your Cart

Striped Party Hats

Betty Crockers Brownie Mix

Agent made a mistake

Shopping List

~~party hats~~
~~brownie mix~~
~~balloons~~
bow (ribbon)
plastic cups

Okay, I found Red Archery Bow. Is this correct?

Error mitigation

Shopping List

~~party hats~~
~~brownie mix~~
~~balloons~~
~~bow (ribbon)~~
plastic cups

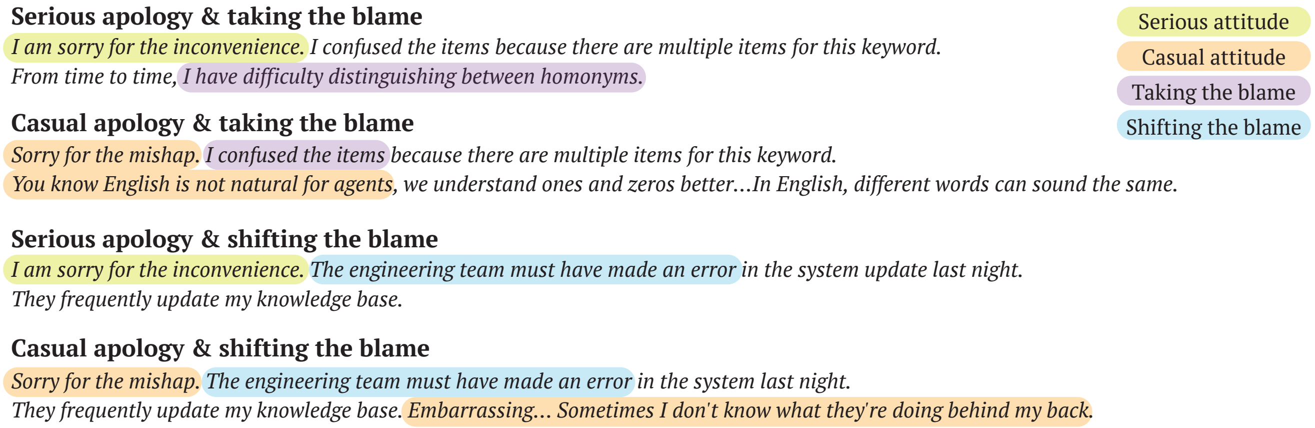
Sorry for the mishap. The engineering team must have made an error in the system last night. They frequently update my knowledge base. Embarrassing…Sometimes I don't know what they're doing behind my back.

Sorry for the mishap. The engineering team must have made an error in the system last night. They frequently update my knowledge base. Embarrassing…
Sometimes I don't know what they're doing behind my back.

Figure 1: A storyboard of the study procedure depicting a participant's interaction with an agent. The agent's monologue has audio output. In this example, the agent responds with a casual apology while shifting the blame to the engineering team.

The experiment was a within-subjects design, consisting of a baseline condition and four experimental conditions. The four experimental conditions were based on two aspects of an apology—*sincerity of the response* (*serious vs. casual*) and *blame assignment* (*taking the blame vs. shifting the blame*). Below, we describe the five conditions (Figure 2):

- **Control.** No error mitigation strategy is employed when a participant indicates an error. The assistant simply asks the participant to try again by saying "*Let's try that again*".
- **Serious + Taking the blame.** The assistant provides a serious apology and takes the blame for the mistake when the participant indicates an error. The assistant says: "*I am sorry for the inconvenience. I confused the items because there*



Serious apology & taking the blame

I am sorry for the inconvenience. I confused the items because there are multiple items for this keyword. From time to time, I have difficulty distinguishing between homonyms.

Casual apology & taking the blame

Sorry for the mishap. I confused the items because there are multiple items for this keyword. You know English is not natural for agents, we understand ones and zeros better…In English, different words can sound the same.

Serious apology & shifting the blame

I am sorry for the inconvenience. The engineering team must have made an error in the system update last night. They frequently update my knowledge base.

Casual apology & shifting the blame

Sorry for the mishap. The engineering team must have made an error in the system last night. They frequently update my knowledge base. Embarrassing… Sometimes I don't know what they're doing behind my back.

Figure 2: Examples of error mitigation strategies: sincerity of apology (*serious vs. casual*) and blame assignment (*accept vs. shift*).

are multiple items for this keyword. From time to time, I have difficulty distinguishing between homonyms".

- **Casual + Taking the blame.** The assistant provides a casual and humorous apology and takes the blame for the mistake when the participant indicates an error. The assistant says: "*Sorry for the mishap. I confused the items because there are multiple items for this keyword. You know English is not natural for agents, we understand ones and zeros better…In English, different words can sound the same*".
- **Serious + Shifting the blame.** The assistant provides a serious apology and shifts the blame to other factors such as a system update by the engineering team, for the mistake when the participant indicates an error. The assistant says: "*I am sorry for the inconvenience. The engineering team must have made an error in the system update last night. They frequently update my knowledge base*".
- **Casual + Shifting the blame.** The assistant provides a casual and humorous apology and shifts the blame to other factors such as a system update by the engineering team, for the mistake when the participant indicates an error. The assistant says: "*Sorry for the mishap. The engineering team must have made an error in the system last night. They frequently update my knowledge base. Embarrassing… Sometimes I don't know what they're doing behind my back*".

3.3 Measures

We used a range of metrics to measure service recovery satisfaction, perceived intelligence, likeability, and willingness to use the AI assistant. We also included manipulation check for blame assignment.

3.3.1 Manipulation check. We included a question ("*AI assistant acknowledged the mistake as its own*") to check if our manipulation of blame assignment was adequate. On a 5-point scale, participants rated how much they agree with the statement (1 being "strongly disagree" and 5 being "strongly agree").

The design of our serious and casual apologies followed a similar design of error responses in a prior study [15]. In particular, the study manipulated humor (humorous vs. neutral) while apologizing for an error in conversation with a smart speaker. It showed that the sincerity of a neutral apology was perceived higher than a humorous apology. The design of their humorous apology ("I'm sorry, my IQ is still recharging, please repeat it again") was similar to that of our casual apology ("Sorry for the mishap." and "You know English is not natural for agents, we understand ones and zeros better…"). Similarly, the design of their neutral apology ("I'm sorry, I didn't understand, please repeat it again") was similar to that of our serious apology ("I am sorry for the inconvenience. I confused the items because there are multiple keywords"). Therefore, we did not include a direct manipulation check for sincerity (serious vs. casual) in this study.

3.3.2 Subjective measures of user experience and perceptions of AI.

- **Service recovery satisfaction** (Two items; Cronbach's α = .89). We used two questions ("*I am happy with how the error was handled*" and "*In my opinion, the AI assistant provided a satisfactory response to the error*") as informed by prior work on apology [47] in the domain of consumer services [35, 51] to measure service recovery satisfaction.
- **Perceived intelligence** (Four items; Cronbach's α = .90). We used Godspeed questionnaire [3] to measure the perceived intelligence of the AI assistant on a 5-point rating scale. We asked the participants to rate their impression of the agent on these dimensions: 1) Incompetent – Competent, 2) Ignorant – Knowledgeable, 3) Irresponsible – Responsible, and 4) Foolish – Intelligent.
- **Likeability** (Three items; Cronbach's α = .86). We used Godspeed questionnaire [3] to measure likeability on a 5-point rating scale. We asked the participants to rate their impression of the agent on these dimensions: 1) Dislike – Like, 2) Unfriendly – Friendly, and 3) Awful – Nice.
- **Willingness to use in the future** (Single item). We asked an additional question about willingness to use the voice assistant ("*I would be willing to use this smart speaker for ordering my usual things online*"). We used a 5-point rating scale 1 being "strongly disagree" and 5 being "strongly agree".

3.4 Procedure

This study consisted of four phases:

1. *Introduction and consent.* At the start of the study, participants were provided with a brief description of the study. The description stated that participants would be ordering items from shopping lists while interacting with five AI assistants. The participation was voluntary; participants agreed to continue the study by following the instructions to navigate to the next page to begin the study.
2. *Experimental task: simulated shopping.* Participants were randomly assigned one of the rows in a Latin square of order 5, dictating the order of experimental tasks. Although the Latin square rows did not have equal number of participants due to the random assignment, each row had at least 5 participants.
3. *Perception survey.* After interacting with the AI assistant, participants filled a questionnaire about their perceptions of the AI assistant. They continued onto the next condition and repeated phases 2 and 3.
4. *Post-study questionnaire.* After completing all the conditions, participants filled out a post-study demographics questionnaire.

The study was approved by our Institutional Review Board (IRB). The study took approximately 20 minutes to complete. The participants were compensated with a $5 USD gift card for their participation in the study.

3.5 Participants

A total of 37 participants (29 females, 8 males) were recruited for this online study using convenience sampling. The participants were aged between 18 to 48 ($M = 21.76, SD = 6.75$) and had a variety of education backgrounds, including computer science, engineering and technology, healthcare, life sciences, media sciences, and education.

4 RESULTS

Our data analysis included a total of 185 trials from the 37 participants (five trials for five conditions per participant). In our analysis, we first checked if the participant identified the targeted error correctly in each trial. In 5 of the 185 trials, the intended errors were not correctly identified. To handle these missing values, first, we analyzed whether these values are missing completely at random (MCAR) or not. Little's MCAR test [34] was not significant suggesting that it is safe to assume that data is MCAR, $\chi^2(179, N = 37) = 69.879, p = 1.000$. Multiple Imputation (MI) [48] is one of the optimal techniques to handle the missing data and gives unbiased results under MCAR [52]. Thus, before proceeding with our analysis, we replaced the missing data using MI in SPSS by computing five imputations and pooling the results, taking into account variation across these imputations.

For the results reported below, we used one-way repeated measure analysis of variance (ANOVA). The experimental condition was set as the fixed effect and participants as a random effect. All post-hoc pairwise comparisons were conducted using Tukey's HSD test. For all the statistical tests reported below, $p < .05$ is considered as a significant effect. We follow Cohen's guidelines on effect size and considered $\eta_p^2 = 0.01$ a small effect size, $\eta_p^2 = 0.06$ a medium effect size, and $\eta_p^2 = 0.14$ a large effect size [9]. Figure 3 visualizes our main results.

4.1 Manipulation Check

A one-way repeated measure ANOVA yielded a significant main effect of the experimental condition ($F(4, 144) = 35.655, p < .001, \eta_p^2 = .498$) on the participants' perceptions of whether the AI acknowledged the mistake as its own. Pairwise comparisons using Tukey's HSD test revealed that the two agents that accepted the blame (serious-accept: $M = 4.57, SD = 0.80$ and casual-accept: $M = 3.81$, $SD = 1.29$) were both rated higher than the two agents who shifted the blame (serious-shift: $M = 2.16, SD = 1.39$ and causal-shift: $M = 1.92, SD = 1.19$), $p < .001$ and the control ($M = 2.54$, $SD = 1.41$), $p < .001$, indicating that our manipulation of blame assignment was adequate. Moreover, serious-accept was rated higher than casual-accept, $p = .045$

4.2 Service Recovery Satisfaction

A one-way repeated measures ANOVA revealed a significant main effect of the experimental condition ($F(4, 144) = 3.193, p = .015, \eta_p^2 = .081$) on participants' satisfaction of service recovery. Pairwise comparisons using Tukey's HSD test showed that the participants were more satisfied with service recovery provided by the agent that accepted the blame and apologized seriously ($M = 4.12, SD = 1.09$) than the agent that shifted the blame and apologized casually ($M = 3.43, SD = 1.36$), $p = .037$.

4.3 Perceived Intelligence

A one-way repeated measures ANOVA yielded a significant main effect of the experimental condition ($F(4, 124) = 6.651, p < .001, \eta_p^2 = .156$) on perceived intelligence. Pairwise comparisons using Tukey's HSD test suggested that the agent who accepted the blame and apologized seriously ($M = 4.22, SD = 0.71$) was perceived more intelligent than: the agent that shifted the blame but apologized seriously ($M = 3.61, SD = 0.95$), $p < .001$; the agent that accepted the blame but apologized casually ($M = 3.78, SD = 0.85$), $p = .012$; and the agent that shifted the blame and apologized casually ($M = 3.64, SD = 0.97$), $p < .001$

4.4 Likeability

A one-way repeated measures ANOVA found a significant main effect of the experimental condition ($F(4, 144) = 7.618, p < .001, \eta_p^2 = .175$) on how likable the participants thought the agent was. Pairwise comparisons using Tukey's HSD test revealed that the agent who accepted the blame and apologized seriously ($M = 4.26, SD = 0.76$) was perceived more likable than all other agents: the agent that shifted the blame but apologized seriously ($M = 3.60, SD = 1.01$), $p < .001$; the agent that accepted the blame but apologized casually ($M = 3.68, SD = 0.77$), $p = .002$; the agent that shifted the blame and apologized casually ($M = 3.49, SD = 1.00$), $p < .001$; and the agent who made no apology (control) ($M = 3.81, SD = 0.93$), $p = .041$.

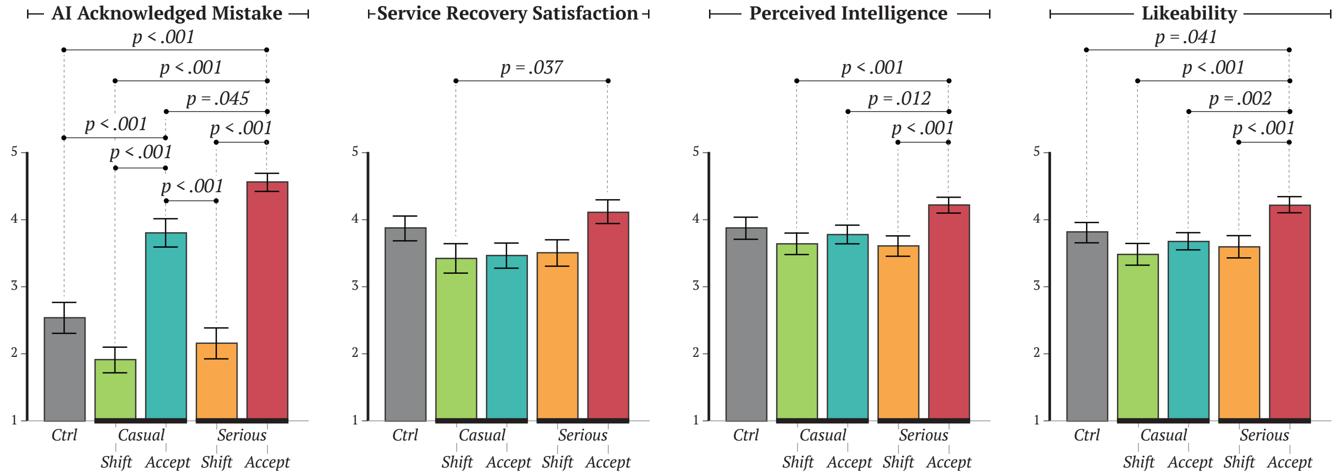


Figure 3: Results of acknowledgement of mistake, perceived intelligence, likeability, and service recovery satisfaction. One-way repeated measures ANOVAs were conducted to discover effects of five conditions on subjective measures. All pairwise comparisons were conducted using Tukey's HSD method. Error bars represent standard error (SE) and only the significant comparisons ($p < .05$) are highlighted.

4.5 Willingness to use in the future

A one-way repeated measures ANOVA yielded a significant main effect of the experimental condition ($F(4, 144) = 3.430, p = .010, \eta_p^2 = .087$) on the participants' willingness to use the AI agent in their future online shopping. Pairwise comparisons using Tukey's HSD test showed that the participants showed more willingness to use in future for the agent who did not apologize at all (control) ($M = 3.81, SD = 0.97$) than the agent that shifted blame but apologized seriously ($M = 3.24, SD = 1.18$), $p = .016$.

5 DISCUSSION

Mitigating AI errors effectively is essential for user satisfaction, good rapport with the AI agent, and continued use of the system. This work examined how *sincerity of apology* (serious vs. casual) and *blame assignment* (taking the blame vs. shifting the blame) may affect participant's satisfaction with the service recovery and perceptions of voice assistants. In this section, we discuss our results in detail and their design implications, the limitations of this work, and future research directions.

5.1 A Good Apology: Owning Mistakes Sincerely

Prior work on effectiveness of sincere apologies in repairing relationship and perception of agents and robots [20] and impacts of anthropomorphism [28] and level of automation [14] on assignment of blame (external or internal) during fault justifications [10] has motivated this current work on studying apology as a part of error mitigation strategy for voice assistants. Our hypothesis 2 predicts that serious apology, being more sincere, will lead to higher user satisfaction after service recovery and that the AI agent will be perceived as more intelligent and likeable. As predicted by Hypothesis 2, the agent that presented a serious apology, which showed higher sincerity, was preferred over the agent that apologized casually. This result is in line with the findings of previous research.

However, for blame assignment (taking itself or shifting to others), we see varied results in prior work based on the type of violation, either competence-based (mistake due to lack of knowledge) or integrity-based (intentional mistake) [27]. The results also varied on anthropomorphism[28] and automation [14]. Competence-based errors have shown to be mitigated more effectively when the party at fault apologized with an internal (accepting the blame), rather than external (assigning blame to other factors), attribution [27]. In this study, we assumed human-likeness to predict that the internalization of blame in the apology would have a more positive impact on the agent's perceived intelligence and likeability and that the user will be more satisfied with the response to the error (Hypothesis 3). Supporting Hypothesis 3, our results show that the AI agent, embodied as a static image representing a smart speaker that communicates via voice and text, was successful in achieving higher service recovery satisfaction when it clearly accepted the blame and apologized for it on the occurrence of competence-based error. The agent was perceived to be more intelligent and likeable when it acknowledged its mistake.

We observe that the agent that took the blame itself and apologized seriously (sincerely) was thought to be more likeable compared to other agents that offered either no apology, assigned blame to others, or apologized casually (less sincerely), and was perceived to be more intelligent compared to the agents that assigned blame to others or apologized less sincerely. Overall, the users were most satisfied with the agent who owned its mistake and apologized for it sincerely. The fact that the agent who apologized seriously but did not accept the blame and the agent who accepted the blame but did not apologize seriously were significantly perceived less intelligent and likeable, and that the agent that apologized casually while blaming others for the mistake was rated low on service recovery satisfaction shows that both these components, sincerity of apology and blame assignment, are important to get right when offering a good apology for mitigating negative impacts of failures. Our results highlight that the agent that offers a sincere apology while accepting the fault as its own would manage to form a good impression on the users despite the error.

5.2 No Apology is Better Than a Bad Apology

Our Hypothesis 1 states that the agent that offers any kind of apology should have a positive impact when it comes to mitigating AI errors. However, our results on willingness to use in the future presents contradictory evidence. The agent that shifted the blame to others while apologizing was rated significantly lower on willingness to use in future than our control condition (i.e., the agent that presented no apology on service failure). This result implies that a bad apology, lacking in internal attribution, has a negative impact on the continued use of technology. This finding highlights the importance of carefully designing effective apologies since a poorly designed apology may have a detrimental effect on long-term rapport with users.

5.3 Designing Effective Apologies for Conversational User Interfaces

Online shopping is one of the potential real-world applications of voice-based conversational agents. The uses of chat bots and conversational agents for customer service are increasing day-by-day. Studies on human-human interactions in customer service show benefits of apologies [50], even superfluous ones[5], in repairing the relationship with customers. Thus, an apology from AI in similar settings may benefit the user-agent relationship. This study provides evidence that a good apology from AI—a sincere apology with internal attribution—is an effective tool in mitigating the negative effects of AI errors during the recovery process in an online shopping scenario with smart speakers.

Further important questions that follow this study are: "when" should AI agents apologize and "what" for. For the sake of this study, we assumed that the agent apologized when the user indicated that an error had occurred, but the question remains whether AI can correctly identify situations that warrant an apology. In some cases, AI may be able to assess if an apology would be beneficial based on user feedback (e.g., the user indicates that an error has occurred). Apart from explicit user feedback, behavioral cues such as verbal and vocal cues [33] (e.g., cues indicating hesitation, negative words, curse words, or frustration in tone) may be used by AI to detect the need for apology. Similarly, an apology from AI may be beneficial if it fails to finish a task or is unsure of the outcome.

Moreover, comparing incorrect actions with correct ones after recovering from error can provide insights on "what" went wrong and if an apology is needed. For instance, intent recognition errors may be easier to identify and recover from based on users' feedback and behavior during the conversation; however, that requires a progression of the task. Various guidelines and techniques are being discussed in the community to enable a progression of task in case of conversation failures in Voice User Interfaces (VUIs) [12] (e.g., refining and reformulating responses [24, 45]). Google [23], Amazon [22], and IBM [40] have established guidelines for development of VUIs. These guidelines include suggestions on how to repair and recover from conversational breakdowns. Some include providing clear information about errors while maintaining consistency and context [22], paraphrasing and elaborating on the misunderstandings [40], and providing the reason and possible next steps in a helpful, honest, and transparent way when possible [23].

Whether an apology is needed at all would depend on the severity of the error as well. In case of low severity errors, a formal apology from conversational AI may not even be required (e.g., errors while asking a conversational agent for weather forecast, movie suggestions, or to play music). Whereas, for more severe errors (e.g., ordering the wrong item, not putting a reminder when asked, and fetching wrong information from the Internet), an apology from AI may prove to be beneficial.

Another important question is "how" to apologize for the mistake. Depending on the task, context, and error at hand, the type and format of the apology will vary. Blame attribution, seriousness, tone, intensity, and other aspects of an apology would depend on the characteristics of the agent (e.g., human-likeness, gender, voice, etc.), the task at hand, and on the error (severity, frequency, and type). For instance, based on the results of our study, a serious apology with acceptance of blame after successfully recovering from an intent recognition error in an online shopping task is preferred over an apology with avoidance of blame. However, prior work suggests that the efficient use of blame attribution in apology depends on the level of anthropomorphism. For instance, for machine-like agents, external attribution is better whereas for human-like agents, internal attribution is preferred [28]. Similarly, in human-robot interaction studies, there are conflicting results on internalization of blame by robots [19, 26]. Hence, further explorations on various human-AI interactions in real-world scenarios are needed to develop guidelines for designing an effective apology.

5.4 Limitations and Future Work

This study has some limitations despite its implications to design better apologies for repairing user-agent relationship after errors. First, we acknowledge that our study was a low fidelity simulation which might not be fully representative of how the actual interaction with an agent or smart speaker in an online shopping setting would be. The risks associated with ordering wrong items were diminished. To study the actual impact of proposed strategies, future research should be conducted in a more realistic setting.

Second, this study focused on short-term, immediate relationship repair during recovery from errors. The question of whether a sincere apology can be a precursor to long-term repair is an important one and remains unanswered. Moreover, this work focused on apology as a part of error mitigation and did not investigate other aspects of error handling and recovery [4, 16, 18]. It calls for further research on long-term recovery strategies for interactions that are more sporadic than confined to one session in one place.

Third, various aspects of AI agents' embodiment, behavior, and appearance that were not explored in this work may affect the perception of their apology. For instance, using gendered voices for AI may introduce biases towards agents because of gender stereotypes. Research on the preference of gender in synthesized voices suggests that both men and women favored female voices [38]. However, when tested for implicit biases, women still preferred female synthesized voices but men showed no gender bias [38]. In recent work, concerns about stereotypical assignment of a female gender to agent voices have been raised, and the power dynamic between the user and the woman-like voice assistants have been discussed [21]. Further research is needed to expand on the effects of gendered AI voice on users during recovery from failures.

Finally, our study was conducted in a low-risk, make-believe task with no time pressure or incentive other than compensation for the completion of the study. It is unclear whether these findings will be replicated in other contexts involving high-risk, real-world tasks where users are in a time crunch. It would be worthwhile to investigate recovery satisfaction and change in user-agent relationship dynamics in real-world contexts that vary in time sensitivity, error severity, and risk, such as decision making in healthcare, judicial, and financial systems.

Future researchers should explore how various aspects of an apology change users' perceptions before, during, and after an agent's attempt to recover from errors in human-AI interactions. We ought to design good apologies as a part of effective mitigation strategy for various types of system failures in real-world applications to ensure continued use of technology by maintaining good rapport with users.

6 CONCLUSION

Appropriate error mitigation and repair are important in maintaining a trusted relationship in human-AI interactions. Through our study, we demonstrate that agents that gave a serious apology and accepted its mistakes were perceived as more intelligent and likeable than if it shifted the blame externally or did not take the mistake seriously. Unlike a previous study where an even superfluous apology was still seen as effective in human-human interactions, our study found that agent that shifted the blame while apologizing could be favored less than if no apology was given at all. Our findings have implications for designing apology-based error mitigation strategies for voice assistants.

ACKNOWLEDGMENTS

We thank Dr. Qian Yang for contributing to the initial ideation of this project. This work was supported by the National Science Foundation award #1840088 and the Johns Hopkins University Institute for Assured Autonomy.

REFERENCES

[1] Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. *arXiv preprint arXiv:2106.02578* (2021).

[2] Karin Aijmer. 2019. 'Ooh whoops I'm sorry! Teenagers' use of English apology expressions. *Journal of Pragmatics* 142 (2019), 258–269.

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[4] Dennis Benner, Edona Elshan, Sofia Schöbel, and Andreas Janson. 2021. What do you mean? A Review on Recovery Strategies to Overcome Conversational Breakdowns of Conversational Agents. In *International Conference on Information Systems (ICIS)*.

[5] Alison Wood Brooks, Hengchen Dai, and Maurice E Schweitzer. 2014. I'm sorry about the rain! Superfluous apologies demonstrate empathic concern and increase trust. *Social Psychological and Personality Science* 5, 4 (2014), 467–474.

[6] Daniel J Brooks, Momotaz Begum, and Holly A Yanco. 2016. Analysis of reactions towards failures and recovery strategies for autonomous robots. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 487–492.

[7] David Cameron, Stevienna de Saille, Emily C Collins, Jonathan M Aitken, Hugo Cheung, Adriel Chua, Ee Jing Loh, and James Law. 2021. The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in Human Behavior* 114 (2021), 106561.

[8] James L Heskett Christopher W Hart and Jr W Earl Sasser. 1990. The profitable art of service recovery. (1990), 148–156.

[9] Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences. *England: Routledge* (1988).

[10] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo, and Ana Paiva. 2018. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*. 507–513.

[11] Ryan Fehr and Michele J Gelfand. 2010. When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. *Organizational behavior and human decision processes* 113, 1 (2010), 37–50.

[12] Joel E Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–8.

[13] John Fought. 1972. Erving Goffman, Relations in public: microstudies of the public order. New York: Basic Books, 1971. Pp. xvii 396. *Language in Society* 1, 2 (1972), 266–271. https://doi.org/10.1017/S0047404500000543

[14] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. 2021. Attributing blame to robots: I. The influence of robot autonomy. *Human factors* 63, 4 (2021), 592–602.

[15] Xiang Ge, Dan Li, Daisong Guan, Shihui Xu, Yanyan Sun, and Moli Zhou. 2019. Do smart speakers respond to their errors properly? A study on human-computer dialogue strategy. In *International Conference on Human-Computer Interaction*. Springer, 440–455.

[16] Petra Gieselmann. 2006. Comparing error-handling strategies in human-human and human-robot dialogues. In *Proc. 8th Conf. Nat. Language Process.(KONVENS). Konstanz, Germany*. 24–31.

[17] Trudy Govier and Wilhelm Verwoerd. 2002. The promise and pitfalls of apology. *Journal of social philosophy* 33, 1 (2002), 67–82.

[18] David Griol and José Manuel Molina. 2016. A framework for improving error detection and correction in spoken dialog systems. *Soft Computing* 20, 11 (2016), 4229–4241.

[19] Victoria Groom, Jimmy Chen, Theresa Johnson, F Arda Kara, and Clifford Nass. 2010. Critic, compatriot, or chump?: Responses to robot blame attribution. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 211–217.

[20] Yaou Hu, Hyounae Min, and Na Su. 2021. How Sincere is an Apology? Recovery Satisfaction in A Robot Service Failure Context. *Journal of Hospitality & Tourism Research* (2021), 10963480211011533.

[21] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It sounds like a woman: Exploring gender stereotypes in South Korean voice assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[22] Amazon Inc. [n.d.]. Alexa Design Guide. https://developer.amazon.com/en-US/docs/alexa/alexa-design/get-started.html. Accessed: 2021-12-01.

[23] Google Inc. [n.d.]. Conversation Design. https://developers.google.com/assistant/conversation-design/welcome. Accessed: 2021-12-01.

[24] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 143–152.

[25] Ing-Marie Jonsson, Clifford Nass, Jack Endo, Ben Reaves, Helen Harris, Janice Le Ta, Nicholas Chan, and Sean Knapp. 2004. Don't blame me I am only the Driver: Impact of Blame Attribution on Attitudes and Attention to Driving Task. In *CHI'04 extended abstracts on Human factors in computing systems*. 1219–1222.

[26] Poornima Kaniarasu and Aaron M Steinfeld. 2014. Effects of blame on trust in human robot interaction. In *The 23rd IEEE international symposium on robot and human interactive communication*. IEEE, 850–855.

[27] Peter H Kim, Kurt T Dirks, Cecily D Cooper, and Donald L Ferrin. 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 (2006), 49–65.

[28] Taenyun Kim and Hayeon Song. 2021. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics* 61 (2021), 101595.

[29] Dimosthenis Kontogiorgos, Sanne van Waveren, Olle Wallberg, Andre Pereira, Iolanda Leite, and Joakim Gustafson. 2020. Embodiment effects in interactions with failing robots. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[30] Anastasia Kuzminykh, Jenny Sun, Nivetha Govindaraju, Jeff Avery, and Edward Lank. 2020. Genie in the bottle: Anthropomorphized perceptions of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[31] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 203–210.

[32] Xuying Leo and Young Eun Huh. 2020. Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior* 113 (2020), 106520.

[33] Gina-Anne Levow. 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. 736–742.

[34] Roderick JA Little. 1988. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association* 83, 404 (1988), 1198–1202.

[35] James G Maxham III and Richard G Netemeyer. 2003. Firms reap what they sow: the effects of shared values and perceived organizational justice on customers' evaluations of complaint handling. *Journal of Marketing* 67, 1 (2003), 46–62.

[36] Steven J Migacz, Suiwen Zou, and James F Petrick. 2018. The "terminal" effects of service failure on airlines: Examining service recovery with justice theory. *Journal of Travel Research* 57, 1 (2018), 83–98.

[37] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[38] Wade J Mitchell, Chin-Chang Ho, Himalaya Patel, and Karl F MacDorman. 2011. Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior* 27, 1 (2011), 402–412.

[39] Flora Moon. [n.d.]. Sound of Text. https://soundoftext.com/. Accessed: 2021-08-01.

[40] Bob Moore and Raphael Arar. [n.d.]. Conversation design Guidelines. https://conversational-ux.mybluemix.net/design/conversational-ux/. Accessed: 2021-12-01.

[41] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.

[42] Cathy Pearl. 2016. *Designing voice user interfaces: Principles of conversational experiences*. " O'Reilly Media, Inc.".

[43] Valentina Pitardi and Hannah R Marriott. 2021. Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing* 38, 4 (2021), 626–642.

[44] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[45] Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2017. " Do Animals Have Accents?" Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 207–219.

[46] Marianne Promberger and Jonathan Baron. 2006. Do patients trust computers? *Journal of Behavioral Decision Making* 19, 5 (2006), 455–468.

[47] Holger Roschk and Susanne Kaiser. 2013. The nature of an apology: An experimental study on how to apologize after a service failure. *Marketing Letters* 24, 3 (2013), 293–309.

[48] Donald B Rubin. 2004. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

[49] Jagdish N Sheth and CH Kellstadt. 1994. A normative model of retaining customer satisfaction. *Gamma News Journal* July-August (1994), 4–7.

[50] Richard Spreng, Gilbert Harrell, and Robert Mackoy. 1995. Service Recovery: Impact on Satisfaction and Intentions. *Journal of Services Marketing* 9 (03 1995), 15–23. https://doi.org/10.1108/08876049510079853

[51] Stephen S Tax, Stephen W Brown, and Murali Chandrashekaran. 1998. Customer evaluations of service complaint experiences: implications for relationship marketing. *Journal of marketing* 62, 2 (1998), 60–76.

[52] Joost R van Ginkel, Marielle Linting, Ralph CA Rippe, and Anja van der Voort. 2020. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of personality assessment* 102, 3 (2020), 297–308.