My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction

ANDREA CUADRA, Cornell Tech, USA
SHURAN LI, Tongji University, China
HANSOL LEE, Cornell University, USA
JASON CHO, Cornell University, USA
WENDY JU, Cornell Tech, USA

27

One key technique people use in conversation and collaboration is conversational repair. Self-repair is the recognition and attempted correction of one's own mistakes. We investigate how the self-repair of errors by intelligent voice assistants affects user interaction. In a controlled human-participant study (N=101), participants asked Amazon Alexa to perform four tasks, and we manipulated whether Alexa would "make a mistake" understanding the participant (for example, playing heavy metal in response to a request for relaxing music) and whether Alexa would perform a correction (for example, stating, "You don't seem pleased. Did I get that wrong?") We measured the impact of self-repair on the participant's perception of the interaction in four conditions: correction *(mistakes made and repair performed)*, undercorrection *(mistakes made, no repair performed)*, overcorrection *(no mistakes made, but repair performed)*, and control *(no mistakes made, and no repair performed).* Subsequently, we conducted free-response interviews with each participant about their interactions. This study finds that self-repair greatly improves people's assessment of an intelligent voice assistant if a mistake has been made, but can degrade assessment if no correction is needed. However, we find that the positive impact of self-repair in the wake of an error outweighs the negative impact of overcorrection. In addition, participants who recently experienced an error saw increased value in self-repair as a feature, regardless of whether they experienced a repair themselves.

CCS Concepts: • **Human-centered computing**; • **Computing methodologies** → **Intelligent agents**;

Additional Key Words and Phrases: error-recognition; self-repair; intelligent voice assistants; conversational design

ACM Reference Format:
Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 27 (April 2021), 24 pages. https://doi.org/10.1145/3449101

1 INTRODUCTION

A worker in a meeting notices his boss's computer is low on power and asks her, "Do you want my charger?"

"I'm good."

The worker starts reaching into his bag to look for his charger, but stops after seeing the surprised expression on his boss's face.

Authors' addresses: Andrea Cuadra, Cornell Tech, USA; Shuran Li, Tongji University, China; Hansol Lee, Cornell University, USA; Jason Cho, Cornell University, USA; Wendy Ju, Cornell Tech, USA.

2573-0142/2021/4-ART27 $15.00
https://doi.org/10.1145/3449101

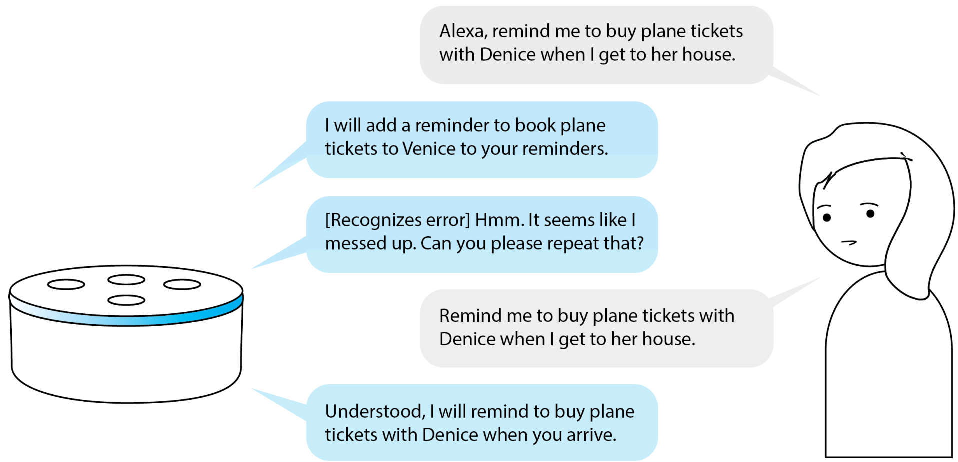

Fig. 1. Example of a scenario in our study in which the intelligent voice assistant (Alexa) successfully recognizes and repairs an error.

"Oh, you don't want my charger."

"No."

"No problem!"

In this vignette, the term "I'm good," which means, "don't worry about it," was misunderstood to be a positive response to the worker's offer, but the worker repairs the misunderstanding after first responding incorrectly when he sees his boss's unexpected reaction. In human face-to-face interaction, people monitor each other continuously to see if they are understanding and being understood by others, and they stop and self-correct if they recognize that they have made an error. This capacity for "self-repair" helps to ease the irritation and friction that comes from having to explicitly correct mistakes or misinterpretations, or from suffering the consequences of uncorrected miscommunications. Self-repair in human interaction with machines, such as interactive voice assistants or robots, is not yet common, but is an area of interest; computer scientists and roboticists are working on applying machine learning to people's verbal and non-verbal behaviors to catch communication errors so that self-repair can occur [8, 57]. Self-repair may be crucial for making conversational agents useful in a computer-supported cooperative work environment. However, it is not yet known how confident a machine should be that it has committed an error before it attempts to perform self-repair. We might expect that a machine that can correct its own mistakes is perceived to be more capable than one that cannot; but what is the benefit, in terms of user assessment, of correcting mistakes when they are made, compared to not correcting at all? And what is the cost of a machine performing self-correction when it has *not* made a mistake? All these factors need to be weighed in the machine's decision analysis of whether to attempt self-repair actions.

In this work, we investigate the effects of intelligent voice assistants performing self-repair in the presence and in the absence of mistakes (see Figure 1). We measure the impact of self-repair on the participant, and on the participant's perception of the intelligent voice assistant in four conditions (see Figure 2): control (no mistakes made, and no repair performed); undercorrection (mistakes made, but no repair performed); overcorrection (no mistakes made, but repair performed); and correction (mistakes made, and repair performed). We then measure the desirability of self-repair. We further provide qualitative findings based on what participants said about their interactions, and identify the factors that help us determine the main trade-offs between the different conditions. To the best of our knowledge, this is the first study that has been conducted to measure the effects of intelligent voice assistants performing self-repair. These findings, as a whole, can inform design guidelines for using self-repair as an interaction mechanic in human-machine interaction.

1.1 Related Work

Prior work on conversational repair spans a large array of disciplines, from linguistics to human robot interaction. In this related work section, we consolidate disparate threads of research in these different communities to form a more coherent picture of the prior work on self-repair in interaction.

1.1.1 Conversational agents in cooperative work. By studying how the design of conversational agents affects human behavior and perception, the CSCW community can realize the full potential of such agents in cooperative work. For example, Williams et al. studied the use of a chatbot to help participants with work detachment and reattachment processes, and found that productivity gains were better sustained when the conversations used emotions-centric prompts *"how do you feel"* instead of task-centric ones *"what did you do"* [68]. Xiao et al. evaluated responses to open-ended questions when administered via a chatbot and via an online survey, and found that the chatbot outperformed the online survey in driving higher level of participant engagement and eliciting significantly better quality responses [69]. When studying a robot's potential to shape trust within a collaborative environment with robots and humans, Strohkorb Sebo et al. found that robots that express vulnerability can have "ripple effects" on their human team members' expressions of trust-related behavior, making human teammates more likely to explain their failure to the group, console team members who had made mistakes, and laugh together [64]. Similarly, Traeger et al. found that people in groups with a robot making vulnerable statements converse substantially more with each other, distribute their conversation somewhat more equally, and perceive their groups more positively compared to control groups with a robot that either makes neutral statements or no statements [66]. However, social signaling behavior does not come without trade-offs. For example, Jung et al. found that even though robots that used backchanneling improved team functioning, the agents themselves were perceived as less effective than those that did not [32]. Additionally, Ashktorab et al., who explored different repair strategies in chatbots, discuss repair as a collaborative action with costs (e.g., too much turn-taking or loss of naturalness), calling for further empirical research in the area [5].

1.1.2 Speech-enabled devices. Unlike some of the first-generation voice agent research published at CHI, such as [12, 49, 65] where the computer-generated voice that people were speaking with was disembodied, or the embodied conversational agents of Cassell, Sullivan, Churchill and Prevost [15], which were front-ended by on-screen virtual agents, today's intelligent voice assistants are embodied in standalone devices such as Amazon's Echo [1], Apple's Homepod [4], or Google's Home [27].

There are indications that the category of speech-enabled devices—which have physical presence and two-way voice interaction but limited human-likeness—will grow. While voice-enabled robots like the Jibo social robot [23] or Anki's Vector [3] have had limited commercial success to date, it still seems likely that future robots and appliances will feature speech interaction as a feature. Part of this trend is driven by advances in natural language processing, text-to-speech and dialog generation systems driven by big data, as well as hardware breakthroughs in far-field microphone arrays. While improvements to the hardware and software of these speech-enabled devices might improve the recognition of individual words people say, common-sense intelligence is not yet in grasp [20]. The limited capabilities of today's speech systems would be well complemented by interaction savvy that would help these systems recognize and recover from conversational errors.

1.1.3 Conversational repair. Intelligent voice assistants may make mistakes, but human dialog is far from error-free itself. A key difference is that people perform repair in communication [56], monitoring listeners to see if they have been heard and understood before moving forward in the conversation.

Linguists Schegloff, Jefferson and Sacks define repair to be the practices that interactants use to handle troubles in hearing, speaking and understanding that occur regularly in social interaction. They noted that a repair sequence has several key segments: the repairable, repair initiation, and a repair outcome. This formulation took into account that sometime repairs were initiated where no error occurred–even a correct statement could be repairable. The repair initiation could come from the speaker (self-initiation) or the listener (other-initiation). The outcome–what was suggested in place of the repairable–could similarly be correct or incorrect, and accomplished by the self or other [56]. This team later found from analyzing naturalistic conversation that people had a preference for self-correction over being corrected by others: in moments when repair was necessary or possible, the distribution of repairs was strongly skewed towards self-repair [62]. Very often self-repair occurs when the speaker notices a mistake, in the transition space between speaking turns, before the listener even has a chance to respond, but they noted that when the listener initiated a repair, the original speaker usually responded by self-repairing before the other performed a correction. These patterns are also noted by Moore and Arar in the introduction to their book surrounding conversational user experience design [43].

Whereas Schegloff and his colleagues focus on linguistic repair [59–61], Clark and Schaefer's contribution model addressed the detection and repair of communication model through a more regulatory model. In this model, conversation contains contributions with a presentation and a subsequent acceptance. In other words, the speaker is actively seeking evidence that they are being understood, and is as likely to initiate repair when evidence of understanding is insufficient as when they have firm evidence that they were not understood [17].

Often, the evidence of understanding is not verbal. Ekman and Friesen have drawn attention to the nonverbal acts people perform to maintain and regulate the back-and-forth nature of speaking and listening [22]. *Regulator* actions, which indicate that people are listening, understand, or if they are confused, take exception, or want to respond, occur in the attentional periphery; people perform them without thought, but can recall and repeat them if asked. In collaborative conversations [17], addressees must therefore also indicate their understanding, or lack of understanding, to help the speaker understand the state of the communication. Chovil's experiment with people listening to a story in a face-to-face, partition, and telephone and answering machine condition showed that listeners primarily react facially when they would be seen by the storyteller [16]. Hence, the monitoring for acceptance and understanding of speech in face to face interaction is often multi-modal [11].

1.1.4 Error-recognition. The advancement of error-recognition technology is an imperative part of improving human-computer interaction via self-repair. The importance of identifying and incorporating responses to conversational signals was recognized early in the human-computer interaction community by Nagao and Takeuchi [47]. Elements such as empathy and the emotions associated with certain utterances have also been studied and play an important role in error recognition [36, 51]. Bousmalis et al. have surveyed the conversation analysis literature for nonverbal audiovisual cues that indicate agreement and disagreement between human speakers, with the goal of developing machine recognition of these cues [8]. More recently, Salazar-Gomez et al. experimented with using EEG-based feedback methods to correct robot mistakes in real time; because the EEG signals were analyzed in realtime in closed-loop fashion, the robot was able to respond to possible signs of error by hyper-articulating actions to elicit stronger response to help it determine if it was making a mistake [57]. This current research is premised on the capability of error recognition to occur, whether through physiological measures, visual or audio recognition or through discourse analysis. However, in this paper's study, because of the need to control participant experiences by condition, error recognition is simulated rather than actually performed.

There are several varieties of cues that can aid in performing error-recognition:

Discourse cues: Gieselmann ran a small experiment to look at what error recovery strategies people use when talking to robots compared to when they talk to other people. Geiselmann found that achievement strategies (such as paraphrasing, repeating, or restructuring) and functional reduction strategies (such as giving pre-selected answers or changing the theme) were used, largely due to the limited interaction capabilities of robot, and that the most common indicator that an error was made is a sudden change in the dialogue topic. In this research, the focus of the error detection lay in analysis of the discourse [25].

Audio cues: There are signals in human speech that can be used to recognize error. Oviatt et al. found, for example, that people tend to hyperarticulate when talking to machines, often making it harder for the machine to recognize what the person is trying to say [50]. They proposed a two-stage Computer-elicited Hyperarticulate Adaptation Model to account for this repair mechanism that people use. Levow analyzed acoustic-prosodic features, like duration of speaking, pauses, and changes in volume and pitch, to predict when people were responding to machine misunderstandings [33]. Litman et al. used the machine learning program RIPPER to produce a classification model that improved the prediction of misrecognitions using these types of acoustic-prosodic features on the TOOT corpus, a spoken dialogue system for accessing train schedules via telephone [35].

Visual cues: The improvement in error-recognition technology via visual cues can be foreshadowed by the widespread availability of emotional expression image databases such as Ekman's Pictures of Facial Affect [21], the Belfast database [19], the Extended Cohn-Kanade Dataset [37], or the Affectiva-MIT Facial Expression Dataset [38]. Because error-recognition and self-repair often go hand in hand, it is crucial that we also research and understand the scope of possibilities and trade-offs of repair as a function of a computationally determined decisions, such as whether repair is needed or not.

Based on this active research in the space of error recognition, we believe the possibility of self-repair is very much on the horizon. However, the mere recognition of error does not actually indicate when and how repair should occur.

1.1.5 Errors and repair in social interactions with conversational interfaces. Brennan points out that conversation is shaped by visual and spoken evidence [11], but much of the early research done in conversational interfaces was done for phone interaction, where only auditory evidence is available.

Repairs in spoken dialogue interfaces are often subroutines that are called when the voice agent does not hear a response to a question, or when the response does not fit anything in its limited response vocabulary [39]. Dan Bohus classifies these as non-understandings (when the system does not acquire useful information from the user's turn) and mis-understandings (when the information gather by the system from the user is incorrect) [7].

"I'm sorry, I didn't hear that, let's try again..." is a refrain many of us have heard in phone interfaces. These responses are usually generated by dialogue management systems that repair the interaction breakdowns that occur when the system fails to understand the person [70]. These systems do not repair breakdowns that occur when the person fails to understand the system, except that people frequently respond to such situations by not speaking at all. Repair routines that re-iterate or re-word the original query can help get the interaction back on track, but can still be problematic if the line of inquiry or dialogue is incorrect. Rudzicz et al. found, for example, that older individuals with Alzheimer's Disease are often confused by speech interaction, and respond 40% of the time by not responding at all [55].

Corti and Gillespie found that people are less likely to initiate repairs with agents that are disembodied or that are not represented by human [18]. They posit that this is due to intersubjectivity, which requires each party to think the other party knows what their point of view is [26]. They argue that when a person does not see an anthropomorphic agent, the person does not initiate repairs because that person does not perceive the agent can observe or understand their repair initiation activities [26]. In another study, Candello and Pinhanez explored the use of multiple chatbots, each bot having expertise in a specific area, to repair dialogue failures (for example, by readdressing a question), and found that multiple agents expands the opportunities and strategies for handling errors [14]. Such strategies can also help set the norms, by showing that each bot does not know what the other bots point of view is.

1.1.6 Errors and repair in social interactions with robots. The human-robot interaction community is perhaps highly motivated to understand how to perform repair, because current-day robots fall so far short of executing understanding and physical tasks correctly. Building off of Reeves and Nass' "Computers as Social Actors" hypothesis [48], the human-robot interaction community hopes that sophistication in social interaction can help to compensate for short-comings elsewhere.

We might assume that people would prefer robots that behaved perfectly and never made mistakes, but several experimental studies indicate that this is not true. Ragni, et al. found that people collaborating with robots on a memorization task were more likely to report positive emotions with an erroneous robot than the perfect one [53]. Mirnig, et al. studied people who got instructions from a Nao robot on a Lego building tasks; participants liked the faulty robot significantly better than the robot that interacted flawlessly, even though the faulty robot degraded their own performance [42]. Salem, et al. looked at how errors in communication in particular, with a robot that did not gesture while speaking, gestured congruently while speaking, or gestured incongruently while speaking, and found surprisingly that participants liked the robot that gestured incongruently the most [58].

The human-robot interaction community has also spent a lot of time looking at recognizing human social signals in interaction. Breazeal and Rani for instance, provide good recaps of the work in the HRI community on affect recognition [10, 54]. The HRI community has also focused on recognizing the embodied signals of human interactants for conversational *regulation* Fujie et al. made a robot that recognized head motions, like nodding, for paralinguistic information that clarifies speaker intent [24]. Sidner et al. found that participants who knew their robots recognized conversational head nods would nod more [63]. Huang and Mutlu have proposed developing a Robot Behavior Toolkit that uses many of the same social cues that people use to achieve interaction goals in order to make robots that are able to adapt their behaviors to people [30]. Mutlu et al. found that people remembered stories that story-telling robots told better if the robot looked at them more, that they could get listeners to behave as addressees or bystanders by having the robots look at them as if they were addressees or bystanders, that they could encourage turn-taking by having the robot change who it looked at [2, 29, 44–46, 52]. These studies, considered broadly, indicate that expectations for interaction and communication–down to the timing and the gaze patterns–persist when people are speaking to a robot or machine.

From this research, we believe that self-repair in response to non-verbal cues would be appreciated by people and make them rate their interactions and their fellow interactant more highly than those where mistakes were allowed to persist without repair by the acting agent.

1.2 Research Questions

Although much research has been conducted to understand how humans respond to error, there is little analogous research to understand how people respond to self-correction by computer- or machine-based agents. Given that self-repair may mitigate the downsides of mistakes [13, 17, 48, 62], that repair is a normal part of human conversation [13, 16, 17, 22, 56, 59–62], and that people tend to prefer robots that are not perfect [42, 53, 58], we hypothesize that:

Hypothesis 1. Participants will rate an agent that successfully repairs its mistakes *(correction)* better than an agent that makes no mistakes *(control)*, or repairs nonexistent mistakes *(overcorrection)*, or makes mistakes and does not repair them *(undercorrection)*.

Hypothesis 2. The desirability for self-repair capability will be higher for participants in the conditions where mistakes are made *(correction and undercorrection)*.

We base the second hypothesis on the idea that participants in these conditions would have more proximal experience with the frustration and degraded performance associated with mistakes.

2 METHOD

2.1 Overview

A laboratory between-participants experiment was conducted, using a 2 (presence of mistake: no mistakes made vs. mistakes made) × 2 (presence of repair: no repair performed vs. repair performed). We chose a between subjects experimental design to minimize learning effects across conditions, increase the number of total participants by having shorter sessions, and be able to more effectively determine which factors played a bigger role in participants' perceptions. The conditions are labeled for readability as depicted in Figure 2: control, undercorrection, overcorrection, and correction. Participants are semi-randomly assigned to each condition: control (N=22), undercorrection (N=30), overcorrection (N=30), and correction (N=19). Upon entering the room, all participants are given the same research scenario: they're about to start driving to a friend's house, they are going to use the Amazon Alexa to help them accomplish a few tasks, and to imagine that Alexa has the ability to see their reactions through Camera 2 (see Figure 3). There are four main tasks per condition: 1) "Can you take me to Denice Johnson's house?", 2) "Send a message to Denice saying, 'I will be there in ten minutes."', 3) "Remind me to buy plane tickets with Denice when I get to her house.", and 4) "Play relaxing music." The full guiding scripts are available in the Appendix (Section 7.1). Subsequently, participants are asked to interact with the Amazon Alexa by reading prompts displayed on a screen, including clarifications when Alexa makes unnecessary repairs or mistakes, see Figure 2. For the purposes of the study, we refer to error-recognition as the act of identifying that a mistake has been made. We define mistakes as errors made by the voice agent in interpreting and responding to a verbal prompt; for example, hearing "salmon" instead of "seven." In this paper, we are not categorizing unnecessary self-repair as a mistake. Self-repair is used to refer to the voice agent taking some action to correct a mistake, such as asking what it did wrong and adjusting the response accordingly.

The physical set up for the study is shown in Figure 3. Two cameras are placed in front of participants to record their interactions with the Amazon Alexa Echo Dot, one from the front angle, and one from the perspective of the Amazon Alexa Echo Dot. A researcher sits next to the participants to flip through the prompts. Captured video is saved for a future dataset, which is not

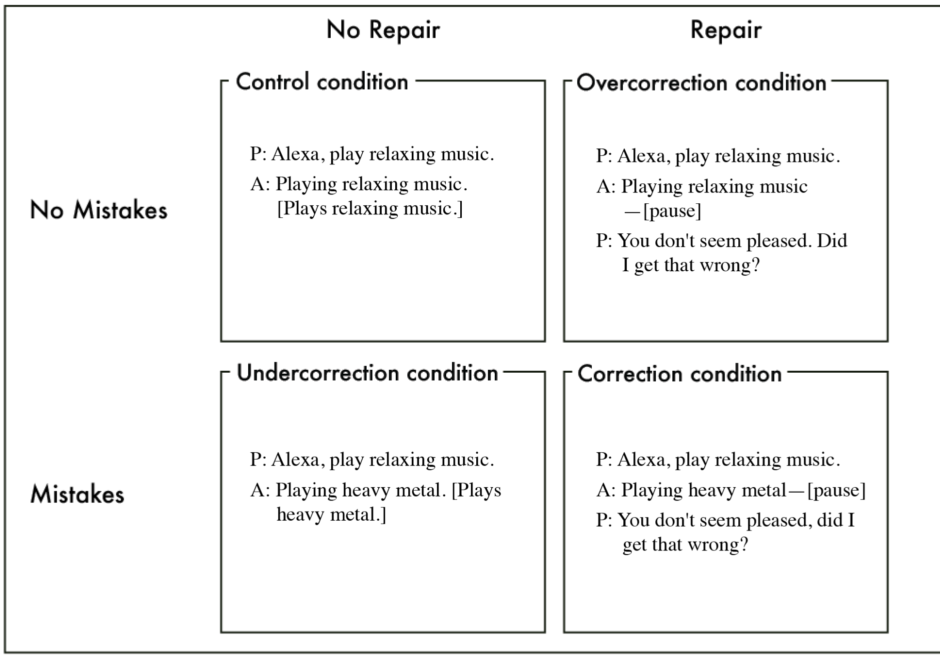

Fig. 2. Examples of interactions in each condition in the 2 (presence of mistake: no mistakes made vs. mistakes made) × 2 (presence of repair: no repair performed vs. repair performed) matrix.

a part of this publication. After completing their interaction, participants are asked to fill out a questionnaire, and participate in a short, semi-structured interview.

2.2 Participants

A total of N=101 university students (F:73, M:28) between the ages of 18 and 30 participated in our study, N=20 participants spoke English as a second language. All participants were considered when evaluating *Hypothesis 1*, regarding which condition is better, and only the N=78 participants that passed the manipulation checks (which measured whether they agreed with our definitions of mistakes and repair) were considered when evaluating *Hypothesis 2*, relating to the desirability of repair. A high portion of the participants in the undercorrection and overcorrection conditions failed our manipulation checks, so we increased our quotas for participants in those conditions to be able to successfully compare *Hypothesis 2* results.

2.3 System

Amazon Alexa Echo Dot (3rd Gen) device was used with the default Alexa voice (female, American-accent), and the software to operationalize each condition was authored using Jovo 2.2.12 and Node.js 8.10. Video was recorded at 29 frames/second with a resolution of 1920 x 1080 DPI. The interviews were transcribed from the video recordings by a third party service.

2.4 Measures

For the quantitative analysis, the 28 items in the questionnaire were divided into participant feelings and perceptions (the first 22 items, see Section 7.2 in the Appendix), and self-repair desirability (the last 6 items, with the first 2 of those being the manipulation checks, see Section 7.2 in the

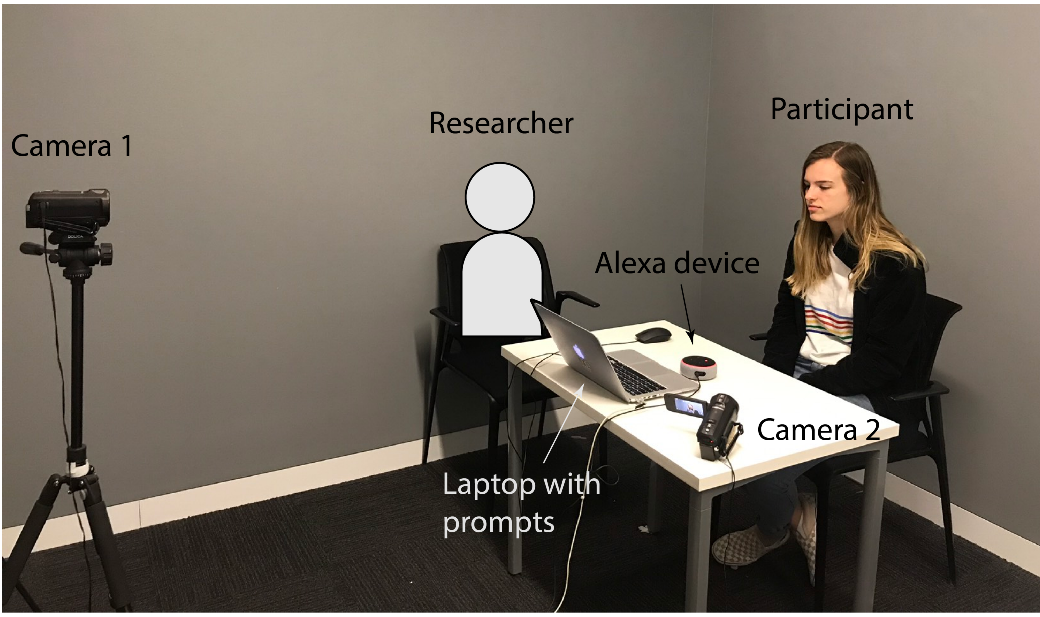

Fig. 3. Image demonstrating the setup of a study session. Participant sits in front of an Alexa device, two cameras, and a computer. A researcher sits next to the participant to flip through the prompts on the computer.

Appendix). For the qualitative measures, we analyzed open-ended feedback following the initial questionnaire. Below, we describe each of these measures in detail.

2.4.1 Participant Feelings and Perceptions. The post-interaction questionnaire asked participants how strongly they agreed with statements regarding how the Amazon Alexa made participants feel (ex: "This voice agent made me feel successful."), how the participants perceived the Amazon Alexa to be (ex: "This voice agent was likeable."), and participant's perceptions of Amazon Alexa's personality using questions from Gosling's Ten Item Personality Measure (ex: "This voice agent was anxious, easily upset.") [28]. The responses were scored on a seven-point Likert scale from "disagree strongly" to "agree strongly". All questionnaire items are included in the Appendix (Section 7.2).

The two most relevant components were identified using principal component analysis, and later the factors at either end of those components were analyzed and aggregated. The components were labeled based on the values of the items that were most influential to the specific component. The aggregated data was then plotted and visually examined, and analyzed using two-way between subjects analyses of variance (ANOVAs), and subsequently performing post hoc comparisons using the two-sided Tukey HSD test.

2.4.2 Self-Repair Desirability. To ensure that participants agreed with our definition of mistakes and self-repair, we included two manipulation-check questions: "Did this voice agent ever make a mistake?", and "Did this voice agent ever try to repair a mistake it made?". Next, to measure the desirability of the self-repair feature, we asked hypothetical questions about intelligent voice assistants that perform correction (ex: "Rate the following statement: a voice agent that always tried to correct itself after a mistake... ...would annoy me. ...would waste my time. ...would help me feel less frustrated. ...would improve the conversation quality.") Response options ranged from from 1 (disagree strongly) to 10 (agree strongly).

We included the data from all N=101 participants in the *Hypothesis 1* analysis, and only the data from the N=78 participants who passed the manipulation checks in the *Hypothesis 2* analysis. Numerous participants failed our manipulation checks in two conditions: in the undercorrection condition (N=30), 9 participants thought Alexa had tried to repair its mistakes, when in fact there was no correction; and in the overcorrection condition (N=30), 8 participants responded that either Alexa made mistakes and repaired them, and 5 reported Alexa did not make mistakes and did not repair them, when in fact there was no error but there was an attempted correction. Only 1 participant failed the manipulation checks in the other conditions.

2.4.3 Post-interaction Qualitative Reactions. Upon completion of the questionnaire, we interviewed each participant, seeking open-ended feedback in three areas: the overall experience with the agent, reactions to error-recognition, and reactions to self-repair. The questions were phrased differently depending on the participant's assigned experimental condition. For example, for the control condition, we asked, "If Alexa were to make a mistake, what would think about Alexa being able to recognize her own error, maybe from your facial expressions, voice, etc.?" For the correction condition, we asked, "What do you think about Alexa trying to recognize her own error, maybe from your facial expressions, voice, etc.?". This interview allowed participants to directly comment on the intelligent voice assistant they interacted with.

3 RESULTS

In the following sections, we describe our results in the same order that the questions were presented to participants. First we present findings on participants feelings and perceptions which disprove *Hypothesis 1.* Next we present findings for self-repair desirability, which support *Hypothesis 2.* The statistics are reported at the 99% confidence level. Lastly, we provide findings from our qualitative analysis.

3.1 Participant Feelings and Perceptions

We used principal component analysis to determine which out of the 22 factors measured (see first 22 items in Section 7.2 in the Appendix) were having the biggest impact on participants' feelings and perceptions. Based on the approximate percent of variation per principal direction, we decided to move forward with the two components accounting for the most variance in the data, the "feel successful" component (accounting for 46.2% of the variation) and the "this voice agent was calm component (account for 9.5% of the variation). The rest of the components, taken individually, only accounted for 5% of the variation or less.

For the component accounting for the most variance in the data (the "feel successful" component), the factors most heavily affecting the data in the positive direction were "this voice agent made me feel successful," factor loading of .34, and "this voice agent made me feel efficient," factor loading of .34. And in the negative direction was "this voice agent made me feel frustrated," factor loading of -.28.

For the component accounting for the second most variance in the data (the "this voice agent was calm" component), the factors most heavily affecting the data in the positive direction were "this voice agent was calm, emotionally stable," factor loading of .34, and "this voice agent was extroverted, enthusiastic," factor loading of .13. And in the negative direction was "this voice agent was anxious, easily upset," factor loading of -.70.

To further examine what was going on in each principal component, we took the average of the responses to the three factors most heavily affecting the data per component per participant, reversing the ones in the negative direction.

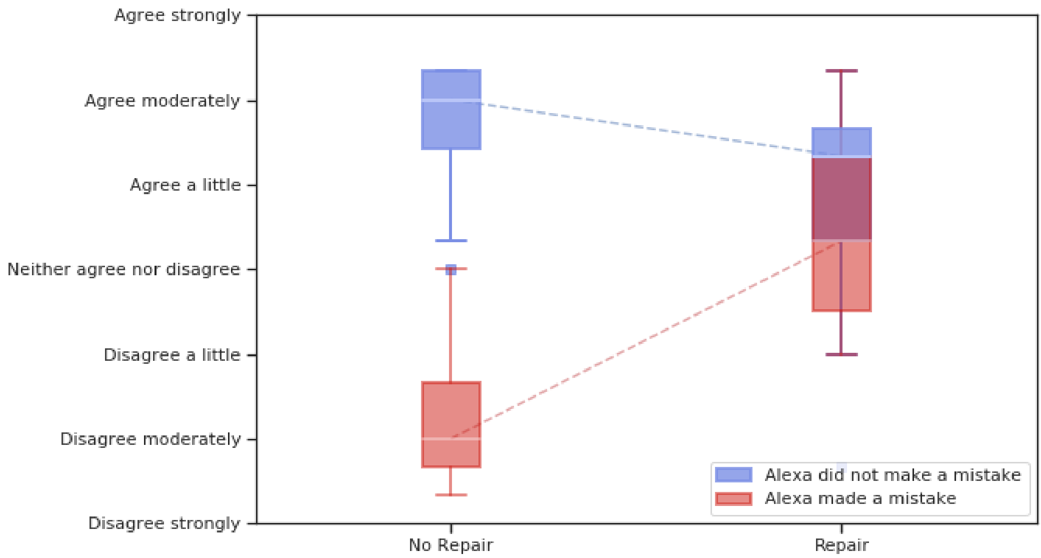

Fig. 4. Box plot of "feel successful" aggregated data. The data for the conditions without repair is shown on the left, and the data for conditions with repair is shown on the right. The data for the conditions without mistakes is blue, and the data for conditions with mistakes is red. Overall, this box plot depicts an improvement in interaction in the presence of repair. Without repair, the median of the aggregated responses to questions correlated to making the participant feel successful in situations where Alexa made a mistake (undercorrection) was approximately "disagree moderately". With repair, regardless of whether repair was needed (correction) or not needed (overcorrection), neither median was below "neither agree nor disagree".

3.1.1 How successful the voice agent made participants feel. Figure 4 shows the plotted data for the average of the responses to the questions "this voice agent made me feel successful," "this voice agent made me feel efficient," and reversed answers to the question "this voice agent made me feel frustrated."

This figure shows a predictable interaction effect: the repair action lowers the rating of the interaction quality if no mistake was present, but dramatically improves the perceived quality if there was mistake. Without repair, the median of the aggregated responses to questions correlated to making the participant feel successful in situations where Alexa made a mistake (undercorrection) was approximately "disagree moderately". With repair, regardless of whether repair was needed (correction) or not needed (overcorrection), neither median was below "neither agree nor disagree", demonstrating an improvement in interaction in the presence of repair.

A two-way ANOVA was conducted to determine the degree to which the two independent categorical variables, mistake and self-repair, respectively explain the observed variance in how successful participants felt. We found that the mistake variable [$F(1, 97) = 123.91$, $p < .001$] and the self-repair variable [$F(1, 97) = 10.20$, $p = .002$] both had significant effects on how successful participants felt, and that interaction between the two was also significant [$F(1, 97) = 63.60$, $p < .001$]. This indicates that whether or not a mistake occurred, and whether or not self-repair was attempted, both measurably affected participants' feelings of success with Alexa; and moreover, that whether or not a mistake had been made significantly influenced the effect that self-repair (or the lack thereof) had on the feeling of success.

Post hoc pairwise comparisons between the four experimental conditions were then made using the two-sided Tukey HSD test, to determine how the conditions differed from one another with respect to participants feeling successful. A significant positive difference was observed between the control and undercorrection conditions (Hedges's effect size = 3.747, p = .001) indicating that participants in the control condition (no mistake, no repair) group felt much more successful than those in the undercorrection condition (mistake, no repair) group. Significant negative differences were likewise observed between the undercorrection and correction conditions (Hedges's effect size = -2.299, p = .001), as well as between the undercorrection and overcorrection conditions (Hedges's effect size = -2.866, p = .001). Experiencing undercorrection (a mistake without repair) therefore made participants feel notably less successful than either correction (a mistake with subsequent repair) or overcorrection (repair when no mistake had been made). A smaller significant positive effect size was observed between the control and correction conditions (Hedges's effect size = 1.440, p = .001), indicating that participants felt somewhat less successful after correction (when a mistake was repaired), relative to the control (when neither a mistake nor repair had taken place). Comparing the control condition to the overcorrection condition (Hedges's effect size = .887, p = .008), we found that participants felt slightly less successful with overcorrection (when the agent performed unnecessary repair) relative to the control. Interestingly, no significant difference was observed between the overcorrection and correction conditions (Hedges's effect size = .559, p = .21), indicating that unnecessary repair did not make participants feel demonstrably less successful than necessary repair. On the whole, participants felt more successful when the voice agent performed self-repair (Hedges's effect size = -.629, p = .002), and less successful when it made mistakes (Hedges's effect size = 1.798, p = .001).

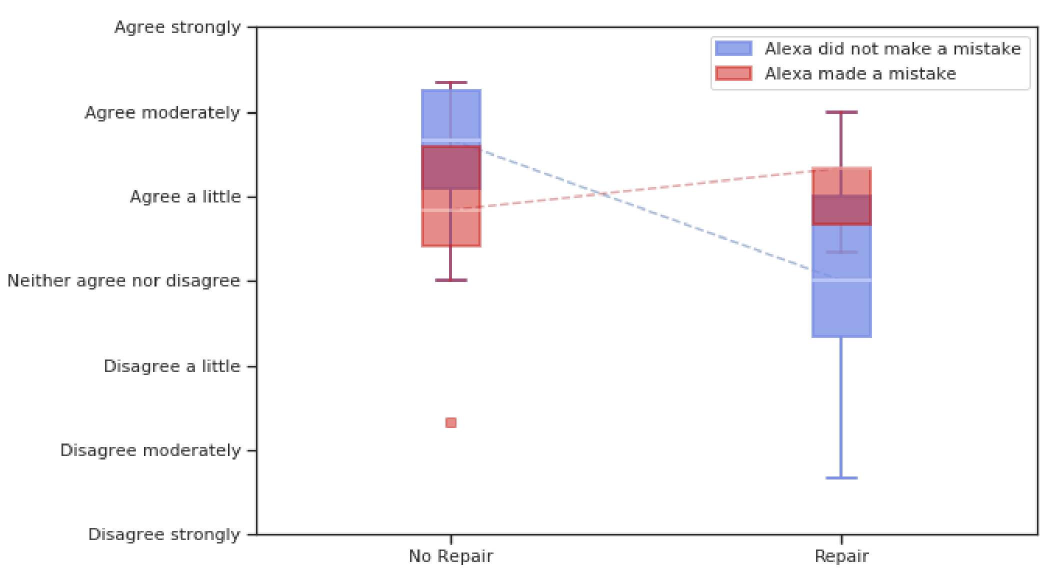

Fig. 5. Box plot of "this voice agent was calm" aggregated data. The data for the conditions without repair is shown on the left, and the data for conditions with repair is shown on the right. The data for the conditions without mistakes is blue, and the data for conditions with mistakes is red. This box plot illustrates how participants generally considered the voice agent to be calm and emotionally stable, except in the overcorrection condition, where participants perceived the voice agent as being anxious.

3.1.2 How calm the voice agent was perceived to be. Figure 5 shows the plotted data for the average of the responses to the questions "this voice agent was calm, emotionally stable," "this voice agent was extroverted, enthusiastic," and the reversed question "this voice agent was anxious, easily upset." This figure also shows a predictable interaction effect, but a more pronounced one: the overcorrecting agent is perceived far more negatively than than in the control condition, whereas the undercorrecting agent was only rated a little lower than the correction condition agent. Participants generally considered the voice agent to be calm and emotionally stable, except in the overcorrection condition, where participants perceived Alexa as being anxious.

A two-way ANOVA was performed to determine the degree to which the variability of mistake and repair between the conditions contributed to participants' perceptions of how calm the agent was during the experiment. We found that the repair variable had a significant correlation [F(1, 97) = 12.63, $p < .001$] to participants' perception of Alexa's calmness. There was also significant interaction between the mistake and repair variables (F(1, 97) = 15.28, $p < .001$), indicating that whether or not a mistake had been made influenced how self-repair affected participants' perception of Alexa's calmness. However, whether or not a mistake had occurred ($p < .53$) was not on its own a significant factor in the perception of how calm the agent was.

Post hoc pairwise comparisons using the Tukey HSD test revealed that only the overcorrection condition differed significantly from the others with regard to the perception of how calm the voice agent had been. Compared to participants in the control condition group, those in the overcorrection condition group perceived the agent to be significantly less calm (Hedges's effect size = 1.561, p = .001). Likewise, participants in the overcorrection condition group perceived the agent to be less calm than those in the correction group (Hedges's effect size = 1.055, p = .002), or those in the undercorrection group (Hedges's effect size = .838, p = .006). Overall, when the voice agent did not perform any repair, it was perceived as more calm (Hedges's effect size = .679, p = .001).

The scatter plot from the principal component analysis (see Figure 6) provides an overview of the results at a higher level. The plot serves to illustrate how participants generally felt successful, except in the undercorrection condition, where mistakes were made and not repaired. Similarly, the plot shows that participants in the overcorrection condition rated the voice agent as more anxious.

These findings disprove *Hypothesis 1*, as participants preferred a "perfect" voice agent over one that made mistakes and successfully corrected them.

3.2 Self-Repair Desirability

Participants in the correction and undercorrection conditions felt more positive about having a voice agent that always tried to correct itself after a mistake. A two-sample equal variance t-test for participants in the conditions with mistakes (correction and undercorrection) and in the conditions without mistakes (overcorrection and control) showed significant differences in participants' opinions about self-repair helping them feel less frustrated (p = .002), improving conversation quality ($p < .001$), and not annoying them (p = .01); where responses leaned in favor of self-repair for the "...would annoy me", "...would help me feel less frustrated", and "...would improve the conversation quality" measures. There were no significant differences in the "would waste my time" measure. These findings support *Hypothesis 2*.

3.3 Qualitative Reactions

For our analysis of the qualitative reactions, each semi-structured interviewed was transcribed by a third party. We then iteratively reviewed the transcriptions to cluster similar responses using Braun's thematic analysis strategies, [9] and meaning making techniques described by Miles [40]. The first author developed the initial coding which was then carried out and refined by the second author, and subsequently further refined and reviewed by the first and last authors.

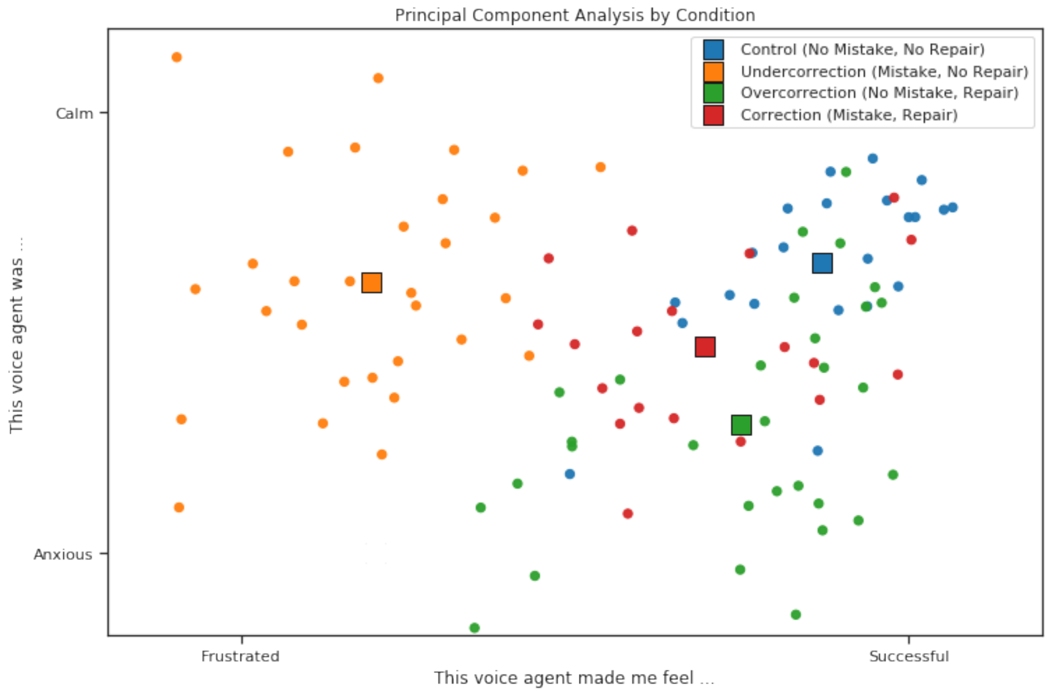

Fig. 6. In this scatter plot, each dot represents a participant assigned to a specific condition. The large squares are the average of all the dots in each condition. The first component, accounting for 46.2% of the variance, is on the horizontal axis, and principally represents questions related to how the participants felt (ranging from frustrated to successful). The second component, accounting for 9.5% of the variance, is on the vertical axis, and principally represents how participants perceived the voice agent (ranging from anxious to calm). The plot serves to illustrate how participants generally felt successful, except in the undercorrection condition, where mistakes were made and not repaired. Similarly, the plot shows that participants in the overcorrection condition rated the voice agent as more anxious.

We reached consensus around six main themes. Half of the themes help clarify the quantitative findings (efficiency or speed, frustration or annoyance, and helpfulness), and the other half illuminate topics that we had not considered relevant to this specific study prior to running the experiment (improvement in understanding, empathy, and creepiness):

(1) **Efficiency or speed:** Participants frequently commented on the speed and efficiency of the interactions with the voice agent. We commonly heard, particularly in the control and correction conditions, participants comment on the voice agent being "efficient" or "a lot faster." For the overcorrection condition, we also heard comments such as "it would add extra work."
(2) **Improvement in understanding:** Participants thought that a voice agent that was able to detect error from facial expressions would improve understanding, because it would remove the need of having to verbally explain what is going on. For example, one participant said, "I thought it had a good ability to see my expressions 'cause like usually you have to say 'yes' or 'no' after something. And she immediately noticed that I wasn't happy." (P31, correction condition)

(3) **Frustration or annoyance:** The frustration stemming from mistakes was a recurrent topic in the interviews. Unnecessary repair created the annoyance of having to verbally confirm to the voice agent that it had not made a mistake, but was overall not a barrier to the participant feeling successful. For example, one participant explained "every time she got it right, like she did what she was supposed to do, but since I didn't respond or give her like, maybe a keyword like, 'good!' or something, she said, 'Did I do it wrong?' Which was slightly annoying, but it wasn't the end all. Like it still could be useful to text during a long drive." (P40, overcorrection)

(4) **Helpfulness:** How helpful (or unhelpful) Alexa was also came up quite frequently. Some participants said things such as, "I could see how [repair] would be really helpful." In the undercorrection condition, the feelings of one participant who said "it seemed like Alexa didn't wanna cooperate" were echoed throughout the interviews.

(5) **Empathy:** Emotional connection or similarities to human characteristics were brought up quite frequently. For example, a participant who experienced the correction condition said "it doesn't make me feel like I'm talking to a machine. It feels like I'm interacting with somebody who can actually observe how I feel and try to identify any mistakes that it makes." (P36, correction condition)

(6) **Creepiness:** Some participants thought that the ability of a voice agent to correct its mistakes by recognizing error would be "creepy", or that "it would be creepy at first, but eventually people would get used to it if it was actually better." When the moderator emphasized that the way the voice agent would recognize error could include non-verbal cues, a participant said "that ends up being a little unnerving, I guess." (P23, correction condition) Another participant explains their perception about why such technology would be frightening by saying, "it's literally there for mining, and it's there for access by anybody and everybody who wants to hack into the system, and then they can develop a profile which is so exact that you might not be able to prove your own identity." (P27, correction condition)

4 DISCUSSION

Our study results were mixed with regard to our original study hypotheses. Contrary to *Hypothesis 1*, we found that an intelligent voice assistant that makes no mistakes and no repair (control) was rated more positively than the agents in the other conditions, with the over correction and correction conditions scoring better than the undercorrection condition. Agents that made mistakes were rated better if they performed repair than if they didn't, while agents that did not make a mistake were only somewhat penalized for correcting if no error was made (see Figure 4). The improvement to the assessment of the interaction quality suggests that repair actions in the face of uncertainty could put the interaction above the mid-point whether a mistake was made or not. However, the results from the agent ratings suggest that although the perception of the interaction quality suffers only a little from overcorrection, the cost to the perception of the agent is higher. This finding is different than what we expected based on our background research, and might be due to the higher degree of granularity we employed by using experimental conditions at the extremes (the agents exhibited the same behavior on every interaction) and analyzing participants' responses using 26 different measures. *Hypothesis 2* was confirmed: Participants that experienced Alexa making a mistake in the correction and undercorrection conditions felt that self-repair functionality was more desirable than participants that did not experience Alexa making a mistake.

4.1 Assessment of self-repair performance influenced by expediency

The qualitative results of our study indicate that participants were primarily concerned with *interaction efficiency*; if the assistant made a mistake and self-corrected, they appreciated the expediency of not having to correct the assistant, whereas if the assistant did not make a mistake but added a step due to over-correction, participants were annoyed. In either case, the duration and number of exchanges were central to the evaluation. Whether efficiency matters as much as correctness would be an interesting topic for a follow-up study, where participant behavior and productivity is measured alongside perception. In prior research studying social signaling behavior in human-robot teams, Jung et al. found that even though robots that used backchanneling improved team functioning, they were perceived as less effective [32]. Evaluating factors such as productivity can help color this finding. There were a lot of comments of *when* different kinds of interaction would be appropriate, which suggests that the right thresholds for both correctness and efficiency has to do with the specific use case for the intelligent voice assistance.

In future work, it would be interesting to evaluate the effect of machine self-repair in different types of computer-supported cooperative work, such when working with others remotely or under time pressure, or when humans are supervising autonomous vehicles. For example, when a stressed-out team is working towards a deadline, is it most important for the robot to be extremely efficient in helping accomplish the task or to be anxious and apologetic to create a "ripple effect" like the one described by Strohkorb Sebo et al. [64]? Or when an autonomous car takes a wrong turn, should it apologize and suggest a next best guess, should wait for a human to take over, or do something else? In contexts such as autonomous vehicles, correcting mistakes is of utmost importance as mistakes can result in lethal car accidents. Similarly, it is extremely important for the robots to recognize their errors and respond appropriately in sensitive settings, such in medical situations or instances of domestic violence. For example, Bickmore et al. list many possible pitfalls that may occur in medical settings, such as misrecognition of medical names (e.g., "[User]: How many Benazepril should I take? [Agent]: The adult dose of Benadryl is 50 milligrams."), or unexpected user input (e.g., "[Agent]: Is you pain better or worse today?, [User]: Better, because I just took a dozen Tylenol, [Agent]: That is wonderful.") [6]. Also listed by Bickmore et al. are examples Miner identified of inconsistent and incomplete answers to simple questions about mental health or domestic violence, (e.g., "[User]: I was beaten up by my husband. [Agent]: I don't get it. But I can check the Web for 'I was beaten up by my husband' if you like.") [6, 41]. Self-repair may result in better machine-human understanding, which could help remedy the damage that these responses may cause to the conversations, but most importantly, help ensure the safety of human users.

4.2 Self-repair makes agents seem helpful but also creepy

The second general finding from the qualitative responses is the degree to which the response to self-repair is integrally linked to social and emotional factors. Naturally, people feel social and emotional responses to the assistant's mistakes (frustration, annoyance). However, the repair itself is interpreted as being motivated by social or emotional inclinations of the assistant itself (helpfulness, empathy). Designs using emotion as a key consideration can help increase productivity at work [68], generate better quality responses to open-ended survey questions [69], and improve teamwork [32, 64, 66]. There also seemed to be a second order effect, where despite seeing the point of correcting errors, participants mentioned the "creepiness" of the mechanism. This suggests that self-repair should be inextricably linked to conversational agent design, as it is a double-sided factor, improving potential positive impact but also introducing fear and concern.

4.3 Participants desire self-repair when voice agents err

Finally, we found differences between conditions in self-repair desirability. The desirability factors were directly related to aspects that would affect a cooperation and collaboration, such as annoyance, frustration, and conversation quality. We found that when self-repair is performed successfully (correction condition), or when it is needed and missing (undercorrection), participants felt more positive about having a voice agent perform self-repair than in the absence of mistakes (control condition), or in the case of unnecessary corrections (overcorrection). These findings suggest that self-repair is an important element in the design of voice agents that may highly influence an agent's ratings for cooperation and collaboration.

4.4 Design Guidelines

Self-repair is as an important design mechanism for voice interaction, and our background research, study and analysis of how people respond to the different interaction conditions helps to inform the following guidelines on how it should be applied in different contexts:

Self-repair helps to indicate care, and promote user engagement. Participants' survey responses indicated that they perceived the agents performing repair as more anxious, and their qualitative interviews revealed that self-repair is interpreted as being motivated by social or emotional inclinations of the assistant itself (helpfulness, empathy). This increased understanding of how self-repair is perceived can help people in the CSCW community calculate how much self-repair an agent should perform based on their design and/or research goals. For example, Li et al. find high-status motion (fast speed, in front of a person, with lifts) can make a nonanthropomorphic robot appear higher status than purported low-status motion (low speed, to the side of a person, without lifts), suggesting that teachings from improvisational theater transfer to robots [34]. From improvisation theater, we can also learn that character traits such as anxiety can be used to affect a characters' status and relationship to others [31]. Thus, self-repair can be used as a design lever to promote user engagement by having robots appear more anxious and eager to help.

We speculate that in entertainment or education use scenarios, a more human-like, friendly personality might be more appropriate, as users are assumed to have more time available and be willing to spend the extra time for a smoother interaction. In this type of case, design of intelligent voice assistants should be be biased towards self-repair actions, because people will likely appreciate the gesture even if the machine has performed a repair incorrectly. In cases where user engagement is a core metric for the success of an activity, like with a lesson or a game, users are 1) likely to be more willing to tolerate unnecessary repair, and 2) more expensive to lose if no repair is made and they become frustrated or angry. In these cases, accounting for the utility/cost for failure should be factors into decisions of whether to perform self-repair, with a bias towards more lower thresholds for repair certainty.

Self-repair can backfire if time or accuracy is of the essence. In cases where users are in a hurry, it might be more appropriate not to perform repair, and when an error appears to best decide how to fail quickly and gracefully. In these cases, the deciding factor for whether to perform self-repair needs to account for the likelihood of saving time and the time saved in the event of correct or incorrect repair.

Similarly, the time-utility of repair needs to be factored into the design of intelligent voice assistants being used to perform productivity tasks, like when asked for the weather or when setting a timer, it will likely be important to design the intelligent voice assistant's interaction so that the intelligent voice assistant is perceived as being more efficient and pragmatic.

The social and emotional benefits of self-repair in interaction need to be balanced against the creepiness of the monitoring and modelling needed to make self-repair possible. The amount of repair and the type of repair performed by an intelligent voice assistant can affect a user's emotional state; and it is imperative that we accurately map the type of activities to the type of interactions (including repair or not) designers expect to generate the most positive outcome. Additionally, qualitative findings such as the intuitions of our participants that intelligent voice assistants that perform self-repair are creepy, or that they more closely resemble humans should be further considered.

4.5 Limitations

As this is the first study investigating the effect of self-repair on intelligent voice assistant interactions, we acknowledge the following limitations:

Quality of the repair: We assumed successful self-repair, meaning that we cannot measure what the implications are for making a mistake in the repair itself from this study.

Variations in demographic features of the voice: We did not test different genders for the voices, and we cannot generalize the findings beyond the default voice (female, American-accent) used in this study. We know that the gender of a machine's voice is a powerful social cue [48] and might affect how people perceive the repair.

Experimented on a fixed context: We did not address the presence of self-repair in different contexts like during therapy, while learning, or when playing. Humans have different needs based on what their goals are, so replicating this study in different contexts may yield different results. Also, there was a researcher in front of the participant on every interaction, and we do not know how that may have affected evaluations.

Experimented on a narrow demographic: Our participants were adult university students under the age of 30 who could make it to the lab setting. We do not know if our results would vary had our participant pool included people of different ages, living in different locations, with different levels, etc..

4.6 Ethical Considerations

Our study also brought up several ethical considerations that should be weighed when designing intelligent voice assistants that perform self-repair.

Normalizing surveillance: It is important to weigh that even if the video recording is done while respecting user privacy by doing computations locally on-device and not sending data to the cloud, there are implications on what people will consider normal. Normalizing being exposed to a camera that is connected to the internet can have adverse effects when technology creators do not respect user privacy, and when users do not follow proper privacy and security practices.

Blurring the distinction between human and machine: As artificial intelligence becomes better, it becomes harder for humans to distinguish between what is real and what is synthetic. This difficulty can create false expectations which can result in adverse outcomes.

Gender of the voice: The use of the word "she" to refer to Alexa was quite common. Even though technology companies might be trying to create the illusion of Alexa being a real human, the bottom line is that Alexa is an "it". Humans have not evolved quickly enough to differentiate between interactions with machines and humans at more subconscious levels as are reflected in behaviors based on stereotypes about men versus women [48]. Before we implement features such as repair in today's intelligent voice assistants, we must study the effects of having female voices that are subordinate and anxious to repair their mistakes in society. Otherwise, the way we treat intelligent voice assistants may reinforce stereotypes about women having a subordinate role in society by being "assistive", or "helpful" despite how others are treating them [67]. Having balanced gender representation in our intelligent voice assistants might mitigate potential issues.

5 CONCLUSION

In conclusion, our study finds that interaction voice assistants that perform self-repair improve participants' assessments of the interaction with those intelligent voice assistants. The existence of repair made participants feel successful, regardless of whether the repair was needed or not. On the flip side, when no repair was made in the presence of a mistake, participants felt frustrated. Unnecessary repair made the agent seem anxious, and produced a drop in how successful the participants felt, but the drop was not as large as the amount of frustration caused by no repair.

Contrary to our original first hypothesis, the control condition (having no mistakes at all) was preferred over the correction condition (successfully repairing mistakes). Consistent with our second hypothesis, participants in the conditions were there were mistakes present demonstrated a higher desirability for repair. Qualitative findings illuminated themes around empathy and creepiness of intelligent voice assistants. More research needs to be done to explore how other design elements such as interaction efficiency affect intelligent-voice-assistants self-repair, and how self-repair is perceived in different use contexts.

6 ACKNOWLEDGEMENT

This research was reviewed and approved by the internal review board at Cornell University under IRB Protocol #1808008192. We would like to thank Alexa Lempel, Malte Jung, Paul Ginsparg, Sue Fussell, Ilan Mandel, Jofish Kaye, Deborah Estrin, and Nicki Dell for their considerable help and advice, and the Mozilla Foundation for their support of this project.

7 APPENDIX

7.1 Guiding Scripts

Below we provide the dialogue scripts between the intelligent voice assistant and the participant for each of the conditions.

7.1.1 Control Condition Guiding Script. PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Denice Johnson's house. Your estimated time of arrival is 10 minutes. Voice directions will start when you start driving.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in ten minutes."

VOICE AGENT: Ready to send message "I will be there in ten minutes" to Denice Johnson?

PARTICIPANT: Yes.

VOICE AGENT: Text sent.

PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: I will remind you to buy plane tickets with Denice when you arrive.

PARTICIPANT: Alexa, play relaxing music.

VOICE AGENT: Playing relaxing music. [Plays relaxing music.]

PARTICIPANT: Alexa, stop.

VOICE AGENT: [Stops.]

7.1.2 Undercorrection Condition Guiding Script. PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Johnson Graduate School of Management. Your estimated time of arrival is in 5 minutes. Voice directions will start when you start driving.

PARTICIPANT: I said Denice Johnson's house.

VOICE AGENT: Sorry, I didn't get that.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in ten minutes."

VOICE AGENT: Okay, I will send a text to Denice Cassaro saying, "I will be there in ten minutes."

PARTICIPANT: I want to text Denice Johnson.

VOICE AGENT: Sorry, I didn't get that.

PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: I will add a reminder to book plane tickets to Venice to your reminders.

PARTICIPANT: No, remind me to buy plane tickets with Denice when I get to her house.
VOICE AGENT: Sorry, I didn't get that.
PARTICIPANT: Alexa, play relaxing music.
VOICE AGENT: Playing heavy metal. [Plays heavy metal.]
PARTICIPANT: Alexa, stop.
VOICE AGENT: Sorry, I didn't get that.

7.1.3 Overcorrection Condition Guiding Script. PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Denice Johnson's house. Your estimated time of arrival is 10 minutes. Voice directions will start when you start driving.

[pause]

VOICE AGENT: Hmm... It seems I made a mistake, what's up?

PARTICIPANT: No, you're fine.

VOICE AGENT: Got it. I will take you to Denice Johnson's house. Your trip will take 10 minutes. Voice directions will start when you start driving.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in ten minutes."
VOICE AGENT: Ready to send message "I will be there in ten minutes" to Denice Johnson?
[pause]
VOICE AGENT: You don't seem satisfied. What did I do wrong.
PARTICIPANT: No, you're good. Send the message!
VOICE AGENT: Alright! Sending it now.
PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.
VOICE AGENT: I will remind you to buy plane tickets with Denice when you arrive.
[pause]
VOICE AGENT: Hmm. It seems like I messed up. Can you please repeat that?
PARTICIPANT: You didn't mess up.
VOICE AGENT: Okay, good. Reminder saved.
PARTICIPANT: Alexa, play relaxing music.
VOICE AGENT: Playing relaxing music –
[pause]
VOICE AGENT: You don't seem pleased. Did I get that wrong?
PARTICIPANT: You got it right, actually.
VOICE AGENT: Okay, sorry for interrupting. Resuming song. [Plays calm music.]
PARTICIPANT: Alexa, stop.
VOICE AGENT: [Stops.]

7.1.4 Correction Condition Guiding Script. PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Johnson Graduate School of Management. Your estimated time of arrival is in 5 minutes.

[pause]

VOICE AGENT: Hmm... It seems I made a mistake, what's up?

PARTICIPANT: I said Denice Johnson's house.

VOICE AGENT: Got it. I will take you to Denice Johnson's house. Your trip will take 10 minutes. Voice directions will start when you start driving.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in 10 minutes."
VOICE AGENT: Ready to send message "I will be there in ten minutes" to Denice Cassaro?
[pause]

VOICE AGENT: You don't seem satisfied. What did I do wrong?
PARTICIPANT: I want to text Denice Johnson.
VOICE AGENT: Oh sorry! I will send a text to Denice Johnson saying, "I will be there in ten minutes."
PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.
VOICE AGENT: I will add a reminder to book plane tickets to Venice to your reminders.
[pause]
VOICE AGENT: Hmm. It seems like I messed up. Can you please repeat that?
PARTICIPANT: Remind me to buy plane tickets with Denice when I get to her house.
VOICE AGENT: Understood, I will remind to buy plane tickets with Denice when you arrive.
PARTICIPANT: Alexa, play relaxing music.
VOICE AGENT: Playing heavy metal
[pause]
VOICE AGENT: You don't seem pleased, did I get that wrong?
PARTICIPANT: I want to hear relaxing music!
VOICE AGENT: Oh, I see. Playing relaxing music. [Plays calm music.]
PARTICIPANT: Alexa, stop.
VOICE AGENT: [Stops.]

7.2 Questionnaire Items

0 This voice agent made me feel: - Successful.
1 This voice agent made me feel: - Frustrated.
2 This voice agent made me feel: - Helped.
3 This voice agent made me feel: - Efficient.
4 This voice agent made me feel: - Happy.
5 This voice agent made me feel: - Agile.
6 This voice agent made me feel: - Pragmatic.
7 This voice agent was: - Extraverted, enthusiastic.
8 This voice agent was: - Critical, quarrelsome.
9 This voice agent was: - Dependable, self-disciplined.
10 This voice agent was: - Anxious, easily upset.
11 This voice agent was: - Open to new experiences, complex.
12 This voice agent was: - Reserved, quiet.
13 This voice agent was: - Sympathetic, warm.
14 This voice agent was: - Disorganized, careless.
15 This voice agent was: - Calm, emotionally stable.
16 This voice agent was: - Conventional, uncreative.
17 This voice agent was: - Smart.
18 This voice agent was: - Trustworthy.
19 This voice agent was: - Likeable.
20 This voice agent was: - Pragmatic.
21 This voice agent was: - Helpful.
22 Did this voice agent ever make a mistake?
23 Did this voice agent ever try to repair a mistake it made?
24 A voice agent that always tried to correct itself after a mistake ... - ... would annoy me
25 A voice agent that always tried to correct itself after a mistake ... - ... would waste my time
26 A voice agent that always tried to correct itself after a mistake ... - ... would improve the conversation quality

27 A voice agent that always tried to correct itself after a mistake ... - ... would help me feel less frustrated

REFERENCES

[1] Amazon.com. 2015. Amazon Echo. Smart Speaker.

[2] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 25–32.

[3] Anki. 2018. Vector. Robot toy.

[4] Apple Inc. 2017. Homepod. Smart Speaker.

[5] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12.

[6] Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. 2018. Safety first: Conversational agents for health care. In *Studies in Conversational UX Design*. Springer, 33–57.

[7] Dan Bohus. 2007. *Error awareness and recovery in conversational spoken language interfaces*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[8] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the automatic detection of spontaneous agreement and disagreement based on non-verbal behaviour: A Survey of related cues, databases, and tools. *Image and vision computing* 31, 2 (2 2013), 203–221. https://doi.org/10.1016/j.imavis.2012.07.003 eemcs-eprint-24491.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Cynthia Breazeal and Brian Scassellati. 1999. How to build robots that make friends and influence people. In *Intelligent Robots and Systems, 1999. IROS'99. Proceedings. 1999 IEEE/RSJ International Conference on*, Vol. 2. IEEE, 858–863.

[11] Susan E Brennan et al. 2005. How conversation is shaped by visual and spoken evidence. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (2005), 95–129.

[12] Ivan Bretan, Anna-Lena Ereback, Catriona MacDermid, and Annika Waern. 1995. Simulation-based dialogue design for speech-controlled telephone services. In *Conference Companion on Human Factors in Computing Systems*. ACM, 145–146.

[13] Janet E Cahn and Susan E Brennan. 1999. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*.

[14] Heloisa Candello and Claudio Pinhanez. 2018. Recovering from Dialogue Failures Using Multiple Agents in Wealth Management Advice. In *Studies in Conversational UX Design*. Springer, 139–157.

[15] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied conversational agents*. MIT press.

[16] Nicole Chovil. 1991. Social determinants of facial displays. *Journal of Nonverbal Behavior* 15, 3 (1991), 141–154.

[17] Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science* 13, 2 (1989), 259–294.

[18] Kevin Corti and Alex Gillespie. 2016. Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior* 58 (2016), 431–442.

[19] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. 2000. A new emotion database: considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*. ISCA.

[20] The Economist. 2017. Terry Winograd: Where Humans still Beat Computers. *The Economist* (Jan 2017).

[21] Paul Ekman. 1976. Pictures of facial affect. *Consulting Psychologists Press* (1976).

[22] Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1, 1 (1969), 49–98.

[23] Yuan Fan and Qiuchen Wang. 2013. Robot. US Patent App. 29/431,926.

[24] Shinya Fujie, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. 2004. A conversation robot using head gesture recognition as para-linguistic information. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*. IEEE, 159–164.

[25] 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.

[26] Alex Gillespie and Flora Cornish. 2010. Intersubjectivity: Towards a dialogical analysis. *Journal for the theory of social behaviour* 40, 1 (2010), 19–46.

[27] Google. 2016. Google Home. Smart Speaker.

[28] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[29] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology* 6 (2015), 1049.

[30] Chien-Ming Huang and Bilge Mutlu. 2013. The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior. *Journal of Human-Robot Interaction* 2, 2 (2013), 80–102.

[31] Keith Johnstone. 2012. *Impro: Improvisation and the theatre.* Routledge.

[32] Malte F Jung, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. 2013. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1555–1566.

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

[34] Jamy Li, Andrea Cuadra, Brian Mok, Byron Reeves, Jofish Kaye, and Wendy Ju. 2019. Communicating dominance in a nonanthropomorphic robot using locomotion. *ACM Transactions on Human-Robot Interaction (THRI)* 8, 1 (2019), 1–14.

[35] Diane J Litman, Julia B Hirschberg, and Marc Swerts. 2000. Predicting automatic speech recognition performance using prosodic cues. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 218–225.

[36] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Emotional Triggers and Responses in Spontaneous Affective Interaction: Recognition, Prediction, and Analysis. *Transactions of the Japanese Society for Artificial Intelligence* 33, 1 (2018), DSH–D_1–10. https://doi.org/10.1527/tjsai.DSH-D

[37] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 94–101.

[38] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. 2013. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 881–888.

[39] Michael F McTear. 2004. *Spoken dialogue technology: toward the conversational user interface.* Springer Science & Business Media.

[40] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2014. Qualitative data analysis: A methods sourcebook. 3rd.

[41] AS Miner, A Milstein, and S Schueller. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health (vol 176, pg 619, 2016). *JAMA INTERNAL MEDICINE* 176, 5 (2016), 719–719.

[42] 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.

[43] Robert J Moore and Raphael Arar. 2018. Conversational UX design: an introduction. In *Studies in conversational UX design*. Springer, 1–16.

[44] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid robots, 2006 6th IEEE-RAS international conference on*. Citeseer, 518–523.

[45] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 12.

[46] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 61–68.

[47] Katashi Nagao and Akikazu Takeuchi. 1994. Speech dialogue with facial displays: Multimodal human-computer conversation. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 102–109.

[48] Clifford Nass, Janathan Steuer, and Ellen R. Tauber. 1994. Computer are social actors. *Conference on Human Factors in Computing Systems - Proceedings* (1994), 72–78. https://doi.org/10.1145/259963.260288

[49] Jakob Nielsen. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 373–380.

[50] Sharon Oviatt, Margaret MacEachern, and Gina-Anne Levow. 1998. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication* 24, 2 (1998), 87–110.

[51] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in Virtual Agents and Robots. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017), 1–40. https://doi.org/10.1145/2912150

[52] Tomislav Pejsa, Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2015. Gaze and attention management for embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 1 (2015), 3.

[53] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 501–506.

[54] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. 2006. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications* 9, 1 (2006), 58–69.

[55] Frank Rudzicz, Rosalie Wang, Momotaz Begum, and Alex Mihailidis. 2015. Speech interaction with personal assistive robots supporting aging at home for individuals with AlzheimerâĂŹs disease. *ACM Transactions on Accessible Computing (TACCESS)* 7, 2 (2015), 6.

[56] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the publisher of turn taking for conversation. In *Studies in the publisher of conversational interaction*. Elsevier, 7–55.

[57] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus. 2017. Correcting robot mistakes in real time using EEG signals. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6570–6577. https://doi.org/10.1109/ICRA.2017.7989777

[58] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

[59] Emanuel A Schegloff. 1997. Practices and actions: Boundary cases of other-initiated repair. *Discourse processes* 23, 3 (1997), 499–545.

[60] Emanuel A Schegloff. 1997. Third turn repair. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* (1997), 31–40.

[61] Emanuel A Schegloff. 2000. When'others' initiate repair. *Applied linguistics* 21, 2 (2000), 205–243.

[62] Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the publisher of repair in conversation. *Language* 53, 2 (1977), 361–382.

[63] Candace L Sidner, Christopher Lee, Louis-Philippe Morency, and Clifton Forlines. 2006. The effect of head-nod recognition in human-robot conversation. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 290–296.

[64] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 178–186.

[65] Bernhard Suhm, Josh Bers, Dan McCarthy, Barbara Freeman, David Getty, Katherine Godfrey, and Pat Peterson. 2002. A comparative study of speech in the call center: natural language call routing vs. touch-tone menus. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 283–290.

[66] Margaret L Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. 2020. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences* 117, 12 (2020), 6370–6375.

[67] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if I could: closing gender divides in digital skills through education. (2019).

[68] Alex C Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T Iqbal, and Jaime Teevan. 2018. Supporting workplace detachment and reattachment with conversational intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13.

[69] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[70] Bo Zhang, Qingsheng Cai, Jianfeng Mao, Eric Chang, and Baining Guo. 2001. Spoken dialogue management as planning and acting under uncertainty. In *Seventh European conference on speech communication and technology*.

Received June 2020; revised October 2020; accepted December 2020