Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns

Zahra Ashktorab
IBM Research AI
Yorktown Heights, NY, USA
zahra.ashktorab1@ibm.com

Mohit Jain
IBM Research
Bangalore, India
mohitjain@in.ibm.com

Q. Vera Liao
IBM Research AI
Yorktown Heights, NY, USA
vera.liao@ibm.com

Justin D. Weisz
IBM Research AI
Yorktown Heights, NY, USA
jweisz@us.ibm.com

ABSTRACT

Text-based conversational systems, also referred to as *chatbots*, have grown widely popular. Current natural language understanding technologies are not yet ready to tackle the complexities in conversational interactions. Breakdowns are common, leading to negative user experiences. Guided by communication theories, we explore user preferences for eight repair strategies, including ones that are common in commercially-deployed chatbots (e.g., confirmation, providing options), as well as novel strategies that explain characteristics of the underlying machine learning algorithms. We conducted a scenario-based study to compare repair strategies with Mechanical Turk workers (N=203). We found that providing options and explanations were generally favored, as they manifest initiative from the chatbot and are actionable to recover from breakdowns. Through detailed analysis of participants' responses, we provide a nuanced understanding on the strengths and weaknesses of each repair strategy.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; **Natural language interfaces**; *Interactive systems and tools*; *Empirical studies in interaction design*; User studies; User interface design;

KEYWORDS

Chatbots, conversational agents, conversational breakdown, repair, grounding

ACM Reference Format:
Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300484

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300484

1 INTRODUCTION

In 1966, Eliza simulated dialogue as a Rogerian psychotherapist [47]. Fast forward to 2016, the MIT Technology Review heralded chatbots as one of the year's breakthrough technologies [33]. Chatbots have made much headway since Eliza's introduction. However, it has become apparent that current conversational technologies are still inadequate at handling all of the complexities of natural language interactions, as manifested by a number of high-profile chatbot failures [2, 34]. Breakdowns in understanding user input happen often, and they can have profound impact on how people perceive and interact with a chatbot. In the worst case, they may abandon the chatbot or the current task. Or, they may need to endure a haphazard trial-and-error process to recover from the breakdown. Both breakdowns and current recovery processes decrease peoples' satisfaction, trust, and willingness to continue using a chatbot [19, 20, 28].

A universal challenge faced by chatbot developers is how to design appropriate strategies that mitigate the negative impact of breakdowns. Previous work [19, 24, 42, 48] studied strategies that aim to alleviate peoples' negative emotional response from agent or robot breakdowns, such as showing politeness and apologetic behaviors. However, in task-oriented settings, such as a chatbot performing information assistance, these strategies may be ineffective if the user still

fails to accomplish the task. In this paper, we focus on strategies that support *repair* – recovering from the breakdown and accomplishing the task goal.

Repair is a ubiquitous phenomenon in human communication. When a breakdown happens in a conversation, people take a variety of actions such as repeating, rephrasing, or clarifying, to repair it. Although chatbot users should be skillful in using similar actions as the *speaker*, the repair task becomes challenging as the *listener* is no longer a fellow human. Two problems often impede the repair process with chatbots: 1) there may be a lack of evidence that a breakdown has occurred, which may either be a limitation of the underlying technology (i.e., unable to recognize a breakdown) or a failure in design to communicate the breakdown; 2) the system's model is unfamiliar for the user to choose an effective way to repair. When talking to another person, repairs are almost subconscious acts, which may include a combination of speech, gesture, and facial expression [6]. Chatbots rely on machine learning algorithms to process a user's input, which are "black boxes" for the user. Though these interfaces are deemed "conversational," they may not be repaired in the same way as talking to another person [32].

In this work, we study repair strategies that a chatbot (listener) could adopt to tackle the above problems – providing evidence for the breakdown and supporting repair towards a desirable direction for the system model. We note that many commercial chatbot products are already adopting repair designs to serve these goals. One example is to ask for *confirmation* when the system has low confidence, which gives a clear signal of a potential breakdown and allows the user to initiate repair without the system mistakenly executing a task. Another example is to provide *options* of tasks that the chatbot can handle based on their proximity to the user's input, which not only indicates that a breakdown occurred, but also drives the interaction to the scope of the system model's capabilities.

This paper makes two contributions. First, we identify a set of repair strategies, informed by communication theories and prior work on conversational agents. In addition, we introduce a group of novel repair strategies that aim to expose the system model, as inspired by recent work in explainable machine learning [35, 41, 46]. These strategies explain *why* a breakdown occurred, such as showing which keywords the system was able/unable to understand, in order to assist a user in effective *self-repair*. These strategies contrast with *system-repair* strategies such as directly providing options. Second, we conducted a scenario-based study with Mechanical Turk workers (N=203) to systematically understand people's preferences for different repair strategies. Our study focuses on text-based chatbots, which are widely used and growing in popularity [21], although some of the repair strategies we examined can be applied to voice-based agents as well.

2 BACKGROUND AND RELATED WORK

Our study is informed by communication theories relevant to conversational breakdown and repair, prior work on repair in human-agent interaction, as well as transparency and explanation of machine learning systems.

Breakdowns and Repairs in Communication

Social scientists have long been interested in studying repairs in human communications, defined as "the replacement of an error or mistake by what is correct" [36]. Schegloff et al. made the distinction between self- and other- repair [36], referring to the correction made by the speaker or the listener, respectively. A distinction is also made between the initiation and the outcome of a repair. The person who initiates a repair is not necessarily the one who completes it. Empirically, Schegloff et al. concluded a preference for self- over other-repair regardless of who initiates it.

Repair is also frequently studied under the framework of *grounding in communication*, proposed by Clark and Brennan [10]. Grounding describes conversations as a form of collective action to achieve common ground or mutual knowledge. As a speaker presents an utterance, *evidence* of understanding, whether explicit or implicit (e.g., a correct response), is expected. If there is a lack of evidence or presence of negative evidence, the speaker may choose to initiate a repair. The theory uses the concept of *cost* to explain why a repair strategy is used, or if the breakdown is ignored without repair. For example, formulation cost predicts that a speaker prefers simple ways of rephrasing (e.g., correcting a partial sentence) over providing a complete new utterance. It also explains the preference for self- over other-repair by minimizing turn-taking cost (number of potential repair turns needed) and fault cost (i.e., being perceived at fault).

In a serial work to adapt the grounding framework for human-computer interaction [5, 7], Brennan highlighted that the understanding models are private to each party, and dialog partners can only estimate how to converge them. When the dialog partner is a machine, its private understanding model is significantly mismatched from the human speaker, posing challenges for grounding or repair. Brennan derived a theory-driven model for a spoken dialog system to explicitly indicate in which state the breakdown happens, such as the attending, recognizing, interpreting, or acting stage.

Repair in Human-Agent Interaction

Recently there has been a growing volume of research on human-agent interaction. A common theme in work studying everyday use of conversational agents is users' struggle with natural language interactions [26–29, 32]. Myers

et al. studied chat logs of a voice based interface (VUI) to identify types of errors and users' coping tactics [29]. They found NLP errors – misunderstanding a user's utterance – to be the most common type of error, and users engaged in a variety of tactics including hyper-articulation, simplification, and providing more information in attempts to repair. Porcheron et al. conducted a field study of user interactions with Amazon Alexa [32] at homes and found that a significant amount of interactions were dedicated to repair. They attributed the challenge of user repair to a lack of indication of trouble in Alexa's error messages: "[Alexa] provides no mechanism for further interaction, and does not make available the state of the system, allying the VUI with notions of a 'black box'". This conclusion echoes a long-standing concern on the limitation of conversational agent interfaces – a lack of transparency on system status and affordance [28, 39].

Besides these studies providing a descriptive account of breakdown, work that suggests design solutions to support the repair process has been limited. A distinction should be made between *agent-initiative* and *user-initiative* systems [17]. In the former case, systems with the dialogue initiative can restrict users' responses by asking close-ended questions. It is in the latter case where breakdowns are common, as users can ask free-form questions, and repairing breakdowns is challenging because users are uncertain about the system's status and capabilities. Popular commercial agents, such as Apple's Siri and Amazon's Alexa, are mostly user-initiative. They are also considered *goal-oriented* because users have an information goal to achieve from the interaction. For *non-goal-oriented* chatbots (i.e., for chit-chat), Yu et al. [50] enumerated a list of strategies such as repeating parts of the user utterance, switching topics, and telling jokes, but they aim to engage users for further interaction instead of supporting repair.

The related human-robot interaction (HRI) community has studied designs to mitigate the negative effects of robot breakdowns. With humanoid robots, the focus has been on social behaviors that make users more tolerant or willing to help. For example, multiple studies explored using politeness and apology strategies to request help when the robot malfunctions [14, 25, 40]. Most relevant to ours is the work by Lee et al. [24]. Using a scenario-based survey, they studied three strategies for a robot to recover from a breakdown: apologies, compensation, and providing options. They found individual differences in repair preferences based on service orientation: those with a relational orientation preferred apologies, while those with a utilitarian orientation (interactions with the bot are purely transactional) preferred compensation [24]. In our study, we borrow the methodology of a scenario-based survey as it provides a means to gather a large quantity of data for our set of repair strategies, and it allows us to strictly control the interaction process and outcomes to evaluate the *perception* of different repair strategies. Different from Lee et al. [24], we adopt a pairwise comparison design to elicit reasons for peoples' preferences between different repair strategies.

Explanation of Machine Learning System

Work reviewed above suggests exposing an agent's underlying model could effectively support repair in user-initiative, goal-oriented conversations. Notably, a recent study introduced an interface that persistently displayed a chatbot's state of understanding to the user [18], and enabled users to edit directly when an error happened. Current chatbots often work in a question-and-answer format relying on an *intent* model [49], which uses machine learning classifiers to map a user utterance to one of many pre-defined intents (e.g., "hello" and "hi" would be classified as the *greeting* intent). However, little work has explored exposing the status of these machine learning classifiers to a chatbot's users.

We draw inspiration from recent work on explanation of machine learning algorithms [15, 16, 46]. For text classifiers, explanations are generated from their features, such as the words used in the documents they classify. A common approach is to highlight keywords in a document that have the highest weights for the classifier's decision – "this document is classified as sports news because it contains the keyword *football*". Stumpf et al. [41] explored peoples' willingness to provide feedback on machine learning systems when explanations for their predictions were provided, including keyword highlighting and rule-based explanations, and they found that people provided rich feedback for improving the systems. We note that keyword extraction can be achieved through various methods for any kind of text classification algorithms [35]; thus, our design of explanation-based repair strategies is agnostic to the actual underlying classifier.

3 REPAIR STRATEGIES & RESEARCH QUESTIONS

We used several concepts from communication theories on grounding [10] and repair [36] to drive the choices of repair strategies we studied. First, we considered the *evidence of misunderstanding* or *initiation of repair* from the agent. Given users' unfamiliarity with the agent's private model, it is necessary for the agent to indicate a potential misunderstanding. However, an HRI study found that users prefer the agent to ignore the uncertainty and carry on an action until the user initiates a correction [14]. Explicitly acknowledging a mistake lowers the likability and perceived intelligence of the agent, and may add friction to the interaction as the user is obliged to respond to the initiation.

Second, we distinguished between *self-repair* and *system-repair*. For a question-and-answer chatbot, users' self-repair

is usually limited to rephrasing the original input. System-repair may diverge from other-repair in human-human conversations given the underlying machine learning model and limited capabilities.

Lastly, we attempted to reduce users' repair cost by exposing details of system's understanding status, so users can engage in *assisted self-repair*. We drew inspiration from work on explainable machine learning and introduce three novel designs of agent explanation strategies.

In our study, we focused on the following eight repair strategies (Figure 1) that have different attributes with regard to the three above factors. We opted out of a factorial design because these factors were either dependent or orthogonal. To initiate system-repair or provide explanation, the agent must acknowledge the potential misunderstanding; engaging in system-repair precludes assisting in users' self-repair. These strategies were also chosen because they can be broadly applied to chatbots that rely on the commonly used intent-based model [49] – a chatbo relies on using a multi-classifier to classify a user utterance to one of many pre-defined intents, triggering a response linked to that intent. Specifically, classification of each intent has a *confidence* score, and the intent with the highest confidence is considered as the recognized intent. With an intent-based model, it is common to define breakdown as when the confidence levels for all intents are below a certain threshold. Our repair strategies are concerned with the immediate action that a chatbot would take *after* recognizing such a breakdown.

Repair Strategies

No evidence of a breakdown.

- **Top response**. Similar to the "ignore" strategy studied by Engelhardt et al. [14], the chatbot gives no evidence of a potential breakdown, but outputs the response to the intent with the highest confidence, even when it is below the threshold. In this scenario, the user would have to initiate a repair after seeing the wrong response.

With evidence of a breakdown.

- **Repeat**. The chatbot recognizes a potential breakdown and explicitly indicates it, then repeats the initial prompt to the user.
- **Confirmation**. The chatbot recognizes a potential breakdown when the top intent falls below the confidence threshold. It then explicitly confirms the top intent (e.g., "sounds like you want to... is that correct?"). This strategy is considered more natural, and similar to how a human listener initiates a repair [36].

With evidence of a breakdown, system-repair.

- **Options**. The chatbot not only indicates a potential breakdown, but also provides options of potential intents in which it has the highest confidence. The system attempts to repair by taking over the dialogue initiative to restrict interaction within its capabilities.
- **Defer**. It is a common strategy for a chatbot to transfer a request it is unable to solve to a human agent. We consider deferring as a type of system-repair as it is a solution for the system to resolve breakdowns via human intervention.

With evidence of a breakdown, assisted self-repair.

- **Keyword highlight explanation**. Inspired by keyword-based explanations for text classifiers [41], we introduce a strategy that reveals why an intent was mistakenly recognized by highlighting keywords in the user's utterance that contribute to the classifier's decision. By exposing the chatbot's understanding mechanism, it is expected to help the user rephrase by avoiding the keywords that the chatbot misunderstood or by using words that are closer to the desired intent.
- **Keyword confirmation explanation**. This strategy is similar to keyword highlighting, but instead of highlighting on the user's original utterance, the chatbot explicitly explains its understanding to the user in a confirmation message. Although it is more natural in a conversational form, it makes a trade-off in that it needs an additional conversational turn.
- **Out-of-vocabulary explanation**. This strategy highlights words that the bot did not understand in order to help the user rephrase. This explanation can be realized by extracting words that are distant or missing from the chatbot's training data or knowledge base.

Research Questions

We addressed the following research questions in our scenario-based study.

- **RQ1**: Which repair strategies are preferred when a conversational breakdown with a chatbot occurs, and why?
  - **RQ1a**: Is it preferable to acknowledge breakdowns?
  - **RQ1b**: Is it preferable to provide system-repair?
  - **RQ1c**: Is it preferable to provide assisted self-repair by explaining system's understanding?
- **RQ2**: How do different individual and task-related factors impact preferences for different repair strategies?

For **RQ2**, there were a number of individual factors we considered, including social orientation with chatbots (i.e., desire for human-like social interactions [26, 27]), service orientation (i.e., viewing service interactions as either transactions or social interactions [24]), prior experience with chatbots, and experience with technology. For task-related factors, we considered scenarios with different repair outcomes (successful or not) and different contexts (shopping, banking, and travel).

4 METHODOLOGY

To answer our research questions, we developed scenarios in which a breakdown happens and then the chatbot adopts one of the eight repair strategies as discussed above (shown in Figure 1). All scenarios started with the same breakdown as shown in the "Initial Prompt.", where a wrong intent was recognized ("add a credit card" instead of "add my daughter to my credit card"). This wrong intent was exposed by the repair strategies except for Top, Repeat and Defer. For example, the Keyword Highlight Explanation strategy highlighted the keywords "add" and "card". After seeing the system's repair action, the user in all scenarios provided the same input (expanded original queries by mentioning "add as authorized user"), and ended the conversation in success in Figure 1.

To answer **RQ2**, we examined different repair outcomes. In half of the scenarios, after the user's second attempt, the chatbot provided the correct answer as shown in Figure 1; in the other half, the user's second attempt led to another breakdown. In chatbot interactions, it is common that a user has to make multiple attempts to repair; thus, it was important to study whether repair strategy preferences differed when the repair interaction did not succeed. We also introduced three different task contexts (shopping, banking, travel) to examine the generalizability of repair strategy preferences. Figure 1 shows the banking scenarios. In the shopping scenarios, the user inquires information about a previous order. In the travel scenarios, the user asks for directions to a tourist attraction. In total, we developed 48 scenarios: 3 (context) × 8 (repair) × 2 (outcome success).

Paired Comparison Experiment

We adopted a pairwise comparison experiment to collect peoples' preferences for repair strategies. Our experiment consisted of tasks in which we randomly showed participants two of the eight repairs, but with the same context (shopping/banking/travel) and outcome (successful/unsuccessful). We asked participants to select which scenario appealed to them more and describe why they had made their selection.

Pairwise experiments are commonly used in various fields of research to determine participant judgments [9, 22]. Pairwise comparisons could yield more realistic results than Likert scales [1] because they take advantage of simple judgments and prioritize a small set of stimuli to learn people's preferences [8, 12]. They also allow us to elicit qualitative responses on the desirable traits of one repair strategy over another. We performed rank analysis of our pairwise comparisons using the Bradley-Terry model [4].

Individual Factors Survey

We are interested in how the following individual factors impact preferences for repair strategies: social orientation toward chatbots, service orientation, prior experience with chatbots, and experience with technology in general. These factors have been shown to impact peoples' preferences and behaviors. All measures were self-reported using 5-point Likert scales.

Social Orientation toward Chatbots: Introduced by Liao et al. [26, 27], this measure reflects a desire to engage in human-like social interactions with chatbots, which is associated with a mental model of an agent system as being a sociable entity. They found that people with a high social orientation desire natural conversation and social designs from the agent while those low in social orientation used chatbots like an information search engine. We used the scale introduced in [26]: *"I like chatting casually with a chatbot"* and *"I think 'small talk' with a chatbot is enjoyable."* Cronbach's α was 0.84 indicating high reliability.

Service Orientation: In Lee et al.'s work studying recovery strategies for robot breakdown [24], they noted a preference difference between those with a utilitarian vs. a relational service orientation. We adapted two items from their work: *"Efficient customer service is important to me"* and *"I found it frustrating when a customer service representative could not immediately give me the information I need."* However, Cronbach's α was 0.38 indicating poor reliability, so we include these items as two separate measures in our analysis: service frustration and service efficiency.

Experience with Chatbots: We assessed self-reported prior experience with chatbots: *"I am familiar with chatbot technologies"* and *"I use chatbots frequently."* Cronbach's α was 0.71 indicating good reliability.

Experience with Technology: We assessed self-reported tech-savviness: *"I consider myself an advanced technology user"* and *"I am eager to try new technologies."* Cronbach's α was 0.70 indicating good reliability.

Participants, Task, and Procedure

Participants were recruited on Mechanical Turk with the requirement of being 18 years or older. In each task, participants performed 10 pairwise comparisons between repair strategies for a given scenario and outcome. Each scenario was presented turn-by-turn, with three-second typing indication pauses in between chat bubbles to simulate the interactive experience of a chat. After reading the first scenario (shown on the left half of the screen), participants clicked a button to show the second scenario (shown on the right half of the screen). After both scenarios were presented, participants were asked to select which chatbot they preferred and give an explanation as to why. Scenarios were selected randomly without replacement so the same participant did not see the same combination of factors twice, and two control scenarios were included as attention checks. The first repeated a previous scenario to see whether the participant

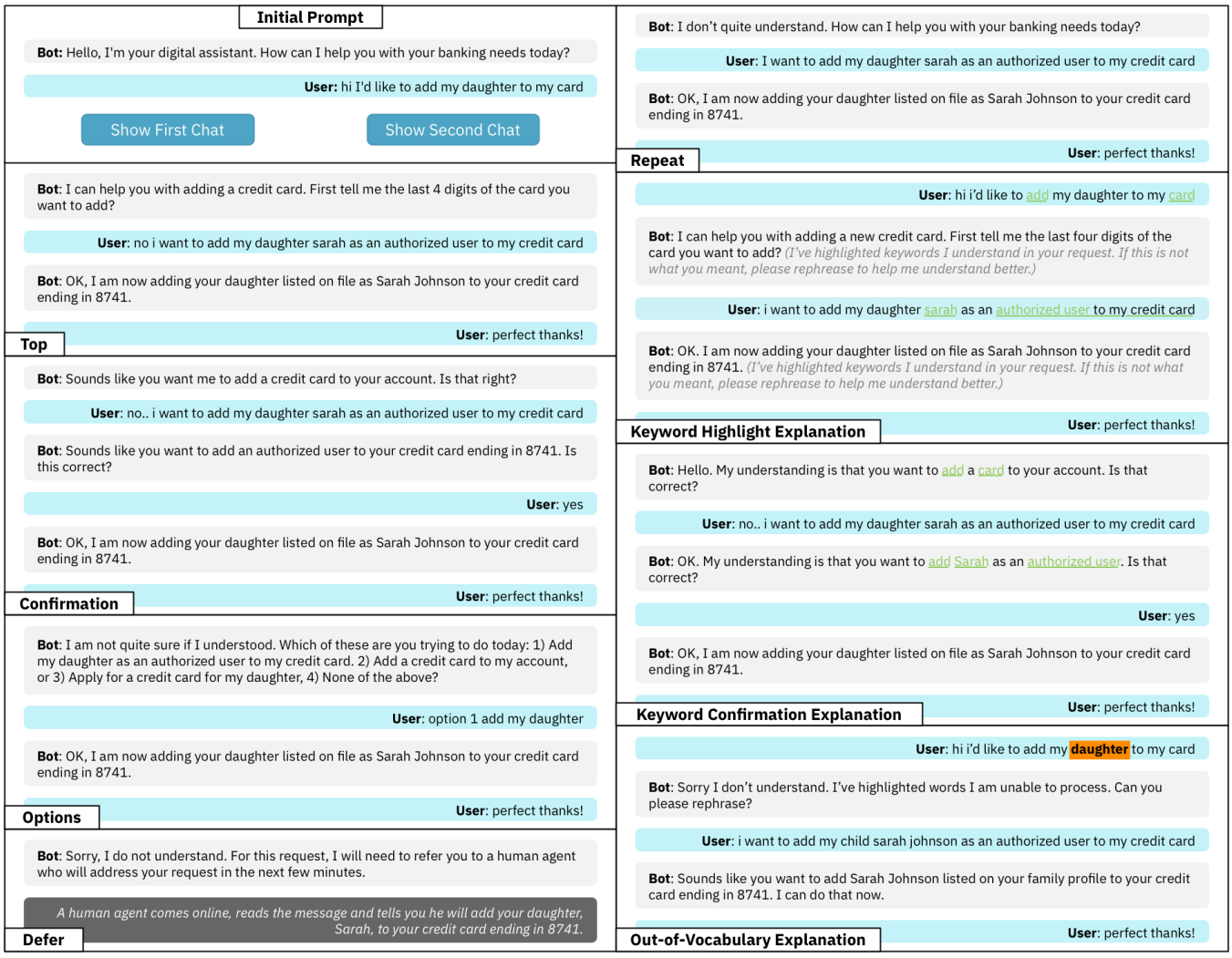

Figure 1: Eight repairs for the successful banking condition. At the top left, we show the initial prompt in all conditions.

gave it the same rating. The second provided a comparison between a chatbot that had successfully repaired the breakdown with one that did not, and participants were expected to express a preference for the one that was able to successfully repair. After finishing all 10 comparisons, participants filled out a survey that collected demographic information and measurements of individual factors as discussed above. The overall task took about 10 minutes to complete, and participants were compensated $1.50 USD for their participation ($9 USD/hr).

We deployed a total of 340 tasks on Mechanical Turk. We filtered out 137 participants (40%) who did not pass the attention checks, yielding a final sample of 203 participants (141 male, 69%) and 1,624 pairwise comparisons. Of these, 124 (61%) held a bachelor's degree, and 28 (14%) held a postgraduate degree. The average age of our participants was 34 years (SD=9 years). Most of our participants spoke English as their native language (N=179, 88%), and other native languages included Hindi (4%), Malay (3%), and Tamil (3%).

5 RESULTS

In this section, we describe participants' preferences for repair strategies and the underlying reasons (RQ1), where we pay attention to preferences with respect to the acknowledgement of breakdowns (RQ1a), system-repair (RQ1b) and assisted self-repair (RQ1c). We then explore how individual and task-related factors impact these preferences (RQ2).

Preferences of Repair Strategies (RQ1)

The Bradley-Terry model [4] is a mathematical model that estimates a vector of "ability scores" for a set of paired object comparisons, which yields an ultimate ranking of all objects.

Preferred repair vs. Rejected repair	p-value
Options vs. Keyword Highlight	0.000**
Options vs. Confirmation	0.000**
Options vs. Repeat	0.000**
Options vs. Top	0.000**
Options vs. Defer	0.000**
Options vs. Keyword Confirmation	0.000**
Options vs. Out-of-Vocabulary	0.002**
Out-of-Vocabulary vs. Confirmation	0.000**
Out-of-Vocabulary vs. Top	0.000**
Out-of-Vocabulary vs. Repeat	0.000**
Out-of-Vocabulary vs. Keyword Highlight	0.000**
Out-of-Vocabulary vs. Defer	0.000**
Out-of-Vocabulary vs. Keyword Confirmation	0.041
Keyword Highlight vs. Top	0.036
Keyword Highlight vs. Confirmation	0.058
Keyword Highlight vs. Keyword Confirmation	0.576
Keyword Highlight vs. Repeat	0.270
Keyword Confirmation vs. Confirmation	0.014
Keyword Confirmation vs. Top	0.008*
Keyword Confirmation vs. Defer	0.061
Repeat vs. Defer	0.855
Repeat vs. Keyword Confirmation	0.094
Defer vs. Keyword Highlight	0.199
Confirmation vs. Defer	0.546
Confirmation vs. Repeat	0.433
Top vs. Defer	0.413
Top vs. Repeat	0.315
Top vs. Confirmation	0.828

Table 1: Significant values, after Bonferroni adjustment (p <0.05/8), are noted with **. Marginally significant values (p <0.1/8) are noted with *.

This model has been used in previous HCI studies that conducted pairwise comparison experiments (e.g. [3, 37]). We use the BradleyTerry2 R package [44] to generate an overall ranking of repair strategies followed by pairwise comparison tests for significance. For each repair, the model conducts a pairwise test that generates a p-value for each other repair to which it is compared. We used a Bonferroni correction [45] to account for the number of individual comparisons made ($p < 0.05/8$ for significance, $p < 0.1/8$ for marginal significance [11]). In Figure 2, we show the overall rankings, as well as separate rankings for when the scenario was successfully or unsuccessfully repaired. In Table 1, we present the p-values for pairwise comparisons.

As seen in Figure 2, the Options repair was unarguably the most favored strategy, preferred in pairwise comparisons over all other strategies (Table 1). Assisted self-repairs – Keyword Highlight, Keyword Confirmation, and Out-of-Vocabulary Explanation – were generally favored, with Out-of-Vocabulary Explanation as the most preferred among the three. For the rest – Defer, Confirmation, Repeat, and Top – preferences were noisier. Part of the reason, as we observe in Figure 2, is that they were ranked differently in scenarios with successful and unsuccessful repair outcomes. Most evidently, Defer was outranked by all other repairs when the repair was successful, but ranked second when the repair was unsuccessful. This difference implies that if a breakdown can be easily repaired, people prefer to resolve it with the chatbot, whereas if the repair fails after an initial attempt, they desire a human agent to be involved, even if the human agent is unable to resolve it immediately (as in the scenario). We also observe that simple strategies – Top and Repeat – were ranked higher in successful than unsuccessful scenarios. This finding suggests that if the breakdown is straightforward enough to repair with one attempt, chatbots that don't offer evidence of breakdown or repair assistance are acceptable.

Reasons for Preferences (RQ1)

Along with collecting preferences, we asked participants to give reasons why they selected one repair strategy over another. The authors individually reviewed this data and used open coding [13] to extract themes in the open-ended answers. Codes were harmonized after two iterations of review and discussion, resulting in the final set of themes shown in Table 2. A few common themes were observed across repair strategies, reflecting general desires for repair design: 1) *efficiency and efficacy* were desired when recovering from the breakdown to accomplish the task goal, as demonstrated by codes such as "faster," "concise" (easy to read), "help to rephrase," and "less typing required"; 2) some strategies increased *perceived intelligence and capability*, especially when the agent demonstrated its understanding through confirmation or explanations, or when it proactively assisted repair via explanation or directly providing options; 3) *politeness* was demonstrated in strategies that presented an understanding before executing a response (e.g. confirmation, explanations); and 4) *naturalness*, in which participants felt that interactions faithfully resembled human conversations, was not felt in strategies that highlighted keywords or provided options. Based on the results shown in Table 2, we focus on addressing our research questions regarding breakdown evidence, system-repair, and assisted self-repair.

*Explicit Acknowledgement of Breakdown (**RQ1a**).* Our ranking results suggest that participants preferred chatbots to

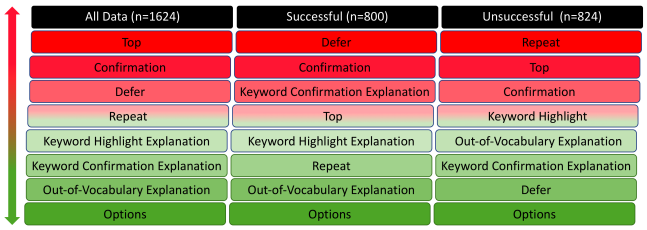

Figure 2: Bradley-Terry rankings of repair strategies. From left to right, rankings for: all data, successful conditions, unsuccessful conditions. From top to bottom: lowest ranked to highest ranked.

Repair	Strengths	Weaknesses
Top	concise with no extraneous questions; simple interaction	began an unwanted process without confirming; lacks resource to resolve breakdown; unfriendly and rude
Repeat	concise; natural; explicit about lack of understanding	appears less intelligent; did not show interest of understanding; lack of resources for user to repair
Confirmation	verify before take an action; show understanding capability; polite; natural	longer conversation to respond to confirmation; appear less competent by repetitively confirming
Options	provide choices to resolve the issue faster; narrow down to what it can do; show understanding capability and intelligence; less typing required by user	complicates with clutter; unnatural; more reading
Defer	interaction with human is faster; human more likely to solve the problem; prefer interacting with a human	wait time and interaction with human slower; human intervention is unnecessary
Keyword Highlight Explanation	show understanding capabilities; help users to rephrase; teach user how to interact with the chatbot; proactively making an effort; intuitive explanation; resolve issue faster with less turns	verbose; repetitive description; highlighting is visually unappealing; less natural
Keyword Confirmation Explanation	show understanding capabilities; help users to rephrase; teach user how to interact with chatbots; proactively making an effort; polite; concise	highlighting is visually unappealing; longer conversation to respond to confirmation; less information provided
Out-of-Vocabulary Explanation	show understanding capabilities; help users to rephrase; teach user how to interact with chatbots ; proactively making an effort; polite; concise; specific about why it fails to understand	appear less competent unable to understand simple words

Table 2: Strengths and weaknesses of repair strategies reported by participants.

explicitly acknowledge a potential breakdown, as Top was generally less favored. Our qualitative data reveals that to proceed with a wrong response is not only unhelpful for resolving the breakdown, but is also perceived as rude, unfriendly, and putting in no effort. Although for scenarios in which the breakdown was resolved in one attempt, participants were more tolerant of the Top strategy, and some also favored its simplicity.

Two similar strategies that acknowledge potential breakdowns – Repeat and Confirmation – had interesting tradeoffs. Participants perceived Confirmation to be more polite (verifying before taking an action) and intelligent (showing its understanding capability) than Repeat, but some found it more burdensome to have to read and respond to the confirmation. Both strategies were considered natural as they resemble ways that a human listener would initiate a repair.

Our results also suggest that while participants like chatbots to acknowledge potential breakdowns, they may be turned away by messages that are redundant and repetitive, such as the current design of Keyword Highlighting, where a prompt about the indication of highlighting was repeated.

*System-Repair (**RQ1b**).* We introduced two distinct strategies for system-repair: Options and Defer. Participants favored Options because it was efficient and required less effort from the user in formulating and typing. They also perceived the chatbot to be more intelligent by taking the dialogue initiative. We note that our scenario-based method may not reflect the real-world success rate of different repair strategies (e.g., Options may not always provide the correct suggestion) However, Options strategy was favored even in the unsuccessful scenarios, and one participant commented that it *"ends the conversation quicker when it doesn't understand instead of stringing me along." (P76, Options vs. Top).* Participants also liked to have the "none of the above" option to explicitly exit a conversation: *' It at least did provide a way to say that it was on the wrong track: i.e. none of the above" (P117, Options vs. Out-of-Vocabulary Explanation).*

As discussed earlier, the status of a breakdown (successful/unsuccessful) affected participants' preferences. When the repair failed, Defer was a preferred strategy as a human agent is more likely to resolve a difficult issue. In contrast, if success can be achieved through a single repair, participants generally found the intervention of a human agent to be unnecessary. *"I liked the fact that the bot continued to try to work out what was being asked rather than immediately referring the user to a human agent, which defeats the purpose of the bot." (P77, Keyword Confirmation vs. Defer).*

*Assisted Self-Repair (**RQ1c**).* Repair strategies that aid with self-repair, by exposing the chatbot's understanding model, were generally ranked highly compared to strategies that provided no evidence of misunderstanding (Top) or simple acknowledgement (Confirmation). The qualitative results revealed several themes shared by these strategies. First, they provide actionable resources for the user to resolve the breakdown, either by avoiding undesirable words or using words more specific to the targeted intent when rephrasing: "*I really like seeing the keywords highlighted since it gives me insight into the logic behind the bot's responses, which will assist me if it does not provide the response I want." (P108, Keyword Highlighting vs. Repeat).*

Second, these strategies make the chatbot to appear more intelligent, not only by exhibiting its understanding capabilities, but also by showing pro-activeness to help repair: "*bot is interactive and appears to have interest in understanding question by asking questions to clarify." (P42, Keyword Confirmation vs. Repeat).*

Lastly, some participants noted an educational aspect, in that the explanations helped them better understand how the chatbot worked by *"teach[ing] you how to speak to the bot" (P151, Out-of-Vocabulary vs. Confirmation).* However, the explanation-based strategies were considered less natural as they did not resemble human conversations due to their use of GUI elements (e.g. highlighting words) that some participants found to be visually unappealing.

By directly highlighting keywords in the user's original utterance, Keyword Highlight Explanation was considered more intuitive in explaining how the underlying algorithm worked. However, the particular design decision of including a repetitive and verbose prompt that described the highlighting – *"I've highlighted keywords in your response..."* – was disfavored. Future work should consider removing description after the first few rounds of interaction. In comparison, Keyword Confirmation was more concise and appeared to be polite by verifying first, but it has the drawback of adding additional turns and user effort in order to respond to the confirmation. While Out-of-Vocabulary Explanation was perceived to be more explicit about its misunderstanding to help the user rephrase, some felt it appeared less intelligent if it could not understand common words.

Impact of Individual and Task Differences (RQ2)

In this section, we explore how the individual factors of social orientation toward chatbots, service frustration and efficiency, experience with chatbots, and experience with technology, as well as task variables of repair outcome (success/failure) and context (shopping/banking/travel), impacted preferences for the eight repair strategies. We rely on a statistical modeling approach. For each repair strategy, we selected all paired comparisons in which it appeared ($N \in [356, 389]$), then built a logistic regression model predicting whether it would be the winner or not by including the individual and task factors as independent variables. Thus, we ran eight logistic regression models. We focus on results that were statistically significant. We also tested preferences by gender and did not find any significant differences.

Social Orientation toward Chatbots. Social orientation reflects individual differences in the tendency to engage in human-like social interactions with chatbots, associated with a difference in mental model, of seeing agents as sociable entities rather than machines [23, 26, 27]. We found that participants with higher social orientation were significantly more likely to favor the Top strategy ($\beta = 0.39$, $SE = 0.11$, $p < 0.001$) and marginally less likely to favor Keyword Confirmation Explanation ($\beta = -0.18$, $SE = 0.10$, $p = 0.07$) or Options ($\beta = -0.22$, $SE = 0.13$, $p = 0.08$). These results are consistent with the notion that people with a high social orientation prefer natural conversations and may have felt the use of options and keywords to be mechanical. While we identified *naturalness* to be a desirable characteristic of repair strategies, it is likely to be preferred more by those with a high degree of social orientation toward chatbots.

Service Frustration and Efficiency. Lee et al. found that people with a utilitarian orientation preferred robot repair that provided instrumental value instead of emotional comfort [24]. In our study, participants with higher service frustration were marginally less likely to favor Keyword Confirmation Explanation ($\beta = -0.20$, $SE = 0.11$, $p = 0.06$), but more likely to favor Keyword Highlight Explanation ($\beta = 0.19$, $SE = 0.11$, $p = 0.10$). The difference between these two strategies is that the latter outputs a response directly and the former takes an additional turn to explain the understanding. Participants who are less patient with service interactions preferred a strategy that resulted in fewer turns, even while it may have appeared more mechanical and less polite.

Experience with Chatbots and Technology. Participants with more prior experience with chatbots were more likely to favor Confirmation ($\beta = 0.32$, $SE = 0.15$, $p = 0.03$), which intuitively makes sense as confirmations are commonly used in existing chatbot services. Participants with a greater level of technological experience were marginally more likely to favor Out-of-Vocabulary Explanation ($\beta = 0.29$, $SE = 0.16$, $p = 0.07$), indicating that designs that expose details of the underlying algorithms may appeal to more tech-savvy users.

Repair Outcome. When repairs were successful, participants were more likely to favor Top ($\beta = 0.45$, $SE = 0.22$, $p = 0.04$) and Repeat ($\beta = 1.38$, $SE = 0.22$, $p < 0.001$), and were less likely to favor Defer ($\beta = -1.37$, $SE = 0.22$, $p < 0.001$) and Keyword Confirmation Explanation ($\beta = -0.45$, $SE = 0.21$, $p = 0.03$). We conclude that simple strategies (Top, Repeat) are more acceptable if a repair can be achieved easily, while more complex repair strategies (Keyword Confirmation) or strategies requiring human intervention (Defer) may be more desirable in more difficult repair situations.

Task Context. We did not find any statistical differences across task context, suggesting that our findings on repair strategy preferences may generalize across different domains.

6 DISCUSSION

We first summarize design recommendations for repair strategies of chatbots. We then revisit the theoretical framework

and discuss how our results contribute to understanding of grounding in the context of human-agent conversation.

Design recommendations

Acknowledging Misunderstanding with Forthrightness and Less Redundancy. Our participants preferred repairs that explicitly acknowledge a breakdown, but complained that the repetitive acknowledgement to be "clutter" and "redundant." We recommend having alternative messages for acknowledging misunderstandings, while carefully setting the uncertainty threshold so that these acknowledgements do not appear overly frequently. For example, for individuals more tolerant of Top strategies, this threshold can be raised.

Explaining Models Naturally, Aesthetically, and Effortlessly. We show that explaining the mechanisms of the underlying models is considered helpful for repair, making the chatbot appear intelligent and teaching users better ways of interaction. While UI elements such as highlighting can be a powerful tool, one should carefully consider how to embed them in conversations so that they do not appear to be "mechanical," "unnatural," "visually unappealing," "hard to read" or "confusing" (some participants confused highlighted keywords with hyperlinks). Meanwhile, utilizing algorithmic inference and rich UI elements are ways to reduce user effort. We found that the Keyword Highlight Explanation was perceived as efficient by highlighting on the users' original utterance, saving a conversational turn. More advanced designs, such as suggesting words to use, may further reduce user effort.

Intelligently Repair with User Control. We show that repair works best when an agent can proactively suggest the correct action. In reality, to achieve such a level of intelligence requires significant effort in implementation, and even so it may fail at times. In the survey responses, some participants noted that the "None of the Above" option provides an explicit "way out" or "reset button." One of the canonical golden rules of user interface design is to provide a user with the control to permit a reversal of actions [38]. It is even more important in intelligent systems to always allow user oversight on system agency. Besides a way to exit, a user may also desire to control the triggering condition of a system repair, even to fine-tune the options (e.g., remove an unlikely option for future interactions).

Adapting to Individuals and Contexts. We observed that preferences for repair strategies are not universal. While it is useful to identify individual and task-related factors that impact preferences, one may also leverage the interactivity of an agent system to adapt to individuals and contexts through data or feedback-driven approaches.

Repair as a Collaborative Action with Costs

To guide the design choices of the repairs we studied, we used grounding in communication as a theoretical framework [10], which views conversations a collaborative action. Our results show that participants increasingly preferred strategies where the system provides increasing level of contribution to the repair process. Specifically, we considered three levels of contribution: 1) evidencing a breakdown; 2) providing resources to assist user-repair; 3) actively taking the initiative to repair.

In line with earlier work that built adaptive dialog systems based on grounding activities [5, 30, 31, 43], our empirical results support the point of view that grounding theory is a robust framework that can be applied from human-human to human-agent conversations. Core concepts such as *collective contribution*, *evidence of understanding*, *cost of repair*, are important to consider in designing repair capabilities of agents. However, the types of cost and their weights may change in the new context of agent conversations, resulting potentially different phenomena in choices of repair strategies. For example, we found that system (other)-repair was preferred over self-repair in our results, contradicting with observations from human-human conversations [10, 36]. One reason could be that fault cost (being perceived at fault), which one would try to minimize when talking to another person, is no longer an issue when interacting with an agent. Moreover, the design we presented, requiring a participant to only click an option, largely reduced formulation (rephrasing) and production (typing) cost compared to all the other repair strategies.

There is a caveat to our study, in that it did not capture all dimensions of cost in actual interactions. While we tried to control for the repair outcome in all conditions, a less capable chatbot may have a low chance of suggesting relevant intents, so a user may spend more effort having to re-try from the beginning for multiple times, than directly engaging in self-repair. This problem is relevant to "start-up cost" and additional "turn-taking cost" that are considered in the original grounding framework, but not captured in our study design.

Cost can also be used to interpret the impact of individual and contextual factors, by considering how they vary the weights of different costs. For example, an individual with high social orientation may consider "loss of naturalness" as an undesired cost, but those low on the orientation may assign little weight to such a cost. This explains why the former group was more likely to appreciate simple, natural repair strategies than the latter.

The notion of different costs can also direct us to consider new designs of repairs. For example, a simple improvement to explanation-based strategies is to allow users to easily

retrieve and edit previous utterances, reducing their production (typing) cost. Based on the idea of reducing turn-taking cost, another improved design is "type-ahead repair," by suggesting a potential breakdown and explanation *before* the user sends out the message.

By considering the dimensions of costs and benefits as the underlying mechanism, and how a specific design embodies them, one may start having a theory-guided framework to understand and predict user preferences for various designs of repair and broader conversational capabilities. While grounding theory enumerates a comprehensive list of costs regarding human communications, our work calls for further empirical investigation to establish a theoretical framework of grounding for human-agent communications.

Limitations

The results of this study are promising in delineating the best repair strategies for human-agent repairs. However, we acknowledge some limitations. First, for a lack of statistical significance, we could not make strong conclusions for how some of the lesser-ranked repairs fare against each other (Top, Repeat, Confirmation, Defer) given their larger p-values. However, by answering research questions guided by the theoretical framework, we believe that we paint an accurate high-level picture of preferred repairs in human-agent breakdowns. Second, limited by using a scenario-based experimental study, our work could not account for how user preferences for repair strategies are affected by nuances in system performance, such as confidence level and performance of the explanation methods. Future work should explore these questions with a real chatbot system. The study was limited by the fact that we only tested scenarios with a one-turn request-response task. Future studies can benefit from evaluating different kinds of user tasks, such as multi-turn conversations. Our study is also limited by our sample of Mechanical Turk workers. Due to the linguistic nature of our task, we desired to have fluent English speakers participate. However, our final sample was biased toward college educated males. Future work is needed to understand how repair strategy preferences differ across languages and cultures, which may have different expectations or norms for how humans ought to interact with conversational agents.

.

7 CONCLUSION

To design repair strategies for breakdowns of conversational agents, we consider key issues based on grounding theory in communication: evidence of breakdown, self- versus other-repair, and cost of repair. We provide a set of eight strategies that capture variances in these dimensions, including a group of novel repair strategies that explain the understanding mechanisms of the underlying model. We conducted a scenario-based study to compare preferences for these repair strategies, and analyzed the reasons behind and individual differences. Our results empirically validate theory-driven guidelines that recommend three levels of contribution from the agent to the collaborative action of repair: acknowledging potential breakdowns, providing resources to assist user repair, and proactively suggesting solutions. As a starting point, we encourage future work to develop a unified framework that guides the choice of repair strategies for different individuals and contexts.

REFERENCES

[1] Alan Agresti. 2003. *Categorical data analysis*. Vol. 482. John Wiley & Sons.

[2] Applied AI. 2016. Epic Chatbot / Conversational Bot Failures (2018 update). Retrieved Sept 10, 2018 from https://blog.appliedai.com/chatbot-fail/

[3] Ahmed Al Maimani and Anne Roudaut. 2017. Frozen suit: designing a changeable stiffness suit and its application to haptic games. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2440–2448.

[4] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[5] Susan E Brennan. 1998. The grounding problem in conversations with and through computers. *Social and cognitive approaches to interpersonal communication* (1998), 201–225.

[6] Bonnie Brinton, Martin Fujiki, Diane Frome Loeb, and Erika Winkler. 1986. Development of conversational repair strategies in response to requests for clarification. *Journal of Speech, Language, and Hearing Research* 29, 1 (1986), 75–81.

[7] 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*.

[8] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. 2009. A crowdsourceable QoE evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 491–500.

[9] Sylvain Choisel and Florian Wickelmaier. 2007. Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *The Journal of the Acoustical Society of America* 121, 1 (2007), 388–400.

[10] Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991 (1991), 127–149.

[11] Duncan Cramer and Dennis Laurence Howitt. 2004. *The Sage dictionary of statistics: a practical resource for students in the social sciences*. Sage.

[12] Herbert Aron David. 1963. *The method of paired comparisons*. Vol. 12. London.

[13] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.

[14] Sara Engelhardt, Emmeli Hansson, and Iolanda Leite. 2017. Better Faulty than Sorry: Investigating Social Recovery Strategies to Minimize the Impact of Failure in Human-Robot Interaction. In *1st Workshop on Conversational Interruptions in Human-Agent Interactions, WCIHAI 2017, Stockholm, Sweden, 27 August 2017*, Vol. 1943. CEUR-WS, 19–27.

[15] Dave Gomboc, Steve Solomon, Mark G Core, H Chad Lane, and Michael Van Lent. 2005. Design recommendations to support automated explanation and tutoring. *Proc. of BRIMS* (2005).

[16] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[17] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.

[18] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 468, 6 pages. https://doi.org/10.1145/3173574.3174042

[19] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 170 (Dec. 2018), 22 pages. https://doi.org/10.1145/3287048

[20] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 895–906. https://doi.org/10.1145/3196709.3196735

[21] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Bogliolo. [n. d.].

[22] Nancy Larson-Powers and Rose Marie Pangborn. 1978. Paired comparison and time-intensity measurements of the sensory properties of beverages and gelatins containing sucrose or synthetic sweeteners. *Journal of Food Science* 43, 1 (1978), 41–46.

[23] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010. Receptionist or information kiosk: how do people talk with a robot?. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 31–40.

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

[25] Yeoreum Lee, Jae-eul Bae, Sona S Kwak, and Myung-Suk Kim. 2011. The effect of politeness strategy on human-robot collaborative interaction on malfunction of robot vacuum cleaner. In *RSS Workshop on HRI*.

[26] Vera Q. Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do?: Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. 264–275.

[27] Vera Q. Liao, Muhammed Masud Hussain, Praveen Chandar, Matthew Davis, Marco Crasso, Dakuo Wang, Michael Muller, Sadat N. Shami, and Werner Geyer. 2018. All Work and no Play? Conversations with a Question-and-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 13.

[28] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297.

[29] 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*. ACM, 6.

[30] Tim Paek and Eric Horvitz. 1999. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In *AAAI Fall Symposium on Psychological Models of Communication, North*.

[31] Tim Paek and Eric Horvitz. 2000. *Grounding criterion: Toward a formal theory of grounding*. Technical Report. MSR Technical Report.

[32] 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*. ACM, 640.

[33] MIT Technology Review. 2016. 10 Breakthrough Technologies. Retrieved Sept 10, 2018 from https://www.technologyreview.com/lists/technologies/2016/

[34] MIT Technology Review. 2016. The Biggest Technology Failures of 2016. Retrieved Sept 10, 2018 from https://www.technologyreview.com/s/603194/the-biggest-technology-failures-of-2016/

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

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

[37] Marcos Serrano, Anne Roudaut, and Pourang Irani. 2017. Visual composition of graphical elements on non-rectangular displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4405–4416.

[38] Ben Shneiderman. 2010. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.

[39] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

[40] Vasant Srinivasan and Leila Takayama. 2016. Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 4945–4955.

[41] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 82–91.

[42] Indrani M Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. 2017. How do you want your chatbot? An exploratory Wizard-of-Oz study with young, urban Indians. In *Proceedings of the International Conference on Human-Computer Interaction (HCI) (INTERACT '17)*. IFIP, 20.

[43] David R Traum. 1999. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*. 124–131.

[44] Heather Turner, David Firth, et al. 2012. Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software* 48, 9 (2012).

[45] Eric W Weisstein. 2004. Bonferroni correction. (2004).

[46] Justin D. Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: Teaching Strategies for Successful Human-Agent Interactions. In *Proceedings of the 2019 ACM International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 12 pages.

[47] Joseph Weizenbaum. 1966. ELIZA - A computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[48] Yorick Wilks. 2010. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues*. John Benjamins Publishing Company, Amsterdam.

[49] Jason D Williams, Nobal B Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia, Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015. Rapidly scaling dialog systems with interactive learning. (2015). https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/iwsds2015.pdf

[50] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander Rudnicky. 2016. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 55–63.