BigBlueBot: Teaching Strategies for Successful Human-Agent Interactions

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

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

Narendra Nath Joshi
IBM Research AI
Cambridge, MA, USA
Narendra.Nath.Joshi@ibm.com

James Johnson
IBM Research AI
Cambridge, MA, USA
jmjohnson@us.ibm.com

Ingrid Lange
IBM Research AI
Cambridge, MA, USA
Ingrid.Lange@ibm.com

ABSTRACT

Chatbots are becoming quite popular, with many brands developing conversational experiences using platforms such as IBM's Watson Assistant and Facebook Messenger. However, previous research reveals that users' expectations of what conversational agents can understand and do far outpace their actual technical capabilities. Our work seeks to bridge the gap between these expectations and reality by designing a fun learning experience with several goals: explaining how chatbots work by mapping utterances to a set of intents, teaching strategies for avoiding conversational breakdowns, and increasing desire to use chatbots by creating feelings of empathy toward them. Our experience, called *BigBlueBot*, consists of interactions with two chatbots in which breakdowns occur and the user (or chatbot) must recover using one or more repair strategies. In a Mechanical Turk evaluation (N=88), participants learned strategies for having successful human-agent interactions, reported feelings of empathy toward the chatbots, and expressed a desire to interact with chatbots in the future.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Natural language interfaces**; *HCI design and evaluation methods.*

KEYWORDS

Explainable AI, conversational agents, Mechanical Turk

ACM Reference Format:
Justin D. Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: Teaching Strategies for Successful Human-Agent Interactions. In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Ray, CA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3301275.3302290

1 INTRODUCTION

Conversational agents have evolved tremendously over the past fifty years. Early visions of how humans would interact with machines in a natural manner [24] led to the first demonstrations of how the application of simple, declarative rules could emulate a Rogerian psychotherapist [41]. More recent improvements in areas such as natural language understanding, natural language generation, dialog management, and intent classification [4, 32, 38, 42, 47] have enabled the creation of commercial platforms for conversational agents [11, 12], enabling anyone and everyone to build their own chatbot.

Despite all of the progress made in AI and machine learning to enable mass-market chatbots, recent work that evaluated the experience of interacting with a chatbot has identified a gap between what people expect chatbots can do and what their technical capabilities actually are [13–15, 23, 25, 39]. People have very high expectations of chatbots, and they want chatbots that can add value to their life by making useful recommendations while being endowed with a sense of humor and possessing emotional sensitivities [39]. Examinations of interaction logs from real-world chatbots [2, 33] reveal that people try to speak to them as if they had human-level understanding. However, in a study of chatbots on Facebook Messenger, Jain et al. [15] found that participants complained that the "brains" of the chatbots were lacking. In other studies of conversational systems, people expressed

IUI '19, March 17–20, 2019, Marina del Ray, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302290

apprehension toward future use due to uncertainties of the chatbot's capabilities and issues with intent classification [18, 22, 25, 34].

As we do not expect the AI technologies that enable chatbots to make sudden improvements and achieve human-level understanding in a rapid timeframe, we believe chatbots and humans need to meet in the middle. We see an opportunity to use chatbots to teach humans how they work, why they fail, and how to best interact with them to achieve success. Given the surge of interest in explainable AI [1, 2, 9], and in chatbots as a new platform for customers to interact with businesses and drive workflows [8], we believe that finding ways to train humans in how to interact with chatbots can be very valuable.

We present *BigBlueBot*, a web experience that teaches people about chatbots, conversational breakdowns, and recovery strategies using two modes: experiential learning [19] through interactions with two chatbots, and conceptual learning [35], by using concrete examples to introduce the concepts of conversational breakdown and repair, followed by formal definition in personalized post-conversation summaries. We evaluated this experience via a Mechanical Turk study (N=88 participants) and found that many participants did learn strategies for how to have successful human-agent interactions. In addition, they reported having feelings of empathy for the chatbots and a desire to interact with chatbots in the future. We believe that by integrating "teachable moments" into the design of actual chatbots, in which a chatbot's understanding (or lack thereof) are explained, people may come to be more accepting of chatbots and have better experiences with them.

2 RELATED WORK

Our work explores a facet of explainable AI, focused on training humans to have better interactions with AI rather than understanding or visualizing low-level technical details of AI systems. We also leverage the notions of conceptual and experiential learning in the design of our learning experience.

Explainable Artificial Intelligence

In recent years, there has been a surge of interest in explainable AI [1, 2, 9], in which people are provided with an explanation for how an intelligent system arrived at a particular decision or conclusion. Prior work suggests that people face difficulties when they are not able to understand how a technology works (e.g., smart thermostats [45] or conversational systems [15]), which limits their usage of the technology to non-complex tasks [25]. In a desire to broaden the adoption of autonomous AI systems, Shneiderman et al. [37] argue that a level of human control ought to be maintained via interfaces that allow users *"to better understand underlying computational processes."* In line with this view, AI researchers have begun organizing their own work on machine learning systems around three dimensions – fairness, accountability, and transparency (FAT) – in order to ensure due process and understandability in decision-making [3].

Focusing on conversational systems, Chen and Wang examined the effect of prior experience and technical knowledge on the perceived usability of Siri [5]. They found that both factors were associated with increased perceptions of usability, with prior experience having the stronger effect. Regarding successful interactions with conversational systems, Porcheron et al. conducted a field study of interactions with Amazon Alexa inside the home and found a significant rate of dialog failures [30]. They attributed these failures to a lack of trouble indicators in Alexa's error messages. They remarked that when an error occurs, *"[Alexa] provides no mechanism for further interaction, and does not make available the state of the system, allying [the interface] with notions of a 'black box'."* This finding echoes a long-standing concern of the limitations of conversational agent interfaces: a lack of transparency of the system's model, status, and affordances leading to difficulties in learning how to use chatbots effectively [25].

There have been a few recent studies that examine transparency from novel perspectives. The Convey system [13] introduced an interface that displayed a chatbot's state of understanding in a persistent view, and it enabled users to directly edit that state when an error happened. Users felt that a chatbot with this level of transparency was easier to use, less mentally demanding, faster, and more intuitive compared to a regular chatbot. Another way to provide transparency is to make it easier for users to discover the capabilities of a chatbot. Valerio et al. catalogued several strategies used by real-world chatbots in order to make their capabilities transparent [40].

Our work approaches improving the transparency and explainability of conversational agents from a different perspective. Rather than explaining the technical details of how chatbots actually work, we instead aim to train users in how to effectively interact with them, especially when errors or misunderstandings occur.

Conceptual and Experiential Learning

Conceptual learning is an educational method in which learners create an understanding of a broad concept or principle via encounters with specific examples of that concept [17, 43]. It focuses on learning how to organize and categorize information over the recall of specific facts or the performance of methodological procedures [28]. Conceptual learning is often applied to science education (e.g., [35]) and has been shown to improve learning outcomes (e.g., [28]). We apply ideas from conceptual learning to the design of our experience by introducing the concepts we wish users to attain, such as the notions of conversational breakdown and repair, via specific examples. For example, Table 3 shows an example conversational snippet from our experience. The first utterance-response pair is an example of a conversational breakdown because the utterance was mapped to the wrong intent (the bot thought the human's question was about entry level positions instead of job qualifications). Thus, the concept of a breakdown is taught first via example; the language of "intent" and "breakdown" are introduced later.

Experiential learning is a process of learning through experience or learning by doing [19]. Experiential education immerses learners in an experience and then encourages them to reflect about the experience in order to develop new skills, attitudes, or ways of thinking [21]. *BigBlueBot* is just that – an experience in which users *learn* about chatbots by *doing* the work of chatting with chatbots (and even pretending to be a chatbot). In addition to the actual conversation, we capture the important aspect of reflection in experiential learning by including a post-conversational summary that explicitly defines concepts to cement learning outcomes.

3 UNDERSTANDING HOW TO WORK WITH CHATBOTS

Researchers have identified numerous reasons why people do not have successful interactions with chatbots [2, 14, 15, 25, 33]. One common reason is that people assume that a chatbot's level of understanding is on par with a human, and they do not understand failures such as when a chatbot is unable to parse an utterance, an utterance doesn't map to an intent, or an utterance maps to an incorrect or undesired intent.

U: I need to know if I would qualify for full time employment. I am a 70 year old male, I am a retired marine, and I have bad knees. A: Most entry level positions require no experience or minimal amounts of experience. Check out the information on our job description page. U: I also suffer from insomnia and don't sleep well at night. Is this a problem? A: <Transfer to human agent>
U: Good morning my name is Samuel and I applied for the baggage inspector position last month and I haven't heard anything yet. A: I'm sorry, I don't know how to answer your question.

Table 1: Fictional conversational breakdowns based on observed data from real-world chat logs [2]. U represents the human user and A represents the conversational agent.

In Table 1, we show two fictional examples of how these problems manifest, based upon actual examples observed in logs of deployed chatbots [2]. In the first conversation, the person has a question about eligibility for full-time employment. The person supplies a wealth of contextual information that would aid a human in answering the question, but for a chatbot, this information is extraneous as it has not been programmed to handle it. In the second example, the person's intent is expressed indirectly, as "I haven't heard anything yet." While a human would understand that this phrase meant that the person wanted to know the status of their job application, the bot was unable to discern that intent and defaulted to an apology response.

As we do not expect the quality of chatbots – both their technical underpinnings and their interaction design – to drastically improve in the near term, we set out to design an experience to teach people how to have better interactions with chatbots. By analogy, when people interact with a search engine and receive poor results, it is often the case that they revise their inquiry and search again. Yet, with chatbots, it seems that blame is placed upon the bot for not understanding [15, 25], rather than the person for a poorly-worded utterance. We believe an experience that can shift mental models of human-agent interactions toward one of cooperative effort (i.e., working with the bot to achieve a result) over pure utility (i.e., the onus of achieving a result always lies with the bot), would be highly valuable in helping improve the quality of chatbot experiences.

The primary concepts we aim to teach through our chatbot experience are: 1) chatbots best understand simple, plain language, with minimum required details, 2) breakdowns occur when a chatbot fails to understand a person's intent, and 3) chatbots have several strategies for recovering from breakdowns, such as showing options, asking to rephrase, or asking a clarifying question. Our selection of recovery strategies is based upon real-world observations of how intent-based chatbots recover from conversational breakdowns [2]. Our research questions focus on gauging the effect our experience has on teaching these lessons, as well as shaping peoples' feelings towards chatbots and their desire to have future interactions with them, especially for people who are not highly experienced in interacting with chatbots.

4 BIGBLUEBOT EXPERIENCE

BigBlueBot is a web experience in which people participate in two interactive conversations – Banking (as a chatbot) and Shopping (as a customer) – in order to learn more about how chatbots work. In each conversation, people are presented with one or more tasks to accomplish in order to achieve success. Each conversation is preceded by an introduction to set context and explain the goal. After each conversation, a summary is shown in which the person's actual conversation is augmented with explanations that highlight the conversational breakdowns that occurred and the strategies used to recover from them. As each conversation can be "won" or

I ran out of carrift last night. My friend paid for hiesard because I lost my perreave and all my carrift and learchink. I think it fell out of my gargemssheen when I sat down in the harlmalt. So, I need to figure out how to pallshoneos.
I ran out of cash last night. My friend paid for dinner because I lost my wallet and all my cash and cards. I think it fell out of my pants when I sat down in the subway. So, I need to figure out how to send him money.

Table 2: Nikita's query in the Banking conversation uses gibberish words to mask the true nature of her question. Underlines are used to show the correspondence between gibberish words and their actual counterparts.

U*: Where are my tee shirts?
A: We have a wide variety of tee shirts on sale! We have a range of simple, colorful tees and a large selection of graphic tees. Which would you like to learn more about?
U*: Something else
A: What can I help you with?
U: Look up my previous order
A: It looks like you want to look up a previous order. Which order would you like to look up? A) Crazy socks ordered three weeks ago. B) Sneakers ordered two weeks ago. C) Graphic tees ordered last week.
U*: The graphic tees
A: This order is scheduled for delivery by tomorrow at 5pm. Tracking #41220384.

Table 3: Example Shopping conversation. The first utterance triggers a breakdown in which Blu guessed the wrong intent, which is then followed up with the player aborting the current conversational sequence and beginning a new one. Asterisks (*) show utterances that were selected using fixed-response options.

"lost," either by completing or not completing the tasks, we refer to people going through the experience as "players."

Banking Conversation

In the Banking conversation, players are cast into the role of a chatbot and must help a customer, Nikita, with a banking question. The genesis of this conversation stems from conversational logs such as those shown in Table 1, in which people speak to chatbots as if they had human-level understanding. We specifically designed the Banking conversation to teach people the strategies that chatbots employ when encountering language they do not comprehend, such as out-of-vocabulary terms or extraneous context. We use role reversal [46], having players take the role of a chatbot, in order to help them develop feelings of empathy toward chatbots in their struggle to satisfy a "difficult" customer. Figure 1 shows screenshots of this conversation.

In order to demonstrate what a chatbot might "feel" when it encounters out-of-vocabulary terms or extraneous context, we desired to induce feelings of *incomprehension* in players. To do this, Nikita's question contains a large number of gibberish words, many of which provide irrelevant contextual details an actualc hatbot would not be able to interpret (Table 2). This conversation simulates how a chatbot might "think about" or "interpret" language outside of its vocabulary. In the UI, thought bubbles are shown above the player's head after each utterance (Figure 1) showing how the player (as a chatbot) might react to Nikita's utterances and what to consider for their next response.

When faced with Nikita's query, players must choose between three fixed-choice conversational responses, which employ a different strategy for handling her query. These strategies were adopted from examinations of how real-world chatbots respond to ambiguous queries [2]:

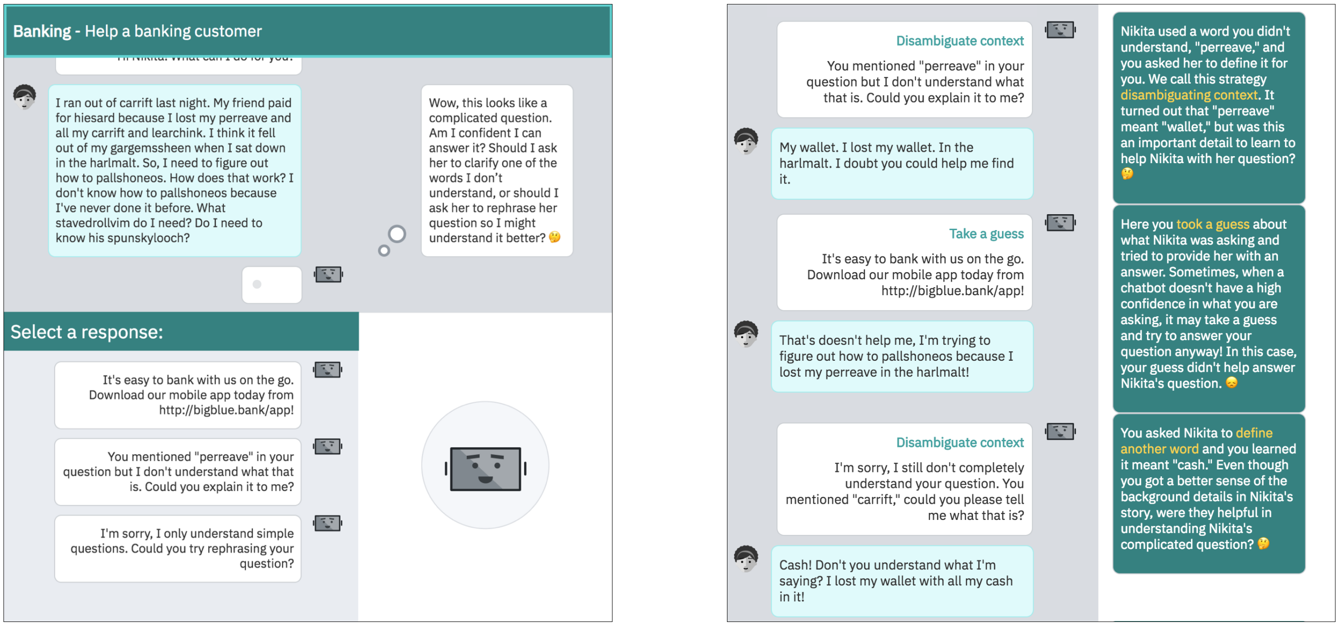
(1) **Take a Guess.** Players can attempt to guess the answer to Nikita's question, akin to how a chatbot might take a guess by choosing the highest-confidence intent that matched an utterance.
(2) **Disambiguate Context.** Players can ask Nikita to define one of the gibberish words in order to better understand the context of her question.
(3) **Simplify.** Players can ask Nikita to rephrase her question using simpler language in order to eliminate extraneous context.

In this conversation, players can select the first two strategies up to three times before Nikita gets frustrated and gives up. By asking Nikita to simplify her question twice [1], the true answer to her question is revealed. We note that although the strategy of disambiguating context does define one of the gibberish words, the words that are defined do not help in interpreting Nikita's question and do not lead to a successful outcome. This decision was deliberate as we wanted to teach players that even if chatbots perfectly understood context, that context may still not be useful for answering the question.

Shopping Conversation

In the Shopping conversation, players must accomplish three tasks by interacting with Blu, a chatbot for a fictitious online store: 1) check if an item is in stock, 2) check if a coupon code is valid, and 3) track a shipment for a previous order.

[1]Originally, players had to ask her to simplify three times, but initial lab-based usability testing showed a reluctance to do that due to feelings that Nikita would be frustrated by a chatbot that kept apologizing and asking her to simplify her query.



(a) Banking conversation. Players act as a chatbot trying to help the customer Nikita. Players must choose from a set of fixed responses during the conversation, which correspond to different strategies for handling Nikita's query.

(b) Banking summary. Players see in-context descriptions of the strategies they selected throughout the conversation, with prompts nudging them to reflect on whether they were (or were not) helpful.

Figure 1: Banking conversation and summary.

This conversation uses a mix of free-text responses and fixed-option responses, and it was implemented using the IBM Watson Assistant platform [11]. Fixed responses were used to force specific kinds of conversational breakdowns to occur, from which the chatbot recovers using different recovery strategies. Table 3 shows an example exchange in which the chatbot misclassified the person's intent. In this conversation, all players were exposed to the following utterance-level breakdowns: 1) the chatbot asked a clarifying question when an utterance was partially understood, 2) an utterance triggered the wrong intent, and 3) the chatbot gave options when it was unsure of which intent to trigger.

In addition to exposing players to conversational breakdowns that occur at the level of a single utterance, we also wished to expose them to breakdowns that occur across multiple conversational turns. To accomplish this, we leveraged the work of Sandbank et al. [33] in detecting "egregious" conversations – problematic conversations that may require human intervention. We implemented a rule-based classifier that tracked four features of egregious conversations. These features were scored after each conversational turn, and the UI displayed the current egregiousness level (low/medium/high). Egregiousness added another layer of difficulty to the conversation, as conversations were ended if they reached a high level of egregiousness.

(1) **Similar, repeated utterances.** When the player or the chatbot makes similar or repeated statements across the conversation, it suggests that the player is struggling in their interactions with the chatbot.
(2) **Asking for a human agent.** When a player explicitly requests intervention from a human agent, that request signals that the chatbot is unable to provide help.
(3) **Utterance does not match an intent.** When an utterance does not match an intent, chatbots typically respond with "I'm sorry, I don't understand" or "I don't know." When a player's input did not match the expected intent in the current node of the dialog tree, our chatbot generated such "I don't know" responses.
(4) **Negative sentiment.** Utterances with a negative sentiment (e.g., expressions of anger or frustration) signal that the chatbot is not performing its duties well.

In addition to a summary screen detailing conversational breakdowns, a separate summary screen was shown detailing how each utterance contributed to ratings of egregiousness.

5 STUDY DESIGN

We conducted an online study to evaluate the effectiveness of *BigBlueBot* in teaching people about chatbots, their limitations, and effective strategies for interacting with them.

Participants

Participants were recruited on Mechanical Turk with the requirement of being 18 years or older. We deployed a total of 162 tasks, but filtered out 74 of these (45.7%) due to participants spending less than 4 minutes on the task (13), skipping one or more of our surveys (54), or participating more than once (7). Our final sample consisted of 88 participants: 50 female (56.8%), 36 male (40.9%), 1 gender variant / non-conforming, and 1 transgender male. The majority of our participants were between the ages of 25-34 (40.9%) and 35-44 (39.7%).

In order to ensure participants spoke English, our task was published for workers in predominantly English-speaking countries (Australia, New Zealand, Canada, U.K., U.S.). All participants reported speaking English as their primary language.

Procedure

Participants began the study by reviewing and signing the informed consent form and filling out an initial survey. Next, they were directed to the *BigBlueBot* experience [10]. Participants completed the two conversations – Banking and Shopping – in a random order. After each conversation, participants reviewed the post-conversation summary, then filled out a survey. At the end of both conversations, participants filled out a final survey. All interaction data with the bots was logged on our server. The study took about 15-20 minutes to complete, and participants were compensated $2.50 USD for their participation (about $7.50 to $10 USD/hr).

Measures

We collected a combination of self-reported measures across four surveys, plus behavioral measures captured in our server logs. Likert scale items were rated on 5-point scales: "Strongly disagree" "Disagree," "Neither disagree nor agree," "Agree," and "Strongly agree." Factor analysis was used to ensure all items loaded on the same factor when building scales.

Experience with Chatbots & Voice Assistants. Overall, participants had more experience in using voice assistants than chatbots: 70 (79.5%) reported ever having used a voice assistant, but only 38 (43.1%) reported ever having used a chatbot. Of the participants who reported having used a voice assistant, 42 (60%) reported using it either 4-10 times (N=17) or 10+ times (N=25) in the past 30 days. For participants who reported ever having used a chatbot, only 7 (18.4%) reported using it either 4-10 times (N=6) or 10+ times (N=1) in the past 30 days. Thus, while many participants had experience with voice-based conversational systems, the majority had very little experience with using chatbots. Thus, our participants matched the types of people we aimed to target with *BigBlueBot*: people having little or no prior experience interacting with chatbots.

Learning. We assessed learning outcomes in multiple ways. First, since participants all had different interactions with the chatbots, we asked questions based on the content present in the post-conversation summaries. For Banking, we asked which strategy helped answer Nikita's question and what information Nikita sought. These questions were scored as either correct or incorrect. For Shopping, we asked about which breakdowns and recovery strategies were experienced, as well as how chatbots know when a conversation becomes egregious. We also assessed whether participants felt like they had learned something from the experience using a small, two question Likert scale: "I learned something from this experience," "I have a better understanding of chatbots from this experience." This scale had good reliability (Cronbach's $\alpha = 0.70$). Finally, we asked participants to explain what they took away from each conversation in their own words via a short, open-ended question they answered after interacting with each chatbot.

Feelings Toward Chatbots. Prior research in HRI has found that humans experience feelings of empathy when interacting with both simulated and physical robots [20, 27, 29, 36]. Generally, empathy is felt when a person has an experience of understanding another person's situation or circumstance [36], or when they share another's emotional state [16]. As role reversal is able to impart feelings of empathy [46], we felt that by casting participants in the role of a chatbot in the Banking conversation, participants would especially experience feelings of empathy toward chatbots due to their experience of the difficulties that chatbots encounter. In addition, we felt that these feelings would also increase participants' desire to interact with chatbots in the future.

We developed a 6-item Likert scale to measure empathy [2] motivated by the description given by Xu et al. [44]: *"The bot should give individualized attention to a user and make them feel valued."* Example items from this scale included, "[The bot] made me feel valued," "I felt sympathetic toward [the bot]," "I understood how [the bot] felt during the conversation." This scale had high reliability for both chatbots (Banking Cronbach's $\alpha = 0.84$, Shopping Cronbach's $\alpha = 0.91$).

In addition to empathy, we also assessed the degree to which participants felt the chatbots were helpful in accomplishing their tasks. If participants felt the chatbots were being helpful, it suggests that the chatbot's use of repair strategies is being noticed. We developed a 3-item Likert

[2]We considered using the scale developed to measure empathy toward robots by Seo et al. [36], but felt that the 24-item scale did not quite fit our experience and added too much response burden.

scale to measure helpfulness based on Xu et al.'s perspective [44]: "[The bot] helped me be successful," "It would have been easier to accomplish the task if this wasn't a conversation" (reverse coded), and "[The bot] provided helpful information." This scale had high reliability (Banking Cronbach's $\alpha = 0.78$, Shopping Cronbach's $\alpha = 0.82$).

We assessed desire to have future interactions with chatbots by having participants rate the item, "To what extent do you desire to interact with chatbots in the future?" on a 4-point scale ("Not at all", "Very little", "Somewhat", "To a great extent").

Enjoyment. To assess enjoyment of the overall experience, we asked participants to rate four Likert scale items: "I enjoyed this experience," "I would recommend this experience to others," "I was confused by this experience" (reverse coded), and "I was frustrated by this experience" (reverse coded). This scale had high reliability (Cronbach's $\alpha = 0.80$).

Behavioral Measures. We collected data on participants' interactions, including time spent on each conversation, outcome of that conversation (success/failure), number of subtasks completed successfully, the content of each conversation (e.g., participant utterance and chatbot response pairs, total number of conversational turns), and time spent reviewing each post-conversation summary.

6 RESULTS

We examine our data using a combination of qualitative and quantitative methods to build a comprehensive understanding of the impact of the *BigBlueBot* experience. In presenting our results, when a participant's comment is in reference to a specific conversation, we specify which one; otherwise it is considered a comment for the experience as a whole.

Enjoyment

Overall, participants expressed positive levels of enjoyment with the experience (M (SD) = 3.91 (0.89) of 5). Interestingly, enjoyment was not correlated with having successful outcomes with either chatbot (Banking $\rho = 0.08$, Shopping $\rho = 0.13$), but it was positively associated with feelings of learning from the experience ($\rho = 0.36, p < .001$).

Many participants made praiseful comments in the final survey when asked to provide optional, free-form feedback about the experience. Though it was an optional question, 19 participants (21.5%) responded with comments indicating they had enjoyed the experience and learned from it.

> *"Answering as a chatbot was more fun than I expected." (P18, F)*

> *"I would just like to thank you for this task. I thoroughly enjoyed it." (P31, F)*

> *"It was enlightening to be on the other end of the conversation." (P49, M)*

> *"It was helpful and I will use the advice [learned here] in the future." (P54, M)*

Task Outcomes

Participants attempted to complete a total of four tasks across the two conversations: answering Nikita's banking question, and getting answers from Blu on three shopping questions. Participants successfully accomplished an average of 2.5 tasks (SD = 1.1 tasks), and 18 participants (20.4%) completed all four tasks successfully. The Banking conversation had a higher success rate (44.3%) than the Shopping conversation (37.5%). Surprisingly, prior experience with chatbots/voice assistants did not predict task success (chatbot F [2,26] = 2.1, p = n.s., voice assistant F [3,26] = 0.88, p = n.s.), which is in contrast to previous results that suggest that prior experience with conversational agents leads to increased task success [5].

Chatbot Interactions

In the Banking conversation, we exposed participants to three different strategies for handling Nikita's question: taking a guess, disambiguating context, and asking to simplify. Taking a guess was the least favored strategy, with only 22 participants (25%) choosing this strategy at least once. Disambiguating context was the most popular strategy, with 64 participants (72.7%) choosing it at least once. Asking to simplify (the winning strategy), was chosen at least once by 61 participants (69.3%).

In the Shopping conversation, participants faced a mix of fixed and free responses when interacting with the chatbot. This conversation needed a minimum of 13 turns to be completed successfully, but because of the egregious classifier, the conversation may have ended sooner. In this conversation, the median conversation lasted 12 turns (M (SD) = 16.5 (16.6) turns).

Analysis of the content of the free responses gives some clues as to why conversations were unsuccessful. In some cases, participants got stuck in a state where they could not determine what to say to get the chatbot to respond appropriately. Signs that participants got stuck included typing random characters, (*"asdf", P87, M*), remarking that the conversation had become egregious (*"this is egregious", P87, M*), asking for a human, or insulting the chatbot (*"Forget this! You're worthless!", P41, M*). Interestingly, when these utterances failed to produce a desirable response from the chatbot, they were often repeated, with or without variation – *"You're stupid!"* was the next utterance after *"You're worthless!"*, and one participant said *"help"* seven times in a row *(P58, M)*.

We note that the design of conversational systems is a difficult task, and even for our own simplistic conversations, we were unable to fully anticipate the breadth of utterances participants would make.

Learning Outcomes

To understand what participants learned from the experience, we first examine how participants answered the four content questions designed to quantify learning. Overall, participants answered an average of 2.2 of 4 questions correctly (SD = 0.80 questions). For the Banking conversation, most participants correctly identified the best strategy to use (82.9%) and the information Nikita sought (76.1%). However, the Shopping conversation questions turned out to be more difficult; fewer numbers of participants correctly identified the two primary breakdowns experienced in the conversation (42%), and even fewer were able to identify the four signals of when a conversation became egregious (15.9%). One explanation is that the Shopping questions asked participants to select multiple options from a list; the first question had 2 correct options among 4, and the second question had 4 correct options among 7. Thus, for the rest of our analysis, we treat each selection of a correct option as contributing to a "learning score."

An examination of learning scores for Shopping (maximum value of 6) shows mean and median scores of 3.7 and 4.0, respectively. Thus, although fewer numbers of participants answered these questions perfectly, many were able to identify at least 3 (70.5%), 4 (55.7%), or 5 (37.5%) of the breakdowns and egregious signals.

Behavioral logs shed additional light as to why some participants were not as successful in answering the content questions. Overall, participants spent an average of 13.2 minutes (SD = 10.7) across the entire experience. Of that time, participants spent more time interacting with both chatbots (M (SD) = 10.7 (6.6) minutes) than in reviewing the post-conversation summary screens (M (SD) = 4.5 (4.3) minutes). Further, 9 participants spent less than 2 minutes across both summary screens, suggesting they gave only a cursory review or skipped those screens entirely. These participants had lower learning scores (M (SD) = 4.1 (1.8)) than participants who spent more time (M (SD) = 5.4 (1.7)), $F[1,86] = 4.7$, $p = .03$. Other factors, such as prior experience with chatbots, demographics, and the first conversation experienced did not significantly account for differences in the number of content questions answered correctly.

Having more interactions with the chatbot (e.g., by taking more conversational turns) did not impact learning in the way we anticipated. For both chatbots, the number of conversational turns was not correlated with the learning score (Banking $\rho = 0.03$, Shopping $\rho = 0.07$).

Despite the varied results in answering content questions, participants generally did walk away from the experience feeling that they had learned something (M (SD) = 4.29 (0.76) of 5).

Feelings Toward Chatbots

Consistent with prior HRI studies that show humans developing feelings of empathy toward robots after interacting with them [20, 36], we found the same result. Participants reported having feelings of empathy toward both the Banking (M (SD) = 3.4 (0.81) of 5) and Shopping (M (SD) = 3.3 (0.92) of 5) chatbots. Participants also felt the chatbots were helpful to them in accomplishing their tasks (Banking M (SD) = 3.1 (1.0) of 5; Shopping (M (SD) = 3.3 (1.1) of 5). Feelings of empathy were strongly and positively correlated with ratings of helpfulness (Banking $\rho = 0.67, p < .001$, Shopping $\rho = 0.60, p < .001$). In addition, helpfulness was strongly correlated with task success (Banking $\rho = 0.63, p < .001$, Shopping $\rho = 0.50, p < .001$).

We created two linear regression models to understand which behavioral measures accounted for empathy ratings, while controlling for the demographic factors. These models included age, gender, which chatbot was experienced first, the amount of time spent interacting with the chatbot, the number of conversational turns, the success outcome of the conversation (Banking: success or not, Shopping: number of tasks completed successfully), the rating of the chatbot's helpfulness, and the learning score. The model for Banking accounted for about half of the variance in the empathy scores (adjusted $R^2 = 0.50$), and the only significant factor was the helpfulness score ($\beta = 0.46, p < .001$). However, the model for Shopping (adjusted $R^2 = 0.45$), which operated more like a real chatbot, showed several significant or marginally significant factors: amount of time spent in the conversation ($\beta = 0.0007, p = 0.03$), number of conversational turns ($\beta = -0.01, p = .05$), helpfulness ratings ($\beta = 0.62, p < .001$), and number of tasks completed successfully ($\beta = -0.18, p = .07$). Although we anticipated that more interaction would lead to greater feelings of empathy, we see that the direction was negative, indicating that more conversational turns yielded weaker feelings of empathy. We believe this effect, though weak, was due to participants getting stuck in conversations and not being able to figure out what to say to move it forward. Thus, the signs of frustration we saw in the interaction logs are consistent with the negative relationship between conversational turns and feelings of empathy.

In addition to reporting feelings of empathy, a majority of participants expressed a desire to interact with chatbots in the future after completing the chatbot experience. Overall, 64 (72.7%) participants expressed the extent of their desire as either "somewhat" (N=48, 54.5%) or "to a great extent"

(N=16, 18.2%). Participants cited lessons they learned from the experience as contributing to their desire for having future chatbot interactions.

> *"I wouldn't mind interacting with one in the future now that I know they are more responsive to simple terms." (P10, M)*

> *"They can be helpful if they are given the right info. I will work with them in the future willingly." (P2, M)*

Lessons Learned from Chatbot Interactions

To create a deeper understanding of what participants learned from the experience, we conducted a qualitative analysis of the relevant open-ended questions in our surveys: what participants learned (in their own words) and how the experience affected their feelings toward chatbots. Two authors individually reviewed participants' responses and used an open coding [7] approach to extract themes. Codes were harmonized after two iterations of review and discussion, resulting in a final set of nine themes across three categories: strategies for having successful human-agent interactions, understanding chatbots' internal functioning, and recognizing a need for attitude change.

Strategies for Successful Human-Agent Interactions. Much of the prior work around chatbots focuses on identifying strategies chatbot designers can use to develop better chatbots [13, 15, 39]. In contrast, after going through the experience, our participants mentioned strategies that chatbot *users* can employ to have more fruitful conversations with chatbots. These strategies included specific lessons we aimed to teach in each conversation, such as using simple language and not providing unnecessary detail, as well as lessons picked up experientially: rephrasing a question to make it more understandable and precisely specifying intent.

Use Simple Language. A majority of the participants (73.8%) stated that asking their query in a concise manner by using simple words increased the likelihood of the chatbot understanding it.

> *"I learned that it's ideal to ask simple, to the point questions." (P4, F, Banking)*

> *"I learned that chatbots can provide helpful information when given simple and clear questions." (P29, M, Shopping)*

In the Banking scenario, participants were cast into the role of a chatbot and experienced how complex language and out-of-vocabulary words are interpreted as gibberish by a chatbot.

> *"To get the best out of a chatbot, it needs to be provided with simple sentences, complex language may lead to miscommunication." (P34, M, Banking)*

> *"It was interesting to see it from the side of the bot! The gibberish words are what the bot hears/reads." (P44, F, Banking)*

Do Not Provide Unnecessary Detail. In the Banking conversation, the initial query by Nikita was very long and contained many unnecessary details. From the chatbot's perspective (i.e., our participant's perspective as they were acting as the chatbot), these extra details were inconsequential in answering Nikita's real question (how to transfer money to her friend's account). Seventeen participants (19.3%) explicitly mentioned that providing extraneous information to a chatbot is unhelpful.

> *"I learned that sharing too much information can just confuse a chatbot." (P11, F, Banking)*

> *"...users can make it hard for the bot to understand even if they think they are helping by providing too much additional information." (P32, M, Banking)*

We are hopeful that these users learned to avoid the pitfalls highlighted in the first conversation listed in Table 1.

Rephrasing. A few participants (10.2%) specifically mentioned that rephrasing a question can make it more understandable.

> *"If you ask the bot a question in a different way sometimes it can better understand your needs." (P53, F, Shopping)*

> *"I can try similar words if I am not being understood." (P88, F, Banking)*

This strategy was explicitly used in both the Banking and Shopping conversations. For example, in the Shopping conversation, a fixed-choice utterance of "TAKE20" (an attempt to look up a coupon code) forces a conversational breakdown in which the bot responds with, "I'm sorry, I don't understand your question. Could you please rephrase it and ask again?" After this forced breakdown, participants were prompted for a free response, forcing them to think precisely about their intent.

Specify Intent Precisely. In the Shopping conversation, we aimed to teach participants explicitly about how chatbots attempt to discern *intent* from an utterance. For example, one intent is about validating a coupon code. Once the intent is correctly identified by the bot, follow-up questions are asked by the bot to drive the rest of the workflow (e.g., asking which coupon code to validate). Many participants (43.1%) commented on how they came to understand this model of chatbot operation. Moreover, participants expressed the notion of intent alongside several related adjectives, such as "concise" (5), "precise" (2), and "specific" (9).

> *"I learn[ed] that I must make my intent clear or else the bot will fail to understand me." (P3, M, Shopping)*

> *"It's best to express intent first and then follow up with the details." (P31, F, Shopping)*

Participants also recognized that the bot's failures were sometimes attributable to the human's inability to define the intent appropriately, and the onus was on the human to alter their interactions with the bot.

> *"I need to try to get to my point better. The bot can have a hard time trying to understand me." (P67, M, Shopping)*

Finally, one participant desired that the bot's vocabulary be made explicit during the interaction.

> *"I learned that there are keywords that can speed my interaction with a chatbot. I wish it would provide them upfront." (P74, F, Shopping)*

This capability is similar to the 'diagnostic' strategy adopted by many chatbots, wherein the chatbot explicitly states its capabilities upfront or when queried [15, 40]. Neither of the two chatbots in our study provided this capability.

Understanding Chatbots' Internal Functioning. The principle of transparency in explainable AI recommends that users have a basic understanding of the internal functioning of a system in order to enhance their experience with that system [1, 3]. In our study, participants stated that they got a glimpse of the internal workings of the chatbot, mainly with respect to how bots "thought" and how bots adopted specific strategies to recover from breakdowns.

Understanding the Thinking Process. As our experience specifically does not delve into the underlying machine learning algorithms behind natural language understanding and intent training and classification, participants' descriptions of what they learned about chatbots' operation used metaphors such as "thinking," "understanding," and "processing information."

> *"I learned how things look from the perspective of the chatbot." (P51, F, Banking)*

> *"[I learned] the process that chat bots use to decipher information that they do not understand." (P56, F, Banking)*

Participants understood the technological limitations of chatbots after going through the experience and recognized a need to adapt their communication style when interacting with them.

> *"I think it's hard to interact with them in some ways, because they can't understand everything that you ask." (P1, F)*

> *"In the future, I'll try to be more clear and use simple words and sentence structure when communicating with them." (P4, F)*

Understanding Breakdowns and Egregiousness. In the Shopping conversation, we explicitly highlighted several ways that conversations may degenerate and become egregious. As not all of these breakdowns were experienced by all participants during the conversation, we provided full descriptions of them in the post-conversation summary. Although few participants perfectly recalled the four ways our conversation may have become egregious (15.9%), a greater number (27.2%) reported in free response that they did learn about conversational failures and the methods to detect egregiousness.

> *"I learned more about how chat bots work and what causes them problems. I learned what causes conversations to break down and become 'egregious'." (P17, M, Shopping)*

> *"I learned that there was an algorithm to detect when a conversation with a bot is not going well... when a human agent needs to step in." (P51, F, Banking)*

> *"I learned a few different strategies that bots can use to prevent bad conversations." (P49, M, Banking)*

Recognizing a Need for Attitude Change. Previous research on conversational agents has identified a mismatch between peoples' expectations of chatbots and their real-world capabilities [2, 15, 25]. Participants recognized a need for adjusting their own attitudes toward chatbots after being exposed to their inner workings by recognizing that they are not human beings, they have limited capabilities, and patience is sometimes required when interacting with them.

A Chatbot is Not a Human. In human conversation, grounding describes the process in which each party establishes a mutual, shared understanding of the conversational context [6]. When a breakdown causes context to be misunderstood or lost, one party can initiate or request a repair in order to re-establish mutual understanding. In the Banking conversation, a majority of participants (72.7%) attempted to repair their impaired understanding of the full conversational context by asking Nikita to disambiguate one of the gibberish words. In human conversation, this strategy is entirely appropriate; when asked if you are feeling "swell," assuming you do not know the meaning of that word, it is natural to ask its meaning before committing to being (or not being) it.

Chatbots do not currently possess the capability for processing, understanding, and using complicated context in conversation that they have not been trained upon (e.g., as shown in the example in Table 1), and participants recognized that chatbots ought not be treated as if they had human-level understanding.

> *"I wanted to know the meaning of the entire request so that I could derive meaning from context. I learned that chatbot intelligence functions differently than human intelligence." (P74, F, Banking)*

Interestingly, in taking the role of a chatbot, one participant expressed a desire for their "human" partner to not speak to them as such.

> *"[Nikita] should not speak to me as if I were human." (P1, F, Banking)*

In addition, one participant felt that man and machine should meet into the middle and learn to speak each others' language.

> *"I supposed it is sort of a 50-50 - chatbots need perfecting and humans need to speak chatbot." (P44, F, Banking)*

Accept Chatbots' Limited Intelligence. Participants commented on the limited intelligence and the limited knowledge of the bots. As both conversations contained dialog failures, participants directly experienced the limitations that chatbots can have. In addition, our use of gibberish language in the Banking conversation showed participants just how limited a chatbot's knowledge base could be.

> *"I learned that these bots do not understand lots of words that we use in our everyday communications." (P55, F, Banking)*

> *"I learned that chatbots can really have a hard time understanding us humans, and it really can be like a whole different language." (P85, M, Banking)*

However, after the experience, 12 participants (13.6%) mentioned they came to accept the limitations of the chatbots and were more willing to work with them to achieve their goals.

> *"I've had great experiences with them so far and will love to use them as they become more and more improved and we learn how to speak to them." (P44, F, Shopping)*

On the other hand, 6 participants (6.8%) stated how they were frustrated with the chatbots, felt chatbots were not yet good enough to be useful, and/or they would prefer to deal with a human instead.

> *"The bot has a long way to go in order to be useful. I think I'd rather just deal with a human and bypass the bot." (P7, F, Shopping)*

> *"I think if my issues are simple I can successfully and regularly use a chatbot. But if my needs are complex I need to be able to get human help right away." (P72, F)*

Have Patience, Don't Get Frustrated. Eleven participants (12.5%) mentioned having patience, or not getting frustrated easily, as being a useful trait to have when interacting with chatbots.

> *"It can take patience, but [Blu] usually gets there." (P6, M, Shopping).*

> *"I learned ... not to get frustrated when it seems like it's not understanding exactly what I want." (P4, F, Banking)*

7 DISCUSSION

Our results suggest that people can be trained to have successful chatbot interactions by experiencing "broken" chatbot conversations. Many participants recalled the concepts they had learned – including those we desired to teach – in their own words in our surveys. Despite the fact that some participants were unsuccessful at accomplishing the prescribed tasks, many participants walked away from the experience with a sense of having gained knowledge, having feelings of empathy toward the chatbots, and having a desire to interact with chatbots in the future.

The Dual Role of Context

We observed a duality in the role that context plays when interacting with a chatbot. In the Banking conversation, a large number of participants (72.7%) opted to use the disambiguate context strategy in order to form a more complete understanding of Nikita's complicated question. This strategy seems to be a rational choice in human-to-human conversation; when establishing common ground, gaps can easily be filled at the request of the self or the other [6]. However, chatbots do not currently possess the ability to grasp the kinds of contextual details woven into Nikita's story. Thus, context seems to be at an impasse: on one hand, people have a desire to ask questions descriptively and with great detail (e.g., the first example in Table 1), but on the other hand, chatbots cannot currently process that information. By making this limitation transparent, participants learned not to include such detail when interacting with a chatbot. However, such self-censoring may reduce human-agent interactions to minimalistic, keyword-based searches, rather than back-and-forth conversations (as observed by Luger and Sellen [26]). Therefore, we propose a middle ground: in chatbot UI design, extraneous context can be highlighted or called out. This emphasis can have the dual purpose of showing users which portions of their query were not processed by the chatbot, as well as allow users to self-repair and rephrase their query with the pertinent portion of their context. In addition, users can also give feedback that their extraneous context *ought* to be understood by the chatbot. The implementation of such context highlighting can be performed using methods such as LIME [31], which determines and shows the features used for making textual classifications.

Impact of Perceived Helpfulness

Feelings of empathy toward the two chatbots were largely determined by the degree to which participants felt the chatbots were being helpful. Unlike other work that showed that perceptions of usability were largely dependent upon prior experience with a technology [5], in our case, prior experience with chatbots was not a predictor of task success, learning outcomes, or feelings of empathy. Although actual helpfulness ratings for both chatbots fell in the middle of the scale, the strong correlation between helpfulness and task success suggests either that participants who were successful felt the chatbots helped them achieve that success, or that participants who felt the bots were more helpful ended up being more successful. Interestingly, participants tended to feel less empathy toward Blu (the Shopping chatbot) when they were more successful. One explanation is that participants who were successful with Blu did not experience as many conversational breakdowns, and thus, saw fewer cases in which Blu "struggled" to repair those breakdowns. Thus, perceived helpfulness may be considered to be a design feature of chatbots. Chatbots may incorporate language suggesting their willingness and ability to help users in order to be perceived as more helpful, and possibly impart stronger feelings of empathy. These feelings, in turn, may increase users' willingness to continue interacting with chatbots when they encounter difficulties, rather than giving up. Future work is needed to understand the efficacy of perceived helpfulness on users' vigilance in completing tasks with chatbots.

Toward Explainable Chatbots

BigBlueBot was implemented as a standalone set of two conversations with chatbots. We propose that the core learning components of our experience – experiential learning from interacting with the chatbots and conceptual learning from exposure to specific examples of conversational breakdowns – can be integrated into real-world conversational agents in order to train users and explain *why* a chatbot responded the way it did. For example, when a chatbot is unable to map an utterance to an intent with a high enough confidence, instead of simply defaulting to an "I don't know" response, it can explain why it was unable to generate a meaningful response. Such explanations may highlight words in the utterance outside of the chatbot's vocabulary or highlight clauses that were not consequential in mapping the utterance to an intent (as previously discussed). Utterances that weakly map to multiple intents, resulting in the chatbot showing a short list of options, can highlight the words that caused each intent option to be identified. We believe these explanations might break down the "fourth wall" by exposing a chatbot's inner workings, making chatbot conversations feel even less like human-to-human conversation. However, as our results indicate that people desired to interact with chatbots after learning more about them, even when they had little prior experience, explanations of how a chatbot's AI works may make users more accepting of their failures.

8 LIMITATIONS

There are a few limitations we must acknowledge in our study. First, our measures focused on assessing participants' attitudes towards chatbots, such as whether they had feelings of empathy toward our chatbots and if they desired to interact with chatbots in the future. However, these attitudinal measures can not ascertain whether our participants really will seek out and engage with conversational services in the future, or whether their interactions with those agents will be more successful because of the lessons learned in this experience.

Second, our participants were drawn from WEIRD countries (Western, Educated, Industrialized, Rich, and Democratic). Although we desired to have native English speakers in our study due to the linguistic nature of our task, we recognize the need for a broader cultural examination before making any claims of the universality of our findings. Future studies ought to examine the role of chatbots in other cultures, especially those with lower literacy levels, in order to understand the applicability of the lessons we aimed to teach about using simple language, conversational breakdowns, and repair.

Finally, our work was developed for textual, intent-based, task-focused chatbots and may not generalize to other modes of interaction (e.g. voice) or other chatbot types (e.g. social or emotional chatbots).

9 CONCLUSION

We designed and implemented an online experience, *BigBlueBot*, to explain how chatbots work and teach strategies for having successful interactions with them. Participants in a Mechanical Turk study interacted with two chatbots and learned lessons such as using simple language, minimizing extraneous context, and making intents clear to have successful conversations. In addition, many participants developed feelings of empathy for the chatbots and expressed a desire to interact with chatbots in the future after viewing the world from a chatbot's perspective. Our study provides an example of how explainable AI need not simply focus on the technical details of how that AI is implemented; rather, another aspect of explainable AI relates to teaching humans strategies for interacting with AI systems to achieve successful outcomes.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[2] 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 (CHI '19)*. ACM, New York, NY, USA, Forthcoming.

[3] Solon Barocas, Sorelle Friedler, Moritz Hardt, Joshua Kroll, Suresh Venkatasubramanian, and Hanna Wallach. 2018. The FAT-ML workshop series on Fairness, Accountability, and Transparency in Machine Learning. (2018). http://www.fatml.org/

[4] Dan Bohus and Alexander I. Rudnicky. 2003. Ravenclaw: dialog management using hierarchical task decomposition and an expectation agenda.. In *INTERSPEECH*. ISCA.

[5] Mei-Ling Chen and Hao-Chuan Wang. 2018. How Personal Experience and Technical Knowledge Affect Using Conversational Agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18 Companion)*. ACM, New York, NY, USA, Article 53, 2 pages. https://doi.org/10.1145/3180308.3180362

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

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

[8] Matt Grech. 2017. The Current State of Chatbots in 2017. Retrieved Jan 5, 2018 from https://getvoip.com/blog/2017/04/21/the-current-state-of-chatbots-in-2017/

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

[10] IBM. 2018. BigBlueBot. (2018). Retrieved September 30, 2018 from http://bigbluebot.mybluemix.net

[11] IBM. 2018. Watson AI Assistant. Retrieved October 3, 2018 from https://www.ibm.com/watson/ai-assistant/

[12] Facebook Inc. 2018. Facebook Messenger Platform. Retrieved October 3, 2018 from https://developers.facebook.com/docs/messenger-platform/

[13] 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

[14] 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

[15] 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

[16] Joris H. Janssen. 2012. A three-component framework for empathic technologies to augment human interaction. *Journal on Multimodal User Interfaces* 6, 3 (01 Nov 2012), 143–161. https://doi.org/10.1007/s12193-012-0097-5

[17] Jacqueline J. Goodnow Jerome S. Bruner and George A. Austin. 1967. *A study of thinking.* Science Editions.

[18] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 506–516. https://doi.org/10.1145/2736277.2741669

[19] David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development.* FT press.

[20] Sonya S Kwak, Yunkyung Kim, Eunho Kim, Christine Shin, and Kwangsu Cho. 2013. What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *RO-MAN, 2013 IEEE*. IEEE, 180–185.

[21] Linda H Lewis and Carol J Williams. 1994. Experiential learning: Past and present. *New directions for adult and continuing education* 1994, 62 (1994), 5–16.

[22] Q. Vera 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)*. ACM, New York, NY, USA, 264–275. https://doi.org/10.1145/2901790.2901842

[23] 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.

[24] J. C. R. Licklider. 1960. *IRE Transactions on Human Factors in Electronics* HFE-1 (March 1960), 4–11.

[25] 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.

[26] 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*. ACM, 5286–5297.

[27] Carlos Garcia Martin. 2014. Factors that might affect empathy towards robots. (2014).

[28] Joseph D Novak. 2002. Meaningful learning: The essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners. *Science education* 86, 4 (2002), 548–571.

[29] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in Virtual Agents and Robots: A Survey. *ACM Trans. Interact. Intell. Syst.* 7, 3, Article 11 (Sept. 2017), 40 pages. https://doi.org/10.1145/2912150

[30] 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.

[31] 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 (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[32] Ronald Rosenfeld, Dan Olsen, and Alex Rudnicky. 2001. Universal Speech Interfaces. *interactions* 8, 6 (Oct. 2001), 34–44. https://doi.org/10.1145/384076.384085

[33] Tommy Sandbank, Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, John Richards, and David Piorkowski. 2018. Detecting Egregious Conversations between Customers and Virtual Agents. In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Vol. 1. 1802–1811.

[34] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[35] Phil Scott, Hilary Asoko, and J Leach. 2007. Student conceptions and conceptual learning. *Handbook of research on science education* (2007), 31–56.

[36] Stela H Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E Young. 2015. Poor Thing! Would You Feel Sorry for a Simulated Robot?: A comparison of empathy toward a physical and a simulated robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 125–132.

[37] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholoas Diakopoulos. 2016. Grand Challenges for HCI Researchers. *interactions* 23, 5 (Aug. 2016), 24–25. https://doi.org/10.1145/2977645

[38] Gabriel Skantze. 2003. Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems. In *Proceedings of the 2003 ISCA Error Handling in Spoken Dialogue Systems (EHSD)*. ISCA, 71–76.

[39] 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.

[40] Francisco AM Valério, Tatiane G Guimarães, Raquel O Prates, and Heloisa Candello. 2017. Here's What I Can Do: Chatbots' Strategies to Convey Their Features to Users. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*. ACM, 28.

[41] 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.

[42] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. *CoRR* abs/1508.01745 (2015). http://arxiv.org/abs/1508.01745

[43] Wikipedia contributors. 2018. Concept learning — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Concept_learning&oldid=822316973 [Online; accessed 5-October-2018].

[44] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3506–3510.

[45] Rayoung Yang and Mark W. Newman. 2013. Learning from a Learning Thermostat: Lessons for Intelligent Systems for the Home. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 93–102. https://doi.org/10.1145/2493432.2493489

[46] Dani Yaniv. 2012. Dynamics of creativity and empathy in role reversal: Contributions from neuroscience. *Review of General Psychology* 16, 1 (2012), 70.

[47] Steve Young. 1996. A review of large-vocabulary continuous-speech. *IEEE Signal Processing Magazine* 13, 5 (Sept 1996), 45–. https://doi.org/10.1109/79.536824