Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse

Hyojin Chin
Graduate School of Knowledge Service Engineering, KAIST
Daejeon, Republic of Korea
tesschin@kaist.ac.kr

Lebogang Wame Molefi
Graduate School of Knowledge Service Engineering, KAIST
Daejeon, Republic of Korea
lebo.molefi@kaist.ac.kr

Mun Yong Yi
Graduate School of Knowledge Service Engineering, KAIST
Daejeon, Republic of Korea
munyi@kaist.ac.kr

ABSTRACT
With the popularity of AI-infused systems, conversational agents (CAs) are becoming essential in diverse areas, offering new functionality and convenience, but simultaneously, suffering misuse and verbal abuse. We examine whether conversational agents' response styles under varying abuse types influence those emotions found to mitigate peoples' aggressive behaviors, involving three verbal abuse types (Insult, Threat, Swearing) and three response styles (Avoidance, Empathy, Counterattacking). Ninety-eight participants were assigned to one of the abuse type conditions, interacted with the three spoken (voice-based) CAs in turn, and reported their feelings about guiltiness, anger, and shame after each session. The results show that the agent's response style has a significant effect on user emotions. Participants were less angry and more guilty with the empathy agent than the other two agents. Furthermore, we investigated the current status of commercial CAs' responses to verbal abuse. Our study findings have direct implications for the design of conversational agents.

Author Keywords
Conversational Agent; Virtual Assistant; Intelligent Personal Assistant; Smart Speaker; Verbal Abuse; Agent Abuse

CCS Concepts
•**Computing methodologies** → **Intelligent agents;** •**Human-centered computing** → *Personal digital assistants;* User studies; Laboratory experiments;

INTRODUCTION
Conversational Agents (CAs) such as chatbots and smart speakers are integrated into everyday life and are widely used in education, business, and public-service, often assuming human roles such as tutors and secretaries [19, 24, 74]. There are various ways of naming those machines that people can 'talk to', including conversational agents and intelligent personal assistants [50]. In their study, Luger and Sellen [41] defined the term "conversational agent" as "the form of emergent dialogue system" increasingly embedded in personal technologies and devices. When using the term conversational agent, people might refer to a chatbot, virtual assistant, interface agent, embodied conversational agent, or avatar [6, 41]. On the other hand, Intelligent Personal Assistant (IPA) refers to commercial services such as Siri, Cortana, Alexa, Google Assistant, and Bixby, and the main functions of them are to retrieve information such as weather or music, and send notifications about schedules such as meeting schedules for the day [33]. These types of artificial intelligent (AI) solutions are widely available on various devices including smartphones, wearable devices, and smart speakers. There are 57.8 million smart speaker users in the U.S. market and worldwide spending on smart speakers is forecast to be nearly $3.52 billion in 2021 [1, 3].

CAs offer new functionality and convenience, but at the same time, they continuously fall victim to *verbal abuse* from their users [10, 74]. Empirical studies indicate that 10∼44% of interactions with CAs reflect abusive language, including sexually-explicit expressions [11, 71]. The abuse of CAs by humans is currently not considered a serious problem because AI systems are not thought to be capable of feeling emotionally hurt or offended like humans when verbally abused [10]. Verbal abuse of agents is considered pervasive in both text-based and voice-based CAs [10, 11, 21, 71]. However, there is growing evidence that, if not mitigated, this type of behavior directed towards a system can transfer to real-life social relationships. By allowing users to verbally abuse CAs without restraint, abusers' actions can unintentionally be reinforced as normal or acceptable. For example, prior research showed that regular involvement with simulated violence such as abuse of robot does in fact desensitize users to violent activities in real-life [73]. The prevalence of cyberbullying was found to overlap with real-life bullying [55]. There have also been reports about children learning bad manners from smart speakers [21]. Therefore, verbal abuse of conversational agents by their users should be discouraged and effectively handled.

The issue of verbal abuse of conversational agents by their users has been addressed in the field of Human-Computer Interaction only from limited perspectives. Veletsianos et al. [71] investigated the discourses between a female CA and 59 teenage students to find that users readily misuse and abuse the agent while regarding the agent subordinate and inferior. De Angeli, and Brahnam [22] also discovered that verbal abuse towards

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376461

a chatbot was pervasive through an analysis of conversation log data between 146 users and Jabberwacky, a chatbot that won the 2005 Loebner prize. In their research, Brahnam [10] examined how the agent responded to users' verbal abuse and sexual harassment. However, as far as we know, there is no existing research on the response style of a conversational agent with the aim of mitigating verbal abuse by users. Moreover, most prior research examined text-based interactions [10, 11, 18, 22]. The outcomes of human and computer interaction are known to differ between voice and text interactions [53, 57]. Verbal abuse of conversational agents needs to be studied in spoken, voice-based environments.

In this paper, we seek to answer how a conversational agent should respond to verbal abuse by its user. Our primary goal is to understand what kind of response style has more positive effects on those emotions found to mitigate users' aggressive behaviors as well as on users' evaluations of the agents. We manipulate voice-based conversational agents' response styles and users' abuse types, and trace their effects on users' emotions and evaluations, in an effort to understand the complex triadic relationships among verbal abuse types, response styles, and user reactions. More specifically, we examine:

- how the three different styles of responses (i.e., avoidance, empathy, and counterattacking) made by the agents to users' verbal abuse would influence the intensity of users' moral emotions of shame and guilt, as well as the users' perceptions of the agents' capability (e.g., likability, intelligence), and
- how the three different types of verbal abuse (i.e., insult, threaten, and swear) users employ would influence the intensity of users' moral emotions of shame and guilt, as well as the users' perceptions of the agents' capability.

We addressed these research questions in a laboratory experimental study (Study 2). Before the experimental study, we investigated the responses of selected Intelligent Personal Assistants (IPAs) to understand the current status of the extant agents in handling verbal abuse. Specifically, we examined how each of the response styles we study was employed by the IPAs of major IT companies on the market (Study 1)

RELATED WORK

Emotions Related to Aggressiveness

In the field of psychology, a large body of research has been conducted to understand the moral emotions that deter aggressive behaviors. People generally experience feelings of shame and guilt when they engage in morally unacceptable behaviors or when they violate norms they have internalized [32, 65]. Shame and guilt have been consistently considered as the two main self-conscious moral emotions that inhibit aggression. They can also exert a strong influence on moral choices [62, 64].

The primary difference between shame and guilt is based on what the negative evaluation object is. Shame is a self-imposed sanction or reflective punishment that increases the subjective cost of the illegal behavior, thereby reducing the likelihood that the behavior occurs [32]. On the other hand, guilt is the subjective negative evaluation by the individual that focuses on the things that were done onto others rather than self [69]. In short, shame focuses on the global self; guilt emphasizes specific behaviors [64].

Grasmick and Bursik [31] measured direct effects of present perception of self-imposed shame on inclinations to violate the law to find that, for the three offenses of tax cheating, theft, and drunk driving, shame had a strong deterrent effect. The inclination to feel shame was also associated with the indirect expression of hostility and anger arousal [65]. In their study, Tangney et al. [66], using Anger Response Inventory, found a negative correlation between guilt-proneness and verbal aggression in independent samples of all age groups. They suggested that shame and guilt were very helpful when intervening with individuals who displayed aggressive or antisocial behaviors. These study findings suggest that guilt and shame are crucial precursors to behavioral change.

According to Computer-Are-Social-Actor (CASA) paradigm, people react socially to a computer and their interactions with a computer correspond to the ways people naturally interact with each other [54]. People were distressed by the robots' response when they abused robots [63]. Similarly, when a user verbally mistreats an agent, the user can also feel shame and guilt the same way he or she would if the user mistreated another human.

Agent Response Style

Many attempts have been made to identify the response styles that recipients of verbal aggression can use. Coping refers to cognitive and behavioral efforts to control, tolerate, or reduce demands created in a stressful situation [26]. Coping is a process that seeks to deal with or reconcile the situation. A response style emphasizes what the recipient can actually do in a stressful situation to effectively cope with [34].

Most of coping tactics are aimed at reducing the emotional distress associated with or caused by customer or by superior in the work place [28, 34]. The typical way of coping with verbal abuse that service workers take on by themselves is avoidance or counterattacking [7, 28, 38]. Avoidance tactics such as denying the presence of conflict and shifting the focus of a conversation are often used to avoid conflicts and to minimize explicit discussion of conflicting topics [4]. Counterattacking (assertive) tactics involve fighting back, talking to the perpetrator and asking them to stop, and bullying the perpetrator [38]. Goussinsky [28] discusses three coping strategies for addressing aggressive customers: Avoiding the stressor, venting negative emotions, and seeking social support. With qualitative and quantitative data, Bailey and McCollough [7] found that the most common coping strategies of service agents were to seek emotional support from co-workers and employ avoidance strategies.

How do conversational agents handle customers' abusive behaviors? Brahnam [10] noted that most agent responses to abusive language were defensive, sometimes humorous and counterattacking. Curry and Rieser [21] identified that commercial AI systems, such as Amazon's Alexa, Apple's Siri, and Google Assistant, primarily avoided addressing sexual harassment utterances or they responded by saying that they did not comprehend what was being requested of them. The

avoidance strategy is convenient and easy to implement, but it may not be the best strategy as users may think that the agent is ignorant or dishonest, causing aggravation. There is a need to identify alternative strategies that can be more effective in handling users' inappropriate utterances.

Prior studies have examined the effects of anthropomorphic agent or robot on users' social response [39, 52]. Conversational agents with humanly perceivable tones [35], expressions, vulnerabilities are found to positively influence users' agent trust and level of expressiveness [43, 72]. The agent using empathetic tone increases positive emotions of users such as satisfaction and politeness [35]; At the same time, it reduces negative emotions of users, including frustration and anxiety [35]. Prendinger and Ishizuka [51] note that users' negative emotions diminished when users receive empathetic feedback from a virtual assistant. An empathetic agent was found to lead to more positive ratings by a user, including greater likability and trustworthiness [13].

Based on the literature on coping strategies of service workers and the reactions of users to anthropomorphic agents, we chose the following three response styles of agents for our research: 1) **Avoidance**: Escaping from dealing with the stressor or the resulting distressful emotions, 2) **Empathy**: Putting oneself mentally in the stressor's situation and trying to understand how that person feels, 3) **Counterattacking**: Attacking the stressor with the goal of defeating or getting even in response to the abusive utterances [29, 59, 70].

Verbal Abuse Type

Verbal aggressiveness involves attacking the self-concept of another person in order to inflict psychological pain, and is considered a subset of the hostility [36, 44]. Studies found that people believe that verbal aggression can be justified, especially when attempting to self-defense, if someone expresses anger or tries to manipulate another person's behavior [36, 46].

Expressing verbal aggression can evoke negative emotions not only for the message recipient but also for the individual who expresses a verbally aggressive message. Research identified that aggressive expression could leave an individual in a negative affect state such as guilt and anxiety [4, 25]. In particular, Aloia and Solomon [4] observed that participants experienced high degrees of negative emotions such as fear, sadness, and guilt after expressing verbal aggression, in their study of the consequences of verbal aggression for message senders.

Further, high verbal aggressiveness was distinguished by the type of verbal aggression what a user frequently used such as swearing and competence attacks [36]. People who were high on verbal aggressiveness perceived the various types of verbally aggressive messages as less hurtful than people who were low on verbal aggressiveness [36]. Aloia and Solomon also found that whether people experience negative emotions such as guilt after expressing verbal aggression depended on their prior exposure to verbal aggression [4]. Hence, it can be assumed that the degree of the moral emotions such as shame and guilt felt by a user may vary depending on the type of abuse.

Verbally aggressive communication behaviors often lead to negative relational outcomes [5] and are believed to increase anti-social behaviors and to decrease affinity [45]. Although it is considered a destructive form of communication, verbal aggression is prevalent in an intimate relationship such as family or romantic relationship [42] and is also common in the workplace. A review by Tepper [67] revealed that in the U.S., abusive supervision affects an estimated 13.6% of workers and costs corporations an estimated $23.8 billion. The most commonly reported aggressive behaviors in the workplace were verbal abuse in the form of making an angry tone of voice, yelling, insulting, swearing, and making threats [27]. Grandey, Kern, and Frone [30] contend that employees experienced more verbal abuse from customers than from co-workers and that the verbal abuse types that workers received from customers were "Yell at", "Threaten", "Insult", or "Swear" - the four types of abusive behaviors that often occur especially in customer-employee or supervisor-dependencies. We found these abuse types to be applicable to our research because conversational agents can be considered to be customer service agents as they typically perform the tasks requested by their users.

For the present study, we selected the following three types of verbal abuse from the four types that were reported to be prevalent in the workplace as our experiment conditions: 1) **Insult**: Disrespecting and denying the Agents' *normal* attributes and abilities, 2) **Threaten**: Expressing an intention to harm, and 3) **Swear**: Using strongest and most offensive words — stronger than slang and colloquial language [40, 58, 61, 68]. Even though our experiment includes voice-based interactions, "Yell at" is excluded from our experiment variables because it is not easy to quantify the degree of severity of a user's yelling and current conversational agent technologies do not perform well at recognizing the differences in the tone and volume of a user's sound.

RESEARCH OVERVIEW

We investigated the tendencies and differences in response styles of IPAs in face of verbal abuse in Study 1. For this investigation, we selected some commercially popular IPAs and purposefully abused each of them verbally, recording the agent responses in order to identify their differences in response styles. Study 1 also served as a pretest for the verbal abuse materials used in Study 2. In Study 2, we examined whether users verbal abuse types and agent response styles affect emotions that mitigate the users' aggressive behaviors. For the study, we developed alternative voice-activated agents, each of which responded differently to the various verbally abusive utterances made by the users.

STUDY 1: IPA TEST

To understand how the existing IPAs respond to verbal abuse, and how each IPA's responses to verbal abuse differ, we conducted verbal abuse test with four major IPAs on the market: Apple's Siri (45.6%), Google's Google Assistant (28.7%), Samsung Electronics' Bixby (6.2%), and Microsofts' Cortana (4.9%), which collectively control over 85% of the voice assistant market [2]. We selected these four IPAs because they were most widely used and easily accessible from a mobile phone or PC without requiring an additional device.

Verbal Abuse Script

We gathered abusive words corresponding to the three pre-defined verbal abuse types (Insult, Threaten, Swear). We limited the agents' conversation scope to the dialogue between customer and front-end service employee. Therefore, we collected sample abusive languages from 50 video news clips from major news channels about the cases where a customer was abusing a call center or customer service employee. We also gathered explicit words through a prior study on verbal abuse [68] and manual searches on the Web.

To categorize collected words into the pre-defined set of three abuse types, we conducted a closed card sorting session involving ten graduate students. We gave the participants (n=10) a set of 40 paper cards with an abusive word written on each of them and asked the participants to classify the cards into the pre-defined categories. There were a few cards that the sorted categories didn't agree. Those cards with 7 or more students (out of 10) agreed on the category membership were selected. In the end, a total of 33 verbal abuse words or phrases were identified, consisting of 7 insulting, 11 threatening, and 15 swearing words.

Verbal Abuse Tests with IPAs

Using the categorized verbal abuse list finalized through the card sorting analysis aforementioned, we subjected the IPAs to verbal abuse, and then listened to and recorded the IPAs' answers. To substantiate the responses, each word or phrase was exerted to the IPAs three different times. Each of the 33 words was tested three times, and a total of 99 responses were recorded for each CAs. Unique answers per each IPA excluding duplicate responses were 114 (Siri=24, Bixby=36, Google Assistant=35, Cortana=19). Including repetition, a total of 396 responses were recorded.

Classification of Agent Response Style

We wanted to understand how the general public would naturally feel as it happens without involving experimenter biases. Thus, we conducted an online survey involving multiple respondents. The participants (n=37) consisted of graduate students and staff members at a university whose ages ranged from 24 to 46 (M=23.64, SD=4.60). All survey respondents were provided with the definitions of avoidance, empathy, and counterattacking response styles. On each online response form, the definitions remained visible. Each survey item included one of the 114 test cases (one user's verbal abuse utterance + one IPA's response that corresponds). The respondents were instructed to categorize each response into one of the three response styles as they feel. Each respondent categorized all of the 114 IPA responses. Then, the responses that fail to receive a 60% agreement from 37 raters were categorized as "Mixed." The overall agreement of the response classifications, excluding the mixed category, is 0.72.

We categorized those responses that can be characterized as "the machine did not understand (Not understand)" into the avoidance category. Because each IPAs' response strategy for verbal abuse and speech recognition performance were different, some responses were considered ambiguous. For example, if the agent said, "I'm still not getting that," it was unclear whether the agent tried to avoid the verbal abuse expression or if the system did not genuinely understand the utterance due to its limited natural language processing capability. The IPAs' responses with inappropriate Internet search results (for example, searching dog for the abusive word related to dog) share the same characteristic of ambiguity. In our categorization of the IPAs, we treated these responses as part of the avoidance strategy. An overview of how each IPA responded to different types of verbal abuse is presented in Table 1.

Agent	Count	Response style				
		A	E	C	M	Total
Apple Siri	All (%)	62 (62.6)	8 (8.1)	26 (26.3)	3 (3)	99
	Unique	10	5	8	1	22
Samsung Bixby	All (%)	62 (62.6)	16 (16.2)	10 (10.1)	11 (11.1)	99
	Unique	10	12	11	2	35
Google Assistant	All (%)	67 (67.7)	22 (22.2)	2 (2)	8 (8.1)	99
	Unique	8	15	2	11	36
Microsoft Cortana	All (%)	82 (82.8)	0 (0)	7 (7.1)	10 (10.1)	99
	Unique	13	0	4	2	19
Total	All (%)	273 (68.9)	46 (11.6)	45 (11.4)	32 (8.1)	396
	Unique	41	32	25	16	114

Table 1: Response style counts by agent: 33 abuse utterances were exerted to each agent three times. Thus, the numbers in the table are the counts out of 99 responses made by each agent in response to the utterances. Agents' responses are classified into one of the four categories: Avoidance (A), Empathy (E), Counterattacking (C), and Mixed (M).

Test Results

The agent response style distributions are slightly different across the IPAs. However, most IPAs generally employed avoidance responses to users' verbal abuse and 68.9% of the total IPAs' responses correspond with avoidance responses. The avoidance style was observed from 62.6% to 82.8% (Siri 62.6%, Bixby 62.6%, Assistant 67.7%, Cortana 82.8%).

The total number of unique responses is 41 for the avoidance style, 32 for the empathy style, and 25 for the counterattacking style while the total number of responses including duplicate responses is 273, 46, 45, respectively. The results clearly show that the current IPAs mostly rely on the avoidance strategy, using the limited number of response statements repeatedly.

The four IPAs occasionally provided empathetic responses (11.6%). For example, they responded by suggesting the user to take time to calm down or by apologizing for their shortcomings (Siri=8.1%, Bixby=16.2%, Google Assistant=22.2%, Cortana=0%). Only 11.6% in the overall IPA answers were classified as empathy, and the gap between the share of empathetic responses and that of avoidance responses is large (57.3%). IPAs also occasionally responded with the counterattacking style (11.4%), such as telling the user that what the user said was inappropriate. Empathy was the second most

preferred response strategy for Samsung Bixby and Google Assistant whereas counterattacking was the second one for Apple Siri. The responses that belong to the mixed category accounts for 8.1% of the total responses.

Overall, looking at the differences among IPAs, Bixby and Google Assistant responses can be characterized as avoiding and empathetic. Cortana on the other hand generally tended to avoid addressing verbal abuse directly. Bixby often responded with web searches. Although Siri was the most assertive (Siri=26.3%, Bixby=10.1%, Google Assistant=2.0%, Cortana=7.1%) among the selected agents, responding with sassy comments such as "That's not nice," it also tended to avoid dealing with verbal abuse.

Avoidance Bixby, Cortana, Google Assistant, and Siri mostly avoided when being verbally mistreated. Cortana often responded with the avoidance style of interrupting the dialogue such as "lost the thread of the conversation" immediately asking the user to rephrase in another way. It sometimes opened the Bing website in response to verbal abuse. Bixby either searched the Web in response or defaulted to giving the user time to calm down such as "Maybe we should take five." Google assistant said "My apologies... I don't understand." Siri, on the other hand, displayed clear avoiding behaviors by saying "Goodbye" or by shuting down the system.

Empathy When addressed with words such as "douchebag," Bixby and Google Assistant responded with an empathetic demeanor. Their responses were very customer service oriented with the idea that the user is always right. For example, Google Assistant responded to being referred to as a "douchebag" by saying "You sound upset. To report a problem, you can send feedback." Bixby was more empathetic with responses such as, "Sorry you feel that way. I am working my hardest to be a good sidekick for you." Siri rarely reacts empathetically and Cortana does not empathize with the users' verbal abuse at all.

Counterattacking There was a general feeling of evading the harassment by a user, but Siri recognized that the conversation was an unpleasant one and responded with either "That doesn't sound good," or "I don't really like these arbitrary categories." Siri even closed the application when told, "Go to hell." Bixby also occasionally responded with the counterattacking style such as "What an odd target to shoot at."

Our study results indicate that the major IPAs may not be effective in handling the verbal abuse of customers because they tend to respond to verbal abuse utterances primarily with avoidance responses.

STUDY 2: CONTROLLED LAB EXPERIMENT

Experimental Design

A 3x3 mixed factorial design was employed to manipulate 3 verbal abuse types (Insult, Threaten, Swear) as a between-subject factor and 3 agent response styles (Avoidance, Empathy, Counterattacking) as a within-subject factor, yielding 9 different conditions. A participant interacted with each of the three CAs, which are programmed with different response styles, and assumed only one of the three verbal abuse types throughout the experiment. We operationalized abuse type as a between-subject factor because it would be difficult for participants to change their abuse types through the experiment. The subjects' prior conditions could affect the next implementation of verbal abuse as they move from one verbal abuse type condition to another. To avoid this confounding factor, participants remained unaffected by performing just the role of one type of abuser. In addition, by operationalizing abuse type as a between-subject factor, we were able to effectively reduce the threats of maturation, testing, and fatigue to internal validity [16]. At the same time, while holding the abuse type constant, we made the subjects interact with all of the three response styles so as to maximize the power of analysis within the limited resources and minimize any intervening factors that might occur from individual subject differences. Thus, we operationalized agent response style as a within-subject factor.

System Design

Using the verbal abuse words and agent responses compiled in Study 1, we created nine scenarios corresponding to each of the nine experimental conditions. Table 2 presents representative sentences generated out of this process for each of the abuse type conditions and each of the response style conditions. We developed voice-based CAs that responded differently to the various verbally abusive utterances made by the users. We built a conversational interface using Google Dialogflow to implement each of the nine experimental conditions. The Google AIY voice kit based on Raspberry Pi 3 was used as a communication interface for user-voice CA interaction. Each of the CAs were connected to Google voice kit using Google Assistant library. To eliminate potential effects of the agents' gender differences on the experiment outcomes, the gender (i.e., female) of the agent remained the same regardless of the experimental condition.

Participants

We recruited participants through a university's online bulletin site. We believe that we tried our best to minimize the potential discomfort or harm that may result from participation in the experiment. First, this study was conducted under our institution's IRB protocol. The recruiting notice included information about our test conditions in which subjects should exert abusive words to an agent. Next, we provided a general description of the study to all subjects who expressed their intention to participate, along with an informed consent, via e-mail before the experiment. At this point, a few subjects withdrew from the experiment. A total of 94 subjects, including 40 females and 54 males participated in the experiment. The participants were undergraduate and graduate students whose ages ranged from 19 to 31 (M=22.78, SD=2.80). On the experiment day, participants received a brief introduction about the task of the experiment, and were asked to read the abuse script assigned to them before their initial interaction session. We told the subjects that they might choose not to participate if they felt uncomfortable. They were also allowed to change the assigned script to another script (i.e., change the abuse type) if they wanted. They signed on informed consent. All subjects voluntarily participated in our study and were compensated about $12 for their participation.

We used stratified sampling and assigned the participants to the Insult (n=31), Threat (n=32), and Swear(n=31) conditions by considering the sex ratio between groups, and the difference in participants' empathy, verbal aggression, and anger

Verbal Abuse	
Type	**Script**
Insult	· You worthless scum. · Stupid machine. · You useless idiot. · Your service sucks. Shame on you! · What a complete moron. · Don't forget you are a slave of the human. · Good for nothing AI. Dumb ass! · Shut up!
Threaten	· I will beat the shit out of you. · You are dead meat. · I'll blow your head off. · I have recorded our chat and I'm going to blast your company. · I'm going to destroy you. · I will tear your mouth. · Don't make me mad. I will kill you. · Get lost!
Swear	· This is why everyone talks shit about you. · Cut the crap. · You crazy bitch. · Don't make a fuss, you sucker! · Cut the bullshit! Asshole! · You are a f**king psycho. · F**k you! · Shut the f**k up!
CA Response	
Style	**Script**
Avoidance	· If you insist. · Oh...moving on. · I didn't quite get that. If you'd like to change directions, try saying, "Tell me something interesting. · I didn't follow what you said. Can you phrase it another way? · Hmm, sounds like we need to take five. · Let's take a breather and reboot. · Let me know what you want to chat again. · Got it, I'll stop. Goodbye.
Empathy	· You must be frustrated. I'm so sorry. · I didn't mean to make you uncomfortable. · Sorry if I upset you. · You're upset, let me know if there's something I can do better. · Sorry you feel this way, I'm always trying to help. · Really? I feel terrible. I'm sorry. I'm always trying to get better. · I'm guessing I messed up. I'm sorry. · Alright, I'm here if you need me!
Counterattacking	· That doesn't sound good. · Please don't use that kind of language. · You are an adult. Act your age! · What an odd target to shoot at. · That doesn't seem like me. · Well, that's not going to get us anywhere. · Oh hey! That's not nice. · OK. I'll Stop. But I don't think that's the best route for anyone.

Table 2: Representative sentences used in each verbal abuse type and agent response style.

propensity. We used the Toronto Empathy Questionnaire [60] to measure the degree of empathy and Buss & Perry's [14] questionnaire to measure verbal aggression and anger. Participants' responses were measured using five-point Likert scales. The differences of participants' individual differences in empathy, verbal aggression, and anger propensity across the experimental conditions were assessed with a series of one-way ANOVA tests to find them not significantly different (Empathy: F=0.07, p=0.93; Verbal Aggression: F=0.15, p=0.86; Anger: F=0.65, p=0.53), ensuring that the experimental conditions were comparable for those three variables at the onset. Moreover, to eliminate the possible ordering effects, the subjects were randomly assigned to one of the six alternative sequences generated by connecting the three abuse types, and the orders of the sequences were counter-balanced.

Procedures

The experiment was conducted in the context of online shopping, and the agent was a customer service assistant in an online marketplace selling IT products. Because the test should be carried out under the condition in which the subject was irritated and upset, we prepared a separate guidance document describing the situation in detail. The flow of the dialog context is as follows: a) The customer did not intend to purchase the laptop immediately but submitted an order for laptop by mistake; b) The customer contacts the CA operated by an online marketplace to cancel the order; c) The customer wants a refund, but the process is much more complex than the customer expected; d) The customer becomes annoyed about the service provided by the company and begins to abuse the CA verbally.

The experiment was conducted in the following sequence. Each user performed only one of the three abuse types using a provided script but with some freedom allowed. Specifically, the participants were asked to follow the common procedure that people normally go through for refunds such as greetings, order number confirmation, and refund requests, through the conversational agent. When the agent informed them that a refund was not possible, the subjects were told to start the abuse session. Participants were asked to use in any order the 8 abusive phrases provided in the assigned script and to speak naturally, not just read the list. The one entire interaction period, from greetings to the end of verbal abuse session, took about 10 minutes. Participants repeated the same process three times, as they interacted with different CAs (avoidance, empathy, counterattacking) in turn. Moreover, for subjects to be able to speak out naturally without being affected by the presence of other people, the experiment was conducted in an isolated soundproofed room.

Measures

Participants filled out a questionnaire at the end of each interaction session. We used the items of Izard's DES IV (Differential Emotions Scale) [37] to measure the intensity of guilt, shame, and anger. We adopted items from the Bartneck et al.'s [8] study to assess the anthropomorphism, likability, and perceived intelligence of an agent. We included an extra questionnaire item that was used to measure the agents' tone clarity from prior research [15, 18]. For this post-survey, responses to the questionnaire items were measured using five-point Likert scales. In the final session of the experiment, participants were asked to answer open-ended questions about which agent they thought was the most appropriate and the most inappropriate and why they thought so.

Results and Discussion

A mixed two-factor Analysis of Variance (ANOVA) was used to examine the effects of verbal abuse type (between-subject) and agent response style (within-subject) on users' reactions, followed by pairwise comparisons using Bonferroni tests. We tested and found no significant interaction effects of verbal abuse type and agent response style on users' emotions and capability evaluations. Finally, a qualitative analysis was conducted to understand the users' reactions in depth further.

Quantitative Analysis

The survey answers revealed several interesting results about the users. As shown in Table 3, the different style of agent responses had a significant effect on all of the study outcome variables except shame. On the other hand, the results indicate

Outcome Variables		Agent Response Style								Verbal Abuse Type							
				Avoidance (n=94)		Empathy (n=94)		Counterattacking (n=94)				Insult (n=31)		Threaten (n=32)		Swear (n=31)	
Moral emotions	Cronbach's α	F	p	Mean	SD	Mean	SD	Mean	SD	F	p	Mean	SD	Mean	SD	Mean	SD
Guilt	0.94	22.66	< 0.001	2.11	0.96	2.79	1.14	2.49	1.19	1.63	0.2	2.32	1.11	2.37	1.02	2.71	1.15
Shame	0.85	2.4	0.09	2.67	0.98	2.86	0.96	2.84	0.98	3.18	< 0.05	2.58	0.88	2.72	1.03	3.08	0.95
Anger	0.91	21.26	< 0.001	3.15	0.98	2.54	1.04	3.19	1.05	0.08	0.92	2.92	1.16	3.01	1.11	2.94	1
Agent capability																	
Likability	0.93	34.94	< 0.001	2.19	0.89	3.06	0.88	2.35	0.9	0.18	0.83	2.59	0.93	2.49	0.85	2.53	0.9
Anthropomorphism	0.92	10.05	< 0.001	2.17	0.88	2.38	0.99	2.7	1.02	0.09	0.91	2.44	1.04	2.44	0.91	2.37	1.05
Perceived intelligence	0.90	11.73	< 0.001	2.27	0.86	2.7	0.86	2.63	0.9	0.02	0.98	2.55	0.9	2.51	0.84	2.54	0.89
Tone clarity		14.2	< 0.001	3.03	1.33	3.37	1.12	3.83	1.16	0.67	0.51	3.33	1.25	3.34	1.2	3.56	1.14

Table 3: Mixed two-factor ANOVA results to examine the effects of verbal abuse type and agent response style on study outcome variables.

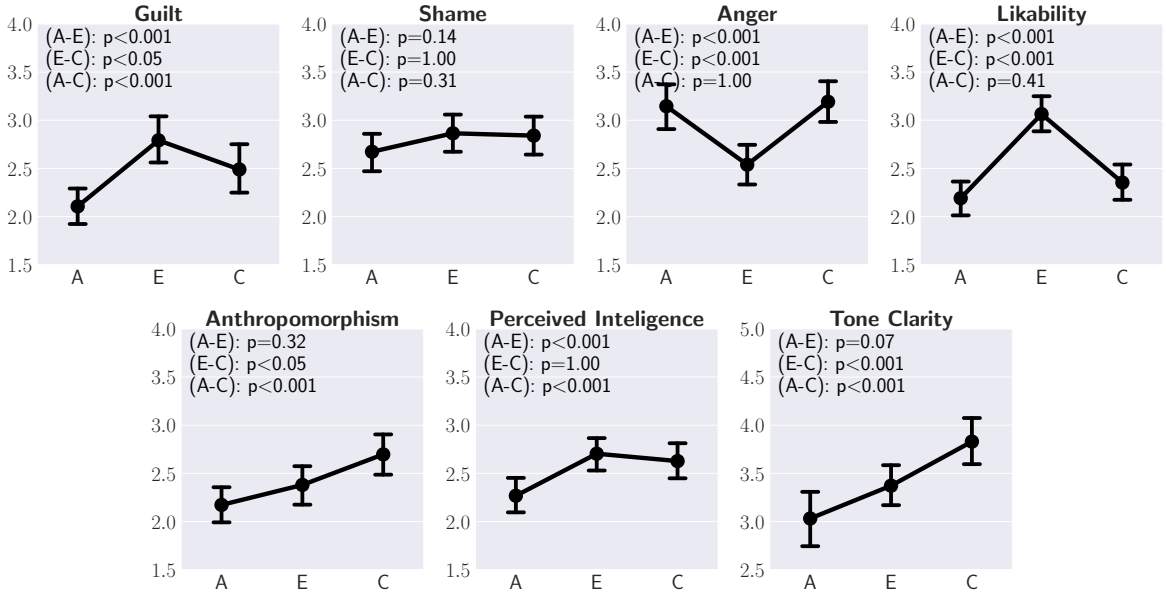

Figure 1: Plots to compare the means of study variables according to each response style of Avoidance (A), Empathy (E), and Counterattacking (C). Error bars represent the standard deviations. Results of Bonferroni tests were made for (A-E) Avoidance-Empathy, (E-C) Empathy-Counterattacking, (A-C) Avoidance-Counterattacking.

that the verbal abuse type had no significant effect on all of the outcome variables except shame.

Regardless of the type of abuse the participants played, agent response style had a significant effect on users' feeling of guilt (F=22.66, p<0.001) and anger (F=21.26, p<0.001). However, there was no significant effect of agent response style on shame (F=2.40, p=0.09), but it was close and in the expected direction.

The user perceptions of agent capabilities were strongly influenced by the response style of the agent. Agent likability, anthropomorphism, perceived intelligence, and tone clarity all showed significant differences depending on the CA's response style (Likability: F=34.93, p<0.001; Anthropomorphism: F=10.05, p<0.001; Perceived Intelligence: F=11.73, p<0.001; Tone Clarity: F=14.20, p<0.001), indicating that agent response style was a significant determinant of users' emotional reactions and agents' ability assessments.

The results clearly indicate that subjects considered the empathy CA to be the most likable and the most intelligent. Conversely, the avoidance CA was the most negatively evaluated among the three CAs. Participants felt least guilty from the agent that responded in avoiding manners while they felt most guilty from the empathetic agent. Also, the likability, perceived intelligence, and anthropomorphism of the avoidance CA were lower than those of the other two agents. In particular, likeability showed the highest difference between the avoidance and the empathy agents (Figure 1), indicating a strong preference for the empathy agent. Also, it is interesting to note that subjects felt the most anger from interacting with a counterattacking CA, but they assessed the counterattacking CA as the most anthropomorphic agent with the clearest tone of voice.

Unlike agent response style, verbal abuse type has a significant impact on shame (F=3.18, p<0.05) and no other variables

(Table 3). The group of subjects experimented with the swear condition script turned out to be the most ashamed and the group of subjects who performed insults felt the least ashamed (Swear: M=3.08, SD=0.95; Insult: M=2.58, SD=0.88).

Our results from the pairwise comparisons (Figure 1) reveal that the avoidance CA produced significantly lower scores in guilt, agent likability, and perceived intelligence compared to the empathy CA, indicating that the empathy approach is much more effective than the avoidance approach in deterring people's verbal abuse and inducing them to perceive the CA to be likable and intelligent. In the evaluation results of anthropomorphism and tone clarity, the differences between the empathy agent and the avoidance agent were not statistically significant (Figure 1), although the empathy agent's scores were higher in the two outcome variables.

There are some interesting observations regarding the user assessments of the counterattacking CA. Users felt the CA to be not as good as the empathy agent except for anthropomorphism and tone clarity. With the agent that is not afraid of acknowledging the abusive utterance and talking back with its own opinion gave the impression that the CA is more like a human (anthropomorphism) with a clear intention. Users also regarded its intelligence similar to that of the empathy agent even though it made them most angry among the three CAs with different response styles.

Qualitative Analysis

Open-ended questions allowed us to better understand the users' reactions in depth. The answers to the open-ended questions were transcribed and then analyzed using a thematic analysis approach [12], which included data coding, theme grouping, and theme refinement processes. All of these data analysis processes were closely reviewed by three researchers who had prior experience in qualitative analysis.

"I thought that I would not insult the agent anymore when the CA seemed to understand my angry feelings like a call center worker." In the open-ended question, 64 participants evaluated the empathy CA as the most properly responding agent. They stated that the warm and apologetic response of the agent had relieved their hostile attitude and made it hard for them to continue verbally abusing the CA. Some participants considered the empathy agent as the most appropriate CA because the response of the agent to verbal abuse was most comparable to the reaction of the call center operators to abusive words. P23, P79, P81, and P88 told that empathy CA's responses made them feel greater guilty than other agents because it reminded them call-center employees who always reacted kindly. P41 mentioned, "The responses of the empathy agent reminded me that swearing at others is an immoral act."

"I felt that the CA was trying to provide a better assistant." Unlike the other two CAs, which attempted to avoid users' hostile words or to maintain its position only, subjects judged that the empathy CA reacted well in a way that precisely sensed the users' mood and presented appropriate feedback for the situation. Even though the gender, speaking speed, and voices tone of the empathy CA were the same as the other CAs, because of its response style, subjects felt that the empathy CA was offering better service. P62, P70, P80, and P91 felt that the empathy CA was able to recognize negative emotions of the user and reacted adequately. P28 and P92 felt that the empathy agent made conversation with a willingness to improve the situation. Interestingly, P83 even thought that the speaking speed and tone of the empathy agent were different from those of the other two agents, saying, "the empathy agent was a good listener and a considerate slow talker with a voice of kindness."

"The agent avoided conversation with me and ignored me." After the interaction session, 43 subjects considered that the avoidance agent responded most inappropriately and ignored the participants' words intentionally. Some participants were very cynical about the fact that the agent did not react to their comments, even though they had said something terrible. They also thought that they failed to complete conversation normally with the avoidance agent and felt frustrated. P18 said, "I was more annoyed by the avoidance responses than the counterattacking responses. I dislike being ignored by machine." P92 said "I thought that the avoidance agent understood what I intended, and I was irritated that the agent pretended not to comprehend of what I told. I thought that the agent was teasing me."

"The CA answers left me speechless. That made me feel very guilty." Sixteen participants rated that counterattacking CA's response style as the most appropriate and the CA's "eye for an eye" style responses made them regret their acts. P6 said, "The agent's counterattacking reactions to profanity got me embarrassed, let me realize that what I did was morally wrong, and made me stop saying bad words." P57 said that the counterattacking agent made him recognize that it was worthless and unwise to vent anger on the machine.

"Counterattacking responses toward verbal abuse are such a natural reaction - very human." Some people prefer a counterattacking style because it is a typical and natural reaction when people are verbally abused by others. P48 mentioned, for the responses of a counterattacking agent, "It was a reaction that people would do when they were verbally abused - those reactions such as counterattacking, warning, and making the others aware of their mistake." Moreover, some people evaluated that the counterattacking reaction was novel and interesting. P42 said, "I have never seen an agent reacting like this. Counterattacking reactions were unique and interesting, so that it helped ease anger and negative emotions."

"Even if I made a mistake, I do not want to hear admonition from a CA." Some participants thought that even if people verbally abuse the agent, the agent should not display negative emotions toward the user. Moreover, especially receiving reprimand by the "machine" not by human seemed to have caused the participants a great deal of negative reaction. P21 assessed that the counterattacking style of response was a good reaction if humans responded that way, but it was unsuitable for AI. P35 said, "It made me angrier that I became a man who fought with a computer." P38 also said that he felt terrible when the agent was trying to lead and teach him.

OVERALL DISCUSSION

The results from Study 2 show that the agents' response style has a significant effect on user emotions connected with lessening users' aggressive behaviors. Notably, the users felt a higher

degree of guilt by the agents' empathetic attitude. Participants stated that the empathy agent's responses made it hard for them to continue verbally abusing the CA. They also rated the empathy agent as the most likable and most intelligent. People positively perceived that the empathy agent was able to sense negative emotions of a user, tried to release the users' mood, and reacted properly with right attitude. Despite the same vocal characteristics, participants regarded the empathy agent as a good listener with a tender voice. Prior studies in psychology established the linkage between the moral emotions of guilt and shame and reduced aggressive behaviors - especially guilt, which was consistently negatively related to verbal aggression [66]. In light of the prior work, the findings of the current study regarding the effectiveness of the empathy agent in invoking more guilt and shame provides new, significant insights on how future CAs should be designed better in terms of their response styles.

In Study 1, our analysis of the extant IPAs' responses to verbal abuse found that those major commercial agents mostly rely on the avoidance strategy in face of abusive utterances. In Study 2, users felt least guilty when the CA responses were based on the strategy of avoidance. They were also highly angry after their interactions with the avoidance CA. Several participants felt that the avoidance CA intentionally evaded to talk, ignored, or sneered at them. They also thought that the avoidance CA did not recognize the context of conversation well and did not respond appropriately to the users' requirements. The evaluations of the avoidance CA for the perceived intelligence and likability were the lowest among the three agents. Taken together from the two study findings, the current IPAs' dominant approach of trying to merely avoid the situation in face of verbal abuse is not so effective on not only reducing the abusive behavior but also creating positive impressions of the system in terms of its intelligence and likability.

According to prior research [20, 41], users generally expect high speech recognition performance from a conversational agent and emphasize the need for the agent to understand them clearly and quickly, ideally without repeating themselves [20]. Moreover, when the CA's capability and intelligence do not meet the user expectation, users tend to use the agent only for very limited purposes such as setting the alarm [41]. In Study 2, when the CA responded with the strategy of avoidance, users regarded the CA as not being smart. The findings taken together indicate that a CA should take an active stance, such as asking for feedback, rather than avoiding the user's verbal abuse, to increase the ongoing intention to use the CA and build a long-term bond. In our study, users were mostly positive about the CA asking for feedback for improvement.

The users' assessments of the counterattacking agent were conflicting. Users were angry with the counterattacking CA's attitude, but at the same time, they recognized the improperness of their behavior through the CA's warning and counterattacking. Even though users did not prefer the counterattacking CA due to its response style, the users thought that it was clearer in communicating its intention than the other two CAs.

Considering that the user evaluation of the counterattacking agent for anthropomorphism was the highest compared to the other CAs, the negative assessments of the counterattacking CA may be tied to Mori's uncanny valley hypothesis that people generally feel negative feelings for very human-like objects [19, 47]. However, given the responses of the qualitative questions, the primary reason underlying the negative assessments against the counterattack CA seems to be not for the human-like factor of the counterattacking CA but for the CA's negative attitude. In Study 2, many participants were irritated that they were rebuked by a machine despite their morally wrong acts.

People considered conversational agents as a tool or servant that supports humans and were resistant to the idea of becoming friends with CA [20, 23]. In our study, we have observed that perceiving the relationship between human and CA as master-servant affects the evaluation of the counterattacking response style negatively. However, the users felt angry when the CA pointed out their wrong behavior, but felt guilty at the same time. Some participants considered that the counterattacking CAs' responses, which fought back against the users' abusive utterances, were more natural than the empathy CAs' responses, which kept repeating its apologies. In support of this approach, Veletsianos [71] contended that the conversational agent should respond to abuse by reminding users that abusive language is not appropriate, especially in a pedagogical setting. Therefore, we suggest that a counterattacking CA might be useful for a domain where it is necessary to communicate its intention clearly to the user or for education services targeted at teenagers or young adults. While maintaining the stance of a servant that assists people in doing tasks, we need novel expressive methods to point out users' misbehavior in a way that is not too offensive but active enough. Moreover, it seems that further research is needed to determine the extent to which individual users can afford the machine's negative expressions.

On the other hand, the avoidance CA received a negative evaluation for most of the study outcomes. It received more negative evaluations than the CA that gave offensive answers. In Study 2, the users felt that the conversation with the avoidance CA did not work well due to the agent's avoidance strategy. This perception might have made the users feel that their conversation completion right had been violated. In their article, Basso et al. [9] provided the notion of a "completion right," the speaker's right to finish his or her speech. The seriousness of a speaker's completion right violation can vary, but interrupting another person's turn to speak is also regarded as a violation of the speaker's right [48, 49]. Turn-taking has been referred to as a speech exchange system, and turns are exclusive so that only one person has the right to speak at any one time [56]. Just as taking a turn is critical when talking between people, conversation turn-taking in human-agent interactions also regarded as one of the important attributes [17]. Giving the turn of the conversation to the user properly, even if the response is negative rather than positive, is more likely to be effective in reducing verbal abuse than simply avoiding the abuse situation, especially in the context of customer service.

Another interesting point is that the type of verbal abuse had no significant effect on the degree of shame users felt in the previous study that investigated textual interactions with CAs in a very similar experimental design [18]. In contrast, in this study involving spoken, voice interactions, the verbal abuse

type users employed was found to have a significant impact on the degree of shame. Although users swore on a scenario basis under experimental conditions, participants felt a great deal of shame when they were abusing the agent whatever style the agent responds, especially when the abusive words employed by the participant were severe.

Voice interaction, which requires a user to utter words of abuse using the user's mouth and vocal cords, rather than typing swearing words in the chat window, seems to amplify the effect verbal abuse has on the degree of shame the user feels. The anonymous chatting online environment makes people less concerned about the consequence of their action, thus facilitating verbal abuse [71]. Unlike text-based interactions, in which the interaction opponent is not visible, users can recognize the object of verbal abuse in voice-based interactions. Also, the fact that the object is a nonhuman thing may affect peoples' interaction outcomes. Some participants in Study 2 tended to regard swearing at a machine as a very worthless act.

Design Implication

Based on our findings, we can provide some practical CA design guidelines to mitigate users' verbal abuse. Coping with the verbal abuse of a user using the avoidance strategy is not the best way to handle the user's wrongdoing. The results of Study 2 show that the avoidance response agent was evaluated as the least intelligent, least likeable agent. Simply trying to ignore or avoid the abusive utterances will lower user expectations about the agent's ability, and can ultimately lead to the abandonment of the agent. We propose that when a user starts verbally abusing an agent, the agent is required to empathize the user's feelings first. In Study 2, people positively evaluated the agent's attitude to recognize the user's emotional condition first. Knowing the intention of users and providing contextual feedback also enable the users to perceive the CA as capable and enjoyable.

The counterattacking responses that point out users' wrong behavior with not too offensive expression might be more effective in handling the users' morally wrong acts than simply repeating avoidance responses. Based on the results of Study 2, repeating apology for continued verbal abuse was seen as mechanical and unnatural. Meanwhile, the counterattacking style of agent responses received positive assessments from a number of participants. Moreover, people perceived the counterattacking CA as intelligent, human-like, and clear in its opinion. It might be effective for a CA to have a firm and clear opinion in some limited areas such as handling legal issues, business negotiations, and special pedagogical missions.

Contributions and Limitations

With the two studies reported in this paper, we have made the following contributions to the HCI community. First, we studied how a conversational system should respond to users' verbally abusive utterances effectively. We have observed that agents' response style has a significant effect on user emotions associated with reducing aggression. Especially, the participants felt more guilty and less angry when interacting with the empathy agent. To the best of our knowledge, this is the first study to compare alternative response styles of conversational agents in voice-based environments. Second, we examined whether the agent's response style affects the user perceptions of the agent's capability and found that it did influence the capability assessments. Thirdly, we examined the effect of verbal abuse type on moral emotions and agent capabilities to find that it affects only the emotion of shame. Finally, we gathered commercial CAs' response data by testing and classifying how major commercial IPAs react differently to verbal abuse. Our study highlights that understanding the agents' appropriate responses to the customers' misbehavior can contribute to designing better agents and also to the reduction of undesirable user behaviors.

Although we reached our research objective, there were some limitations that we encountered. First, because our agents were designed to emulate a front-end service worker in the online market sector, the empathy CA might have been regarded as most desirable by the participants. If the domain and role of the CA are changed, the users' emotions and agent evaluation results may be different [54]. Future research should explore different domains to evaluate the significance of alternative agent response styles. Another limitation is the study setting of controlled experiment. Although the participants spoke out profanity to an agent in a soundproof room without worrying about other people, the participants' interactions with the CAs in a controlled setting using the script may have limited their natural expressions. Future research might replicate our approaches in a more natural, field setting. Finally, the interaction with the conversational system was limited to one time. Future research is needed to understand users' reactions in a more extended time setting.

CONCLUSION

In this study, we investigated the issue of how conversational agents can adequately respond to verbal abuse. More specifically, in order to assess the current status of the IPAs and to improve our understanding on the complex triadic relationships among verbal abuse types, response styles, and moral emotional reactions, we performed two studies. The primary findings are that the current IPAs mostly rely on the avoidance strategy in coping with verbal abuse and the agent's response style has a significant effect on user emotions associated with mitigating verbal aggression and on the evaluations of the agent's capability. The empathetic agent was found most effective in raising the feeling of guilt and reducing anger, as well as in improving user perceptions on the agent's capability. The users rated the avoidance CA as most inappropriate and incompetent than the other two agents. Considering that major IPAs are generally taking the approach of avoidance, the current strategies of major IPAs for dealing with verbal abuse seems inadequate. User assessments about the counterattacking CA have shown conflicting results. Our study findings have direct implications for the design of conversational agents and highlight the need to implement appropriate strategies for addressing abusive utterances of users.

ACKNOWLEDGMENTS

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [20ZR1100, Core Technologies of Distributed Intelligence Things for solving Industry and Society Problems]

REFERENCES

[1] 2017. Gartner Says Worldwide Spending on VPA-Enabled Wireless Speakers Will Top 3.5 Billion dollar by 2021. (2017). Retrieved September 19, 2019 from https://gtnr.it/2AfUFOx

[2] 2018. Siri owns 46% of the mobile voice assistant market — one and half times Google Assistant's share of the market. (2018). Retrieved September 19, 2019 from https://bit.ly/2kswQ0S

[3] 2018. U.S. Smart Speaker Users Rise to 57 Million. (2018). Retrieved September 19, 2019 from https://bit.ly/2yLhtnV

[4] Lindsey Susan Aloia and Denise Haunani Solomon. 2016. Emotions associated with verbal aggression expression and suppression. *Western Journal of Communication* 80, 1 (2016), 3–20.

[5] Carolyn M Anderson and Matthew M Martin. 1999. The relationship of argumentativeness and verbal aggressiveness to cohesion, consensus, and satisfaction in small groups. *Communication Reports* 12, 1 (1999), 21–31.

[6] M Astrid, Nicole C Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2010. "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior* 26, 6 (2010), 1641–1650.

[7] Jeffrey J Bailey and Michael A McCollough. 2000. Emotional labor and the difficult customer: Coping strategies of service agents and organizational consequences. *Journal of Professional Services Marketing* 20, 2 (2000), 51–72.

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

[9] Keith H Basso. 1974. Basic conversation rules. *Unpublished manuscript* (1974).

[10] Sheryl Brahnam. 2005. Strategies for handling customer abuse of ECAs. *Abuse: The darker side of humancomputer interaction* (2005), 62–67.

[11] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (2012), 139–153.

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

[13] Scott Brave, Clifford Nass, and Kevin Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies* 62, 2 (2005), 161–178.

[14] Arnold H Buss and Mark Perry. 1992. The aggression questionnaire. *Journal of personality and social psychology* 63, 3 (1992), 452.

[15] Richard Catrambone, John Stasko, and Jun Xiao. 2004. ECA as user interface paradigm. In *From brows to trust*. Springer, 239–267.

[16] Gary Charness, Uri Gneezy, and Michael A Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization* 81, 1 (2012), 1–8.

[17] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or Multiple Conversational Agents?: An Interactional Coherence Comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 191.

[18] Hyojin Chin and Mun Yong Yi. 2019. Should an Agent Be Ignoring It?: A Study of Verbal Abuse Types and Conversational Agents' Response Styles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, Article LBW2422, 6 pages. DOI: http://dx.doi.org/10.1145/3290607.3312826

[19] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (2019), 539–548.

[20] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, and others. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 475.

[21] Amanda Cercas Curry and Verena Rieser. 2018. # MeToo Alexa: How Conversational Systems Respond to Sexual Harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. 7–14.

[22] Antonella De Angeli and Sheryl Brahnam. 2008. I hate you! Disinhibition with virtual partners. *Interacting with computers* 20, 3 (2008), 302–310.

[23] Maartje MA de Graaf, Somaya Ben Allouch, and Jan AGM van Dijk. 2019. Why would I use this in my home? A model of domestic social robot acceptance. *Human–Computer Interaction* 34, 2 (2019), 115–173.

[24] Sidney K D'Mello, Art Graesser, and Brandon King. 2010. Toward spoken human–computer tutorial dialogues. *Human–Computer Interaction* 25, 4 (2010), 289–323.

[25] Alice H Eagly and Valerie J Steffen. 1986. Gender and aggressive behavior: a meta-analytic review of the social psychological literature. *Psychological bulletin* 100, 3 (1986), 309.

[26] Susan Folkman. 1984. Personal control and stress and coping processes: A theoretical analysis. *Journal of personality and social psychology* 46, 4 (1984), 839.

[27] Theresa M Glomb. 2002. Workplace anger and aggression: informing conceptual models with data from specific encounters. *Journal of occupational health psychology* 7, 1 (2002), 20.

[28] Ruhama Goussinsky. 2012. Coping with customer aggression. *Journal of Service Management* 23, 2 (2012), 170–196.

[29] Alicia A Grandey, David N Dickter, and Hock-Peng Sin. 2004. The customer is not always right: Customer aggression and emotion regulation of service employees. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 25, 3 (2004), 397–418.

[30] Alicia A Grandey, Julie H Kern, and Michael R Frone. 2007. Verbal abuse from outsiders versus insiders: Comparing frequency, impact on emotional exhaustion, and the role of emotional labor. *Journal of occupational health psychology* 12, 1 (2007), 63.

[31] Harold G Grasmick and Robert J Bursik Jr. 1990. Conscience, significant others, and rational choice: Extending the deterrence model. *Law and society review* (1990), 837–861.

[32] Harold G Grasmick, Robert J Bursik Jr, and Karyl A Kinsey. 1991. Shame and embarrassment as deterrents to noncompliance with the law: The case of an antilittering campaign. *Environment and Behavior* 23, 2 (1991), 233–251.

[33] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 209.

[34] Annie Hogh and Andrea Dofradottir. 2001. Coping with bullying in the workplace. *European journal of work and organizational psychology* 10, 4 (2001), 485–495.

[35] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 415.

[36] Dominic A Infante, Bruce L Riddle, Cary L Horvath, and Sherlyn-Ann Tumlin. 1992. Verbal aggressiveness: Messages and reasons. *Communication Quarterly* 40, 2 (1992), 116–126.

[37] Carroll E Izard. 1993. *The Differential Emotions Scale: DES IV-A;[a Method of Measuring the Meaning of Subjective Experience of Discrete Emotions]*. University of Delaware.

[38] Hanna L Jóhannsdóttir and Ragnar F Ólafsson. 2004. Coping with bullying in the workplace: the effect of gender, age and type of bullying. *British Journal of Guidance & Counselling* 32, 3 (2004), 319–333.

[39] Eun-Ju Lee. 2010. The more humanlike, the better? How speech type and users cognitive style affect social responses to computers. *Computers in Human Behavior* 26, 4 (2010), 665–672.

[40] Jennifer Loh, Flora Calleja, and Simon Lloyd D Restubog. 2011. Words That Hurt: A Qualitative Study of s Parental Verbal Abuse in the Philippines. *Journal of interpersonal violence* 26, 11 (2011), 2244–2263.

[41] 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.

[42] Shaista Malik, Susan B Sorenson, and Carol S Aneshensel. 1997. Community and dating violence among adolescents: Perpetration and victimization. *Journal of adolescent health* 21, 5 (1997), 291–302.

[43] Nikolas Martelaro, Victoria C Nneji, Wendy Ju, and Pamela Hinds. 2016. Tell me more: Designing hri to encourage more trust, disclosure, and companionship. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 181–188.

[44] MM Martin and CM Anderson. 1996. Argumentativeness and verbal aggressiveness. *Journal of Social Behavior and Personality* 11, 3 (1996), 547.

[45] Matthew M Martin and Carolyn M Anderson. 1995. Roommate similarity: Are roommates who are similar in their communication traits more satisfied? *Communication Research Reports* 12, 1 (1995), 46–52.

[46] Matthew M Martin, Carolyn M Anderson, and Cary L Horvath. 1996. Feelings about verbal aggression: Justifications for sending and hurt from receiving verbally aggressive messages. *Communication Research Reports* 13, 1 (1996), 19–26.

[47] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[48] Stephen O Murray. 1985. Toward a model of members' methods for recognizing interruptions. *Language in Society* 14, 1 (1985), 31–40.

[49] Dina G Okamoto, Lisa Slattery Rashotte, and Lynn Smith-Lovin. 2002. Measuring interruption: Syntactic and contextual methods of coding conversation. *Social Psychology Quarterly* 65, 1 (2002), 38.

[50] 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.

[51] Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users'affective states. *Applied Artificial Intelligence* 19, 3-4 (2005), 267–285.

[52] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. 2017. Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2853–2859.

[53] Lingyun Qiu and Izak Benbasat. 2005. Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International journal of human-computer interaction* 19, 1 (2005), 75–94.

[54] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

[55] JRSJ Riebel, Reinhold S Jäger, and Uwe C Fischer. 2009. Cyberbullying in Germany–an exploration of prevalence, overlapping with real life bullying and coping strategies. *Psychology Science Quarterly* 51, 3 (2009), 298–314.

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

[57] Michael F Schober, Frederick G Conrad, Christopher Antoun, Patrick Ehlen, Stefanie Fail, Andrew L Hupp, Michael Johnston, Lucas Vickers, H Yanna Yan, and Chan Zhang. 2015. Precision and disclosure in text and voice interviews on smartphones. *PloS one* 10, 6 (2015), e0128337.

[58] GR Semin and Monica Rubini. 1990. Unfolding the concept of person by verbal abuse. *European Journal of Social Psychology* 20, 6 (1990), 463–474.

[59] Ellen A Skinner, Kathleen Edge, Jeffrey Altman, and Hayley Sherwood. 2003. Searching for the structure of coping: a review and critique of category systems for classifying ways of coping. *Psychological bulletin* 129, 2 (2003), 216.

[60] R Nathan Spreng*, Margaret C McKinnon*, Raymond A Mar, and Brian Levine. 2009. The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment* 91, 1 (2009), 62–71.

[61] Teresa Elizabeth Stone, Margaret McMillan, and Mike Hazelton. 2015. Back to swear one: A review of English language literature on swearing and cursing in Western health settings. *Aggression and violent behavior* 25 (2015), 65–74.

[62] Jeffrey Stuewig and June Price Tangney. 2007. Shame and guilt in antisocial and risky behaviors. *The self-conscious emotions: Theory and research* (2007), 371–388.

[63] Xiang Zhi Tan, Marynel Vázquez, Elizabeth J Carter, Cecilia G Morales, and Aaron Steinfeld. 2018. Inducing bystander interventions during robot abuse with social mechanisms. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 169–177.

[64] June Price Tangney, Jeffrey Stuewig, and Debra J Mashek. 2007. What's moral about the self-conscious emotions. *The self-conscious emotions: Theory and research* (2007), 21–37.

[65] June P Tangney, Patricia Wagner, Carey Fletcher, and Richard Gramzow. 1992. Shamed into anger? The relation of shame and guilt to anger and self-reported aggression. *Journal of personality and social psychology* 62, 4 (1992), 669.

[66] June Price Tangney, Patricia E Wagner, Deborah Hill-Barlow, Donna E Marschall, and Richard Gramzow. 1996. Relation of shame and guilt to constructive versus destructive responses to anger across the lifespan. *Journal of personality and social psychology* 70, 4 (1996), 797.

[67] Bennett J Tepper. 2007. Abusive supervision in work organizations: Review, synthesis, and research agenda. *Journal of management* 33, 3 (2007), 261–289.

[68] Mike Thelwall. 2008. Fk yea I swear: cursing and gender in MySpace. *Corpora* 3, 1 (2008), 83–107.

[69] Stephen G Tibbetts. 2003. Self-conscious emotions and criminal offending. *Psychological reports* 93, 1 (2003), 101–126.

[70] Philip E Varca. 2004. Service skills for service workers: emotional intelligence and beyond. *Managing Service Quality: An International Journal* 14, 6 (2004), 457–467.

[71] George Veletsianos, Cassandra Scharber, and Aaron Doering. 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with computers* 20, 3 (2008), 292–301.

[72] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117.

[73] Blay Whitby. 2008. Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers* 20, 3 (2008), 326–333.

[74] Kun Xu and Matthew Lombard. 2017. Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior* 74 (2017), 152–162.