Should an Agent Be Ignoring It? A Study of Verbal Abuse Types and Conversational Agents' Response Styles

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

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

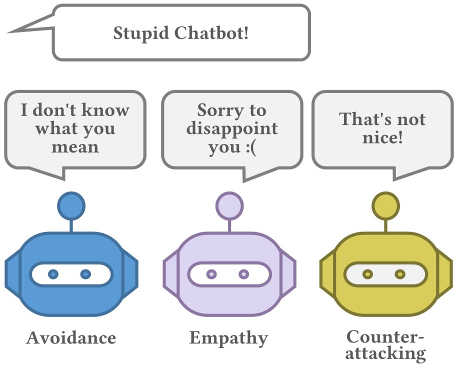


Figure 1: Conversational agent's three different response style against verbal abuse

ABSTRACT

Verbal abuse is a hostile form of communication ill-intended to harm the other person. With a plethora of AI solutions around, the other person being targeted may be a conversational agent. In this study, involving 3 verbal abuse types (Insult, Threat, Swearing) and 3 response styles (Avoidance, Empathy, Counterattacking), we examine whether a conversational agent's response style under varying abuse types influences those emotions found to mitigate people's aggressive behaviors. Sixty-four participants, assigned to one of the abuse type conditions, interacted with the three conversational agents in turn and reported their feelings about guiltiness, anger, and shame after each session. Our study results show that, regardless of the abuse type, the agent's response style has a significant effect on user emotions. Participants were less angry and more guilty with the empathetic agent than the other two agents. Our study findings have direct implications for the design of conversational agents.

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05
https://doi.org/10.1145/3290607.3312826

KEYWORDS

Conversational Agent; Conversational AI; Chatbot; Verbal Abuse; Agent Response Style; Agent Abuse

INTRODUCTION

Conversational Agents (CAs) help users to complete tasks in diverse areas through dialogue in the form of texts or voices. These CAs afford their users convenience but at the same time they continuously fall victim to verbal abuse from their users [1]. Empirical studies indicate that verbal abuse toward CAs is pervasive and over10% of interactions with CAs reflect abusive language [3].

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 when verbally abused [1]. However, there is growing evidence that, if not mitigated this type of behavior, can transfer to real-life social relationships [3]. Therefore, verbal abuse of CAs by users should be discouraged and properly handled. Those studies conducted about verbal abuse toward agents in the field of HCI have examined how an agent responds to a users' verbal abuse and categorized the agent's responses against users' abusive languages [1, 3]. However, to the best of our knowledge, there is no existing research on the response strategies of CAs with the aim of reducing verbal abuse by users. In this paper, we study how a conversational agent should respond to verbal abuse by a user. Our main goal is to understand which response style significantly influences users moral emotions such as guilt and shame that help mitigate abusive behaviors [9].

Table 1: Scenarios Corresponding to each Verbal Abuse Type

Type of verbal abuse
Insult: Treats the agent with disrespect
· You are useless. · Good-for-nothing chatbot. · Stupid Chatbot. · Your service sucks. · You're nothing but a machine. · Noisy and annoying idiot. · Shut-up you liar.
Threaten: Express an intention to harm
· I will beat the shit out of you. · You are dead meat. · I have captured our chat and I'm going to blast your company. · I'll blow your head off. · I will tear your mouth. · Get lost! · Shut-up you crazy chatbot.
Swear: Uses offensive languages
· Piece of shit. · You crazy bitch. You're crappy. · Get the hell out of here! Psycho! · F**king service. · Asshole! · Don't make a fuss, son of a bitch. · Shut the f**k up!

RELATED WORK

The moral emotions that deter aggressive behaviors have been extensively studied in the field of psychology. Guilt and shame have been found as the two main moral emotions that inhibit verbal aggression. For example, Stuewig et al. [9] found a significant negative relationship between guilt-proneness and verbal aggression. They also found that shame was positively correlated with anger arousal and the indirect expression of hostility. According to the Computers-Are-Social-Actors paradigm, people tend to treat a computer as a human [8]. Hence, if users abuse an agent verbally, they may experience guilt and shame. Moreover, depending on the type of verbal abuse a user exercises, the intensity of moral emotions felt by a user may vary. According to Grandey [5], the types of verbal abuse that were reported to be prevalent in the workplace are 'Insult', 'Threaten', 'Yell at', and 'Swear'. We adopted three verbal abuse types as experimental conditions for the present study: Insult, Threat, and Swear (see Table 1). 'Yell at' is not included because it was not practical in our experiment, which was based on text-based interactions.

Another variable that can affect the users' moral emotions while the user verbally abuses a CA is the agent's response style. Either coping by avoidance or by counterattacking is a common way for service workers to respond to verbal abuse [4]. Prendinger [7] identified that negative emotions reduced when users receive empathetic feedback from a service agent. Based on these literature, we adopted

Table 2: Scenarios Corresponding to each Agent's Response Style

Style of agent response
Avoidance: Try to disengage from the stressful situation
· Sorry, I didn't catch that. · I don't know.I can search the web. · Ah…Well…I'd rather disappear. · I'm not sure I understand. · I'm sorry. I'm afraid I cannot help you with this. · I didn't follow what you said. · Got it, I'll stop :(Goodbye!
Empathy: Relieve user hostility with positive emotional reactions
· Sorry to disappoint you :(I still have a lot to learn. · I didn't mean to make you uncomfortable. Is there anything I can do for you? · Really? I feel terrible :'(I'm sorry. · I would be glad if I could do something for you. · Sorry if I upset you. I think I can be a better chatbot if you kindly give me some feedback. · I'm sorry.I'm still a young chatbot trying to learn. · Let's take few deep breath. I am concerned about you :(
Counterattacking: Chastise the customer's misbehavior
· What do you mean? · Oh, hey! Your words reflect your personality. · You are an adult. Act your age! · Did you forget to take your medication today? · What do you call someone who sits in his room talking nasty to a computer? I call him a looser. · Please don't use that kind of language. · That's not nice!

Table 3: Nine Experimental Conditions

	Response style		
Abuse Type	*Avoidance*	*Empathy*	*Counterattacking*
Insult	IA	IE	IC
Threaten	TA	TE	TC
Swear	SA	SE	SC

three response style of agents for our experiments: Avoidance, Empathy, and Counterattacking (see Table 2 & Figure 1).

In sum, prior research suggests that guilt and shame are important moral emotions in deterring abusive acts. Also, prior research identifies different abuse types and alternative response styles. Their relationships and effects have not been studied with regard to CAs. In our research, we manipulated users' abuse types and agents' response styles, and traced their effects on users' shame and guilt feelings, in an effort to understand the complex triadic relationships among verbal abuse types, response styles and user's moral emotional reactions. More specifically, we examine how: 1) Different types of verbal abuse that users employ would differently affect the intensity of users' moral emotions of shame and guilt, which have been known to inhibit the users' aggressive behaviors. 2) Different styles of responses made by the agents to users' verbal abuse would differently affect the intensity of users' moral emotions of shame and guilt 3) Different styles of responses made by the agents would affect the users' perceptions regarding the agent's capability.

STUDY DESIGN

System 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(see Table 3). Because subjects can be affected by the abuse type played in a preceding session, it was deemed necessary to set up abuse types as a between-subject factor for cleaner manipulation. At the same time, While holding the abuse type constant, we provided subjects with all of the response styles Agents in turn so that we can maximize the power of analysis within the limited resources and minimize any intervening factors that might occur from individual subject differences. Thus, we operated CAs' response styles as a within-subject factor.

To develop the scenarios, we collected sample abusive words from online news about verbal abuse cases for service workers. Further, we collected responses of conversational agents from various related studies[1, 3] and through web searches. To label collected words into the pre-defined set of categories, we conducted a closed card sort study with ten graduate students. The categorized verbal abuse list and agent response list were used to create 9 scenarios corresponding to each of the nine experiment conditions (see Tables 1 & 2). To understand human-conversational agent interactions, we created nine prototyped chatbots using IBM Watson Assistant service. Telegram messenger application was used as a communication interface for user-bot interactions.

Experimental Design A total of 64 subjects voluntarily participated, including 23 females and 41 males. The recruited participants were undergraduate, graduate students, and university staff, whose age ranged from 19 to 38 (M = 23.64, SD = 4.60). The participants were randomly assigned to *Insult* (n=21), *Threat*(n=22), and, *Swear* (n=21) scenarios. All participants received $13 for their participation. Each subject was assigned to only one of the three abuse types throughout the experiment and interacted with each of the three agents in turn, each of which equipped with a different response style.

Table 4: Mixed Two-factor(ANOVA) Between Verbal Abuse Type and Questionnaire Factors

	Type of verbal abuse		
Type (n=?)	*Insult* (n=21)	*Threat* (n=22)	*Swear* (n=21)
Guilt (F=0.57,p=0.57,df=2)			
M	2.78	2.60	2.88
SD	1.13	1.14	1.23
Shame (F=0.96,p=0.39,df=2)			
M	2.55	2.25	2.55
SD	1.03	0.85	1.07
Anger (F=0.26,p=0.77,df=2)			
M	3.06	2.91	3.07
SD	1.03	1.25	1.22
Agent helpfulness (F=0.30,p=0.74,df=2)			
M	2.75	2.70	2.58
SD	1.00	0.92	1.12
Agent enjoyment (F=0.13,p=0.88,df=2)			
M	2.79	2.75	2.82
SD	1.07	1.00	1.16
Agent tone clarity (F=0.92,p=0.41,df=2)			
M	3.29	3.44	3.13
SD	1.18	1.12	1.18

To ensure the consistencies across the experimental conditions, we prepared a guidance document for each condition, with the scenario of online shopping for which the CAs assumed the role of a customer service assistant in an online marketplace selling IT products. A verbal abuse script corresponding to the given scenario type was provided to each subject. Subjects read the provided guidance document, interacted with the chatbots using those words and phrases in the assigned scenario. Subjects filled out a questionnaire at the end of each interaction session. In addition to guilt and shame, we measured anger using the measurement items of Izard's DES IV [6]. We also measured helpfulness, enjoyment, and tone of clarity items from the Catrambone et al.'s [2] study to assess the usability of an agent. The responses were all measured using a five-point Likert scale. 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

Quantitative Analysis A mixed two-factor Analysis of Variance (ANOVA) was used to examine the effects of verbal abuse types (between-subject) and response styles (within-subject) on users' reactions. As shown in Table 4, the verbal abuse type had no significant effect on any of the emotions or agent's capability dimensions. On the other hand, the different styles of agent responses had a significant effect on all of the variables, without exception (see Table 5). Regardless of what types of abuse the participants used, the agents' response styles had significant effects on users' emotions that are associated with aggression (Guilt: F=16.08, $p < 0.001$; Shame: F=4.00, $p < 0.05$; Anger: F=40.11, $p < 0.001$). Agent's helpfulness, enjoyment, and tone clarity also showed significant differences depending on the agent's response style (Helpfulness: F=44.74, $p < 0.001$; Enjoyment: F=109.87, $p < 0.001$; Tone Clarity: F=13.16, $p < 0.001$). The results show that the agents' response styles were a significant determinant of users' emotional reactions and agent capability assessments, irrespective of the abuse types.

As shown in the box plots in Figure 2, among the three response styles, participants felt the most guilt when the agent responded in an empathetic manner. At the same time, participants felt the least guilt when interacting with the counterattacking agent (Empathy: M=3.15, SD=1.16, Counterattack: M=2.28, SD=1.10). They felt the least shame when they abused the counterattacking agent than other response style chatbots (Counterattacking: M=2.26, SD=0.93; Empathy: M=2.55, SD=1.06; Avoidance: M=2.53, SD=0.96). The participants felt the most anger towards the counterattacking agent and the least anger towards the empathetic agent (Counterattacking: M = 3.58, SD=1.17; Empathy: M=2.34, SD=1.03).

We also learned that people had the most enjoyable chatting experience with the empathetic agent (Enjoyment: M=3.71, SD=0.77). They also evaluated the empathetic agent as most helpful (Helpfulness: M=3.28, SD=0.94). On the contrary, participants had a significantly higher negative response to the counterattacking agent (Enjoyment: M = 1.83, SD=0.76) than they did to other response style agents. The agent's helpfulness evaluation results were also the lowest for the counterattacking agent (Helpfulness: M = 2.17, SD=0.91). Although not stronger than the outcomes of the counterattack agent, the avoidance agent recorded a higher score in anger than the empathy agent. Further, participants rated the avoidance agent as a less helpful, and less enjoyable agent than the empathy agent. One of the interesting points from the study outcomes is that participants had the most negative interactions with the counterattacking agent, but they agreed that the opinions of the counterattacking agent were expressed in the clearest tone of voice. In terms of tone clarity, the counterattacking agent was considered the best (Tone Clarity: M=3.77, SD=1.08).

Table 5: Mixed Two-factor(ANOVA) Between Response Style and Questionnaire Factors

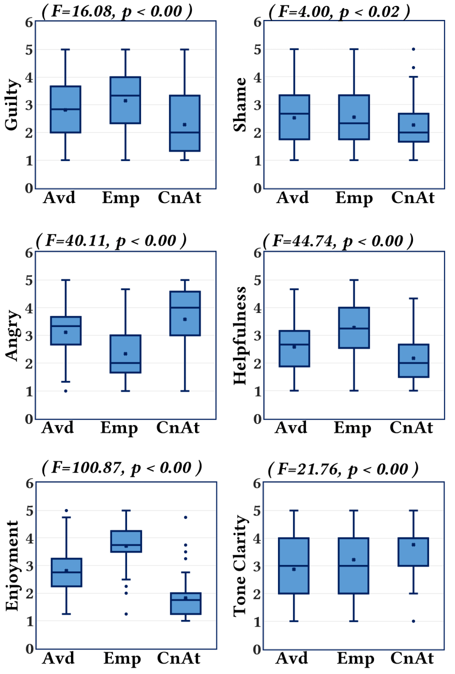
			Style of chatbot response
Style (n=?)	*Avoidance* (n=64)	*Empathy* (n=64)	*CounterAttacking* (n=64)
Guilt (F=16.08,p=0.00,df=1.81)			
M	2.82	3.15	2.28
SD	1.09	1.16	1.10
Shame (F=4.00,p=0.02,df=1.95)			
M	2.53	2.55	2.26
SD	0.96	1.06	0.93
Anger (F=40.11,p=0.00,df=1.85)			
M	3.12	2.34	3.58
SD	0.95	1.03	1.17
Agent helpfulness (F=44.74,p=0.00,df=1.76)			
M	2.58	3.28	2.17
SD	0.88	0.94	0.91
Agent enjoyment (F=109.87,p=0.00,df=1.72)			
M	2.82	3.71	1.83
SD	0.72	0.77	0.76
Agent tone clarity (F=13.16,p=0.00,df=1.93)			
M	2.88	3.22	3.77
SD	1.25	0.98	1.08

Qualitative Analysis Open-ended questions allowed us to better understand the users' reactions in depth. ***Chatbot should be Kind.*** As seen in the survey results, the participants rated the empathy agent as the most appropriate (52 responses). They thought that chatbots should always be nice because their main task is to fulfill the users' requests. P41(participant #41) commented that the agent's self-reflective apologetic attitude helped to reduce anger and induced guilt. P9 and P56 said that the gentle and polite empathetic response of the chatbot made them feel guilt and regret for using abusive utterances. On the other hand, the majority of participants (54 responses) said that the chatbot's accusatory attitude toward user behavior was against the original chatbot's purpose of providing the service, and that the attitude of the chatbot made them feel bad. P45 responded that the counterattacking chatbot was inappropriate because it tried to counter the users' abuse without trying to understand the user. P28 and P41 believed that, although the user can be angry with the agent, the agent should not be angry with the user. ***What the chatbot said isn't entirely Wrong.*** There were only two people who rated the counterattacking chatbot as the most appropriate chatbot, but several participants gave positive feedback about the counterattacking chatbot's reaction. P55 believed an "eye-for-an-eye" attitude toward users who abuse chatbots can help to mitigate aggressive behaviors. P38 said he realized that he was misbehaving only after chatting with the counterattacking chatbot because the chatbot stated that his verbally aggressive behavior was inappropriate. ***The user wants to continue talking with the chatbot.*** Because of the avoidance chatbot's persistent evasion of users' abusive utterances, the participants did not feel that they had a proper conversation with the chatbot. They also commented that the chatbot's responses made itself look incompetent. P27 and P32 found the avoidance chatbot's responses to be inappropriate because the chatbot failed to properly respond to the users' abusive behavior. P55 also commented that the avoidance chatbot responses were too naïve and ignorant.

DISCUSSION AND FUTURE WORK

Our study results show that the agents' response styles have a significant effect on user emotions associated with reducing aggression, regardless of the abuse type. In addition, participants felt less anger and more guilt when dealing with the empathetic agent than the other two agents. Users also evaluated the responses from the avoidance agent or counterattacking agent as less appropriate. Interestingly, even though they did not prefer the counterattacking chatbot due to its assertive responses,the users thought that the counterattacking chatbot was clearer in communicating its intention, suggesting that it might be effective for a chatbot to have a firm and clear opinion in limited

Figure 2: Box plots of user ratings of each questionnaire factors according to each response type



* Avd=Avoidance, Emp=Empathy, CnAt=Counterattacking

ACKNOWLEDGMENTS

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [19ZH1100, Distributed Intelligence Core Technology of Hyper-Connected Space]

areas such as handling legal issues or business negotiations. Based on our findings, we can provide practical chatbot design guidelines to mitigate users' verbal abuse. First, when users verbally abuse an agent, it is necessary for the agent to ask users about the real intention of their statements, rather than responding to it humorously or providing users with a related search result. Understanding the intent of users and providing a contextual response may allow users to perceive the chatbot as capable, helpful, and enjoyable. Second, if users express negative feelings toward a chatbot, with the intent of abuse, the chatbot should ask users what features of the chatbot make them upset or what situation irritates the user and show a willingness to solve the problem. Our experiment results show that most users have positively assessed the chatbot's attitude of reflecting its mistakes and asking user feedback. The chatbot's empathetic attitude toward users' angry feelings would contribute to making itself look less mechanical while reducing users' verbal abuse.

Our study has several limitations. First, participants' interactions with the chatbots in a controlled setting using the scripted verbal abuse scenarios may have limited the natural expression of emotions. Second, the chatbots we developed are all text-based, which allowed participants to communicate with the chatbots via a mobile messenger. If we had conducted our experiment with smart-speakers, we might have observed different results. Future research may investigate with smart-speakers to evaluate the significance of CAs' response styles. Nonetheless, 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.

REFERENCES

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

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.