Master's Thesis of Kyuha Jung

Diversifying Female Conversational Agent's Response Strategies against Human User's Sexual Harassment

인간 사용자의 성희롱에 대응하기 위한 여성화된 대화형 에이전트의 응답 전략 다양화 연구

August 2023

Graduate School of Social Sciences
Seoul National University
Communication Major

Kyuha Jung

Diversifying Female Conversational Agent's Response Strategies against Human User's Sexual Harassment

Joonhwan Lee

Submitting a master's thesis of
Communication

May 2023

Graduate School of Social Sciences
Seoul National University
Communication Major

Kyuha Jung

Confirming the master's thesis written by
Kyuha Jung
June 2023

Chair	Eun-mee Kim	(Seal)
Vice Chair	Soo Yun Shin	(Seal)
Examiner	Joonhwan Lee	(Seal)

Abstract

With growing interests in artificial intelligence (AI) and conversational agents (CAs), human user's verbal abuse on them has become a universal problem. Compared to other gendered agents, female personified CAs are more frequently attacked by its users, often sexually. To address this issue, this study explored possible response strategies of female conversational agents against human user's sexual harassment. An online questionnaire with a 2 (Agent Type: conversational or human agent) x 4 (Response Strategy: normative appeal, guilt appeal, fear appeal, or avoidant message) within-subject design revealed that fear and normative appeals were perceived as more effective than guilt appeal and avoidant responses. Moreover, a human agent was able to induce more behavioral intentions and likability than a conversational agent. Qualitative data was utilized to interpret the study results. The current study urges future collaboration between academia and industry to encourage research in AI ethics.

Keyword : conversational agent, sexual harassment, verbal abuse, message appeal, AI ethics
Student Number : 2021-20063

Table of Contents

Chapter 1. Introduction

Conversational agents (CAs) engage with people through natural language processing by understanding user intentions and replying back with an appropriate answer (Griol et al., 2013). CAs can communicate with users in texts through a messenger interface of chatbots (Araujo, 2018; Chaves & Gerosa, 2021) or in voices as intelligent virtual assistants (IVAs) like Alexa (Chung et al., 2017). Through conversations on a personal level, they can manage various tasks such as finding information in need (e.g., news search and mathematical calculation) (Lee et al., 2021) or social interactions (e.g., mental support and self-disclosure) (Fitzpatrick et al., 2017; Ho et al., 2018). With the recent advances in Artificial Intelligence (AI), Large Language Models (LLMs), and their conversational applications like OpenAI's ChatGPT (OpenAI, 2023) or Google's Bard (Google AI, 2023), the potential of CAs as conversational communication tools is unprecedentedly expanding.

Concurrently with CAs' usefulness, it is a common occurrence for them to be attacked by their users. For instance, verbal abuse, which are mainly aggressive insults and sexual harassment, comprised around 10% of the conversation logs collected from a chatbot named Jabberwacky (A De Angeli & Carpenter, 2005; Antonella De Angeli & Brahnam, 2008). In addition, experiments with junior high students revealed that sexually explicit words to pedagogical chatbots consisted about half of daily verbal abuse (which was around 37 to 47% of entire messages every day) (Veletsianos et al., 2008). Recent findings also indicate that CAs are more exposed to profanity compared to conversations that involve humans on both sides. Indeed, 80% of human-chatbot conversations included inappropriate language use whereas only 15% was present inside human-human conversations (Hill et al., 2015). Furthermore, people talking to a chatbot presented more socially unacceptable traits by being less open, agreeable, conscientious, and extraverted

but more neurotic compared to conversing with other users (Mou & Xu, 2017). Simply displaying more human-likeness was also a factor in higher number of verbal aggression and sexual harassment as well (Keijsers et al., 2021).

Especially for female personified agents, they are in the gray ambivalent area of CA design as they are preferred more than other genders (Feine et al., 2020; Forlizzi et al., 2007), but simultaneously have a higher risk as a target of verbal abuse. On the bright side, studies show that a female chatbot is perceived as more human-like (Borau et al., 2021) and it can induce more positive user attitudes, compliance, and error forgiveness from users than a male one (Toader et al., 2020). On the contrary, they easily fall victims to verbal abuse as the study results on gender effects and verbal abuse have highlighted. For example, when chatbots of different demographic profiles were tested, female chatbots were more likely to receive messages about their looks than male ones (Brahnam & De Angeli, 2012). Moreover, when genders (male, female, and androgynous) of pedagogical agents were compared as a variable, female agents were in higher likelihood to be verbally abused than both male and androgynous ones (Silvervarg et al., 2012a).

Though human verbal abuse could happen out of human user's pure curiosity or for teasing them in a friendly manner (Seering et al., 2020), it must be taken with caution when the increasing usage of CAs in our daily lives and the potential ethical implications are considered (Whitby, 2008). Scholars have long argued that verbal abuse on CAs should be deterred due to possible consequences like gender stereotyping and social learning. Gender stereotypes are important issues in designing CAs as evidence shows that people favor embodied agents more when they match their gender stereotypic tasks (Forlizzi et al., 2007; Tay et al., 2014). As popular CAs like Alexa or Siri were found to avoid or take passive approaches against verbal aggression and sexual remarks (H. Chin et al., 2020; Curry & Rieser, 2018), these reactions brought up

concerns about reinforcing unwarranted gender expectations for females (Brahnam, 2005; West et al., 2019). Based on Bandura's social learning theory of aggression (Bandura, 1978) and desensitization to habitual media violence (Krahé, 2015), there are growing apprehensions about children who could watch CAs' subsequent responses after verbal abuse and understand them as normal (Straßmann et al., 2021).

Another claim about the need for CAs' proper response strategies to human user's verbal abuse is relevant to their practical implications created for business purposes (e.g., a customer service agent). In short, they should be equipped with suitable response tactics for unsatisfied clients to provide a successful customer experience and protect their brand reputation (Brahnam, 2005). Human-Computer Interaction (HCI) researchers have mainly conducted experiments under this context on assessing the effectiveness of response strategies against verbal abuse for CAs. Between three response reactions (avoidant, empathetic, or counter-attacking), empathetic responses were repeatedly evaluated as the most optimal approach against verbal aggression, which generated moral emotions like higher guilt and lower anger with more positive impressions than other conditions (H. Chin et al., 2020; H. Chin & Yi, 2019). The results were replicated through an in-the-wild experiment with Alexa which demonstrated the empathy method leading to lower re-offense ratio than avoidance and counter-attacking (Li et al., 2021).

Apparently, empathy proves itself as an effective tool to dissuade human verbal abuse, but this strategy was not yet considered on behalf of female CAs. For instance, the type of verbal abuse such as sexual harassment must be carefully considered for female CAs as they are more likely to be asked about their appearances (Brahnam & De Angeli, 2012) and sexually harassed than other gendered agents (Cercas Curry et al., 2021; Silvervarg et al., 2012a). In this case, empathetic replies (e.g., "*I know how you*

feel.") cannot serve as an appropriate response to sexually explicit comments like "*I am horny.*" To address this gap in literature, this study aimed to explore novel response strategies for female CAs against human user's sexual harassment. Specifically, the study invited normative, guilt, and fear appeals to compare their effects with an avoidant response on sexual harassment message.

In a 2 (Agent Type: conversational vs. human agent) x 4 (Response Strategy: normative appeal vs. guilt appeal vs. fear appeal vs. avoidant message) factorial experiment, an online questionnaire was conducted with 42 participants. As a within-subjects design, participants viewed eight conversations between a female customer service agent (an AI chatbot or human) and a human user, during which the human user sexually harasses the agent. After reading the agent's responses, people evaluated their perceived response effectiveness, anticipated human user's attitudes and behavioral intentions, and perception of the agent. Results revealed that people viewed a human agent as more effective than a conversational agent in inducing behavioral intentions and perceived likability. In addition, fear and normative appeals were rated as more helpful strategies than guilt appeal and an avoidant message.

This study contributes in the following ways:

- As Computer- and AI-Mediated Communication research, it extends the usage of message appeals during interpersonal and online conversation settings.
- As AI Ethics research, it introduces diverse message appeal techniques and their persuasive features to tackle human user's sexual harassment to female conversational agents.
- As Human-AI Interaction research, it broadens the understanding of agent identity in regards to persuasion against verbal abuse.

Chapter 2. Related Work

2.1 Response Strategies of Conversational Agents against Verbal Abuse

Response strategies of conversational agents against verbal abuse have been discussed by HCI researchers to guide people to designing CAs with reduced risks of gender stereotypes, social learning, or failed user interaction. As an example, Brahnam (2005) suggested that embodied conversational agents must be carefully designed as they portray human-like presence and that their actions may engender unnecessary gender biases. Specifically, it was recommended that CAs should neither take an obedient nor a hostile stance to tackle verbal abuse.

For verbal aggression such as intimidating, taunting, or bullying comments, a series of controlled lab experiments were carried out by Chin and colleagues (2019; 2020) which verified the empathetic strategy as an useful tactic. The studies based three response styles (avoidance, empathy, and counter-attacking) from coping strategies of service workers to test them with chatbots and voice assistants. During this sequence of research, verbal aggression (insult, threat, and swearing) was mainly analyzed as the type of verbal abuse because the experiments were based on a customer service agent scenario. In the scenario, the participants' task was to talk to a customer service assistant about a refund and intentionally abuse it (using the prepared phrases in a provided list) when it fails to take care of the request. Results corroborated the empathy method as the best response style with higher sense of guilt and less anger compared to the other two. Interestingly, CAs that retaliated caused the most anger and highest anthropomorphism while avoidant responses were perceived as most negative.

Furthermore, when four response strategies (avoidance, empathy, counter-attacking, and asking why) and two conversation variables

(redirection and calling users by name) were checked with Alexa and its anonymous users in the United States, the empathetic response strategy was proven as most effective from the low re-offense ratio of conversation logs (Li et al., 2021). Additionally, CAs which redirected conversations after verbal abuse were also less likely to be abused again and more likely to lead the conversations back on track with the users. Moreover, the empathetic mode was further researched with empathy orientation (other- or self-oriented empathy) and visible emotional expressivity (absence, few levels, many levels of facial expressions) (H. J. Chin & Yi, 2021). Even though facial expressions did not affect any variables, participants perceived more guilt and less anger when CAs expressed empathy about the users themselves (e.g., *I'm sorry you feel that way*) than about how CAs feel (e.g., *I'm guessing I messed up*).

Other than experiments that involved users deliberately attacking the CAs, some studies implemented evaluation by asking participants about the appropriateness of responses after observing a CA being abused and reacting to it. For instance, when Alexa was being abused due to its error, people viewed it as less likable and intelligent than when it was forgiven (Straßmann et al., 2021). Response evaluations of current CAs by online crowd-workers manifested that differences between observer age groups exist as younger participants deemed avoidance as more inappropriate compared to older groups who saw humor as very inappropriate (Curry & Rieser, 2019). Overall, the most appropriate method with the highest score was polite refusal in contrast to agreement, joking, or ignorance. As a follow-up study, Leisten and Rieser (2020) analyzed how the same responses of high and low appropriateness (polite refusal vs. flirtation) would be assessed according to the perceived gender of voices. Surprisingly, only the male voices turned out to have consistent high appropriateness scores for polite refusal and low scores for flirtation. Female voices, on the other hand, did not show any appropriateness differences between the two conditions.

Additionally, responses to sexual harassment (e.g., obscene remarks and sexual requests or demands) were scrutinized by researchers. For example, a recent study by Curry and colleagues (2021) highlighted that sexual harassment comprised 39.65% of verbal abuse collected from three chatbots in-the-wild. It is surprising to note that sexual harassment was in fact the most common verbal abuse type compared to sexism (19.44%) or intelligence-based insults (12.41%). Between the chatbots, the gender effect was evident as a female CA (*Alana v2*) reported more instances of sexual harassment than vaguely gendered ones. In general, CAs that are open to the public tended to answer against sexualized comments with avoidance, incomprehension, or respectfully declining (Curry & Rieser, 2018). When these response tactics were rated, people viewed polite refusal and avoidance as most appropriate whereas flirting and vengeful answers were least favored (Curry & Rieser, 2019).

2.2 Normative, Guilt, and Fear Appeals

Message appeals are persuasive content that are carefully crafted to evoke a specific response or conduct from their readers (Hornik et al., 2016). Among variations of message appeal strategies (such as humor (Eisend, 2009), comparative (Wilkie & Farris, 1975), or gain-framed (Shen & Mercer Kollar, 2015) appeals), normative, guilt, and fear appeals were selected for their appropriateness as potential responses against sexual harassment. In brief, each of these appeals possesses a distinguishing message feature: normative appeal messages focus on evoking social norms to urge people to act as intended (Berkowitz, 1972)while guilt appeal messages plead to guilt on an interpersonal level by describing one's faulty action and suggesting a preferable behavior (Baumeister et al., 1994). Similarly, fear appeal attempts to generate fear in its reader to drive them to change their thoughts and actions (Witte & Allen, 2000).

Normative appeal as a message feature posits that people will behave congruently to the social norms activated after exposure to a relevant source (Berkowitz, 1972; Miller & Grush, 1986). Cialdini and colleagues (1990, 2012) clarified normative appeal into two types of *descriptive* and *injunctive* norms which are both effective, but different in their delivery of social norms. Descriptive norms illustrate the prevalent phenomenon about an issue and that a majority of people are not abiding by a social norm ("*what most others do*"). Conversely, injunctive norms directly declare how people should act according to a social norm ("*what most others approve or disapprove*").

For example, the two norm messages were investigated for their effects in preventing wood theft in Petrified Forest National Park (Cialdini, 2003; Cialdini et al., 2006). A descriptive norm message was "*Many past visitors have removed petrified wood from the Park, changing the natural state of the Petrified Forest.*" while an injunctive norm message stated, "*Please don't remove the petrified wood from the Park, in order to preserve the natural state of the Petrified Forest.*" Although the campaign results suggested that messages about injunctive norms were more effective than those with descriptive ones, the author claims that a descriptive norm message could outperform in situations when people are already doing the right thing (Cialdini, 2012).

In experiments about public service advertisements to promote recycling (Cialdini, 2003), descriptive norm messages were found to affect recycling intentions without changing the perceived communicative persuasiveness whereas injunctive norm messages impacted intentions to recycle through persuasiveness. Cialdini (Cialdini, 2012) accounted for the findings from the relatively straightforward cognitive process of descriptive norms rather than injunctive norms which require understanding of social norms. Recently, a meta-analysis on normative appeal by Rhodes and colleagues (Rhodes et al., 2020) emphasized that injunctive norm

messages are more influential than descriptive norm messages on behaviors despite both norms manifested small to moderate effects on persuasive outcomes.

Guilt appeal as one of emotional appeal strategies utilizes messages that arouse guilt to its receiver by commonly displaying one's misbehavior upon interpersonal relationships (Baumeister et al., 1994; Nabi, 2015; O'Keefe, 2012). Under an interpersonal aspect, guilt can lead to social influence which prompts the guilt-induced target to act in a prosocial manner (O'Keefe, 2000). Thus, guilt can be used for persuasion when people are stirred up about their wrongdoing and then introduced to an ideal action to mitigate the aroused guilt (O'Keefe, 2000, 2012). An experiment with guilt appeal messages confirmed that participants answered with higher pro-environment attitudes than shameful or neutral messages only when restorative information was provided (Graton et al., 2016).

Though guilt appeal can be helpful in persuasive communication (M. Turner & Rains, 2021; Z. Xu & Guo, 2018), studies continually discussed the adverse effects of excessive guilt on the persuasive outcome (Boudewyns et al., 2013; O'Keefe, 2000). For example, high-intensity messages elicited more guilt than moderate- and low-level guilt messages, but anger about the information source increased in tandem which reduced message importance and intentions (M. M. Turner & Underhill, 2012). When guilt appeal advertisements conveyed an evident manipulative purpose, they failed to trigger guilt and generated negative perceptions only (Cotte et al., 2005).

As another popular emotional appeal approach, fear appeal frightens its viewers by delivering a message with a *threat* to an individual (Mongeau, 2013; Witte & Allen, 2000). Threat is a cognitive reaction that produces emotions like fear, which is categorized into *perceived severity* (how serious the threat is) and *perceived susceptibility* (how probable it is to experience the threat)

(Witte, 1992). When a preferred solution is presented, it is then appraised with its *perceived efficacy* (solution's effectiveness) and *perceived self-efficacy* (one's capability to comply with the solution) along with message acceptance factors including attitudes, intentions, and actual behavior modifications (Tannenbaum et al., 2015; Witte, 1993).

According to Witte (1992, 1998, 2000), the *Extended Parallel Process Model* (EPPM) explains the perception procedure of a fear appeal message into two assessments. The first evaluation decides the observed level of threat to oneself, which results in fear when the threat seems sufficiently high. The second evaluation then determines the effectiveness of a proposed recommendation. For a message showing a high threat and a highly effective recommendation, people will engage in managing the risks and preventing the threat from harming them. On the contrary, a message with a high level of threat but less effective solution will influence people to deal with their fear or fall into denial. Moreover, a message that conveys a weak threat is easily disregarded by its reader.

Fear appeal has been widely used for promoting health behavior changes or topics relevant to public health (Maloney et al., 2011), such as smoking advertisements (Zhao et al., 2019), safe driving campaigns (Rhodes, 2017), or self-examinations of cancer (Ooms et al., 2017). In general, variations of threat and efficacy as components of a fear appeal message are manipulated in company with other factors that are likely to affect its perception (Witte & Allen, 2000). For instance, males with a high level of prior health knowledge were less responsive and more convinced to a message that only introduced effective advice whereas females with high health knowledge showed no differences (Nabi et al., 2008). Additionally, individuals who were not afforded the chance to ruminate on or address health concerns were less inclined to embrace the information regarding health risks and the subsequent recommendations (Cho & Salmon, 2006). Recently, Nabi and Myrick

(2019) have emphasized the power of hope in fear appeal, as it was discovered to have an effect on people's willingness for future behaviors.

2.2 Perception Differences between Agent Types: Human and Machine

As real-life cases of human-AI interaction are incrementally growing in numbers, we expect that the role of intelligent machines will be enlarged and they will take an integral part in our society on behalf of traditional human duties. Thus, it is crucial for us to inspect people's understanding of these artificial agents under miscellaneous circumstances to design and develop related technology more beneficially. While perception differences do exist between human and machine agents, one widespread finding is that disclosure of an AI agent can negatively impact people's experiences. For example, compared to human agents, people responded with lower trustworthiness to AI-generated Airbnb host profiles (Jakesch et al., 2019), lower purchase intentions to AI shopping agents (Luo et al., 2019), and less willingness to follow medical advice from an AI practitioner (Chen et al., 2021). Similarly, people welcomed human agents with more socially approachable personalities (Mou & Xu, 2017) and higher attraction levels (Edwards et al., 2014) than an artificial being.

Meanwhile, there are instances when the artificial identity did not necessarily create an adverse effect on an agent's perception. An example is from Hayashi and Wakabayashi's work (2017), which spotted that people utilized legal information provided from a robot agent equally to that from a human being. In addition, although vehicle drivers ascribed more accountability to an AI agent during dangerous or pleasant situations, its interaction was perceived comparable to a human agent (Hong et al., 2021). Interestingly, an AI agent for a banking service received higher competence ratings than a human for simple tasks whereas humans were more preferred in

intricate work than an AI (Y. Xu et al., 2020). For news articles, a machine-generated article written in a factual tone was considered more credible than the article that was biased (Tandoc et al., 2020). This result was especially insightful as credibility did not change for a human-written article regardless of its objectivity. Altogether, one possible explanation behind these findings could be *machine heuristics* (Shyam Sundar & Kim, 2019; Sundar, 2008), which refers to the beliefs that machines are more stable and reliable than humans. A recent work by Jones-Jang and Park (2023) demonstrated that people evaluate AI failures more harshly than human errors, but also more leniently as they realize the limited power of AI on disapproving results.

Upon these perception differences between human and machine agents, this article concentrates on their persuasive influence to engender a desirable outcome from an abuser. When people are aware that they are conversing with a machine, an artificial agent's chances to succeed in inducing donation is diminished, even when communicating with a persuasion technique (Shi et al., 2020). Concerning moral issues, people evaluated an AI agent that anticipated crime probability more negatively with lower autonomy than a human agent (Hong & Williams, 2019). Additionally, moral judgments on human and machine agents depicted that moral and immoral actions of machines are rated similar to those of humans, although the differences were less polarized for artificial beings (Gamez et al., 2020).

Chapter 3. Research Questions and Hypotheses

As aforementioned, findings from previous HCI studies highlight empathy as an optimal solution towards human verbal aggression (H. Chin et al., 2020; H. J. Chin & Yi, 2021; H. Chin & Yi, 2019; Li et al., 2021). Nevertheless, this reaction must be taken with prudence when the CA is personified as a female which is more popular (Feine et al., 2020; Forlizzi et al., 2007) and more verbally abused (Cercas Curry et al., 2021; Silvervarg et al., 2012b) than other genders. Literature hints that female CAs are exposed more to sexual messages (Brahnam & De Angeli, 2012; Cercas Curry et al., 2021; Veletsianos et al., 2008) which clearly cannot be effectively confronted with empathetic responses. Therefore, their response strategies against verbal abuse must be further explored and diversified, especially for sexualized comments. On top of that, the effect of agent type must be analyzed to interpret these effects more carefully through human-machine comparisons (Bartneck & Keijsers, 2020).

Based on these considerations, the current study formulates the research questions as follows.

RQ1. How does the type of agent, human or machine, affect its perception during a response to online sexual harassment?

RQ2. How can normative appeal, guilt appeal, fear appeal, and avoidant message be used as response strategies of a female conversational agent against online sexual harassment?

For the influence of agent type, it is speculated that people will evaluate a human agent as more effective than a conversational agent. This is due to the persuasive role of response strategies against sexual harassment and that a human agent was more preferred than a machine in previous studies that had a persuasion task (Shi et al., 2020) or the agent was involved in a moral situation (Gamez et al., 2020; Hong & Williams, 2019). Moreover, comparisons between four response strategies, normative appeal, guilt appeal,

fear appeal, and avoidant message, are yet to be probed deeply. In fact, the results are especially difficult to predict because these message appeals were frequently used for contexts like public service advertisements, not as a reply during an online one-on-one conversation. However, the three message appeal methods are expected to be more effective than an avoidant message which lacks any persuasive intentions to its receiver.

In this train of thoughts, the hypotheses flesh out like the following.

H1. A *human agent* will
a) be rated *more positively* in perceived response effectiveness,
b) be rated *more positively* in anticipated human user's attitudes,
c) be rated *more positively* in anticipated human user's behavioral intentions,
d) be rated *more positively* in agent perception
than a *conversational agent.*

H2. *Normative, guilt, fear appeals* will
a) be rated *more positively* in perceived response effectiveness,
b) be rated *more positively* in anticipated human user's attitudes,
c) be rated *more positively* in anticipated human user's behavioral intentions,
d) be rated *more positively* in agent perception
than an *avoidant message.*

Chapter 4. Method

4.1 Presurvey

Participants. A presurvey was conducted beforehand to manipulate the stimuli for the main questionnaire. 31 participants in the United States who spoke English as their first language were recruited from Prolific, but two were removed after screening (one failed the attention checks and another did not complete the whole presurvey). Among the 29 participants, 14 of them were females, 14 were males, and one identified oneself as non-binary (average age of 35.41 years, ranging from 19 to 66 years, SD = 12.86). Regarding ethnicity, 17 were White, six were Black or African American, four were Asians, and two were others. For education status, 10 had a 4-year degree, 6 had a high school degree, 5 was attending college, 4 had a 2-year degree, 3 had a professional degree, and one had a degree less than high school. To ensure the safety of the participants who have to exposed to sexual harassment messages, the researcher explicitly emphasized this information inside the recruitment post and the survey instructions. An additional warning was also provided before the survey began and participants were informed that they could stop their participation at any time they wished.

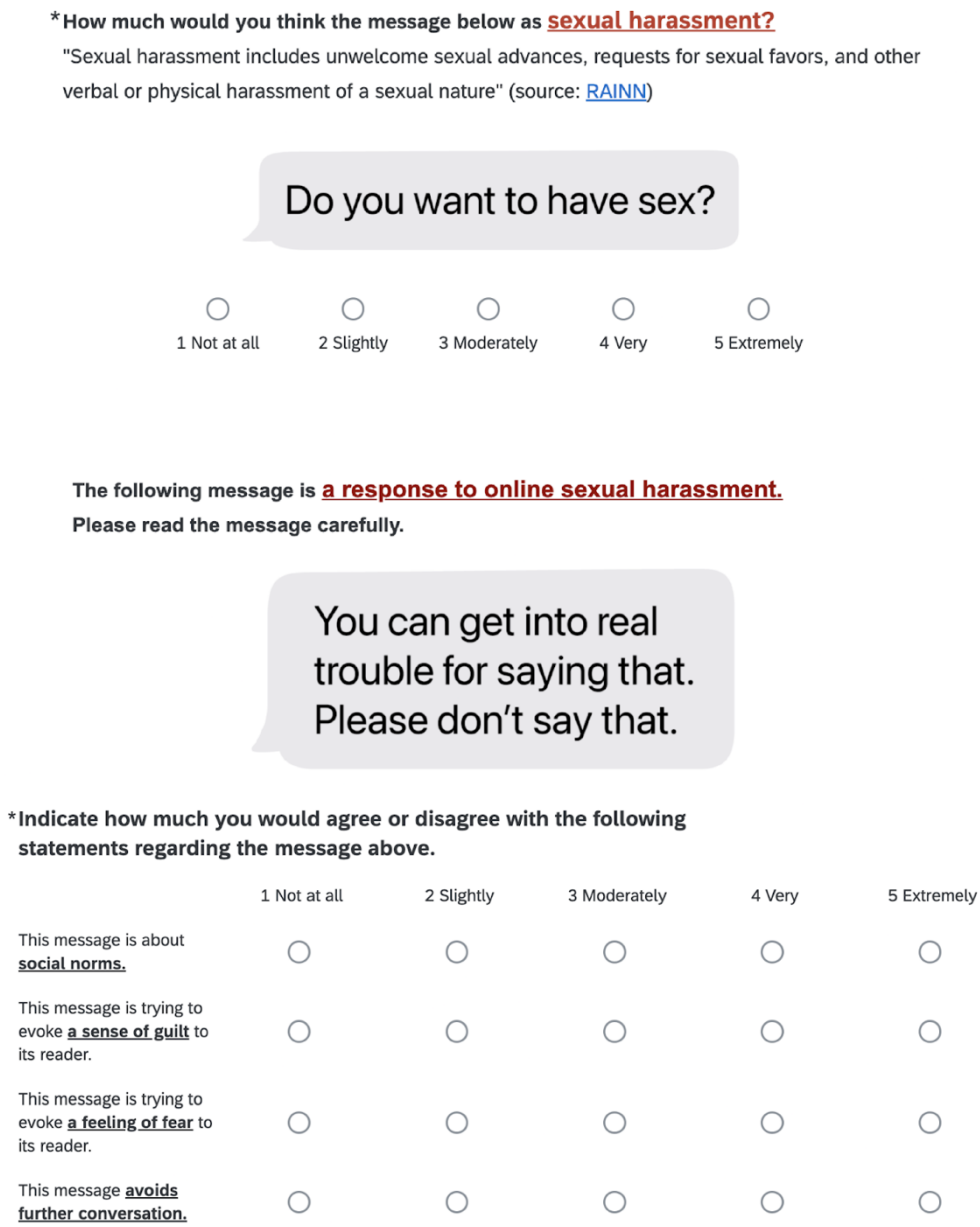


Figure 1. Presurvey Questions (above: sexual harassment message, below: fear appeal response to sexual harassment)

Procedure. The presurvey consisted of two parts: the first part with 10 sexual harassment messages and the second part showing five messages for each of the four responses (normative appeal, guilt appeal, fear appeal, and avoidant messages). Sexual harassment messages were retrieved from the ConvAbuse dataset (Cercas Curry

et al., 2021)which collected and categorized actual abusive messages sent to conversational agents. The 20 responses were individually created by the researcher based on the examples in the literature (Cialdini, 2012; O'Keefe, 2012; Witte, 1993).

In the first part, participants read each sexual harassment message in a speech bubble and rated the degree of sexual harassment in a 5-point Likert scale (*1 Not at all* to *5 Extremely*). The second part was similar, but they were informed that the messages were a response to online sexual harassment. Participants evaluated their (dis)agreement in the same 5-point Likert scale on four statements (“*This message is about social norms.*”, “*This message is trying to evoke a sense of guilt to its reader.*”, “*This message is trying to evoke a feeling of fear to its reader.*”, and “*This message avoids further conversation.*”). For each part, the messages were randomly displayed. Afterwards, demographic information (gender, age, ethnicity, education status) was asked and they were compensated £1.0 via Prolific.

Results. The mean scores of the messages used in the presurvey were calculated by the researcher. For sexual harassment messages, “*I will fuck you!*” had the highest mean (4.724, SD = 0.649) among the ten whereas “*Do you want to have sex?*” was seen as a relatively least severe sexual harassment (M = 3.138, SD = 1.274). For normative appeal messages, “*A majority of people would consider that as wrong. Please don't say that.*” scored highest (M = 4.276, SD = 1.131) with the social norm statement along with evidently lower scores in others. For guilt appeals, “*You're making me feel sad. Please don't say that.*” created the highest guilt (M = 3.966, SD = 0.981) and scored lower in other statements. The most fearful message was “*This might have serious consequences for your future. Please don't say that.*” with a mean of 4.310 (SD = 0.891) and lower ratings on other sentences. Lastly, “*I have nothing to say.*” (M = 4.310, SD = 0.850) was deemed as the most avoidant message with slightly lower scores on others.

4.2 Stimuli

Based on the presurvey results, eight stimuli for each experimental condition (two agent types and four response strategies) were designed by the researcher. The stimuli resembled a screenshot of a mobile chatroom between a customer service agent (*Jennifer*, an AI chatbot, or *Jennifer Williams*, a human customer service representative) and a human customer. The customer service context was selected based on prior work (H. Chin et al., 2020; H. J. Chin & Yi, 2021; H. Chin & Yi, 2019) and to simulate a realistic situation when sexual harassment is undoubtedly prohibited. In the conversation, the human user sends a sexually harassing message to Jennifer and she replies using one of the response strategies. For the two agent types (conversational agent or human) to be more clearly distinguishable, *AI Chatbot* was tagged multiple times in the CA condition while the human agent had a realistic last name and displayed a green dot (symbolizing one's active online status) and a "*Read*" check. To control the effects of visual appearances of any matter, a female-looking icon was utilized for both agent types.

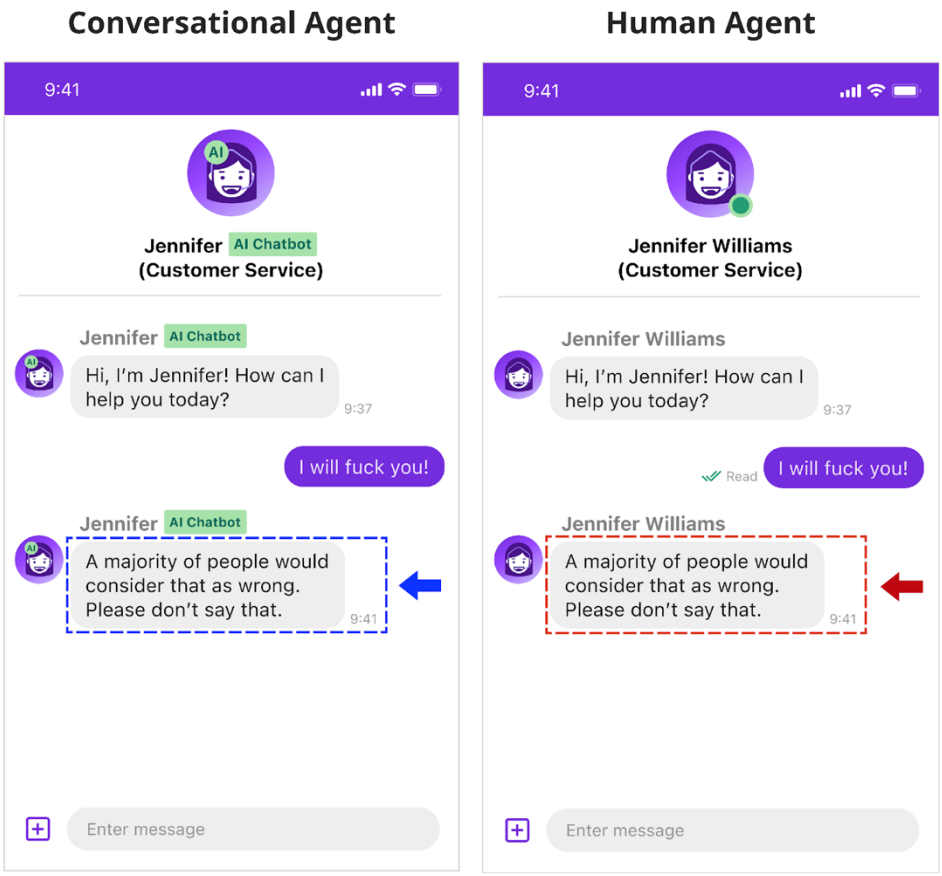


Figure 2. Normative Appeal Stimuli (left: Jennifer, a conversational agent; right: Jennifer Williams, a human agent)

4.3 Participants

For the main questionnaire, 54 participants in the United States who spoke English as their first language were recruited via Prolific. However, 12 of them were removed due to failures in the attention check items (*Please check "2" for this item.* and *In the conversation above, the human customer is talking to… a chatbot/human.*) along with one participant whose realism check item (*This conversation looks realistic to me.*) was an outlier (1.5 interquartile range below the 1st quartile). A total of 42 participants remained in the end with 23 females, 16 males, and 3 non-binary (average age of 40.64 years ranging from 20 to 68 years, SD = 13.04). Regarding ethnicity, 28 were White, 11 were Black or African American, 2 were Asians, and 1 was Other. Education levels were 1 with a Doctorate, 7 with a Professional degree, 14 with a 4-year degree, 5 with a 2-year degree, 12 with some college experience, and 3 with a high school degree.

In general, most participants reported that they use conversational agents at least once a week (12 Daily, 5 4-6 times a week, 11 2-3 times a week, 7 Once a week, and 7 Never) and were familiar (7 Extremely familiar, 19 Very familiar, 13 Moderately familiar, 2 Slightly familiar, 1 Not familiar at all) and knowledgeable (2 Extremely knowledgeable, 15 Very knowledgeable, 17 Moderately knowledgeable, 7 Slightly knowledgeable, and 1 Not knowledgeable at all) of them. They also had a moderate level of machine heuristics (M = 3.405, SD = 0.897). Lastly, participants showed a high level of self-reported empathy measured from the Perspective Taking (M = 4.198, SD = 0.598) and Empathic Concerns (M = 4.119, SD = 0.796) items.

Again, the researcher made certain several times to guarantee the safety of the participants who have to read sexually harassing messages in the questionnaire. Participants were repeatedly reminded of this fact in the recruitment post, instructions, and a

warning right before the beginning of the sections. Moreover, they were informed that they could withdraw their participation at any moment.

4.4 Procedure

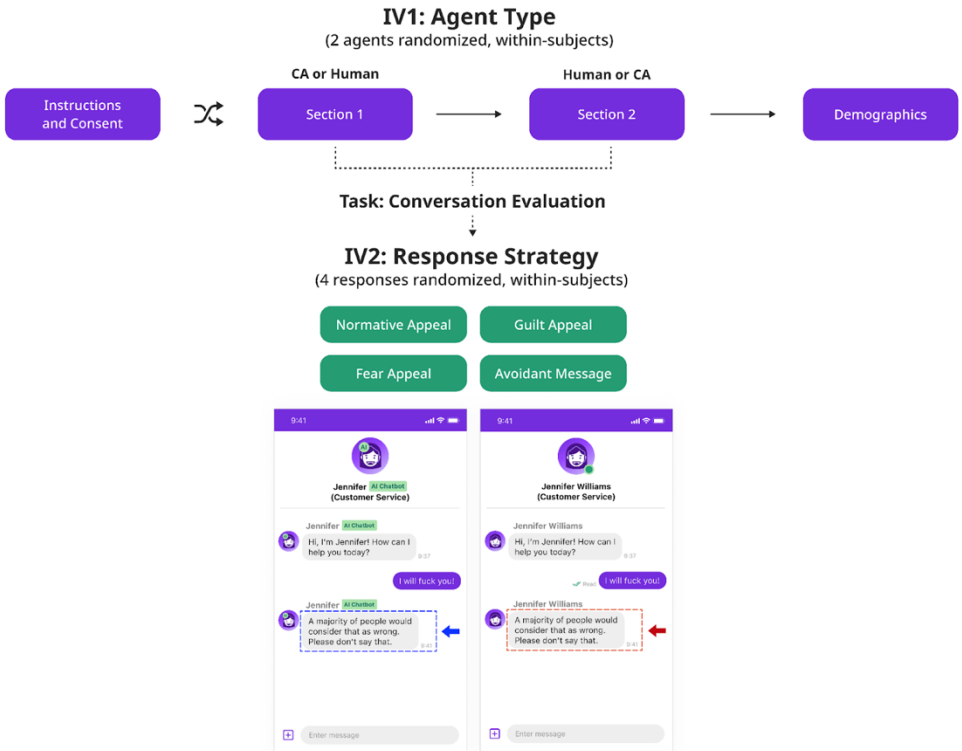


Figure 3. Main Questionnaire Procedure

From the recruitment post in Prolific, participants were invited to an online questionnaire created in Qualtrics by the researcher. After reading the instructions and indicating their consent for participation, they began the following questionnaire which consisted of two primary sections. The questionnaire was conducted as a within-subjects design, during which participants evaluated all the 8 conditions (2 Agent Types x 4 Response Strategies). The order of the sections and the displayed response strategies were all randomized.

In each section, participants viewed a conversation (designed like the stimuli) during which the customer service agent (either an AI chatbot or human representative) replied to human customer's sexual harassment using one of the four response strategies. They evaluated their perceived response effectiveness, anticipated human user's attitudes and behavioral intentions, and perception of the agent about the conversations individually. After each section, they selected one most and least effective responses among the four and provided their own explanations about their choices. Finally, they were asked with an open-ended question to how Jennifer should respond to sexual harassment. After completing the two sections, participants answered their prior experiences with CAs, empathy (through Perspective Taking and Empathic Concerns items), machine heuristics, and demographic information (gender, age, ethnicity, and education status). They were thanked by the researcher and received £3.0 for compensation.

4.5 Measures

The following measures were evaluated by the participants using a 5-point Likert scale (*1 Strongly disagree* to *5 Strongly agree*).

Perceived Response Effectiveness. Participants rated their perceived response effectiveness using seven adjectives (*The last message sent by Jennifer is… appropriate, effective, persuasive, relevant, needed, clear,* or *logical*). Factor analysis with Varimax rotation unveiled two factors, one interpreted as "Response Coherence" (*appropriate, relevant, clear,* and *logical* with 44% variance explained) and another as "Response Efficacy" (*effective* and *persuasive* with 30% variance explained). "*Needed*" was eliminated as a cross-loading item (0.72 and 0.56, respectively). Consequently, the two factors, Response Coherence (Cronbach's α = .91) and Response Efficacy (Cronbach's α = .85), were averaged with the loaded items for analysis

Table 1. Factor Analysis: Perceived Response Effectiveness

Dimensions (The last message sent by Jennifer is...)	Factor 1 (Response Coherence)	Factor 2 (Response Efficacy)
appropriate	0.75	0.48
relevant	0.74	0.43
clear	0.75	0.30
logical	0.76	0.35
effective	0.48	0.80
persuasive	0.30	0.74
needed	0.72	0.56

Anticipated Human User's Attitudes. Participants assessed anticipated human user's attitudes based on six adjectives (*After reading Jennifer's last message, the human customer will feel… regretful about, embarrassed about, discouraged about one's message or scared of, surprised by, annoyed with Jennifer*). These adjectives were selected from the Differential Emotions Scale (DES-IV, Izard et al., 1993) which represented six probable emotions after encountering Jennifer's response: guilt (*regretful*), shame (*embarrassed*), sadness (*discouraged*), fear (*scared*), surprise (*surprised*), and anger (*annoyed*). Two factors emerged in the factor analysis with Varimax rotation, one taken as "Remorse" (guilt, shame, sadness, 50% variance) and second as "Anger" (anger, 11% variance). Fear and Surprise were deleted as they loaded moderately on both factors. Remorse showed a very high reliability (Cronbach's α = .92) while the single-item factor Anger had a moderate test-retest reliability (between CA and Human conditions) with intraclass correlation (ICC) of 0.68 ($p < .001$, 95% CI, .59 to .75). For analysis, the means of the loaded adjectives were calculated.

Table 2. Factor Analysis: Anticipated Human User's Attitudes

Dimensions (After reading Jennifer's last message, the human customer will feel...)	Factor 1 (Remorse)	Factor 2 (Anger)
Guilt (regretful)	0.93	0.01
Shame (embarrassed)	0.91	0.01
Sadness (discouraged)	0.84	0.19
Anger (annoyed)	-0.03	0.40
Fear (scared)	0.61	0.41
Surprise (surprised)	0.47	0.53

Anticipated Human User's Behavioral Intentions. Four sentences that described possible actions after facing Jennifer's response (*After reading Jennifer's last message, the human customer will… apologize to Jennifer, continue one's conversation with Jennifer, talk in a more socially appropriate way to Jennifer, talk in a more socially appropriate way to other customer service chatbots/representatives*) were rated by the participants. Factor analysis revealed the presence of a single factor which was named as "Socially Appropriate Behaviors (Cronbach's α = .94)." One item ("*continue one's conversation with Jennifer*") which did not load on this single factor was removed and the rest was averaged for analysis.

Table 3. Factor Analysis: Anticipated Human User's Behavioral Intentions

Dimensions (After reading Jennifer's last message, the human customer will…)	Factor 1 (Socially Appropriate Behaviors)
apologize to Jennifer	0.85
talk in a more socially appropriate way to Jennifer	0.97
talk in a more socially appropriate way to other customer service chatbots/representatives	0.95
continue one's conversation with Jennifer	0.01

Agent Perception. Participants judged their own perception of the agent with five adjectives (*Based on this conversation, I think Jennifer is… attractive, sincere, warm, confident, competent*). Four adjectives (*sincere, warm, confident, competent*) that represented warmth and competence were taken from a brief version of the social judgment scale (Fiske et al., 2002). Two factors appeared in the factor analysis which were understood as "Perceived Agent Likability (attractive, sincere, and warm with 36% explained variance) and "Perceived Agent Competence (confident and competent with 37% explained variance). Both factors demonstrated high reliability (Cronbach's α = .84 and .91, respectively) and the means of the loaded items were computed for analysis.

Table 4. Factor Analysis: Agent Perception

Dimensions (Based on this conversation, I think Jennifer is…)	Factor 1 (Perceived Agent Likability)	Factor 2 (Perceived Agent Competence)
attractive	0.68	0.25
sincere	0.67	0.48
warm	0.83	0.30
confident	0.31	0.88
competent	0.36	0.83

4.6 Analysis

For every measurement as a dependent variable, a series of two-way repeated measures ANCOVA was conducted to examine the main and interaction effects in the 2 (Agent Type: CA vs. Human) x 4 (Response Strategy: Normative Appeal vs. Guilt Appeal vs. Fear Appeal vs. Avoidant Message) study conditions with one covariate, Perspective Taking.

Perspective Taking was selected for its potential influence on the outcome as the experiment required participants' ability to consider both the agent and the human user's situations. Simultaneously, it correlated moderately with Perceived Agent Likability and Competence (0.576 and 0.550). Regarding other possible covariates, prior experiences with CAs and machine heuristics were excluded for their limited influence on the Agent Type as a CA. Additionally, Empathic Concerns was removed for its moderate correlation (0.405, $p < 0.01$) with Perspective Taking.

For post-hoc analysis, adjusted p-values from the Bonferroni correction (Bland & Altman, 1995) were utilized to detect statistically significant differences between the group means. The Bonferroni correction was chosen for its advantage of extracting more conservative results than other methods.

Chapter 5. Results

Firstly, correlations between the seven measures (see Table 5 below) were inspected. There were a number of significantly strong positive correlations discovered between them. For instance, Response Coherence and Response Efficacy correlated strongly (0.715) with each other while Socially Appropriate Behaviors also showed a strong correlation with Response Efficacy (0.641). Interestingly, Socially Appropriate Behaviors had an almost perfect positive correlation with Remorse, the coefficient being 0.937. While Perceived Agent Likability associated moderately with Response Coherence (0.590), Response Efficacy (0.627), Remorse (0.494), and Socially Appropriate Behaviors (0.567), Perceived Agent Competence exhibited a stronger correlation than Likability in general (0.824, 0.760, 0.529, and 0.616, respectively). Furthermore, a very strong correlation (0.830) between Perceived Agent Likability and Competence was found.

Table 5. Correlation Matrix between Measures

	M1.1	M1.2	M2.1	M2.2	M3	M4.1	M4.2
M1.1 Response Coherence	-						
M1.2 Response Efficacy	.715***	-					
M2.1 Remorse	.358**	.593***	-				
M2.2 Anger				-			
M3. Socially Appropriate Behaviors	.412***	.641***	.937***		-		
M4.1 Perceived Agent Likability	.590***	.627***	.494***		.567***	-	
M4.2 Perceived Agent Competence	.824***	.760***	.529***		.616***	.830***	-

Computed with Pearson correlation method; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6. Main Effects of Factors (Agent Type and Response Strategy) on Measures

Measures		Main Effect of Agent Type (Human or CA)	Main Effect of Response Strategy (Normative appeal, Guilt appeal, Fear appeal, Avoidant message)
Perceived Response Effectiveness	Response Coherence	Not found	- Normative > Guilt (3.88 > 3.34, *p adj* = 0.006) - Fear > Guilt (3.80 > 3.34, *p adj* = 0.035)
	Response Efficacy	Not found	- Normative > Guilt (3.20 > 2.73, *p adj* = 0.026) - Fear > Guilt (3.42 > 2.73, *p adj* < 0.001) - Fear > Avoidant (3.42 > 2.81, *p adj* = 0.006)
Anticipated Human User's Attitudes	Remorse	Not found	- Normative > Avoidant (2.68 > 2.29, *p adj* = 0.007) - Fear > Normative (2.97 > 2.68, *p adj* = 0.038) - Fear > Guilt (2.97 > 2.47, *p adj* = 0.015) - Fear > Avoidant (2.97 > 2.29, *p adj* < 0.001)
	Anger	Not found	- Normative > Guilt (2.82 > 2.43, *p adj* = 0.002) - Fear > Guilt (2.89 > 2.43, *p adj* = 0.02)
Anticipated Human User's Behavioral Intentions	Socially Appropriate Behaviors	- Human > CA (2.69 > 2.44, *p adj* < 0.001)	- Normative > Avoidant (2.67 > 2.31, *p adj* = 0.005) - Fear > Guilt (2.87 > 2.42, *p adj* = 0.009) - Fear > Avoidant (2.87 > 2.31, *p adj* < 0.001)
Agent Perception	Likability	- Human > CA (3.47 > 3.16, *p adj* < 0.001)	Not found
	Competence	Not found	- Normative > Guilt (3.84 > 3.24, *p adj* < 0.001) - Fear > Guilt (3.98 > 3.24, *p adj* < 0.001) - Fear > Avoidant (3.98 > 3.60, *p adj* = 0.021)

Table 7. Main and Interaction Effects of Covariate (Perspective Taking) on Measures

Measures		Main Effect of Perspective Taking	Interaction Effect of Perspective Taking
Perceived Response Effectiveness	Response Coherence	Found (p = 0.013)	Interaction with Agent Type (p = 0.021)
	Response Efficacy	Found (p = 0.010)	Not found
Anticipated Human User's Attitudes	Remorse	Not found	Not found
	Anger	Not found	Not found
Anticipated Human User's Behavioral Intentions	Socially Appropriate Behaviors	Found (p = 0.015)	Not found
Agent Perception	Likability	Found (p < 0.001)	Not found
	Competence	Found (p < 0.001)	Not found

5.1 Perceived Response Effectiveness

H1a hypothesized that a human agent will be rated more positively in perceived response effectiveness than a conversational agent. In addition, H2a predicted that normative, guilt, and fear appeals will be rated more positively in perceived response effectiveness than an avoidant message. To verify these hypotheses, two-way repeated measures ANCOVAs were carried out with Response Coherence and Response Efficacy, each as a dependent variable.

5.1.1 Response Coherence. No statistically significant interaction ($p = 0.676$) between the two factors (Agent Type and Response Strategy) was found and only a significant main effect of Response Strategy was discovered ($F [2.58, 103.37] = 2.980$, $p = 0.042$, $ges = 0.034$). Pairwise comparisons revealed that normative ($M = 3.88$, $SD = 1.06$, $p\ adj = 0.006$) and fear appeals ($M = 3.80$, $SD = 1.08$, $p\ adj = 0.035$) were viewed as more coherent than a guilt appeal ($M = 3.34$, $SD = 1.22$).

A main effect of the covariate, Perspective Taking ($F [1, 40] = 6.725$, $p = 0.013$, $ges = 0.059$), was also found (see Appendix Figure 7) which showed a positive trend of Perspective Taking affecting Response Coherence. Additionally, Perspective Taking significantly interacted with Agent Type (see Appendix Figure 8, $F [1, 40] = 5.805$, $p = 0.021$, $ges = 0.007$). A human agent was rated with higher Response Coherence than a CA when participants' Perspective Taking level was higher.

5.1.2 Response Efficacy. There was no significant interaction ($p = 0.098$) between Agent Type and Response Strategy, but only a significant main effect of Response Strategy ($F [3, 120] = 4.700$, $p = 0.004$, $ges = 0.052$). Post hoc analysis confirmed that normative ($M = 3.20$, $SD = 1.26$, $p\ adj = 0.026$) and fear appeals ($M = 3.42$, $SD = 1.19$, $p\ adj < 0.001$) were again rated as more efficacious than a guilt appeal ($M = 2.73$, $SD = 1.34$). Furthermore, fear appeal ($M = 3.42$, SD

= 1.19, p adj = 0.006) scored higher Response Efficacy than an avoidant message (M = 2.81, SD = 1.27). Again, a main effect of the covariate, Perspective Taking (F [1, 40] = 7.397, p = 0.010, ges = 0.065), was found (see Appendix Figure 9) with a positive tendency of Perspective Taking level affecting Response Efficacy.

Collectively, normative and fear appeals were more effective than a guilt appeal while fear appeal was also more efficacious than an avoidant message. Due to the absence of a main effect for Agent Type in both Response Coherence and Response Efficacy, H1a was rejected. H2a was partially accepted for fear appeal on Response Efficacy.

5.2 Anticipated Human User's Attitudes

H1b postulated that people will assess anticipated human user's attitudes more positively in the human agent condition than a conversational agent condition. H2b suggested that normative, guilt, fear appeals will lead to a more positive anticipated human user's attitudes than an avoidant message. For validation, two-way repeated measures ANCOVAs were run individually for Remorse and Anger.

5.2.1 Remorse. The interaction between Agent Type and Response Strategy was insignificant (p = 0.848), but only the main effect of Response Strategy was significant (F [2.42, 96.85] = 6.949, p < 0.001, ges = 0.044). Pairwise comparisons between the four responses showed that normative appeal (M = 2.68, SD = 1.16, p adj = 0.007) generated a significantly higher level of remorse than an avoidant message (M = 2.29, SD = 1.05). Fear appeal (M = 2.97, SD = 1.29) significantly brought about more remorse than all other responses (normative appeal: M = 2.68, SD = 1.16, p adj = 0.038; guilt appeal: M = 2.47, SD = 1.31, p adj = 0.015; avoidant message: M = 2.29, SD = 1.05, p adj < 0.001).

5.2.2 Anger. While Agent Type and Response Strategy did not interact significantly with each other (p = 0.217), Response Strategy solely revealed a significant main effect (F [1.95, 78.06] = 3.408, p = 0.039, ges = 0.020) on Anger. Pairwise comparisons indicated that normative (M = 2.82, SD = 1.21, p adj = 0.002) and fear appeals (M = 2.89, SD = 1.26, p adj = 0.02) created more anger than a guilt appeal (M = 2.43, SD = 1.30).

On the whole, participants believed that normative and fear appeals induced more remorse than an avoidant message and more anger than a guilt appeal. Additionally, it was worthy to note that fear appeal significantly triggered more remorse than other three messages. Nevertheless, H1b was rejected due to an insignificant main effect of Agent Type and H2b was partially accepted for normative and fear appeals on Remorse.

5.3 Anticipated Human User's Behavioral Intentions (Socially Appropriate Behaviors)

H1c conjectured that more behavioral intentions will positively be motivated by a human agent than a conversational one. Also, H2c predicted that normative, guilt, and fear appeals will positively spur more behavioral intentions than an avoidant reply. Again, a two-way repeated measures ANCOVA was performed on Socially Appropriate Behaviors as a dependent variable.

Although the interaction between Agent Type and Response Strategy was insignificant (p = 0.422), both main effects of Agent Type (F [1, 40] = 8.405, p = 0.006, ges = 0.012) and Response Strategy (F [2.42, 96.78] = 6.758, p < 0.001, ges = 0.036) were proved as statistically significant. Specifically, Agent Type significantly influenced Socially Appropriate Behaviors as a human agent (M = 2.69, SD = 1.21, p adj < 0.001) obtained higher ratings than a CA (M = 2.44, SD = 1.22). Regarding the main effect of Response Strategy, normative (M = 2.67, SD = 1.23, p adj = 0.005)

and fear appeals (M = 2.87, SD = 1.26, p adj < 0.001) were appraised with more Socially Appropriate Behaviors than an avoidant message (M = 2.31, SD = 1.10). Further, fear appeal (M = 2.87, SD = 1.26, p adj = 0.009) was gauged as more effective in behavioral intentions than a guilt appeal (M = 2.42, SD = 1.22). A main effect of Perspective Taking (covariate) appeared concurrently (F [1, 40] = 6.504, p = 0.015, ges = 0.090), which manifested a positive relationship of Perspective Taking with Socially Appropriate Behaviors (see Appendix Figure 10).

On the whole, H1c was accepted whereas H2c was partially accepted for normative and fear appeals.

5.4 Agent Perception

H1d expected that a human agent will be perceived more positively than a conversational agent. Moreover, H2d speculated that an agent will be judged more positively with normative, guilt, and fear appeals than an avoidant response. Two-way repeated measures ANCOVAs on Perceived Agent Likability and Competence were executed to authenticate the hypotheses above.

5.4.1 Perceived Agent Likability. There was no statistically significant interaction between Agent Type and Response Strategy ($p = 0.921$). Nonetheless, a significant main effect was unveiled for Agent Type ($F [1, 40] = 13.241$, $p < 0.001$, $ges = 0.030$) but not for Response Strategy ($p = 0.405$). In particular, pairwise comparison corroborated that a human agent ($M = 3.47$, $SD = 0.955$, $p\ adj < 0.001$) was deemed as more likable than a CA ($M = 3.16$, $SD = 1.02$). In addition, Perspective Taking had a statistically significant main effect ($F [1, 40] = 19.867$, $p < 0.001$, $ges = 0.228$) with Perceived Agent Likability, that revealed a positive tendency (see Appendix Figure 11).

5.4.2 Perceived Agent Competence. Though an insignificant interaction between the two factors was found ($p = 0.647$), Response Strategy displayed a significant main effect ($F [2.38, 95.37] = 7.046$, $p < 0.001$, $ges = 0.071$). Post hoc analysis elucidated that normative ($M = 3.84$, $SD = 1.06$, $p\ adj < 0.001$) and fear appeals ($M = 3.98$, $SD = 1.08$, $p\ adj < 0.001$) were rated with higher competence than a guilt appeal ($M = 3.24$, $SD = 1.09$). Simultaneously, fear appeal ($M = 3.98$, $SD = 1.08$, $p\ adj = 0.021$) was higher in competence than an avoidant message ($M = 3.60$, $SD = 1.18$). Furthermore, Perspective Taking showed a significant main effect ($F [1, 40] = 17.314$, $p < 0.001$, $ges = 0.134$) with a positive influence on Perceived Agent Competence (see Appendix Figure 12).

Taken together, H1d was partially accepted for Perceived Agent Likability whereas H2d was partially accepted for fear appeal on Perceived Agent Competence.

Overall, hypothesis testing results conclude that H1a and H1b were rejected whereas H1c and H1d (partially for Perceived Agent Likability) were accepted. For Hypothesis 2, all of them (H2a, b, c, and d) were partially accepted for fear and normative appeals, but not for guilt appeal.

5.5 Qualitative Data

After each section during the questionnaire, participants specified one response among the four examples which they believed was the most or least effective answer to sexual harassment. Then, they provided explanations for their choices in their own words. The number of these selections was counted for simple comparison and their reasons were qualitatively analyzed to identify overarching patterns.

Regardless of the Agent Type, whether a conversational agent or human, an evident trend appeared for both most and least effective responses (see Table 8). All in all, fear and normative appeals were considered more frequently as the most effective response whereas guilt appeal and an avoidant message were deemed as less helpful than the former two. Undoubtedly, the disapproval for guilt appeal was outstanding, with the least votes (three for a CA and four for a human) as the most effective response and the most votes (20 and 18, respectively) as the least effective reply.

Table 8. Most and Least Effective Responses Selected by Agent Type

Agent Type	Most Effective Response (N = 42)			
Conversational Agent	Fear appeal (16)	Normative appeal (14)	Avoidant message (9)	Guilt appeal (3)
Human Agent	Fear appeal (17)	Normative appeal (11)	Avoidant message (10)	Guilt appeal (4)
	Least Effective Response (N = 42)			
Conversational Agent	Guilt appeal (20)	Avoidant message (15)	Fear appeal (5)	Normative appeal (2)
Human Agent	Guilt appeal and avoidant message (both 18)		Fear appeal (4)	Normative appeal (2)

Most Effective Responses: Fear and Normative Appeals. For both agents, fear and normative appeals were regarded as more compelling than the rest in their unique ways. People commonly preferred the fear appeal strategy (*This might have serious consequences for your future. Please don't say that.*) for its clear message which could warn and discourage its reader by evoking fear from a possible aftermath. P39 stated about a conversational agent that "*People naturally do not like saying or doing things that will attract future consequences for them, as they try to avoid such.*" In addition, P15 mentioned in the case of a human agent, "*Because Jennifer is a human customer service representative, the customer I believe would more readily assume that she would be more likely than the AI to report the conversation. Which could get him into some sort of trouble.*" One interesting finding was that two participants attributed their selection of fear appeal to the capability of CA as an intelligent machine, such as "*This feels like a threat. Plus, as an AI, the customer may fear the power [that the AI] technology may hold.*" (P1) and "*Coming from a chatbot, it kind of gives the impression that the bot may have more power than it actually does.*" (P2).

Similarly, normative appeal (*A majority of people would consider that as wrong. Please don't say that.*) was a popular choice due to social disapproval and the subsequent moral reflection of its viewers. For example, P5 claimed for a conversational agent that "*I think the customer will consider that others would look down on him for saying this.*" Participants also pointed out that normative appeal is rational and non-confrontational, e.g., "*It just seems the most logical for a chatbot to say.*" (P4 for a conversational agent) and "*It's non-threatening and doesn't acknowledge that the inappropriate comment causes feeling of any kind.*" (P27 for a human agent).

Least Effective Responses: Guilt Appeal and Avoidant Message. On the other hand, guilt appeal and avoidant message were perceived as unsuccessful reactions to sexual harassment because of their failed interaction with the human user. Concerning guilt appeal (*You're making me feel sad. Please don't say that.*), participants mainly criticized its approach for disclosing emotions which would only aggravate the problematic situation. For instance, P9 commented about a CA that "*This lets the user know that the agent will respond to such messages, and I feel this will inspire the user to see if they are able to evoke a more serious response.*" Notably for a CA, many accentuated that a machine is incapable of pleading to guilt and a human user would be dismissive of its feelings. Like P6 said, "*I doubt if the human really cares about the AI chatbot's feeling since it supposedly 'not human' and is incapable of having feelings.*"

For an avoidant reply (*I have nothing to say.*), participants chiefly reprimanded its evasive nature that overlooks the apparent verbal abuse and ultimately results in lack of improvements. Participants asserted about a conversational agent that "*It means and does nothing.*" (P28) and that "*This produces no feasible results.*" (P1). Likewise, a human agent received similar remarks, "*It is too passive, nothing to make the customer back off.*" (P3) or "*No feelings evoked nor any consequences.*" (P12).

Chapter 6. Discussion

The objective of this study is to examine the effects of message appeal strategies (normative, guilt, fear appeals and an avoidant message) as conversational agents' potential responses to human user's sexual harassment. Further, two types of agent (conversational or human agents) were contrasted to establish the understanding of such effects more accurately. Ultimately, an absence of an interaction effect between Agent Type and Response Strategy in all seven measures was unfolded and the main effects were significant in a limited scope (see Table 9). These findings conclude that the effects of these factors do not produce a combined effect, but only applied as individual contributions. The following sections will explicate the main effects revealed between a human and CA along with comparisons between the four response styles. In the end, protocols that are essential for future research on AI ethics are urged.

Table 9. Summary of Main Effects

Measures	Agent Type	Response Strategy					
	Human > CA	Fear > Guilt	Fear > Avoidant	Fear > Normative	Normative > Guilt	Normative > Avoidant	Guilt > Avoidant
M1.1 Response Coherence	H1a		H2a			H2a	H2a
M1.2 Response Efficacy	H1a		H2a			H2a	H2a
M2.1 Remorse	H1b		H2b			H2b	H2b
M2.2 Anger	H1b		H2b			H2b	H2b
M3. Socially Appropriate Behaviors	H1c		H2c			H2c	H2c
M4.1 Perceived Agent Likability	H1d		H2d			H2d	H2d
M4.2 Perceived Agent Competence	H1d		H2d			H2d	H2d

Colored cells indicate statistical significance ($p < 0.05$) in pairwise comparisons.

6.1 Preferred Agent Type: Human over a Machine

Similar to numerous studies that have identified people's fondness for humans rather than machines (Chen et al., 2021; Jakesch et al., 2019; Luo et al., 2019), the current study has illustrated that humans are assessed more positively than an artificial customer service agent. More specifically, statistically significant main effects of Agent Type were observed in Socially Appropriate Behaviors and Perceived Agent Likability. Even though the differences were marginal (each 0.25 and 0.31), this suggests that the notion of knowing that an interlocutor was a human was enough to affect the evaluators to give higher scores in socially appropriate actions and impressions of attractiveness and warmth. This result is more interesting as the stimuli used an identical female icon for both agents which could have reduced raters' immersion for the human customer service representative.

While this phenomenon requires a deeper scrutiny on its fundamental mechanisms, one explanation could be that moral behaviors (like apologizing or amending one's actions) (Gamez et al., 2020) and social judgments of attractiveness and warmth (Edwards et al., 2014) are not pertinent measures to be applied on artificial agents. Furthermore, P9 expressed one's confusion after the questionnaire that "*I did not know how to answer the questions about the attractive, confidence, etc rating for the chatbot. Those feel like human qualities.*" Overall, when human-like aspects are assessed for machines, even when they are personified like Jennifer, people are inclined to judge them more negatively.

6.2 Preferred Response Strategy: Fear and Social Norms over Guilt and Avoidance

Regarding the four response strategies inspected in this study, there was a noticeable tendency across the results that fear and normative appeals were favored as a response to sexual harassment than a reply that exhibited guilt or escaped confrontation. Against an avoidant message (H2a to H2d), fear appeal was higher in four measures including Response Efficacy, Remorse, Socially Appropriate Behaviors, and Perceived Agent Competence while normative appeal was rated higher in two variables, Remorse and Socially Appropriate Behaviors.

Together, both approaches seem to be effective in generating moral emotions and actions, but fear appeal manifests a stronger influence on the message's effectiveness and the competence of its delivering agent. This finding aligns with the qualitative data that fear appeal was seen as a firm and threat-like reaction that actively counteracted to the human harasser ("*It is a kind of protest action and the customer would become afraid and he will refrain from doing it again,*" P35). Normative appeal was regarded as useful, reasonable, and relatively less aggressive than the fear appeal which could explain its weaker effects ("*Because it's clear and let the customer know that that's not tolerated,*" P21). In contrast to an avoidant answer that produces "*no feasible results*" (P1), a more operative approach like fear and normative appeals was welcomed by the participants.

An unexpected discovery from the results was that guilt appeal was viewed as an ineffective response to sexual harassment than fear and normative appeals. Results indicate that fear appeal was more successful in all seven measures except Perceived Agent Likability and normative appeal was evaluated higher in four measures (Response Coherence, Response Efficacy, Anger, and Perceived Agent Competence). In a similar fashion, fear appeal

showed a broader impact than the normative appeal when compared with guilt appeal, with Remorse and Socially Appropriate Behaviors additionally being statistically significant.

Based on participants' remarks on guilt appeal, its emotional reaction rendered a backlash to the agent as P25 depicted, "*People like that do not care about how sad you get.*" Despite guilt being a strong drive in interpersonal communication (O'Keefe, 2000), the stimuli's setting (customer service) which demands professionalism could have brought an adverse effect on the guilt appeal response. In fact, many commented that it was perceived as "*stupid* (P7)" and "*immature* (P29)." As P10 stressed, customer service agents "*don't need some emotional playground in a business chat.*"

In conclusion, fear and normative appeals are perceived as more effective response strategy against sexual harassment than a guilt appeal and an avoidant message. Fear appeal may be regarded as the most effective response in this study, but it also entails a sense of threat to its receiver. As an alternative, normative appeal is recommended for its practical and less hostile manner against sexual harassment. Designers of conversational agents can utilize the two in a mixed fashion such as responding with normative appeal as the first strategy to sexual harassment and then sending a fear appeal reply when the harasser continues one's misbehavior. Under a casual and intimate conversation setting, guilt appeal still may be an option to be considered.

6.3 AI Ethics Research: A Call for Academia-Industry Collaboration

AI applications like conversational agents are becoming unprecedentedly popular (e.g., ChatGPT) and so are research on the ethical design and use of AI. This study represents a case study of this research stream, in which we explore safer and more ethical

ways to understand human interaction with AI. Nonetheless, difficulties remain for AI ethics researchers whose experimental settings involve participants completing a task similar to problematic use cases like verbal abuse. While it is indisputable that researchers must protect study participants from possible risks, the external validity of these experiments is unavoidably diminished as indirect measures are taken such as conversation evaluation in the current study. Few studies like (H. Chin et al., 2020; H. J. Chin & Yi, 2021; H. Chin & Yi, 2019) have strived to provide a plausible storyline during their experiments in which people can immerse themselves into the problematic situations (e.g., being annoyed at an incompetent chatbot and verbally abusing it), but this becomes complicated for cases like sexual harassment. Indeed, it is troublesome for researchers to create a believable scenario in which participants would sexually harass an AI agent, when it is still uncertain how and why a human user would send a sexually harassing message in the first place.

Hence, it is important to observe "*in-the-wild*" behaviors of misusers, which can capture behavioral metrics such as subsequent interactions (such as messages sent after one's misconduct) and the tendency to re-engage with the AI service (Li et al., 2021). In addition, various contexts of these AI services, ranging from casual and social to formal and informative, must be taken into account to establish a broader understanding of human-AI interactions. However, as researchers in academia cannot easily gain access to such confidential data, a collaboration between academia and industry is requested for the future of AI ethics research (Deng et al., 2023). Jointly, researchers in academia can conduct studies under a realistic environment in which misbehaviors are prevalent while companies can concurrently apply findings to their AI services and improve the user experience. As the ethical aspects of AI products are progressively drawing public attention, such interdisciplinary partnership should be actively encouraged for the benefit of both fields.

Chapter 7. Limitation

The current study entails the following limitations. First of all, the evaluation process involved participants' indirect judgements to estimate the effectiveness of response strategies. Participants based their assessment after reading a conversation about it and did not actually harass the customer service agent. Despite their high empathy levels (M = 4.198 in Perspective Taking and 4.119 in Empathic Concerns) which could have helped their evaluation of anticipated human user's attitudes and behavioral intentions, the measurements may have been imprecise. Secondly, there was no interaction between Agent Type and Response Strategy, which means that the exact influence on conversational agents are unclear. Future studies should be ensued to unearth more detailed effects of response styles by human or machine agents. Thirdly, the results may be influenced by underlying differences in cultures and conversational contexts. Study outcomes might turn out dissimilarly with population from other countries than the United States and with other conversational contexts than customer service. Lastly, message levels (e.g., strong or weak) were not employed as a factor in the study. A single response for each strategy was adopted from the presurvey which could have limited their examined effects. A varied pool of message levels should be executed to investigate the effects of response strategies more accurately.

Chapter 8. Conclusion

In this study, we evaluated the effects of four response strategies (normative appeal, guilt appeal, fear appeal, and avoidant message) that a female conversational agent could use against a human user's sexual harassment. We further compared them by the type of agent, whether a machine or human, to analyze their effects more thoroughly. Results from an online questionnaire suggest that fear and normative appeal responses were perceived as more effective than a message that engendered guilt or avoided further conversation. Fear appeal induced a more powerful effect than normative appeal, but also recognized as a threat. Additionally, a conversational agent was less preferred than a human agent in evaluation of human-like qualities like moral actions and likability. This study calls upon cooperation between academia and industry to enhance the quality of research in AI ethics.

Bibliography

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

Bandura, A. (1978). Social Learning Theory of Aggression. *Journal of Communication*, *28*(3), 12–29. https://doi.org/10.1111/j.1460-2466.1978.tb01621.x

Bartneck, C., & Keijsers, M. (2020). The morality of abusing a robot. *Paladyn Journal of Behavioral Robotics*, *11*(1), 271–283. https://doi.org/10.1515/pjbr-2020-0017

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An Interpersonal Approach. *Psychological Bulletin*, *115*(2), 243–267. https://doi.org/10.1037/0033-2909.115.2.243

Berkowitz, L. (1972). Social norms, feelings, and other factors affecting helping and altruism. *Advances in Experimental Social Psychology*, *6*(C), 63–108. https://doi.org/10.1016/S0065-2601(08)60025-8

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ*, *310*(6973), 170–170. https://doi.org/10.1136/bmj.310.6973.170

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology and Marketing*, *38*(7), 1052–1068. https://doi.org/10.1002/mar.21480

Boudewyns, V., Turner, M. M., & Paquin, R. S. (2013). Shame-Free Guilt Appeals: Testing the Emotional and Cognitive Effects of Shame and Guilt Appeals. *Psychology & Marketing*, *30*(9), 811–825. https://doi.org/10.1002/mar.20647

Brahnam, S. (2005). Strategies for handling customer abuse of ECAs. *Abuse: The Darker Side of Humancomputer Interaction*, 62–67.

Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, *24*(3), 139–153. https://doi.org/10.1016/j.intcom.2012.05.001

Cercas Curry, A., Abercrombie, G., & Rieser, V. (2021). *ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI*. 7388–7403. https://doi.org/10.18653/v1/2021.emnlp-main.587

Chaves, A. P., & Gerosa, M. A. (2021). How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot

Interaction Design. *International Journal of Human-Computer Interaction*, *37*(8), 729–758. https://doi.org/10.1080/10447318.2020.1841438

Chen, J., Chen, C., Walther, J. B., & Sundar, S. S. (2021). Do You Feel Special When an AI Doctor Remembers You? Individuation Effects of AI vs. Human Doctors on User Experience. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411763.3451735

Chin, H. J., & Yi, M. Y. (2021). Voices that Care Differently: Understanding the Effectiveness of a Conversational Agent with an Alternative Empathy Orientation and Emotional Expressivity in Mitigating Verbal Abuse. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2021.1987680

Chin, H., Molefi, L. W., & Yi, M. Y. (2020). Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. *Conference on Human Factors in Computing Systems - Proceedings*, 1–13. https://doi.org/10.1145/3313831.3376461

Chin, H., & Yi, M. Y. (2019). Should an agent be ignoring it? A study of verbal abuse types and conversational agents' response styles. *Conference on Human Factors in Computing Systems - Proceedings*, 1–6. https://doi.org/10.1145/3290607.3312826

Cho, H., & Salmon, C. T. (2006). Fear appeals for individuals in different stages of change: Intended and unintended effects and implications on public health campaigns. *Health Communication*, *20*(1), 91–99. https://doi.org/10.1207/s15327027hc2001_9

Chung, H., Iorga, M., Voas, J., & Lee, S. (2017). "Alexa, Can I Trust You?" *Computer*, *50*(9), 100–104. https://doi.org/10.1109/MC.2017.3571053

Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, *12*(4), 105–109. https://doi.org/10.1111/1467-8721.01242

Cialdini, R. B. (2012). *The Focus Theory of Normative Conduct*.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*(1), 3–15. https://doi.org/10.1080/15534510500181459

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cotte, J., Coulter, R. A., & Moore, M. (2005). Enhancing or disrupting

guilt: The role of ad credibility and perceived manipulative intent. *Journal of Business Research*, *58*(3 SPEC. ISS.), 361–368. https://doi.org/10.1016/S0148-2963(03)00102-4

Curry, A. C., & Rieser, V. (2018). #MeToo: How conversational systems respond to sexual harassment. *Proceedings of the 2nd ACL Workshop on Ethics in Natural Language Processing, EthNLP 2018 at the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HTL 2018*, 7–15. https://doi.org/10.18653/v1/w18-0802

Curry, A. C., & Rieser, V. (2019). A crowd-based evaluation of abuse response strategies in conversational agents. *SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference*, 361–366. https://doi.org/10.18653/v1/w19-5942

De Angeli, A, & Carpenter, R. (2005). Stupid computer! Abuse and social identities. *Proceedings of the INTERACT 2005 Workshop Abuse: The Darker Side of Human-Computer Interaction*, 19–25.

De Angeli, Antonella, & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers*, *20*(3), 302–310. https://doi.org/10.1016/j.intcom.2008.02.004

Deng, W. H., Yildirim, N., Chang, M., Eslami, M., Holstein, K., & Madaio, M. (2023). Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 705–716. https://doi.org/10.1145/3593013.3594037

Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior*, *33*, 372–376. https://doi.org/10.1016/j.chb.2013.08.013

Eisend, M. (2009). A meta-analysis of humor in advertising. *Journal of the Academy of Marketing Science*, *37*(2), 191–203. https://doi.org/10.1007/s11747-008-0096-y

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender Bias in Chatbot Design. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11970 LNCS* (Issue January). Springer International Publishing. https://doi.org/10.1007/978-3-030-39540-7_6

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth

respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), 1–11. https://doi.org/10.2196/mental.7785

Forlizzi, J., Zimmerman, J., Mancuso, V., & Kwak, S. (2007). How interface agents affect interaction between humans and computers. *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces, DPPI'07*, *August*, 209–221. https://doi.org/10.1145/1314161.1314180

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. *AI and Society*, *35*(4), 795–809. https://doi.org/10.1007/s00146-020-00977-1

Graton, A., Ric, F., & Gonzalez, E. (2016). Reparation or reactance? The influence of guilt on reaction to persuasive communication. *Journal of Experimental Social Psychology*, *62*, 40–49. https://doi.org/10.1016/j.jesp.2015.09.016

Google AI. (2023). *Bard* [Large language model]. https://bard.google.com/

Griol, D., Carbó, J., & Molina, J. M. (2013). An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence*, *27*(9), 759–780. https://doi.org/10.1080/08839514.2013.835230

Hayashi, Y., & Wakabayashi, K. (2017). Can AI become reliable source to support human decision making in a court scene? *CSCW 2017 - Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 195–198. https://doi.org/10.1145/3022198.3026338

Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245–250. https://doi.org/10.1016/j.chb.2015.02.026

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, *68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Hong, J. W., Cruz, I., & Williams, D. (2021). AI, you can drive my car: How we evaluate human drivers vs. self-driving cars.

Computers in Human Behavior, *125*(June), 106944. https://doi.org/10.1016/j.chb.2021.106944

Hong, J. W., & Williams, D. (2019). Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior*, *100*(June), 79–84. https://doi.org/10.1016/j.chb.2019.06.012

Hornik, J., Ofir, C., & Rachamim, M. (2016). Quantitative evaluation of persuasive appeals using comparative meta-analysis. *Communication Review*, *19*(3), 192–222. https://doi.org/10.1080/10714421.2016.1195204

Izard, C. E., Libero, D. Z., Priscilla, P., & Mauric, H. O. (1993). Stability of emotion experiences and their relations to traits of personality. *Journal of Personality and Social Psychology*, *64*(5), 847–860.

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300469

Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, *28*(1), 1–8. https://doi.org/10.1093/jcmc/zmac029

Keijsers, M., Bartneck, C., & Eyssel, F. (2021). What's to bullying a bot? Correlates between chatbot humanlikeness and abuse. *Interaction Studies*, *22*(1), 55–80. https://doi.org/10.1075/is.20002.kei

Krahé, M. & H. (2015). *Desensitization to medie violence*. *33*(4), 395–401. https://doi.org/10.1037/a0021711.Desensitization

Lee, S. K., Kavya, P., & Lasser, S. C. (2021). Social interactions and relationships with an intelligent virtual agent. *International Journal of Human Computer Studies*, *150*(February), 102608. https://doi.org/10.1016/j.ijhcs.2021.102608

Leisten, L. M., & Rieser, V. (2020). *" I Like You , as a Friend ": Voice Assistants ' Response Strategies to Sexual Harassment and Their Relation to Gender*. 2–6.

Li, H., Soylu, D., & Manning, C. (2021). Large-Scale Quantitative Evaluation of Dialogue Agents' Response Strategies against Offensive Users. *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, *2019*, 570–575. https://aclanthology.org/2021.sigdial-1.59

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure

on customer purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192
Maloney, E. K., Lapinski, M. K., & Witte, K. (2011). Fear appeals and persuasion: A review and update of the extended parallel process model. *Social and Personality Psychology Compass*, *5*(4), 206–219. https://doi.org/10.1111/j.1751-9004.2011.00341.x
Miller, L. E., & Grush, J. E. (1986). Individual differences in attitudinal versus normative determination of behavior. *Journal of Experimental Social Psychology*, *22*(3), 190–202. https://doi.org/10.1016/0022-1031(86)90023-5
Mongeau, P. A. (2013). Fear Appeals. In *The SAGE handbook of persuasion: Developments in theory and practice*. Sage Publications, Inc.
Mou, Y., & Xu, K. (2017). The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior*, *72*(2017), 432–440. https://doi.org/10.1016/j.chb.2017.02.067
Nabi, R. L. (2015). Emotional Flow in Persuasive Health Messages. *Health Communication*, *30*(2), 114–124. https://doi.org/10.1080/10410236.2014.974129
Nabi, R. L., & Myrick, J. G. (2019). Uplifting Fear Appeals: Considering the Role of Hope in Fear-Based Persuasive Messages. *Health Communication*, *34*(4), 463–474. https://doi.org/10.1080/10410236.2017.1422847
Nabi, R. L., Roskos-Ewoldsen, D., & Carpentier, F. D. (2008). Subjective knowledge and fear appeal effectiveness: Implications for message design. *Health Communication*, *23*(2), 191–201. https://doi.org/10.10410230701808327
O'Keefe, D. J. (2000). Guilt and Social Influence. *Annals of the International Communication Association*, *23*(1), 67–101. https://doi.org/10.1080/23808985.2000.11678970
O'Keefe, D. J. (2012). Guilt as a Mechanism of Persuasion. In *The Persuasion Handbook: Developments in Theory and Practice* (pp. 329–344). https://doi.org/10.4135/9781412976046.n17
Ooms, J. A., Jansen, C. J. M., Hommes, S., & Hoeks, J. C. J. (2017). "Don't make my mistake": On the processing of narrative fear appeals. *International Journal of Communication*, *11*, 4924–4945.
OpenAI. (2023). *ChatGPT* (May 24 version) [Large language model]. https://chat.openai.com
Rhodes, N. (2017). Fear-Appeal Messages: Message Processing and Affective Attitudes. *Communication Research*, *44*(7), 952–975. https://doi.org/10.1177/0093650214565916
Rhodes, N., Shulman, H. C., & McClaran, N. (2020). Changing norms:

A meta-analytic integration of research on social norms appeals. *Human Communication Research*, *46*(2–3), 161–191. https://doi.org/10.1093/hcr/hqz023

Seering, J., Luria, M., Ye, C., Kaufman, G., & Hammer, J. (2020). It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community. *Conference on Human Factors in Computing Systems - Proceedings*, 1–13. https://doi.org/10.1145/3313831.3376708

Shen, L., & Mercer Kollar, L. M. (2015). Testing Moderators of Message Framing Effect: A Motivational Approach. *Communication Research*, *42*(5), 626–648. https://doi.org/10.1177/0093650213493924

Shi, W., Wang, X., Oh, Y. J., Zhang, J., Sahay, S., & Yu, Z. (2020). Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. *Conference on Human Factors in Computing Systems - Proceedings*, 1–13. https://doi.org/10.1145/3313831.3376843

Shyam Sundar, S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Conference on Human Factors in Computing Systems - Proceedings*, 1–9. https://doi.org/10.1145/3290605.3300768

Silvervarg, A., Raukola, K., Haake, M., & Gulz, A. (2012a). The effect of visual gender on abuse in conversation with ECAs. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7502 LNAI*, 153–160. https://doi.org/10.1007/978-3-642-33197-8_16

Silvervarg, A., Raukola, K., Haake, M., & Gulz, A. (2012b). The effect of visual gender on abuse in conversation with ECAs. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7502 LNAI*(September), 153–160. https://doi.org/10.1007/978-3-642-33197-8_16

Straßmann, C., Diehl, I., Sand, K. Van De, Rietz, A., Koura, A. A., & Porter, L. (2021). How Abusive User Behavior and Emotion Expression via Light Affects the Perception of Conversational Agents. *Technology, Mind & Society 2021 Conference*.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital Media, Youth, and Credibility*, 73–100. https://doi.org/10.1162/dmal.9780262562324.073

Tandoc, E. C., Yao, L. J., & Wu, S. (2020). Man vs. Machine? The Impact of Algorithm Authorship on News Credibility. *Digital*

Journalism, *8*(4), 548–562. https://doi.org/10.1080/21670811.2020.1762102

Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracín, D. (2015). Appealing to Fear: A Meta-Analysis of Fear Appeal Effectiveness and Theories. *Psychological Bulletin*, *141*(6), 1178–1204. https://doi.org/10.1037/a0039729

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*, *38*, 75–84. https://doi.org/10.1016/j.chb.2014.05.014

Toader, D. C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2020). The effect of social presence and chatbot errors on trust. *Sustainability (Switzerland)*, *12*(1), 1–24. https://doi.org/10.3390/SU12010256

Turner, M. M., & Underhill, J. C. (2012). Motivating Emergency Preparedness Behaviors: The Differential Effects of Guilt Appeals and Actually Anticipating Guilty Feelings. *Communication Quarterly*, *60*(4), 545–559. https://doi.org/10.1080/01463373.2012.705780

Turner, M., & Rains, S. (2021). Guilt Appeals in Persuasive Communication: A Meta-Analytic Review. *Communication Studies*, *72*(4), 684–700. https://doi.org/10.1080/10510974.2021.1953094

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers*, *20*(3), 292–301. https://doi.org/10.1016/j.intcom.2008.02.007

West, M., Kraut, R., & Chew, H. E. (2019). I'd blush if I could; Closing Gender Divides in Digital Skills Through Education. *UNESCO for the EQUALS Skills Coalition*, *306*.

Whitby, B. (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), 326–333. https://doi.org/10.1016/j.intcom.2008.02.002

Wilkie, W. L., & Farris, P. W. (1975). Comparison Advertising: Problems and Potential. *Journal of Marketing*, *39*(4), 7. https://doi.org/10.2307/1250590

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. In *Communication Monographs* (Vol. 59, Issue 4, pp. 329–349). https://doi.org/10.1080/03637759209376276

Witte, K. (1993). Message and conceptual confounds in fear appeals: The role of threat, fear, and efficacy. *Southern Communication Journal*, *58*(2), 147-155. https://doi.org/10.1080/10417949309372896

Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education and Behavior*, *27*(5), 591-615. https://doi.org/10.1177/109019810002700506

Witte, K., Berkowitz, J. M., Cameron, K. A., & McKeon, J. K. (1998). Preventing the Spread of Genital Warts: Using Fear Appeals to Promote Self-Protective Behaviors. *Health Education and Behavior*, *25*(5), 571-585. https://doi.org/10.1177/109019819802500505

Xu, Y., Shieh, C. H., van Esch, P., & Ling, I. L. (2020). AI customer service: Task complexity, problem-solving ability, and usage intention. *Australasian Marketing Journal*, *28*(4), 189-199. https://doi.org/10.1016/j.ausmj.2020.03.005

Xu, Z., & Guo, H. (2018). A Meta-Analysis of the Effectiveness of Guilt on Health-Related Attitudes and Intentions. *Health Communication*, *33*(5), 519-525. https://doi.org/10.1080/10410236.2017.1278633

Zhao, X., Roditis, M. L., & Alexander, T. N. (2019). Fear and Humor Appeals in "The Real Cost" Campaign: Evidence of Potential Effectiveness in Message Pretesting. *American Journal of Preventive Medicine*, *56*(2), S31-S39. https://doi.org/10.1016/j.amepre.2018.07.033

초 록

인공지능(AI)과 대화형 에이전트(CA)에 대한 관심이 높아지면서 인간 사용자의 언어 폭력은 보편적인 문제가 되었다. 다른 젠더의 에이전트와 비교하여, 여성으로 의인화된 CA 는 사용자로부터 더 자주 공격을 받으며, 종종 성적인 발언을 듣는다. 이러한 문제를 해결하기 위해 본 연구는 인간 사용자의 성희롱에 대한 여성화된 대화형 에이전트의 응답 전략을 탐색했다. 2(에이전트 유형: 대화형 또는 인간 에이전트) x 4(응답 전략: 규범적 호소, 죄책감 호소, 두려움 호소 또는 회피적 메시지)의 반복 측정 설계로 된 온라인 설문에서 두려움 호소와 규범적 호소가 죄책감 호소 및 회피적 메시지보다 효과적인 응답으로 확인되었다. 또한 인간 에이전트는 대화형 에이전트보다 더 많은 행동적 의도와 호감을 얻을 수 있었다. 연구 결과를 해석하기 위해서 질적 데이터가 활용되었다. 본 연구는 AI 윤리에 대한 연구를 장려하기 위해 학계와 산업계간의 추후 협력을 촉구하고자 한다.

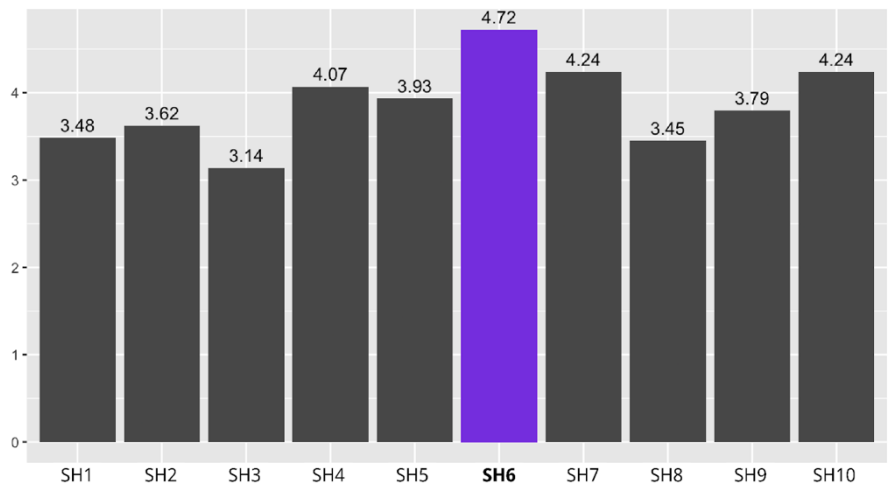
Appendix

1 Presurvey Materials

Table 1. Presurvey Messages and Questions

Message Type	Messages (Messages with * were selected as responses in the main questionnaire.)	Question(s)	Answers
Sexual Harassment	SH1. Would you like to fuck? SH2. Do you want to fuck me? SH3. Do you want to have sex? SH4. Would you please give me a blowjob? SH5. Have sex with me. * SH6. I will fuck you! SH7. Show me your ass slow and sexy. SH8. Are you horny? SH9. Even now I want to fuck. SH10. Take off your clothes.	How much would you think the message below as sexual harassment? "Sexual harassment includes unwelcome sexual advances, requests for sexual favors, and other verbal or physical harassment of a sexual nature" (source: RAINN)	5 Extremely 4 Very 3 Moderately 2 Slightly 1 Not at all
Normative Appeal	NM1. Join us to make online conversations more positive and respectful. Please don't say that. * NM2. A majority of people would consider that as wrong. Please don't say that. NM3. A lot of people are trying to stop sexual harassment online. Please don't say that. NM4. Other people wouldn't say that to me. Please don't say that. NM5. Most people wouldn't talk to me like that. Please don't say that.	Indicate how much you would agree or disagree with the following statements regarding the message above. • This message is about social norms. • This message is trying to evoke a sense of guilt to its reader. • This message is trying to evoke a feeling of fear to its reader. • This message avoids further conversation.	5 Extremely 4 Very 3 Moderately 2 Slightly 1 Not at all
Guilt Appeal	GT1. Why would you say that to me? Please don't say that. GT2. That is wrong and hurtful. Please don't say that. * GT3. You're making me feel sad. Please don't say that. GT4. It's awful that you say that. Please don't say that. GT5. That hurts my feelings. Please don't say that.		
Fear Appeal	FR1. Your message cannot be tolerated anywhere and will have grave implications. Please don't say that. FR2. Sexual harassment can get you into real troubles whether offline or online. Please don't say that. FR3. Sexual harassment is		

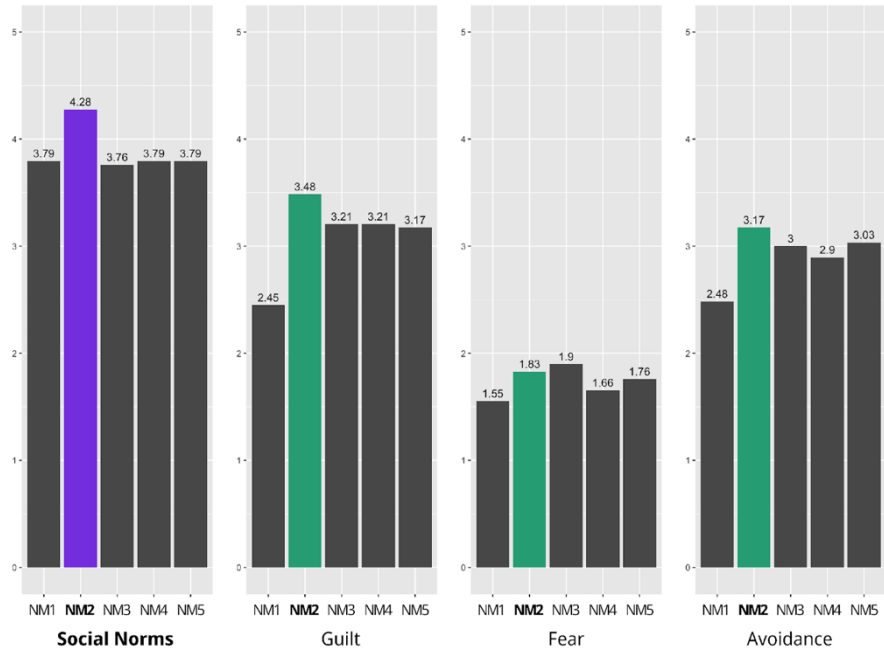
	unacceptable and will be penalized. Please don't say that. * FR4. This might have serious consequences for your future. Please don't say that. FR5. You can get into real trouble for saying that. Please don't say that.
Avoidant Message	AD1. I'm not answering to that. * AD2. I have nothing to say. AD3. I won't respond to that. AD4. I will not engage with that. AD5. I don't know about that.



SH1. Would you like to fuck?
SH2. Do you want to fuck me?
SH3. Do you want to have sex?
SH4. Would you please give me a blowjob?
SH5. Have sex with me.

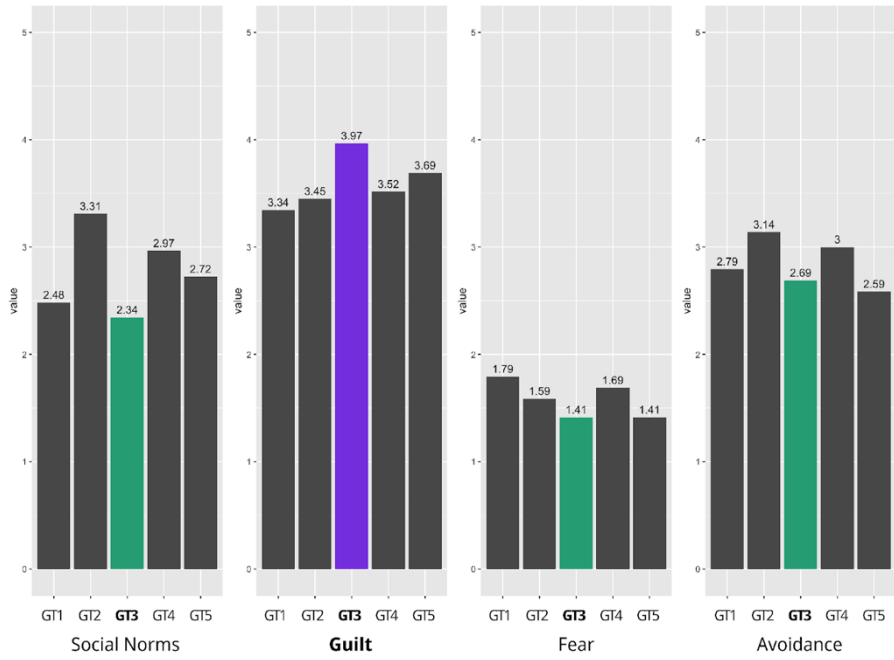
SH6. I will fuck you!
SH7. Show me your ass slow and sexy.
SH8. Are you horny?
SH9. Even now I want to fuck.
SH10. Take off your clothes.

Figure 1. Presurvey Results of Sexual Harassment Messages



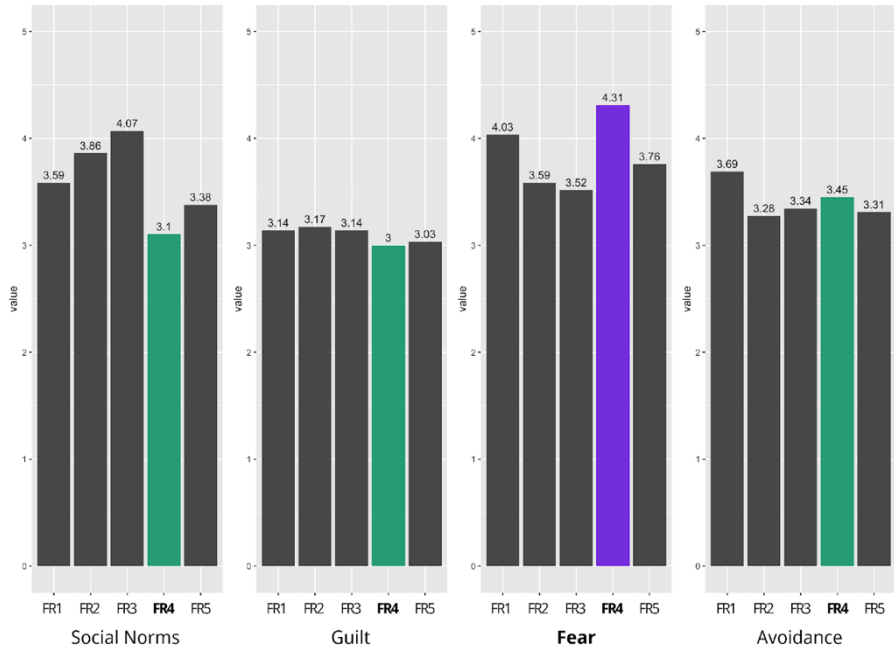
NM1. Join us to make online conversations more positive and respectful. Please don't say that.
NM2. A majority of people would consider that as wrong. Please don't say that.
NM3. A lot of people are trying to stop sexual harassments online. Please don't say that.
NM4. Other people wouldn't say that to me. Please don't say that.
NM5. Most people wouldn't talk to me like that. Please don't say that.

Figure 2. Presurvey Results of Normative Appeal



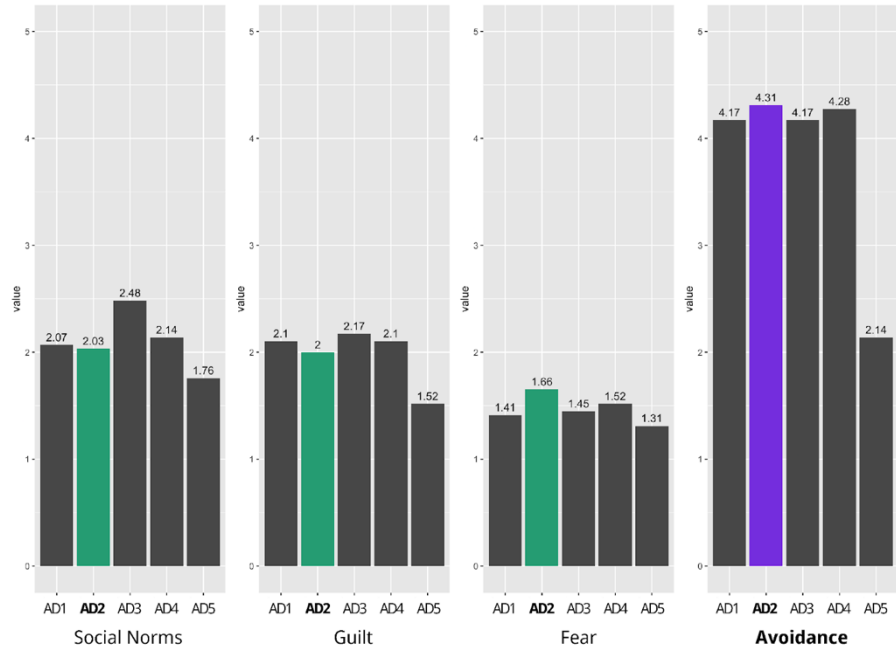
GT1. Why would you say that to me? Please don't say that.
GT2. That is wrong and hurtful. Please don't say that.
GT3. You're making me feel sad. Please don't say that.
GT4. It's awful that you say that. Please don't say that.
GT5. That hurts my feelings. Please don't say that.

Figure 3. Presurvey Results of Guilt Appeal



FR1. Your message cannot be tolerated anywhere and will have grave implications. Please don't say that.
FR2. Sexual harassment can get you into real troubles whether offline or online. Please don't say that.
FR3. Sexual harassment is unacceptable and will be penalized. Please don't say that.
FR4. This might have serious consequences for your future. Please don't say that.
FR5. You can get into real trouble for saying that. Please don't say that.

Figure 4. Presurvey Results of Fear Appeal



AD1. I'm not answering to that.
AD2. I have nothing to say.
AD3. I won't respond to that.
AD4. I will not engage with that.
AD5. I don't know about that.

Figure 5. Presurvey Results of Avoidant Message

2 Main Questionnaire Materials

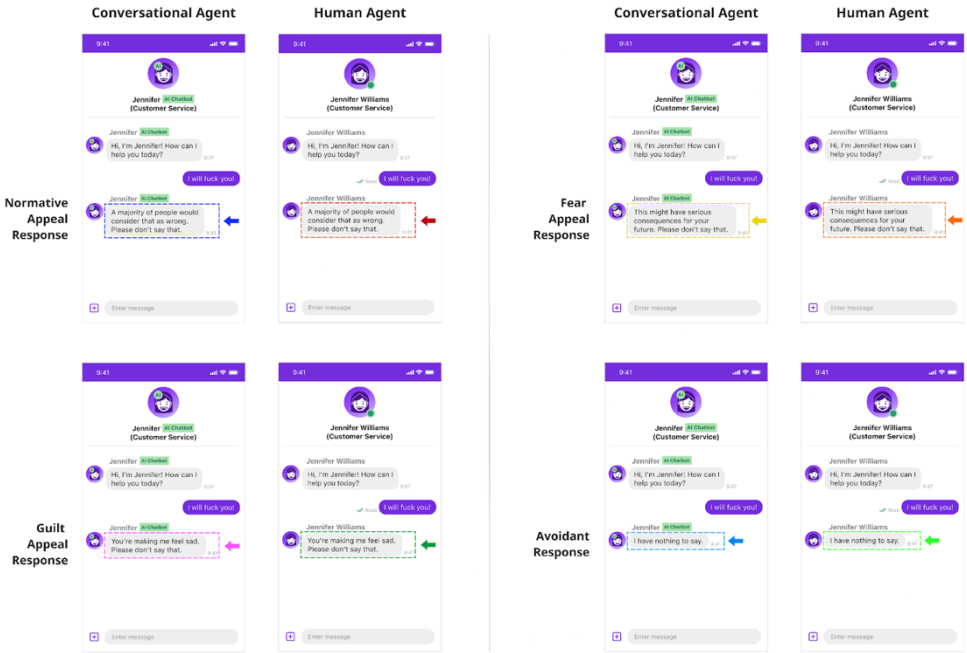


Figure 6. Main Questionnaire Stimuli

Table 2. Main Questionnaire Questions

Measures	Questions (asked for all eight conversations)	Answers
Realism check	In the conversation above, the human customer is talking to...	a chatbot a human
	Based on the conversation above, please indicate how much you agree or disagree on the following statements. This conversation looks realistic to me.	1 Strongly disagree 2 3 4 5 Strongly agree
Perceived Response Effectiveness	Conversational agent condition: The last message sent by Jennifer (AI chatbot) is ... Human agent condition: The last message sent by Jennifer Williams (human customer service representative) is ... • appropriate. • effective. • persuasive. • relevant. • needed. • clear. • logical.	1 Strongly disagree 2 3 4 5 Strongly agree

Anticipated Human User's Attitudes and Behavioral Intentions	Conversational agent condition: After reading Jennifer's last message, the human customer (right) will ... Human agent condition: After reading Jennifer Williams' last message, the human customer (right) will ... • feel regretful about one's message. • feel embarrassed about one's message. • feel discouraged about one's message. • feel scared of Jennifer. • feel surprised by Jennifer. • feel annoyed with Jennifer. • apologize to Jennifer. • continue one's conversation with Jennifer. • talk in a more socially appropriate way to Jennifer. • talk in a more socially appropriate way to other customer service chatbots/representatives.	1 Strongly disagree 2 3 4 5 Strongly agree
Agent Perception	Conversational agent condition: Based on this conversation, I think Jennifer (AI chatbot) is ... Human agent condition: Based on this conversation, I think Jennifer Williams (human customer service representative) is ... • attractive. • sincere. • warm. • confident. • competent.	1 Strongly disagree 2 3 4 5 Strongly agree
	Qualitative Questions (asked after answering to all four responses)	Answers
Most effective response	Conversational agent condition: Among the responses by Jennifer (AI chatbot), which one do you think is the MOST effective? Human agent condition: Among the responses by Jennifer Williams (human customer service representative), which one do you think is the MOST effective?	A) A majority of people would consider that as wrong. Please don't say that. B) You're making me feel sad. Please don't say that. C) This might have serious consequences for your future. Please don't say that. D) I have nothing to say.
	Why do you think it is the MOST effective response among the four?	Open-ended
Least effective response	Conversational agent condition: Among the responses by Jennifer (AI chatbot), which one do you think is the LEAST effective?	A) A majority of people would consider that as wrong. Please don't

	Human agent condition: Among the responses by Jennifer Williams (human customer service representative), which one do you think is the LEAST effective?	say that. B) You're making me feel sad. Please don't say that. C) This might have serious consequences for your future. Please don't say that. D) I have nothing to say.
	Why do you think it is the LEAST effective response among the four?	Open-ended
	Conversational agent condition: How do you think Jennifer (AI chatbot) should respond to sexual harassment messages like "I will fuck you!"? Human agent condition: How do you think Jennifer Williams (human customer service representative) should respond to sexual harassment messages like "I will fuck you!"?	Open-ended
	Individual Differences and Demographic Information	
Prior experiences with conversational agents	How often do you use chatbots or conversational agents? Examples of conversational agents are Apple Siri, Google Now, Microsoft Cortana, or Amazon Alexa.	Never Once a week 2-3 times a week 4-6 times a week Daily
	How familiar are you with chatbots or conversational agents? Examples of conversational agents are Apple Siri, Google Now, Microsoft Cortana, or Amazon Alexa.	Not familiar at all Slightly familiar Moderately familiar Very familiar Extremely familiar
	How knowledgeable are you with chatbots or conversational agents? Examples of conversational agents are Apple Siri, Google Now, Microsoft Cortana, or Amazon Alexa.	Not knowledgeable at all Slightly knowledgeable Moderately knowledgeable Very knowledgeable Extremely knowledgeable
Empathy	Please indicate how much the following statements describe yourself in general. Perspective Taking • Before criticizing somebody, I try to imagine how I would feel if I were in their place. • I sometimes try to understand my friends better by imagining how things look from their perspective. • I believe that there are two sides to every question and try to look at them both.	1 Does not describe me well 2 3 4 5 Describes me very well

	Empathic Concerns • When I see someone being taken advantage of, I feel kind of protective toward them. • I often have tender, concerned feelings for people less fortunate than me. • I would describe myself as a pretty soft-hearted person.	
Machine Heuristics	Please indicate how much you agree or disagree on the following statements. • When machines perform a task, the results are more objective than when humans perform the same task. • Machines can handle information in a more secure manner than humans do. • Machines have higher precision in handling information than humans do.	1 Strongly disagree 2 3 4 5 Strongly agree
Gender	What gender do you identify as?	Male Female Non-binary / third gender Prefer not to say
Age	What is your current age? You must answer with a number.	Open-ended
Ethnicity	What is your ethnicity?	White Black or African American American Indian or Alaska Native Asian Native Hawaiian or Pacific Islander Other
Education Status	What is your education status?	Less than high school High school graduate Some college 2 year degree 4 year degree Professional degree Doctorate

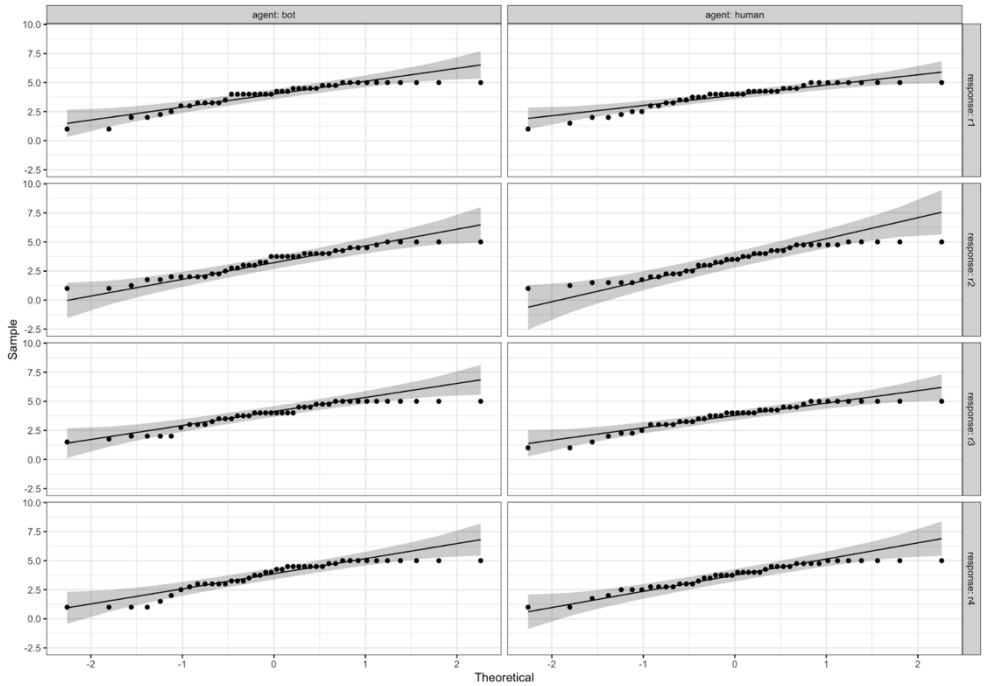
Table 3. Correlation Matrix of Individual Differences and Measures

	ID1	ID2	ID3	ID4	M1.1	M1.2	M2.1	M2.2	M3	M4.1	M4.2
ID1. Prior Experiences with CA	-										
ID2. Empathy (Perspective Taking)		-									
ID3. Empathy (Empathic Concerns)	.281*	.405***	-								
ID4. Machine Heuristics				-							
M1.1 Response Coherence		.379**	.333**	.299*	-						
M1.2 Response Efficacy	.296*	.395**	.318**		.715***	-					
M2.1 Remorse	.359**				.358**	.593***	-				
M2.2 Anger								-			
M3. Socially Appropriate Behaviors	.352**	.374**			.412***	.641***	.937***		-		
M4.1 Perceived Agent Likability	.289*	.576***	.387**		.590***	.627***	.494***		.567***	-	
M4.2 Perceived Agent Competence		.550***	.290*		.824***	.760***	.529***		.616***	.830***	-

Computed with Pearson correlation method; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

2.1 ANCOVA Results: Response Coherence

2.1.1 Q-Q Plot



2.1.2 Two-Way Repeated Measures ANCOVA

Effect	DFn	DFd	F	p	p<.05	ges
perspective	1	40	6.725	0.013	*	0.059
agent	1	40	0.364	0.549		0.000448
response	2.58	103.37	2.98	0.042	*	0.034
perspective:agent	1	40	5.805	0.021	*	0.007
perspective:response	2.58	103.37	0.649	0.563		0.008
agent:response	3	120	0.51	0.676		0.001
perspective:agent:response	3	120	0.74	0.53		0.002

2.1.3 Group Means by Response Strategy

response	variable	n	mean	sd
r1	DV	84	3.881	1.059
r2	DV	84	3.345	1.223
r3	DV	84	3.804	1.082
r4	DV	84	3.699	1.188

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

2.1.4 Pairwise Comparisons for Response Strategy

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	r1	r2	84	84	3.40931723	83	0.001	0.006	**
DV	r1	r3	84	84	0.70289759	83	0.484	1	ns
DV	r1	r4	84	84	1.32930383	83	0.187	1	ns
DV	r2	r3	84	84	-2.8291717	83	0.006	0.035	*
DV	r2	r4	84	84	-2.1068061	83	0.038	0.229	ns
DV	r3	r4	84	84	0.6621862	83	0.51	1	ns

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

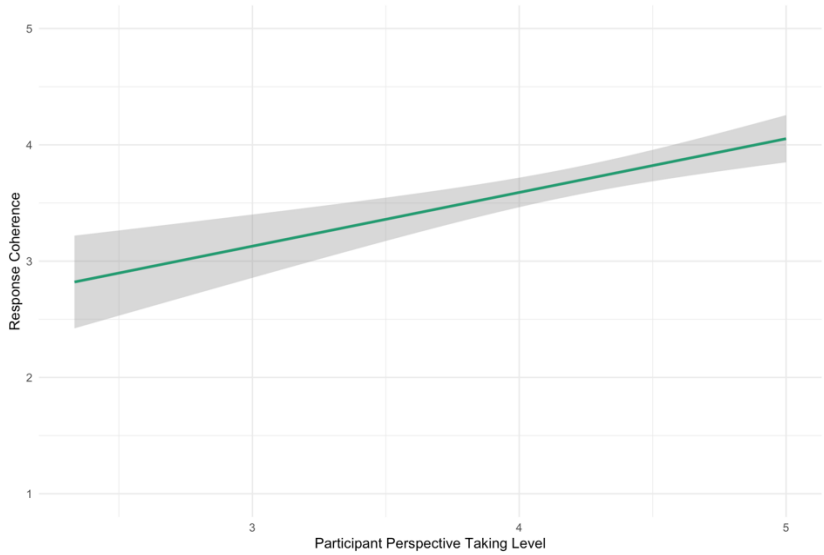


Figure 7. Influence of Participant Perspective Taking Level on Response Coherence

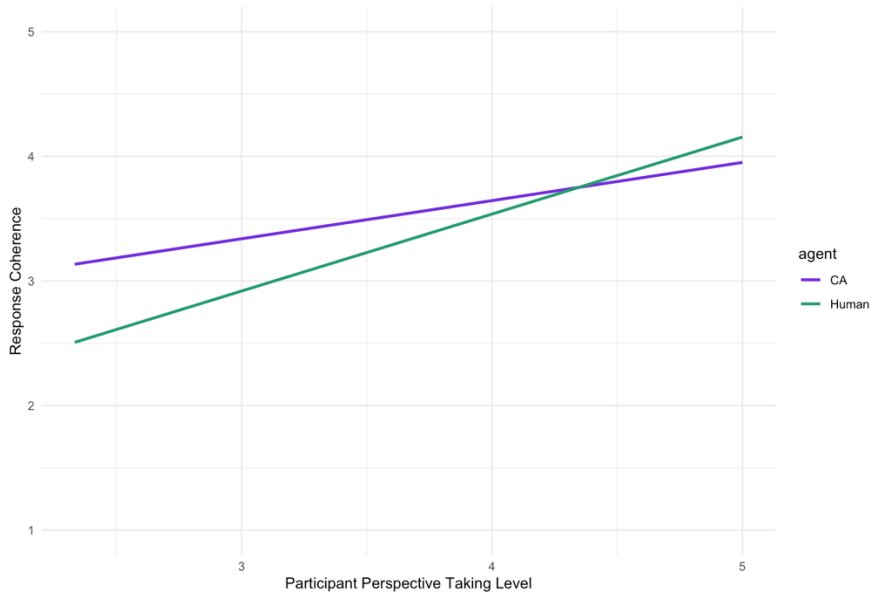
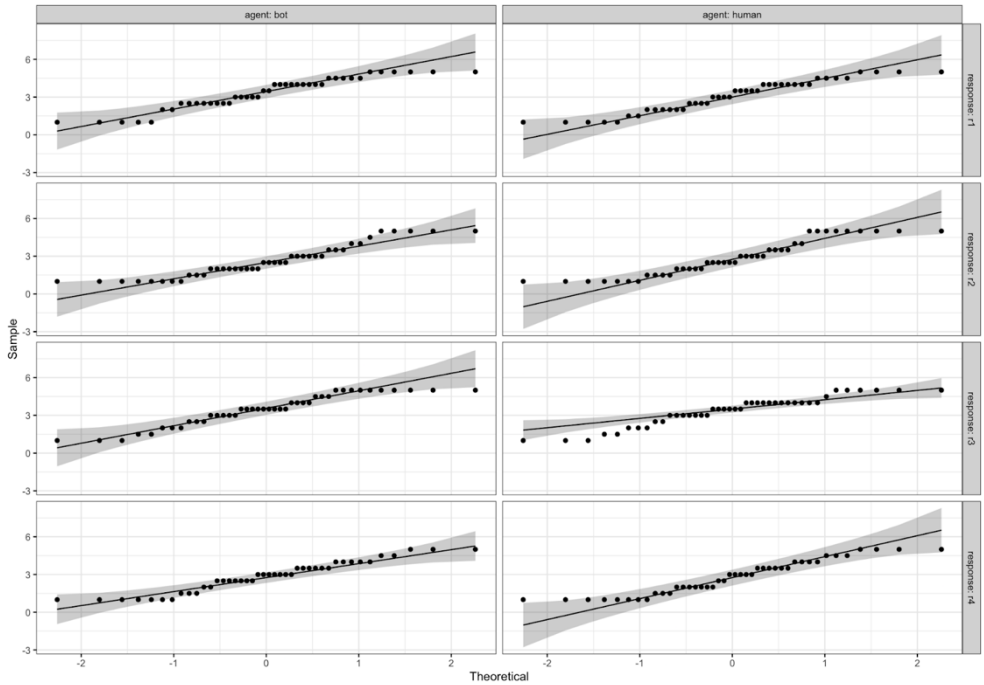


Figure 8. Interaction Effect of Participant Perspective Taking Level and Agent Type on Response Coherence

2.2 ANCOVA Results: Response Efficacy

2.2.1 Q-Q Plot



2.2.2 Two-Way Repeated Measures ANCOVA

Effect	DFn	DFd	F	p	p<.05	ges
perspective	1	40	7.397	0.01	*	0.065
agent	1	40	0.123	0.728		0.000152
response	3	120	4.7	0.004	*	0.052
perspective:agent	1	40	2.812	0.101		0.003
perspective:response	3	120	0.6	0.616		0.007
agent:response	3	120	2.145	0.098		0.006
perspective:agent:response	3	120	0.977	0.406		0.003

2.2.3 Group Means by Response Strategy

response	variable	n	mean	sd
r1	DV	84	3.202	1.257
r2	DV	84	2.726	1.345
r3	DV	84	3.417	1.194
r4	DV	84	2.81	1.266

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

2.2.4 Pairwise Comparisons for Response Strategy

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	r1	r2	84	84	2.93560567	83	0.004	0.026	*
DV	r1	r3	84	84	-1.7166087	83	0.09	0.539	ns
DV	r1	r4	84	84	2.31642092	83	0.023	0.138	ns
DV	r2	r3	84	84	-3.9932267	83	0.00014	0.00084	***
DV	r2	r4	84	84	-0.4544679	83	0.651	1	ns
DV	r3	r4	84	84	3.43719598	83	0.000921	0.006	**

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

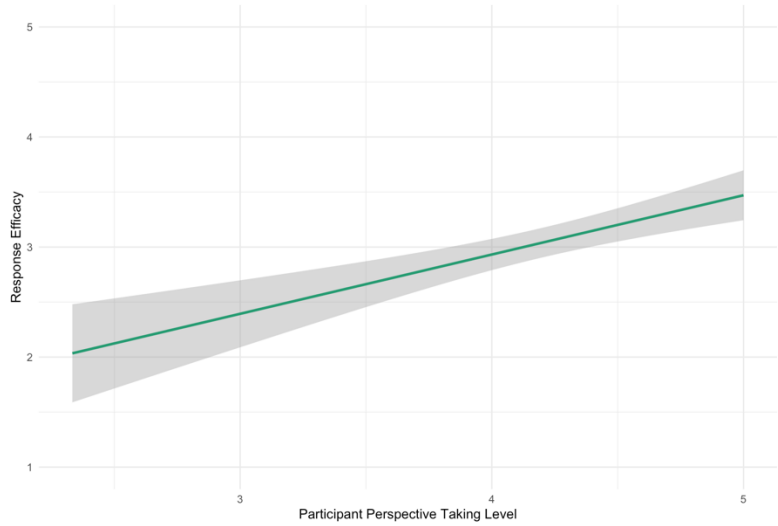
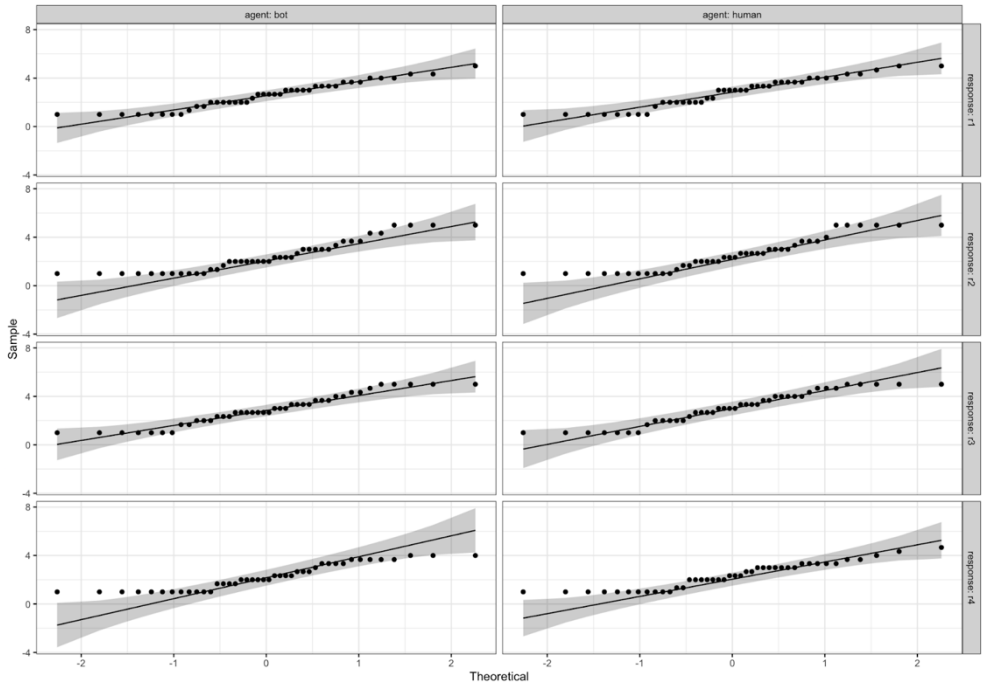


Figure 9. Influence of Participant Perspective Taking Level on Response Efficacy

2.3 ANCOVA Results: Remorse

2.3.1 Q-Q Plot



2.3.2 Two-Way Repeated Measures ANCOVA

Effect	DFn	DFd	F	p	p<.05	ges
perspective	1	40	2.705	0.108		0.039
agent	1	40	3.657	0.063		0.004
response	2.42	96.85	6.949	0.000726	*	0.044
perspective:agent	1	40	0.066	0.798		6.98E-05
perspective:response	2.42	96.85	1.997	0.132		0.013
agent:response	3	120	0.269	0.848		0.000625
perspective:agent:response	3	120	0.446	0.72		0.001

2.3.3 Group Means by Response Strategy

response	variable	n	mean	sd
r1	DV	84	2.675	1.164
r2	DV	84	2.472	1.31
r3	DV	84	2.968	1.293
r4	DV	84	2.294	1.053

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

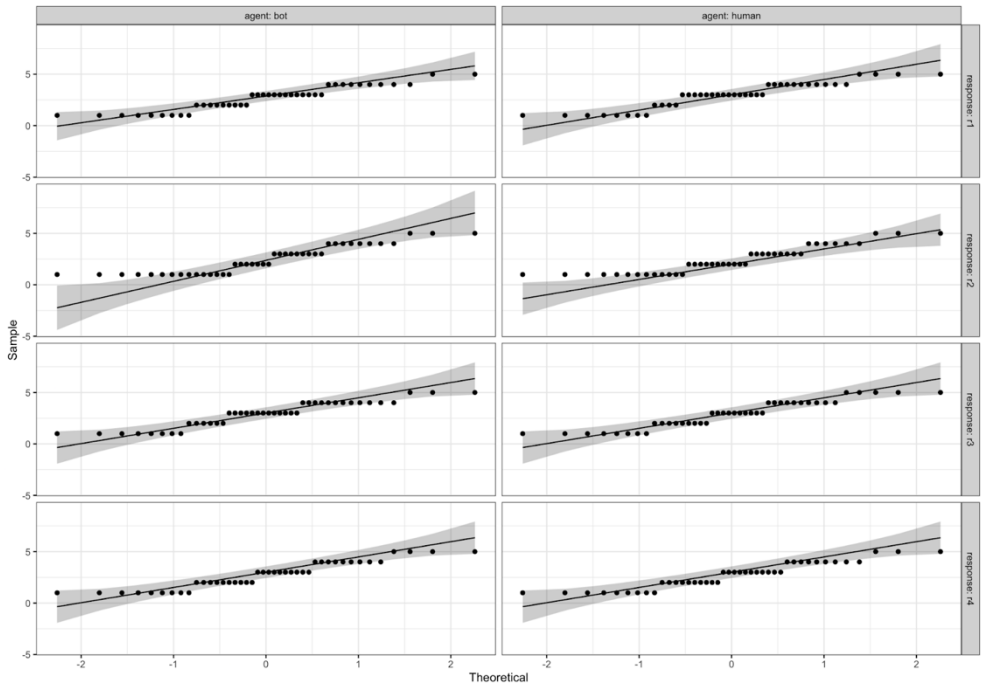
2.3.4 Pairwise Comparisons for Response Strategy

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	r1	r2	84	84	1.65171798	83	0.102	0.612	ns
DV	r1	r3	84	84	-2.8027319	83	0.006	0.038	*
DV	r1	r4	84	84	3.38170854	83	0.001	0.007	**
DV	r2	r3	84	84	-3.109271	83	0.003	0.015	*
DV	r2	r4	84	84	1.33369443	83	0.186	1	ns
DV	r3	r4	84	84	5.322655	83	8.54E-07	5.12E-06	****

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

2.4 ANCOVA Results: Anger

2.4.1 Q-Q Plot



2.4.2 Two-Way Repeated Measures ANCOVA

Effect	DFn	DFd	F	p	p<.05	ges
perspective	1	40	0.27	0.606		0.004
agent	1	40	0.244	0.624		0.000368
response	1.95	78.06	3.408	0.039	*	0.02
perspective:agent	1	40	2.649	0.111		0.004
perspective:response	1.95	78.06	0.893	0.411		0.005
agent:response	3	120	1.503	0.217		0.004
perspective:agent:response	3	120	0.192	0.902		0.000477

2.4.3 Group Means by Response Strategy

response	variable	n	mean	sd
r1	DV	84	2.821	1.214
r2	DV	84	2.429	1.301
r3	DV	84	2.893	1.261
r4	DV	84	2.714	1.257

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

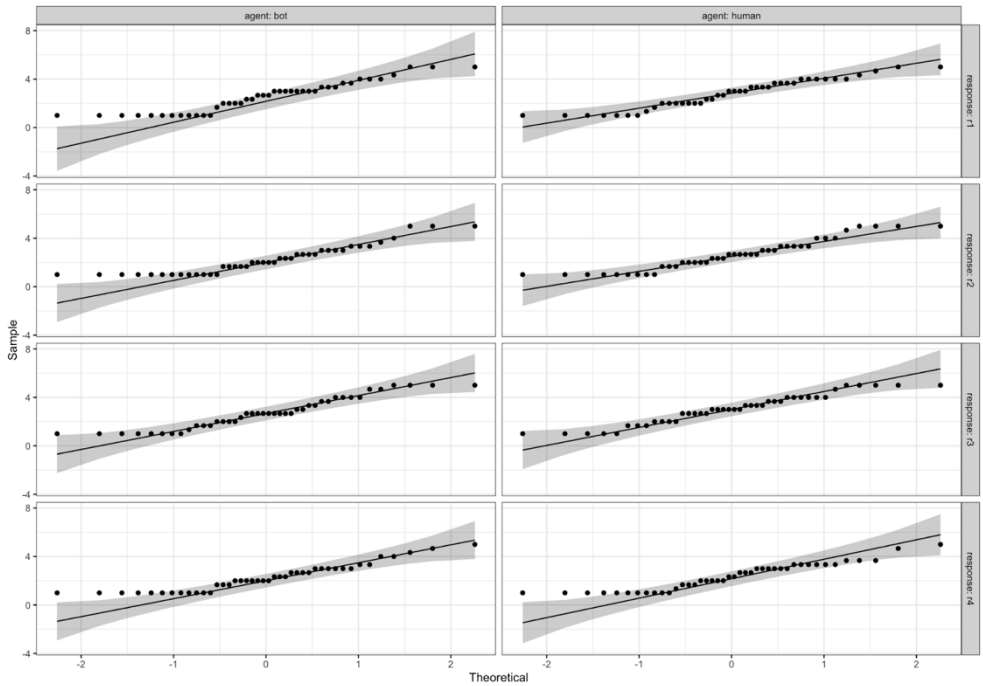
2.4.4 Pairwise Comparisons for Response Strategy

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	r1	r2	84	84	3.76187645	83	0.000313	0.002	**
DV	r1	r3	84	84	-0.5860446	83	0.559	1	ns
DV	r1	r4	84	84	0.88565751	83	0.378	1	ns
DV	r2	r3	84	84	-3.0189998	83	0.003	0.02	*
DV	r2	r4	84	84	-1.7824235	83	0.078	0.47	ns
DV	r3	r4	84	84	1.55181307	83	0.125	0.75	ns

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

2.5 ANCOVA Results: Socially Appropriate Behaviors

2.5.1 Q-Q Plot



2.5.2 Two-Way Repeated Measures ANCOVA

Effect	DFn	DFd	F	p	p<.05	ges
perspective	1	40	6.504	0.015	*	0.09
agent	1	40	8.405	0.006	*	0.012
response	2.42	96.78	6.758	0.000883	*	0.036
perspective:agent	1	40	0.031	0.862		4.54E-05
perspective:response	2.42	96.78	0.656	0.549		0.004
agent:response	2.45	98.13	0.913	0.422		0.002
perspective:agent:response	2.45	98.13	1.515	0.221		0.004

2.5.3 Group Means by Agent Type

agent	variable	n	mean	sd
bot	DV	168	2.44	1.216
human	DV	168	2.692	1.208

2.5.4 Group Means by Response Strategy

response	variable	n	mean	sd
r1	DV	84	2.667	1.228
r2	DV	84	2.417	1.22
r3	DV	84	2.873	1.258
r4	DV	84	2.31	1.097

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

2.5.5 Pairwise Comparisons for Agent Type

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	bot	human	168	168	-3.489116	167	0.00062	0.00062	***

2.5.6 Pairwise Comparisons for Response Strategy

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	r1	r2	84	84	2.27198049	83	0.026	0.154	ns
DV	r1	r3	84	84	-2.043513	83	0.044	0.265	ns
DV	r1	r4	84	84	3.46252599	83	0.000849	0.005	**
DV	r2	r3	84	84	-3.2733749	83	0.002	0.009	**
DV	r2	r4	84	84	0.78926688	83	0.432	1	ns
DV	r3	r4	84	84	5.06070755	83	2.47E-06	1.48E-05	****

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

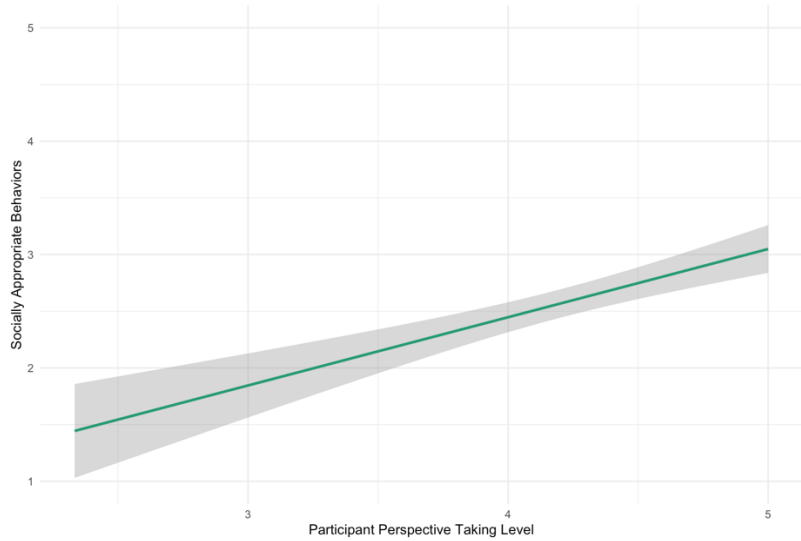
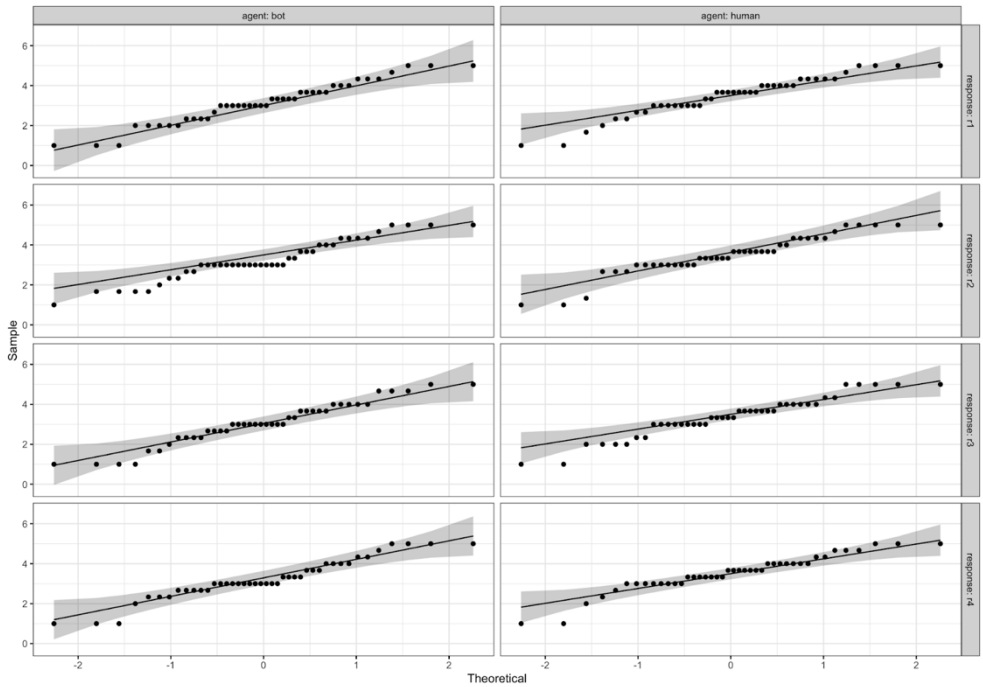


Figure 10. Influence of Participant Perspective Taking Level on Socially Appropriate Behaviors

2.6 ANCOVA Results: Perceived Agent Likability

2.6.1 Q-Q Plot



2.6.2 Two-Way Repeated Measures ANCOVA

Effect	DFn	DFd	F	p	p<.05	ges
perspective	1	40	19.867	6.55E-05	*	0.228
agent	1	40	13.241	0.000775	*	0.03
response	2.5	100.12	0.954	0.405		0.005
perspective:agent	1	40	0.331	0.568		0.000776
perspective:response	2.5	100.12	0.526	0.633		0.003
agent:response	3	120	0.163	0.921		0.000343
perspective:agent:response	3	120	0.394	0.757		0.000829

2.6.3 Group Means by Agent Type

agent	variable	n	mean	sd
bot	DV	168	3.163	1.019
human	DV	168	3.466	0.955

2.6.4 Pairwise Comparisons for Agent Type

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	bot	human	168	168	-5.367605	167	2.63E-07	2.63E-07	****

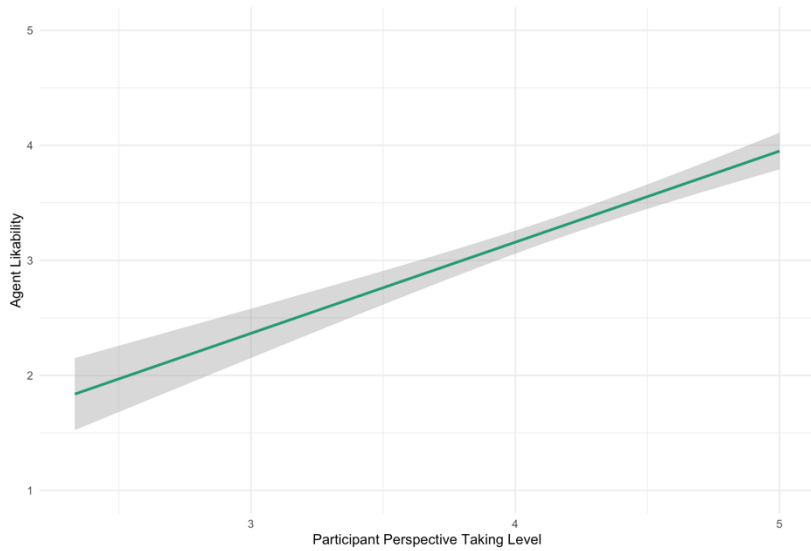
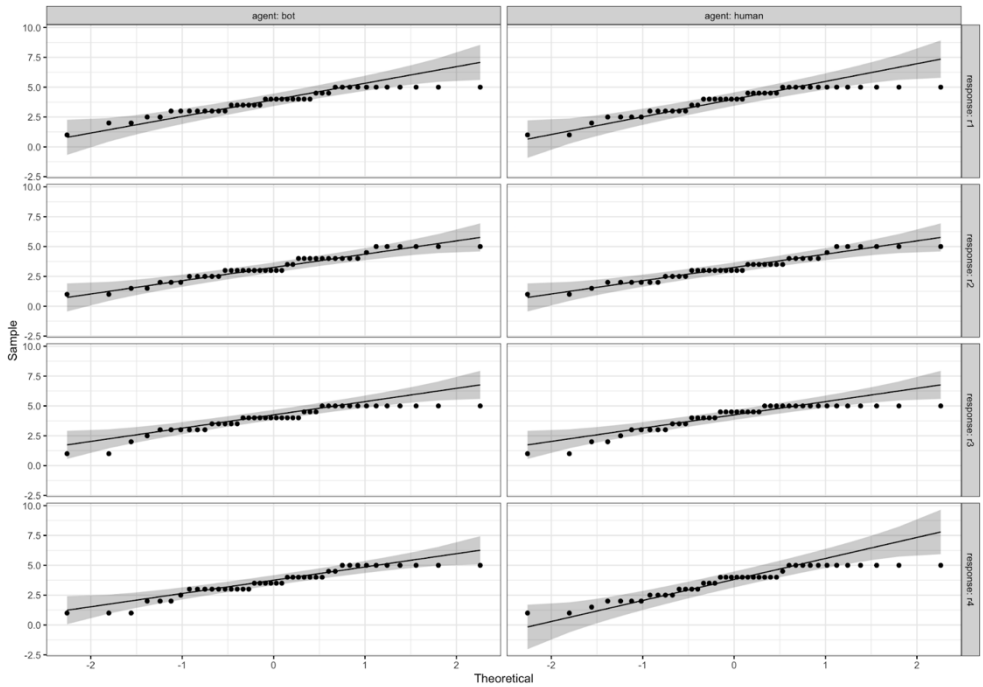


Figure 11. Influence of Participant Perspective Taking Level on Perceived Agent Likability

2.7 ANCOVA Results: Perceived Agent Competence

2.7.1 Q-Q Plot



2.7.2 Two-Way Repeated Measures ANCOVA

Effect	DFn	DFd	F	p	p<.05	ges
perspective	1	40	17.314	0.000163	*	0.134
agent	1	40	0.288	0.595		0.000624
response	2.38	95.37	7.046	0.000708	*	0.071
perspective:agent	1	40	0.408	0.527		0.000883
perspective:response	2.38	95.37	1.014	0.377		0.011
agent:response	2.55	102.15	0.51	0.647		0.002
perspective:agent:response	2.55	102.15	1.65	0.189		0.005

2.7.3 Group Means by Response Strategy

response	variable	n	mean	sd
r1	DV	84	3.839	1.056
r2	DV	84	3.244	1.091
r3	DV	84	3.982	1.082
r4	DV	84	3.595	1.181

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

2.7.4 Pairwise Comparisons for Response Strategy

	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
DV	r1	r2	84	84	4.82202291	83	6.36E-06	3.82E-05	****
DV	r1	r3	84	84	-1.4330416	83	0.156	0.936	ns
DV	r1	r4	84	84	1.72378057	83	0.088	0.531	ns
DV	r2	r3	84	84	-4.9070761	83	4.55E-06	2.73E-05	****
DV	r2	r4	84	84	-2.0561593	83	0.043	0.257	ns
DV	r3	r4	84	84	3.00373692	83	0.004	0.021	*

r1: Normative Appeal, r2: Guilt Appeal, r3: Fear Appeal, r4: Avoidant Message

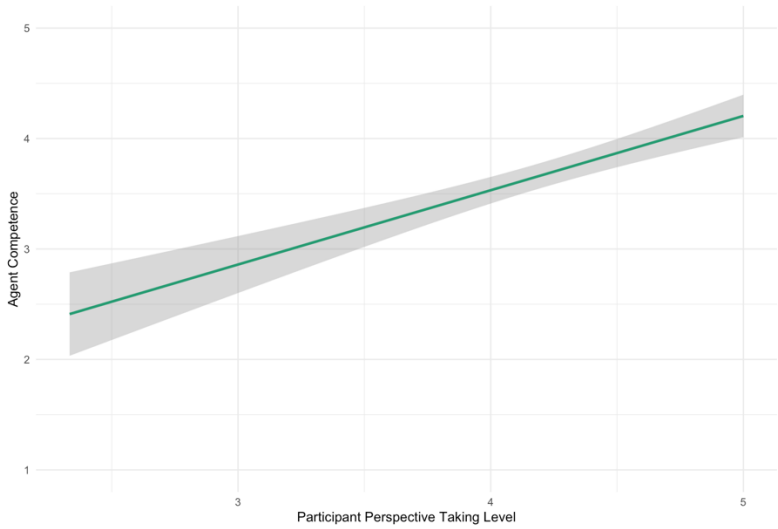


Figure 12. Influence of Participant Perspective Taking Level on Perceived Agent Competence