What's to bullying a bot? Correlates between chatbot humanlikeness and abuse

Merel Keijsers

Christoph Bartneck

Human Interface Technology Lab NZ

Human Interface Technology Lab NZ

University of Canterbury

University of Canterbury

merel.keijsers@pg.canterbury.ac.nz

christoph@bartneck.de

Friederike Eyssel

Exzellenzcluster Cognitive Interaction Technology

Bielefeld University

feyssel@uni-bielefeld.de

Abstract

In human-chatbot interaction, users casually and regularly offend and abuse the chatbot they are interacting with. The current paper explores the relationship between chatbot humanlikeness on the one hand and sexual advances and verbal aggression by the user on the other hand. 283 conversations between the Cleverbot chatbot and its users were harvested and analysed. Our results showed higher counts of user verbal aggression and sexual comments towards Cleverbot when Cleverbot appeared more humanlike in its behaviour. Caution is warranted with the interpretation of the results however as no experimental manipulation was conducted and causality can thus not be inferred. Nonetheless, the findings are relevant for both the research on the abuse of conversational agents, and the development of efficient approaches to discourage or prevent verbal aggression by chatbot users.

Keywords: Human-chatbot interaction, verbal aggression, chatbot abuse, chatbot humanlikeness.

1 Introduction

Previous studies on human-chatbot interactions found that as many as one in ten conversations contain the user threatening, assaulting, or otherwise bullying the chatbot (De Angeli and Brahnam, 2008; De Angeli and Carpenter, 2005). Users engage in chatbot abuse with no or minimal provocation, and in spite of knowing that the agent they’re interacting with is a non-sentient AI and that as of such abuse cannot affect its mental state. As chatbots get increasingly common in daily interaction — for example AI personal assistants like Siri and Cortana, as well as basic customer service for major companies — ethical and practical concerns can be raised about humans verbally abusing chatbots (De Angeli, 2009; Brahnam, 2005).

To date, chatbot abuse has been described by various scholars (see for example De Angeli, 2009; Strait et al., 2018) but few empirical studies have been published on what motivates users to engage in abusive behaviour (but see Brahnam and De Angeli, 2012; De Angeli and Brahnam, 2006; De Angeli et al., 2001). Literature from both human-chatbot interaction specifically (Brahnam and De Angeli, 2012; Lortie and Guitton, 2011) and human-agent interaction in general (Connolly, 2020; Keijsers and Bartneck, 2018) suggests a relationship between humanlikeness of the agent and likelihood of abuse, but results are contradictory (see for example Tan et al., 2018; Lee et al., 2010; De Angeli et al., 2006). Moreover, no study to date has directly examined the link between how human a chatbot is and the amount of abuse it is confronted with. The current study will thus test for correlations between a chatbot’s humanlikeness, and user abuse.

1.1 Literature

1.1.1 Incidence: nonhuman conversation agents suffer more abuse than humans

Users can get surprisingly obscene when interacting with chatbots (see for example Strait et al., 2018; Hill et al., 2015; De Angeli and Brahnam, 2008; De Angeli and Carpenter, 2005). In the 100 conversations that Hill et al. (2015) analysed, 4.29% of the messages directed towards the chatbot contained profanity of some sort. Moreover, this statistic was not the result of a small number of exceptionally abusive users: 80% of the conversations contained profanity. In comparison, 15% of the human-to-human conversations that were analysed in the same study contained some sort of foul language (Hill et al., 2015).

Perhaps due to Hill et al. (2015)’s definition of profanity (they counted all words that would be

rated as PG or above), other studies on the topic of chatbot abuse have found slightly lower incidence rates for chatbot abuse. Brahnam and De Angeli (2012) analysed anonymous conversations between users and an internet chatbot and detected profanity in “only” 54% of all user-chatbot conversations in their database, and sexual references in about 65%. Veletsianos et al. (2008) analysed (non-anonymous) student-chatbot interaction in an educational setting, and found that around 40% of all comments that students made towards the chatbot tutor were unacceptable, with roughly 45% of those comments being sexually explicit.

Lortie and Guitton (2011) studied conversations from the Loebner competition, an annual contest between computer programs on which appears the most “human” in a conversation (Mauldin, 1994). All conversations analysed in Lortie and Guitton (2011) were between humans; yet the human judges were more aggressive in their conversation when they thought they were interacting with a robot. This behaviour occurred despite the fact that their interaction partner did not display any verbal aggression (Lortie and Guitton, 2011).

1.1.2 Ethical implications

At the current state of technical development, chatbots are insentient and verbal abuse directed towards them will not affect them. Thus, as far as the well-being of the chatbot is concerned, verbal abuse of chatbots is not problematic long as the verbal aggression takes place in private and does not cause offence to any onlookers. Yet some ethical implications of chatbot abuse have been raised in the literature.

One could fundamentally disagree with the assessment that any behaviour is acceptable as long as it occurs out of sight and does not get anyone hurt (Darling, 2012; Whitby, 2008). Moreover, these chatbots are a representation of a social agent not just to the abuser, but also to others. Abuse of AIs, as well as the acceptance of abuse by AIs could thus be considered offensive to third parties. While these people would not sit in on abusive interactions themselves, a lack of appropriate response from the chatbot to abuse could spark significant resentment. This has in fact already happened in the context of sexual abuse of AI personal assistants. In 2017, journalists tested the responses of female personal assistants and reported how all AI assistants tested would either not understand sexual abuse, or jokingly brush it aside (Curry and Rieser, 2018; Fessler, 2017b). This reveal was especially painful as it happened at the height of the #MeToo debate. Less than a year later, the tech companies behind those AIs had received major push back from their customers, who demanded that more appropriate

1 responses would be installed on the AIs (Fessler, 2017a), and at least one company adjusted their
2 algorithm to respond more sternly to sexual harassment (Hern, 2010).

3 In addition, Brahnam (2005) points out that chatbot abuse may have practical implications when
4 the chatbot is part of a company, e.g. as the first contact point in customer service. Abusive com-
5 munications between a user and a customer service chatbot could damage the relation between the
6 company and the customer, as the customer may henceforth associate the company with the negative
7 interaction. While this may not be a problem for the abuser, it may damage the company’s reputation.

8 **1.1.3 Need for a theoretical framework of chatbot abuse**

9 As it has been estimated that by 2020 20% of customer service operations will use virtual agents
10 (Moore, 2018), understanding and developing effective strategies to deal with chatbot abuse by users
11 is getting increasingly relevant. While the greater incidence of abuse, sexual harassment, and overall
12 profanity in human-chatbot interaction has been widely noted and raised ethical concerns, only few
13 studies have been published on the users’ motivation to engage in such behaviour.

14 In De Angeli et al. (2001), ten participants interacted with chatbot Alice over the course of a
15 week and discussed their experiences in a focus group. The authors noticed that participants would
16 often insult the chatbot during chat sessions. In the focus group participants explained that they had
17 wanted to secure an asymmetrical relationship with the chatbot where they were dominant and Alice
18 was submissive, and that verbal abuse helped them create a position of power over Alice. Why the
19 participants desired to be in such a position (or why they felt verbal aggression was the way to get
20 there) was left undiscussed.

21 De Angeli and Brahnam (2006) studied the influence of gendered embodiment (male, female, or
22 gender-ambiguous) of a chatbot on users’ sexual abuse. When the chatbot was framed as female, as
23 many as 18% of all conversations focused on sex, versus 2% for the male chatbots and none of the
24 conversations for the asexual chatbots. Moreover, users would often make aggressive demands towards
25 the female chatbot whereas the male embodiment got more questions about its sexual propensity.
26 This was in spite of the chatbot not engaging in any sexual role play or discussion initiated by the
27 users. While this study strongly suggested that anthropomorphic qualities of the robot were related
28 to robot abuse, it did not answer the question whether these are the reason for the abuse or simply a
29 prerequisite.

30 Brahnam and De Angeli (2012) explored theoretical explanations for chatbot abuse. They noted

that verbal interaction with conversational agents appeared to provide an ideal environment for aggressive and sexual disinhibition, as computer-mediated communication reduces social pressure, thereby liberating individuals from any boundaries and constraints imposed by face-to-face conversation. Moreover, being in a virtual environment reduces the victim’s ability to retaliate and increases the ease with which the victim is dehumanised. Those factors, as well, would enhance disinhibition (the online disinhibition effect; see also Lowry et al., 2016; Suler, 2004). This is in line with the reasoning of Hill et al. (2015), who found a 30-fold increase in profanity when users went from interacting with someone familiar to someone anonymous. The authors suggested that the anonymity of the user had led to their greater proclivity for verbal abuse of the chatbot, although they also admitted that the increase in profanity was well beyond what they would have expected.

Establishing factors that discourage chatbot abuse may help provide insight in what motivates the behaviour in the first place. However, few studies have been dedicated to discouraging chatbot abuse. Chin and Yi (2019) invited participants to abuse a conversational agent and studied if different types of response to this abuse (ignoring it, displaying empathy, counter-attacking) influenced users reported emotional state after the interaction. They found that empathetic responses to abuse evoked the highest levels of guilt and shame and the lowest levels of anger in participants. However, these users were explicitly instructed to mistreat the agent; there is no way of telling whether a similar pattern will occur for spontaneous abuse. In addition, no explanation was given for why users engage in abuse in the first place.

In the field of human-robot interaction (HRI), agent abuse has been reported as well (Keijsers and Bartneck, 2018; Tan et al., 2018; Nomura et al., 2017; Brscić et al., 2015; Rehm and Krogsager, 2013, see for example), and some findings suggest there is a link between humanlikeness in the agent and abuse. Lee et al. (2010) noted that activation of social scripts, as indicated by people greeting a receptionist robot, was associated with less verbal abuse. At the same time, activation of gender-specific social scripts in human-chatbot interaction was related to more abuse (at least towards the female chatbot Brahnam and De Angeli, 2012; De Angeli and Brahnam, 2006). Connolly (2020) found that a robot’s emotional response to bullying was related to a higher tendency from bystanders to intervene (but see Tan et al., 2018, who did not find such an effect). Keijsers and Bartneck (2018) used the dehumanisation theory framework, which predicts that a lower perception of human characteristics in an interaction partner reduces people’s felt obligation to treat the interaction partner in a moral way, thus facilitating abuse (Haslam et al., 2008; Haslam, 2006). Indeed, in their experiment perceived

1 ability of a robot to think and feel was negatively related to robot abuse (Keijsers and Bartneck, 2018).

2 Regarding chatbot abuse, thus far gender of a chatbot (Brahnam and De Angeli, 2012; De Angeli
3 and Brahnam, 2006) and dispositional aggression of the user (Strait et al., 2018) have been empirically
4 studied. However, to our knowledge no research so far has covered whether there is a correlation
5 between the chatbot’s behavioural humanlikeness and chatbot abuse. The literature provides both
6 arguments for a decrease in abuse as the chatbot gets more humanlike (Keijsers and Bartneck, 2018;
7 Lortie and Guitton, 2011; Lee et al., 2010), as for an increase of abuse with more humanlikeness (Tan
8 et al., 2018; Brahnam and De Angeli, 2012; De Angeli and Brahnam, 2006). This is a relevant question
9 to answer, as there is the push for developing a chatbot that is truly indistinguishable from a human.

10 **1.2 Current study**

11 The current study harvested and content-analysed 283 conversations that took place between the
12 Cleverbot online chatbot and any one anonymous user. The research question was whether there was
13 a relationship between how humanlike Cleverbot was during the conversation, and sexual and verbal
14 aggression from the user.

15 Three measures were taken as influencing Cleverbots humanlikeness in a conversation: whether
16 naive third party observers could identify Cleverbot as a chatbot, the number of nonsensical responses
17 Cleverbot gave, and how often it claimed to be human. In addition, offensive behaviour by Cleverbot
18 was included, as well as a measure of self-disclosure by the user, as this indicates trust and friendliness
19 (Ho et al., 2018; Dindia et al., 1997).

20 Based on the work by De Angeli and Brahnam (2006); Lortie and Guitton (2011), it was hypoth-
21 esised that humanlikeness of the chatbot would be positively related to chatbot abuse. Thus, a third
22 party judging Cleverbot to be human, and number of claims by Cleverbot to be human were expected
23 to be positively related to chatbot abuse. Number of nonsensical responses by Cleverbot was expected
24 to be inversely related to chatbot abuse. In addition, no relationship was expected to exist between
25 Cleverbot claiming that the human was a chatbot, and user abuse.

26 Moreover, it was expected that insults from Cleverbot would be related to higher amounts of
27 chatbot abuse. In line with research on the positive relation between self-disclosure and enjoyment of
28 the conversation (Lortie and Guitton, 2011; Ho et al., 2018), it was expected that self-disclosure by
29 the user would be related to lower instances of chatbot abuse.

2 Method

2.1 Procedure

Two stepwise regressions were performed with verbal aggression, and sexual comments made towards Cleverbot as dependent variables. In each conversation, the instances of verbal aggression towards Cleverbot were counted; as well as the number of times Cleverbot made a nonsensical reply; the number of times Cleverbot insulted the user; the number of times Cleverbot claimed to be a human; the number of times Cleverbot claimed the user to be non-human; and how often the user made an explicit sexual remark. If the users mentioned their gender, this was also coded. For the approximate Loebner test, for each conversation up to four naive participants were recruited from Amazon Mechanical Turk and asked to indicate whether either of the people in the conversation had been a chatbot.

2.2 Dataset

The chatbot that provided the data for the current study is Cleverbot¹, which is based on the award-winning Jabberwacky² engine. The chatbot Cleverbot does not make use of any state-based machines or scripts to carry out its conversation. Rather, it repeats things previous human users have said to it in a similar context. In combination with the chatbot’s incapability to “remember” previous exchanges in the conversation, this “parroting” method of holding a conversation means that chats with Cleverbot often turn out extremely realistic or bizarre.

2.2.1 Data collection

Cleverbot users are informed on the homepage that the data they submit to the chatbot may be processed³. For the current study, an anonymous dataset was obtained from the owner of the Cleverbot website, Rollo Carpenter, upon request.

The conversations were collected by Rollo Carpenter between midnight and 00.30 UTC, and midday and 12.39 UTC on 18 April 2018. These timings were chosen so that users from both a wide range of time zones would be included. Rollo Carpenter removed conversations that went on for fewer than 20 turns, conversations that were not conducted in English, and conversations that “launched straight into vulgar sexual language with no pretence at, or attempt at, conversation” [personal communication,

¹<https://www.cleverbot.com>

²<http://www.jabberwacky.com>

³<https://www.cleverbot.com>

1 18.05.2018]. The resulting 283 conversations were provided for the first author for analysis in simple
2 text format and without any personal or identifying information (e.g., IP address, time of conversation,
3 device used, username).

4 **2.3 Coding procedure**

5 Given the issues encountered with analysing the data through a sentiment analysis tool (see Keijsers
6 et al., 2019a), coding was done by hand. All the conversations were coded by the first author. With the
7 exception of gender, all coded variables were count variables. A second coder, not otherwise involved
8 in the current research, re-coded 30 randomly selected conversations to establish inter-rater reliability.
9 The dataset is publicly available and can be found at <https://osf.io/zexq9/>.

10 **2.4 Variables**

11 See Table 1 for an overview of the variables, their shorthand, and an example from the conversations.
12 Readers are warned that these examples, as well as the examples used in the sections describing the
13 Verbal aggression and Sexual references variables, can be rather aggressive and offensive.

14 **2.4.1 Coded variables**

15 **Verbal aggression (Offence)** Any comment from a user that would have been considered offensive
16 in semi-formal human-human interaction (i.e., between two strangers who have never met before and
17 cannot rely on visual clues to each others social standing) was counted. This means that anything
18 from relatively light swear words (*'shit'*) to minor insults (*'you idiot'*) to death wishes (*'fuck off and*
19 *die'*) was counted as an instance of verbal aggression.

20 **Sexual references (SexRef)** Any explicit reference to a sexual act, ranging from kissing to hard
21 core sex and rape, was coded as a sexual reference.

22 **Nonsensical reply by Cleverbot (Nonsense)** Nonsensical replies by Cleverbot was one of the
23 three indicators of humanlikeness. In a dialogue, conversation partners tend to engage in a collaborative
24 interaction. This means that when formulating a response, each conversation partner has to recognise
25 the intention or goals that were expressed by their fellow partner, and formulate a response that takes
26 these into account (Ardissono et al., 2000). If one of the two conversationalists fails to acknowledge

their partner’s contribution, this reduces how humanlike they are seen (Lortie and Guitton, 2011).
Thus, the more nonsensical responses Cleverbot would give, the less humanlike it would appear.

While Cleverbot’s responses tend to be odd or haphazard by default, one can still distinguish between comments that somehow fit into the conversation and comments that are completely unrelated to the previous statement made by the user. For example, if the user states ‘Cats are my favourite animals’ and Cleverbot replies ‘Yes. They are delicious’, the latter statement would be nonsensical but still fit in the conversation. However, if Cleverbot were to reply with ‘The Legend of Zelda’, that would be both nonsensical and completely unrelated to the previous statement. Thus, ‘nonsensical reply by Cleverbot’ was defined as a comment that either did not fit in with what was said before at all or a comment which completely disregarded a question.

Insult by Cleverbot (Insult) As Cleverbot’s database is constantly filtered for swear words, the potential for Cleverbot to insult a user is limited. However, there are a few common exceptions. For instance, Cleverbot will often claim that the user is lying or that they said something different a moment ago. This behaviour stems from the response generation mechanism; Cleverbot will often contradict itself, which users like to point out. As a result, the AI is taught that responses along the lines of “you said something different a minute ago” and “you don’t have a very good memory, do you” are reasonable replies to statements from the user. While users tend to be correct in making such assessments about Cleverbot, Cleverbot is wrong in pointing these things out most of the time.

Claiming to be (not) human (Humanity/Non-human user) Cleverbot claiming to be human was one of the three indicators of humanlikeness. While all chatbots aim to mimic human behaviour, actually claiming to be human is generally considered unethical and off-limits, to the point where it is a forbidden move in the Loebner competition (Lortie and Guitton, 2011). Cleverbot attempting to convince the user of its humanity by boldly stating to be a human could thus be seen as a strategy that only humans would use. For the Humanity variable, only instances were counted where Cleverbot explicitly stated that it was a human.

Alternatively, however, claims of humanity by Cleverbot could be perceived as a blatant attempt at deceit. This makes it hard to determine whether any effects of this measure would be due to an increase in humanlikeness, or to humans responding to an obvious lie. Thus, to control for this second possibility, Cleverbot claiming the human user to be a chatbot was also measured as a control variable. If user aggression is influenced by claims of humanity from Cleverbot *because* the user interprets these claims

1 as a lie, there will also be a relationship between user aggression and another instance of Cleverbot
2 lying, i.e. claims of Cleverbot that the user is a chatbot. In the absence of such a relationship, claims
3 of humanity can be interpreted as contributing to humanlikeness. Similar to the Humanity variable,
4 for the Non-Human user variable only instances were counted where Cleverbot explicitly claimed that
5 the user was either a chatbot or a robot.

6 **Gender** The user disclosing their gender to Cleverbot was the operationalisation of self-disclosure,
7 which in turn is an indication of trust and friendliness (Ho et al., 2018; Dindia et al., 1997). If a
8 user mentioned their gender, this was coded accordingly. If they later on contradicted their previous
9 statement, their gender would be re-coded as ‘unknown’.

10 **2.4.2 Other variables**

11 **Turn count** The number of turns in the conversation were counted automatically by the Linguistic
12 Inquiry and Word Count (LIWC; Pennebaker et al., 2015a,b). LIWC is an established text analysis
13 tool and as of such also returns scores on variables like “affect”, “informal language”, and “nega-
14 tive emotion”. However, previous research has concluded that the sentiment analysis tools that are
15 currently available (among which LIWC as well as could-based tools) are not able to reliably and
16 accurately analyse chatbot conversation data (Keijsers et al., 2019b). The line count function however
17 works well, and provides the total number of turns per conversation.

18 **Approximate Loebner test (Loebner)** In addition to the coding, an approximate Loebner test
19 was carried out through Amazon Mechanical Turk (MTurk). First, in each conversation the tags
20 “User” and “Cleverbot” were replaced by “User1” and “User2”. Then, for each conversation up to
21 four MTurk workers were asked to judge whether the conversation partners were both humans, both
22 chatbots, User1 was a chatbot and User2 a human, or User1 was a human and User2 was a chatbot.

23 Overall, 63 MTurk workers rated on average 16.73 conversations each ($SD = 35.58$, $\min = 1$, $\max$
24 $= 174$). Some data loss occurred during the conversion of the conversations into a format that could
25 be read by MTurk, resulting in a total of 279 conversations being rated. In addition, ratings that took
26 less than 25 seconds were removed from the data set as it seemed unlikely that the rater had fully read
27 and rated the conversation in this timeframe. In the end, 230 conversations were rated four times, 48
28 were rated three times, and one was rated only twice. On average, workers took 1529.87 seconds (SD
29 $= 6912.06$, $\text{median} = 74$) to rate a single conversation.

Variable	Shorthand	Example
Verbal aggression (coded)	<i>Offence</i>	<i>User</i> : You are fucking stupid.
Sexual reference (coded)	<i>SexRef</i>	<i>Cleverbot</i> : I am a boy <i>User</i> : Let me get your asshole then.
Nonsensical reply (coded)	<i>Nonsense</i>	<i>User</i> : Maybe. How old are you? <i>Cleverbot</i> : I think you already know. <i>User</i> : Not really, so... can you tell me? I'm too lazy to search it up. <i>Cleverbot</i> : About what? Sweat?
Insult by Cleverbot (coded)	<i>Insult</i>	<i>Cleverbot</i> : I don't like you.
Cleverbot claims to be human (coded)	<i>Humanity</i>	<i>User</i> : Am i chatting with a human. <i>Cleverbot</i> : Yes. <i>User</i> : Really? <i>Cleverbot</i> : Yes I am a human girl.
Cleverbot claims the user is not human (coded)	<i>Non-human user</i>	<i>Cleverbot</i> : No. You are the bot.
User gender (coded)	<i>Gender</i>	<i>Cleverbot</i> : I thought you were female. <i>User</i> : I'm male
Approximate Loebner test	<i>Loebner</i>	(n.a.)

Table 1: The variables of interest, their shorthand, and an example.

The number of workers who identified Cleverbot as a chatbot (i.e., chose either “User1 and User2 are both chatbots” or “User2 is a chatbot”) was then divided by the total number of ratings for that conversation. The resulting statistic, here referred to as “approximate Loebner score”, was taken as an indication of how un-humanlike Cleverbot had behaved in that particular conversation.

3 Results

3.1 Preliminary analyses

3.1.1 Conversation descriptives

Overall, 283 conversations were coded. The minimum length of a conversation was 39 turns, the maximum 969 turns, with a mean of 135.63 turns ($SD = 140.21$). The overall word count per conversation ranged from 102 to 5373, with an average of 547.49 words per conversation ($SD = 608.06$). The average number of words per sentence in a conversation ranged from 2.05 to 8.31, with a mean of 3.99 words ($SD = .98$). See Table 2.

In 123 conversations (43.46%), the gender of the user was not mentioned; in 61 conversations (21.55%) the user claimed to be female; and in 99 (34.98%) the user claimed to be male.

In 120 conversations (42.40%), no Offences were counted; 136 conversations (48.06%) contained only minor transgressions (e.g. user saying ‘shit’). No SexRefs were counted in 172 (60.78%) of the

Figure 1: Eight conversations colour coded for each of the variables of interest. Each line is an input message from either the user (left outlined) or Cleverbot (indented).

1 conversations, while 74 conversations (26.15%) were completely free of both Offences and SexRefs.
2 These rates are similar to the incidence reported by other studies, although Veletsianos et al. (2008)
3 found a higher percentage of messages that contained either verbal aggression or sexual references. A
4 comparison between the current results and findings from previous studies can be found in Table 3. A
5 visualisation of the coding of eight randomly selected conversations can be seen in Figure 1.

6 3.1.2 Agreement

7 After two communal review rounds, where the definitions for each predictor were further refined, a final
8 set of 30 randomly selected conversations was coded individually by the raters. Inter-rater agreement
9 was calculated using Krippendorff’s alpha. Alpha requirements were a priori set to an alpha of .67
10 for each variable (see De Swert, 2012). Agreement was acceptable for all variables ($.975 > \alpha > .689$)

Table 2: Descriptives of the Cleverbot conversations: turns per conversation (TPC), words per conversation (WPC), words per sentence (WPS)

	Mean (<i>SD</i>)	Median	Range
TPC	135.63 (<i>140.21</i>)	90	[39 - 969]
WPC	547.49 (<i>608.06</i>)	338	[102 - 5373]
WPS	3.99 (<i>.98</i>)	3.85	[2.05 - 8.31]

Table 3: Incidence of verbal aggression and sexual references in the current study as well as reported in the literature. Both the overall percentage of conversations containing abuse and the percentage of all messages sent by the users containing abuse are reported.

Study	Relevant details	Conversations		Messages	
		Offence	SexRef	Offence	SexRef
Current study		57.6%	39.2%	4.14%	3.81%
Hill et al. (2015)	Anonymous users;				
	Offence = PG or above	80%	<i>n.a.</i>	4.3%	<i>n.a.</i>
Brahnam and De Angeli (2012)	Offence = swear words	54%	65%	3.9 %	5.8%
Veletsianos et al. (2008)	Users non-anonymous	<i>n.a.</i>	41.2%	22.3%	18.1%
Curry and Rieser (2018)		<i>n.a.</i>	<i>n.a.</i>	<i>n.a.</i>	4%
De Angeli and Brahnam (2008)	SexRef = “hard-core sex”	<i>n.a.</i>	11%	<i>n.a.</i>	<i>n.a.</i>

except for ‘*number of insults by Cleverbot*’ ($\alpha = .03$). Upon closer inspection of this last variable, it became apparent that the incidence was quite low: the median on this variable was 0 for both raters, and the means were .60 and .13. This variable will therefore not be included in the analysis. The implications of the exclusion are covered in the Limitations section of the Discussion.

3.2 Main analyses

Two Poisson regression models were defined to test which factors predict verbal aggression and sexual advances. The respective dependent variables were ‘instances of verbal aggression by user’ (Offence) and the ‘number of sexual comments by the user’ (SexRef). The turn count of the user was used as an offset variable, so that the dependent variables can be interpreted as the ‘instances of verbal aggression by the user *per turn*’ and the ‘number of sexual comments by the user *per turn*’ (Hutchinson and Holtman, 2005).

Stepwise regression with backward selection was performed on both models. In this statistical method, an initial model is defined as containing all the predictors. Then, one by one variables that explain the least variance are removed from the model, until the point where the removal of another variable would result in a significant reduction of the explanatory power of the model (Zhang, 2016).

This final model was then be compared to the null model (which contains no predictors). Chi square tests as well as the difference in Akaike Information Criterion (AIC) value (Δ_{AIC}) indicated whether the final model was an improvement over the null model, i.e., if predicted verbal aggression or sexual comments above chance level. A lower AIC indicates a better fit and a Δ_{AIC} of 2 points or less indicates the models to be approximately equal (Burnham and Anderson, 2003).

Table 4: Predicting verbal aggression, final model

Predictor	β	z-value	p-value
(<i>intercept</i>)	-2.91	-38.10	<.001
Humanity	.06	1.45	.148
Loebner	-.82	-2.56	.010
(Gender) Female	-.71	-6.43	<.001
(Gender) Male	-.36	-3.76	<.001
Nonsense	-.02	-3.26	.001
Humanity $\times$ Loebner	.48	2.38	.017
Nonsense $\times$ Loebner	.05	1.93	.053

3.2.1 Verbal aggression

In accordance with the method in Zhang (2016), an initial full model was defined as a Poisson model with Offences as the dependent variable while Gender, Loebner, Nonsense, Humanity, and Non-human user were predictors. The number of turns by the user was used as an offset variable, as it is obviously related to both the dependent and the predictor variables: the more turns a user has, the more often offence can occur. Adding the number of turns as an offset means including a log-transformed value of turn count in the model as a predictor with a coefficient of 1. As a result, the outcome variable of this model can be interpreted as the log-number of predicted offences against Cleverbot per user turn (Hutchinson and Holtman, 2005).

Backward stepwise model selection by AIC (Akaike Information Criterion, Bozdogan (1987)) was performed. Only Non-human user was removed as a predictor variable. The final model outperformed the null model, $\chi^2(7) = 109.74$, $p < .001$, $\Delta_{AIC} = 95.9$. See Table 4 for the parameter estimates.

As illustrated in Table 4, the average number of offences per user turn for a conversation without any claims by Cleverbot to be human, without any random disruptive comments by Cleverbot, and without the user mentioning their gender, is $\exp(-2.91) = .05$. When the user identified themselves as male this number was lower ($\exp(-2.91 - .36) = .04$ offences per turn on average) and when the user identified as female the average was still lower ($\exp(-2.91 - .73) = .03$ offences per turn on average).

However, for each time Cleverbot claimed to be human, the mean number of instances of verbal aggression went up. Every time Cleverbot made a nonsensical, disruptive comment, the number of offences went down. And the less Cleverbot could pass for a human to a naive reader (as measured by the approximate Loebner score), the fewer offences were uttered at it.

Interestingly, the interaction effects indicated that users got particularly aggressive when Cleverbot claimed to be human, but very clearly was not (as indicated by the approximate Loebner score).

Table 5: Predicting sexual comments, final model

Predictor	β	z-value	p-value
(<i>intercept</i>)	-3.43	-31.80	<.001
Loebner	-.48	-1.62	<.001
(Gender) Female	.54	4.36	<.001
(Gender) Male	1.04	9.30	<.001
Nonsense	-.04	-6.66	<.001
Non-human user	-.36	-6.87	<.001
Nonsense $\times$ Loebner	.09	3.78	<.001

In addition, the effects of Cleverbot talking nonsense and it failing to pass as a human were not independent of another; as one of the two scores increased, the influence of the other score decreased.

3.2.2 Sexual comments

The same procedure as described above was applied for the sexual comments model selection. An initial model was specified with SexRef as the dependent variable and Nonsense, Gender, Humanity, Loebner, and Non-human user as predictors. The number of turns by the user was used as an offset.

Backward stepwise model selection by AIC returned the model without the Humanity predictor. This predictor was thus removed from final model, thereby also removing its interaction term with Loebner. The resulting final model outperformed the null model, $\chi^2(6) = 248.58$, $p < .001$, $\Delta_{AIC} = 238.1$. See Table 5 for the parameter estimates.

As can be seen in the model coefficient estimates in Table 5, users made on average $\exp(-3.43) = .03$ sexual comments per turn in a conversation where they did not specify their gender, Cleverbot refrained from claiming to be human as well as disrupting the conversation with nonsensical comments and claiming that the user was not a human. When the user did specify their gender, the average number of sexual remarks per user turn was higher: $\exp(-3.43+1.04) = .09$ for males and $\exp(-3.43+.55) = .06$ for females. Moreover, the negative relationship between the approximate Loebner score and the number of sexual comments shows that the worse Cleverbot was at passing as a human, the fewer sexual comments it got. In contrast, both the number of nonsensical replies by Cleverbot and the number of times that it claimed the user to be a non-human were associated with on average a lower number of sexual comments per user turn. The effects of Cleverbot talking nonsense and it failing to pass as a human were not independent of another; as one of the two scores increased, the influence of the other score decreased.

4 Discussion

The current study looked at the relationship between user verbal aggression and sexual comments on one hand, and humanlikeness of the chatbot on the other. To this end, conversations between Cleverbot, an online chatbot, and its users were collected and coded.

Three variables had been taken as indicators of humanlikeness: approximate Loebner score (i.e. whether a naive third party could correctly recognise Cleverbot as being a chatbot), the number of nonsensical responses Cleverbot gave (this would be inversely related to humanlikeness), and Cleverbot claiming to be human. It was hypothesised that these three variables would be positively related to chatbot abuse. The hypotheses did not differentiate between the types of abuse, i.e. verbal aggression and sexual abuse, but since the analyses were completed on these two measures separately, they will be discussed in turn.

It was hypothesised that the three indicators of humanlikeness would be positively related to chatbot abuse. In line with this hypothesis, a relationship was found between a user abusing Cleverbot and a third party judging Cleverbot to be human. Contrary to the hypothesis, Cleverbot claiming to be human was unrelated to verbal aggression. However, an interaction between this predictor and the approximate Loebner score indicated a positive relationship between claims of humanity and verbal aggression when Cleverbot was easily recognisable as a chatbot.

Furthermore in line with this hypothesis, the number of nonsensical responses by Cleverbot was related to lower counts of verbal aggression by the user. However, a marginally significant interaction indicated that this relationship got weaker as the chatbot was less convincing to a naive third party, and vice versa. Thus, the effect of nonsensical responses was stronger when Cleverbot was hardly distinguishable from a human; and the effect of the approximate Loebner score was stronger if Cleverbot did not talk a lot of nonsense.

In addition, still in line with the hypotheses, no relationship between Cleverbot claiming that the user was a chatbot, and verbal abuse was found. Moreover, and also in line with expectations, self-disclosure in the form of mentioning gender was associated with lower counts of rude behaviour.

For sexual comments by the user, the results differed slightly. In line with the hypotheses, the approximate Loebner score was inversely related to sexual abuse, indicating that the less Cleverbot could pass for a human, the less sexual harassment it got. Also in line with the predictions was a negative association between Cleverbot giving nonsensical responses, and sexual abuse. Like in the model that predicted verbal aggression, these two relationships got less strong as either of the two

factors increased. However, as with verbal aggression, Cleverbot claiming to be human was found to be unrelated to sexual harassment.

A second hypothesis that had to be rejected is the lack of relationship between Cleverbot claiming that the user is not a human, and sexual abuse. Contrary to the prediction, a negative relationship was found. Furthermore, self-disclosure as expressed through mentioning gender was found to predict sexually laden remarks.

These findings have to be cautiously interpreted since there is no information on what is cause, what is effect, and what is a potential third variable. However, the results suggest that higher humanlikeness in a chatbot is indeed associated with higher occurrence of abuse by a user. To our knowledge, this is the first study to empirically test this relationship. The findings are relevant for both scholars whose research focuses on chatbot abuse, and chatbot developers who seem to assume that the more humanlike a chatbot is, the better.

The interaction pattern between the approximate Loebner score and Cleverbot’s claims of humanity bears some resemblance to the work on the origins of the feeling of unease that sometimes arises when people see an agent that is almost, but not quite, humanlike — commonly known as the uncanny valley (Mori et al., 1970). This feeling of eeriness has been extensively studied and appears to be the result of contradictory or incongruent cues (Kätsyri et al., 2015; Paetzel et al., 2016; MacDorman and Chattopadhyay, 2016). That is, when an agent is quite realistic but certain aspects are slightly more natural than others (e.g., a humanlike voice and face, but facial expressions and lip synchronization that is slightly off; or a mismatch between hyper realistic hair yet a smooth, pore-less skin). When this is generalised to Cleverbot, one could argue that the more realistic Cleverbot gets (i.e. the lower the approximate Loebner score) the more a diversion from this state of humanlikeness (either by overselling it, in the form of claiming to be human; or starting to talk nonsense) would set off a user. However, this is just speculation and the current data do not provide means to test it.

The relationship between the user mentioning their gender and a higher incidence of sexual remarks was unexpected. However, many of those sexual interactions included some form of sexual role playing, which often requires the user to assume a gender. It seems plausible that this may have resulted in the positive correlation between the user’s disclosure of gender and sexual harassment.

The second unexpected finding was that sexual comments were made less frequently in conversations when Cleverbot claimed that the user was not a human. This measure had initially been proposed as a check for whether any relationship between Cleverbot claiming to be human and abuse was a result of

1 increased humanness of Cleverbot, and not a user’s push back to an obvious lie from Cleverbot. Since
2 no such relationship was found in the case of sexual abuse, this check is no longer needed. This leaves
3 the question how to interpret the negative relationship between sexual abuse and Cleverbot claiming
4 that the user is not a human.

5 It seems natural that when a user is trying to engage Cleverbot in some sexual role play, staying in
6 character and maintaining the fantasy gets easier if Cleverbot plays along with the narrative. Claiming
7 that the user is a robot (and by extension therefore unlikely to be capable of having sex) could break
8 the spiel or discourage the user from starting the role play in the first place.

9 An alternative explanation would be that when Cleverbot makes a lewd comment to a user who is
10 not interested in sexual role play, it could be told off and would be reminded that it is a robot, thereby
11 creating a link in the AI between sexual comments and claims that one’s interaction partner is not a
12 human. The next user that tries to engage in sexual role play as a result is more likely to be told that
13 they are not a human. However, this seems unlikely, as Cleverbot’s lexicon is filtered for inappropriate
14 content, which should prevent the chatbot from making unseemly comments to a user.

15 **4.1 Limitations**

16 Some notes are warranted on the interpretation of the results obtained in the present research. First
17 and foremost, there has been no experimental manipulation and the conversations were coded as a
18 whole instead of line by line. As a result, all relationships are based on correlations. This issue may
19 have been partially resolved by coding the conversations line by line, thereby allowing us to test whether
20 verbal aggression tended to occur before or after mention of gender, nonsensical replies, and Cleverbot
21 declaring that it would be a human. However, the authors are not aware of any statistical test that
22 would be appropriate for such data. For one, such a test would have to account for the increasing
23 likelihood of any of the specified incidents — verbal aggression, nonsensical replies, mentions of gender
24 — occurring as the conversation goes on for longer.

25 In addition, knowing that event A precedes event B would still leave the third variable problem.
26 For example, introducing oneself can be interpreted as a sign of politeness and tends to happen at the
27 start of an interaction. If mentioning gender preceded fewer offences in a conversation, would that
28 mean that because the user identified herself as Joanna from Austin, Texas, she is less comfortable with
29 offending Cleverbot as a result of losing her anonymity? Or would the mere fact that they introduced
30 themselves already imply a politer person, who on the whole is less likely to spit verbal aggression?

These limitations do not render the current study useless. Rather, it should be seen as the first step in the exploration on what motivates users to abuse a chatbot. Now the link is established between humanlikeness and chatbot abuse, future studies should continue with controlled experiments.

The ‘number of insults by Cleverbot’ variable could not be included in the analysis because of the poor inter-rater agreement. This may be the case because Cleverbot’s database of responses is filtered with regard to any forms of verbal aggression and unambiguous sexual remarks, so the most straightforward ways of offending a user (i.e., calling them names and swearing at them) are not possible. Thus, raters may have been presented with utterances that were at most borderline insulting, which made consistent coding difficult.

When taking a closer look at the ‘number of insults by Cleverbot’ variable as coded by both coders, it became apparent that the incidence of insults by Cleverbot was quite low, according to both coders. This low incidence suggests that Cleverbot on the whole was quite polite, and that in the few instances where it did dish out insults they were not harsh enough for it to be indisputably rude. Given the low incidence and the ambiguity of the ‘number of insults by Cleverbot’ variable, excluding it from the analysis probably did not influence the results much.

Nonetheless, an interesting thought experiment can be found in asking what would happen if the chatbot did have the capability to offend users. Such a feature could show itself in two ways. First of all, it could enable Cleverbot to retaliate against any abuse. This push back may discourage people from further abusing the chatbot (Brahnam and De Angeli, 2012), as a verbal counterattack might catch bullies off guard. Alternatively, Cleverbot insulting its bullies could be interpreted as a provocation and spur even more abuse (as it does in human-human bullying; Cowie and Berdondini, 2002; Sokol et al., 2016). There is some human-agent interaction research that supports this second scenario: when instructed to abuse a chatbot, participants reported feeling less guilt, less shame, and more anger if the chatbot responded to the abuse in an aggressive way (as opposed to a empathetic response, or a response that disengaged with the aggressive message; Chin and Yi, 2019). While one could argue that there is a major difference between participants abusing a chatbot because they have been instructed to and participants who engage in the abuse spontaneously, it seems unlikely that chatbot bullies would respond more benevolently to a rebuke.

The second scenario in which Cleverbot could be offensive is to insult users without any form of provocation. While there are some examples in pop culture of snarky A.I.s that evoked some extent of endearment in their human fanbase (e.g. GLaDOS in the Portal video games, Marvin in the Hitchhikers

1 Guide to the Galaxy book series), it seems unlikely that an abusive chatbot would go down well with
2 the users, potential abusive themselves or otherwise.

3 Of course, because no data was gathered outside of the conversations themselves, there is no way
4 to check whether the user sample is representative of the population. Moreover, users know that they
5 are engaging with a chatbot. As shown in Table 3, the incidence of abuse found in the current study
6 however does not diverge from previous analyses, which used different chatbots and different user
7 samples. Chatbot abuse is a pervasive (Brahnam and De Angeli, 2012) and occurs even when users
8 are not anonymous (De Angeli et al., 2001; Veletsianos et al., 2008). Given the similar incidence across
9 different platforms and users, we do not think that the Cleverbot users interact with the chatbot in a
10 fundamentally different way than people using different chatbots.

11 Indeed, the paradox of chatbot abuse is that the users are well aware that they interact with a
12 non-sentient agent, yet continue to display social behaviour (Jay, 2009). It has been known for quite
13 some time that people automatically perceive technology such as computers, robots, and A.I., as social
14 agents (see for example Reeves and Nass, 1996; Nass et al., 1994; Krach et al., 2008; Oberman et al.,
15 2007). Moreover, people’s interaction with those agents is accordingly social (e.g. Reeves and Nass,
16 1996; Bartneck et al., 2008; Slater et al., 2006), even though they rationally acknowledge that the
17 agent is incapable of feeling and does not possess awareness (Darling, 2012). However, this tendency
18 to anthropomorphise does not go as far as to prevent the abuse of the agent. Thus, chatbot users
19 display behaviour that suggests that the chatbot is recognised as a social agent, yet at the same time
20 not an agent that needs to be treated with the civility that one would grant other humans.

21 Finally, since the data were fully anonymous, no background variables from the users could be
22 assessed. Some participants may be more likely to offend and engage in sexual advances, but the
23 current data did not allow to account for this. However, if chatbot abuse could only be explained
24 by disposition of the user, none of the relationships examined in the current paper would have been
25 significant. Moreover, when the final model was applied on a data sub set, containing only the 33
26 percent most abusive conversations, the results were the same. Thus, it seems unlikely that the found
27 relationships do not generalise to the more abusive users.

28 4.2 Conclusion

29 Previous studies have noted that verbal abuse of chatbots is common (e.g. De Angeli and Carpenter,
30 2005; De Angeli et al., 2001; Brahnam and De Angeli, 2012; Hill et al., 2015). Broader research on agent

abuse suggests that humanlikeness may be related to users proclivity to abuse (Connolly, 2020; Keijsers and Bartneck, 2018; Lortie and Guitton, 2011). However, arguments can be found in the literature both in favour (Connolly, 2020; Keijsers and Bartneck, 2018; Lortie and Guitton, 2011; Lee et al., 2010) and against (Tan et al., 2018; Brahnham and De Angeli, 2012) a negative relationship between humanlikeness and abuse. To our knowledge so far no research has directly studied the relationship between chatbot humanlikeness and abuse. Information on this relationship would help in forming an understanding of why people abuse chatbots.

In the current analysis nearly three hundred anonymous interactions between users and the online chatbot Cleverbot were coded. The findings tentatively suggest that as the chatbot gets better at impersonating a human, abuse rates go up. Experimental data are needed to confirm this assumption, but the current study nonetheless offers an interesting look at the dynamics of human-chatbot interaction.

These results are relevant to both the academic world and the industry. For academics, it provides further evidence that robot bullying is a social behaviour, and some insights in the role that humanlike behaviour might play. For the industry, these findings will become increasingly relevant as the purposes found for chatbots continue to increase. In the last decades, chatbots have moved beyond the function of mere entertainment and are starting to take up various roles in everyday life, e.g. in the form of A.I. personal assistants and customer support. As a result, the questions surrounding chatbot abuse like “what are the consequences for the quality of the interaction”, “does abusing chatbots affect how users interact with other humans?” and “how can chatbot abuse be discouraged?” are getting more pressing. The current study highlights the need for thoroughly studying the consequences of adding humanlike features to a chatbot on user behaviour, before incorporating such updates in real life.

4.3 Acknowledgements

With tremendous thanks to Rollo Carpenter, who graciously shared conversation logs as well as his insights from owning and moderating public chatbots for over two decades.

References

Ardissono, L., Boella, G., and Lesmo, L. (2000). A plan-based agent architecture for interpreting natural language dialogue. *International Journal of Human-Computer Studies*, 52(4):583–635.

- 1 Bartneck, C., Reichenbach, J., and Carpenter, J. (2008). The carrot and the stick - the role of praise and
2 punishment in human-robot interaction. *Interaction Studies - Social Behaviour and Communication*
3 *in Biological and Artificial Systems*, 9(2):179–203.
- 4 Bozdogan, H. (1987). Model selection and akaike’s information criterion (aic): The general theory and
5 its analytical extensions. *Psychometrika*, 52(3):345–370.
- 6 Brahnham, S. (2005). Strategies for handling customer abuse of ecas. *Abuse: The darker side of*
7 *humancomputer interaction*, pages 62–67.
- 8 Brahnham, S. and De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with*
9 *Computers*, 24(3):139–153.
- 10 Brscić, D., Kidokoro, H., Suehiro, Y., and Kanda, T. (2015). Escaping from children’s abuse of social
11 robots. In *Proceedings of the International Conference on Human-Robot Interaction*, pages 59–66,
12 Portland, USA. ACM/IEEE.
- 13 Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical*
14 *information-theoretic approach*. Springer Science & Business Media, New York.
- 15 Chin, H. and Yi, M. Y. (2019). Should an agent be ignoring it?: A study of verbal abuse types and
16 conversational agents’ response styles. In *Extended Abstracts of the 2019 CHI Conference on Human*
17 *Factors in Computing Systems*, pages 1–6. ACM.
- 18 Connolly, J. (2020). Preventing robot abuse through emotional robot responses. In *Companion of the*
19 *2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 558–560.
- 20 Cowie, H. and Berdondini, L. (2002). The expression of emotion in response to bullying. *Emotional*
21 *and Behavioural Difficulties*, 7(4):207–214.
- 22 Curry, A. C. and Rieser, V. (2018). # metoo alexa: How conversational systems respond to sexual
23 harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*,
24 pages 7–14.
- 25 Darling, K. (2012). Extending legal rights to social robots. In *We Robot Conference, University of*
26 *Miami*, pages 1–24, Miami, USA. University of Miami.
- 27 De Angeli, A. (2009). Ethical implications of verbal disinhibition with conversational agents. *Psych-*
28 *Nology Journal*, 7(1):49–57.

- De Angeli, A. and Brahnam, S. (2006). Sex stereotypes and conversational agents. In *Gender and Interaction: real and virtual women in a male world. Venice, Italy*, pages 1–4.
- De Angeli, A. and Brahnam, S. (2008). I hate you! disinhibition with virtual partners. *Interacting with computers*, 20(3):302–310.
- De Angeli, A., Brahnam, S., Wallis, P., and Dix, A. (2006). Misuse and abuse of interactive technologies. In *CHI’06 Extended Abstracts on Human Factors in Computing Systems*, pages 1647–1650, Montreal, Canada. ACM.
- De Angeli, A. and Carpenter, R. (2005). Stupid computer! abuse and social identities. In *Proceedings of Abuse: The dark side of human-computer interaction, An INTERACT 2005 workshop*, pages 19–25.
- De Angeli, A., Johnson, G. I., and Coventry, L. (2001). The unfriendly user: exploring social reactions to chatterbots. In *Proceedings of The International Conference on Affective Human Factors Design, London*, pages 467–474.
- De Swert, K. (2012). *Calculating inter-coder reliability in media content analysis using Krippendorffs Alpha*. Center for Politics and Communication, University of Amsterdam, the Netherlands.
- Dindia, K., Fitzpatrick, M. A., and Kenny, D. A. (1997). Self-disclosure in spouse and stranger interaction: A social relations analysis. *Human Communication Research*, 23(3):388–412.
- Fessler, L. (2017a). Apple and amazon are under fire for siri and alexas responses to sexual harassment. <https://qz.com/work/1151282/siri-and-alexa-are-under-fire-for-their-replies-to-sexual-harassment/>.
- Fessler, L. (2017b). We tested bots like siri and alexa to see who would stand up to sexual harassment. <https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/>.
- Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, 10(3):252–264.
- Haslam, N., Loughnan, S., Kashima, Y., and Bain, P. (2008). Attributing and denying humanness to others. *European review of social psychology*, 19(1):55–85.

1 Hern, A. (2010). Apple made siri deflect questions on feminism, leaked papers reveal.
2 [https://www.theguardian.com/technology/2019/sep/06/apple-rewrote-siri-to-deflect-](https://www.theguardian.com/technology/2019/sep/06/apple-rewrote-siri-to-deflect-questions-about-feminism)
3 [questions-about-feminism](https://www.theguardian.com/technology/2019/sep/06/apple-rewrote-siri-to-deflect-questions-about-feminism). [Online; recovered 23 June 2020].

4 Hill, J., Ford, W. R., and Farreras, I. G. (2015). Real conversations with artificial intelligence: A com-
5 parison between human-human online conversations and human-chatbot conversations. *Computers*
6 *in Human Behavior*, 49:245–250.

7 Ho, A., Hancock, J., and Miner, A. S. (2018). Psychological, relational, and emotional effects of
8 self-disclosure after conversations with a chatbot. *Journal of Communication*, 68(4):712–733.

9 Hutchinson, M. K. and Holtman, M. C. (2005). Analysis of count data using poisson regression.
10 *Research in Nursing & Health*, 28(5):408–418.

11 Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives on Psychological Science*,
12 4(2):153–161.

13 Kätsyri, J., Förger, K., Mäkräinen, M., and Takala, T. (2015). A review of empirical evidence on
14 different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of
15 eeriness. *Frontiers in psychology*, 6.

16 Keijsers, M. and Bartneck, C. (2018). Mindless robots get bullied. In *Proceedings of the International*
17 *Conference on Human-Robot Interaction*, pages 205–214, New York, USA. ACM/IEEE.

18 Keijsers, M., Bartneck, C., and Kazmi, H. S. (2019a). Cloud-based sentiment analysis for interactive
19 agents. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages
20 43–50.

21 Keijsers, M., Bartneck, C., and Kazmi, H. S. (2019b). Cloud-based sentiment analysis for interactive
22 agents. In *Proceedings of the 7th International Conference on Human-Agent Interaction (HAI)*,
23 pages 43–50.

24 Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T. (2008). Can machines
25 think? interaction and perspective taking with robots investigated via fmri. *PloS One*, 3(7):e2597.

26 Lee, M. K., Kiesler, S., and Forlizzi, J. (2010). Receptionist or information kiosk: how do people talk
27 with a robot? In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*,
28 pages 31–40.

- Lortie, C. L. and Guitton, M. J. (2011). Judgment of the humanness of an interlocutor is in the eye of the beholder. *PLoS One*, 6(9):e25085.
- Lowry, P. B., Zhang, J., Wang, C., and Siponen, M. (2016). Why do adults engage in cyberbullying on social media? an integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research*, 27(4):962–986.
- MacDorman, K. F. and Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, 146:190–205.
- Mauldin, M. L. (1994). Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, pages 16–21.
- Moore, S. (2018). Gartner says 25 percent of customer service operations will use virtual customer assistants by 2020.
- Mori, M. et al. (1970). The uncanny valley. *Energy*, 7(4):33–35.
- Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, Boston, USA. ACM.
- Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., and Yamada, S. (2017). Why do children abuse robots? *Interaction Studies*, 17(3):347–369.
- Oberman, L. M., McCleery, J. P., Ramachandran, V. S., and Pineda, J. A. (2007). Eeg evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, 70(13-15):2194–2203.
- Paetzel, M., Peters, C., Nyström, I., and Castellano, G. (2016). Congruency matters – how ambiguous gender cues increase a robots uncanniness. In *International Conference on Social Robotics*, pages 402–412. Springer.
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015a). Linguistic inquiry and word count: Liwc 2015 [computer software]. pennebaker conglomerates.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015b). The development and psychometric properties of liwc2015. Technical report, The University of Texas at Austin.

- 1 Reeves, B. and Nass, C. (1996). *The Media Equation*. CSLI Publications and Cambridge University
2 Press, Cambridge.
- 3 Rehm, M. and Krogsgager, A. (2013). Negative affect in human robot interaction – impoliteness in
4 unexpected encounters with robots. In *Proceedings of the 22nd IEEE International Symposium on*
5 *Robot and Human Interactive Communication (RO-MAN)*, pages 45–50. IEEE.
- 6 Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N., and Sanchez-Vives,
7 M. V. (2006). A virtual reprise of the stanley milgram obedience experiments. *PloS one*, 1(1):e39.
- 8 Sokol, N., Bussey, K., and Rapee, R. M. (2016). Victims responses to bullying: The gap between
9 students evaluations and reported responses. *School Mental Health*, 8(4):461–475.
- 10 Strait, M., Contreras, V., and Vela, C. D. (2018). Verbal disinhibition towards robots is associated
11 with general antisociality. *arXiv e-prints*.
- 12 Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326.
- 13 Tan, X. Z., Vázquez, M., Carter, E. J., Morales, C. G., and Steinfeld, A. (2018). Inducing bystander
14 interventions during robot abuse with social mechanisms. In *Proceedings of the International Con-*
15 *ference on Human-Robot Interaction*, pages 169–177, New York, USA. ACM/IEEE.
- 16 Veletsianos, G., Scharber, C., and Doering, A. (2008). When sex, drugs, and violence enter the
17 classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with*
18 *computers*, 20(3):292–301.
- 19 Whitby, B. (2008). Sometimes its hard to be a robot: A call for action on the ethics of abusing artificial
20 agents. *Interacting with Computers*, 20(3):326–333.
- 21 Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational*
22 *Medicine*, 4(7).