HUMAN TECHNOLOGY

ISSN: 1795-6889

www.humantechnology.jyu.fi Volume 15(1), February 2019, 30–54

HELP! IS MY CHATBOT FALLING INTO THE UNCANNY VALLEY? AN EMPIRICAL STUDY OF USER EXPERIENCE IN HUMAN–CHATBOT INTERACTION

Marita Skjuve
SINTEF
Norway

Ida Maria Haugstveit
SINTEF
Norway

Asbjørn Følstad
SINTEF
Norway

Petter Bae Brandtzaeg
SINTEF
Norway

Abstract: *Advances in artificial intelligence strengthen chatbots' ability to resemble human conversational agents. For some application areas, it may be tempting not to be transparent regarding a conversational agent's nature as chatbot or human. However, the uncanny valley theory suggests that such lack in transparency may cause uneasy feelings in the user. In this study, we combined quantitative and qualitative methods to investigate this issue. First, we used a 2 x 2 experimental research design (*n *= 28) to investigate effects of lack in transparency on the perceived pleasantness of the conversation in addition to perceived human likeness and affinity for the conversational agent. Second, we conducted an exploratory analysis of qualitative participant reports on these conversations. We did not find that a lack in transparency negatively affected user experience, but we identified three factors important to participants' assessments. The findings are of theoretical and practical significance and motivate future research.*

Keywords: *Chatbot, human–chatbot interaction, uncanny valley, user experience.*

DOI: https:/doi.org/10.17011/ht/urn.201902201607

VERTAISARVIOITU KOLLEGIALT GRANSKAD PEER-REVIEWED www.tsv.fi/tunnus

INTRODUCTION

Chatbots can be surprisingly humanlike. When typing "hello" to Mitsuku, one of the most humanlike chatbots in the world (Milton-Barker, 2016), she may respond by saying, "It's my old friend Marita! How's it going?" Sounds just like a human, right? As described by social psychologist Paladiono, quoted by Mone (2016, p. 19), chatbots "are not human, but they're not exactly machines…. They are a different entity."

The interest in chatbots—that is, software applications capable of communicating through natural language[1] (Dale, 2016)—has skyrocketed in recent years (Piccolo, Mensio, & Alani, 2018), driven by, among other things, Microsoft's and Facebook's 2016 launches of frameworks for the integration of chatbots in their messaging platforms. The expectations for chatbots are particularly high within the area of customer service (Hyken, 2017), especially when coupled with machine learning for improved ability to tend to customer requests (Xu, Liu, Guo, Sinha, & Akkiraju, 2017). Alibaba and Domino Pizza are among the businesses that have deployed chatbots to help customers (Følstad, Nordheim, & Bjørkli, 2018). However, customer service is not the only place where chatbots can be found. Chatbots such as Mitsuku or Replika can become users' social companions,[2] and therapist chatbots like Woebot and Wysa can answer questions related to mental health.[3] Meanwhile, chatbots in programs such as Duolingo and Mondly can help in language acquisition.[4] Chatbots are thus used within a variety of different contexts and it is not hard to understand the interest this type of technology has generated.

The availability of mobile messaging platforms makes chatbots convenient for users and may be a promising alternative for service providers (Klopfenstein, Delpriori, Malatini, & Bogliolo, 2017). The three biggest messaging platforms (Facebook Messenger, WhatsApp, and WeChat) have more than a billion users each (Statista, 2018). Several major messaging platforms provide chatbot integrations and facilities for chatbot developers. At Facebook Messenger, 300,000 chatbots were available in 2018 (Johnson, 2018).

Developments within artificial intelligence and machine learning enable chatbots to mimic human behavior (Candello, Pinhanez, & Figueiredo, 2017) through sophisticated interaction skills. Candello et al. (2017) suggested that, due to these advancements in technology, users soon might have to "remind themselves that they are interacting with machines not people" (p. 3476). In the field of customer service, suppliers such as DigitalGenius work toward providing a seamless handover between chatbot and human customer representatives. This development creates situations where automated and human conversational agents participate on nearly equal footing.

In the near future, it will be important to know more about human–chatbot interaction as part of human technology as a research field. What happens when automated conversational agents become so sophisticated that users have a hard time distinguishing a chatbot conversation from a human conversation? Designing chatbots that resemble human conversational agents is certainly intriguing, but should this be a goal in itself? These questions remain unanswered. However, following Mori, the robotics researcher who developed the theory of the uncanny valley, the answer to the latter question above might be "no" (Mori, Macdorman, & Kageki, 2012).

According to Mori et al. (2012), humans' affinity—that is, humans' liking, attraction, or sense of kinship—toward a robot depends on its perceived human likeness, its perceived resemblance to human beings. However, Mori argued that if robots become too humanlike, they are at risk of inducing an uncanny feeling in users: a sense of dislike, unease, unpleasantness, or

eeriness in the face of entities that almost, but not quite, resemble healthy humans (Figure 1). The more humanlike a robot is perceived to be, the more affinity users will feel toward it, up to the point where the robot becomes too humanlike. When the robot reaches this point, there will be an abrupt shift in affinity, where users will experience the interaction as unpleasant or eerie. Such an abrupt shift in users' affinity is referred to as "the uncanny valley." This phenomenon seems to apply not only to humanlike robots but also to other humanlike artifacts, such as a theater mask or a prostatic arm, or diseased humans and corpses (Mori et al., 2012). Extending Mori's metaphor, we use the image that an entity that is almost, but not quite, humanlike runs the risk of being pushed into the uncanny valley, that is, to induce uncanny feelings in the user.

The uncanny valley effect has been argued to be one of the reasons why movies such as *Final Fantasy* have failed (Pollick, 2010). Others have found that users exhibit less desire to interact with an object that was perceived as uncanny (Strait, Vujovic, Floerke, Scheutz, & Urry. 2015), which illustrates that the human likeness of an artifact may become a liability rather than an asset.

The notion of an uncanny valley is important to investigate in response to the increasing use of chatbots, seen as developers all over the world currently strive toward sophisticated and humanlike interaction skills in chatbots. However, although the theory of the uncanny valley often is used to understand the implications of the physical appearance of an automated agent, little research is available on the aspects that are characteristic of chatbots. Relevant humanlike aspects of chatbots may include the agent's ability to reason and understand language (Lortie & Guitton, 2011), as well as its capability of humanlike responses indicating, for instance, politeness

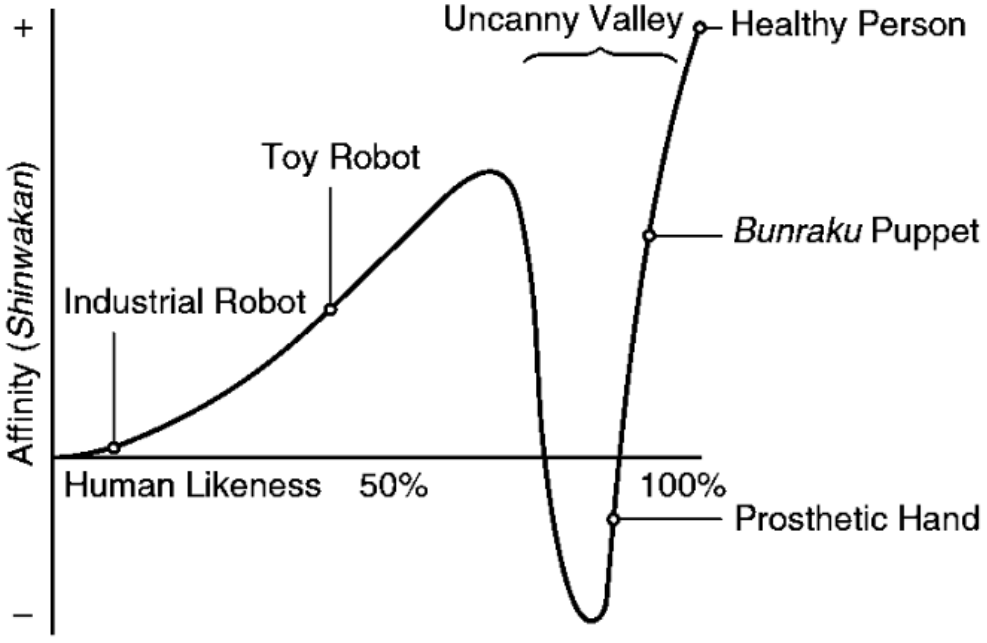

Figure 1. The graph presents the relation between users' affinity to entities at different levels of human likeness. As human likeness approaches that of a human being there is a sudden drop in affinity referred to as "the uncanny valley" (Mori et al., 2012).

friendliness, sociability, and humor (Thies, Menon, Magapu, Subramony, & O'Neill, 2017). In line with the notion of an uncanny valley, such aspects of chatbots may potentially elicit negative emotions or unease because these make the chatbot almost, but not quite, humanlike. The notion of an uncanny valley also echoes within popular culture and is assumed relevant for artifacts such as chatbots and virtual assistants (Waddell, 2017).

In recent years, the distinction between human and automated conversational agents in social media services such as Twitter has become blurred, where a large proportion of automated accounts pose as human conversational agents (Varol, Ferrara, Davis, Menczer, & Flammini, 2017). Sophisticated automated conversational agents in a social media context introduce the likely future situation where users at times are unsure whether they are interacting with a human or a machine (Ferrara, Varol, Davis, Menczer, & Flammini, 2016). The same tendency might be prominent in customer service.

The more sophisticated the chatbots become, the more tempting it may be for service owners to downplay the chatbots' identity as automated agents. As Hyken (2017) noted, "The best ones deliver a customer experience (CX) in which customers cannot tell if they are communicating with a human or a computer." In those situations, companies may see it as unnecessary to reveal to their customers whether they are chatting with a human or a chatbot when seeking support. In a situation in which a user is uncertain about the identity of a conversational agent, the user may perceive a chatbot as more humanlike than the user otherwise would have done if the chatbot is initially believed to be a human. Likewise, the user may perceive a human conversational agent as less humanlike if the human is believed initially to be a chatbot. In such a scenario, both the human conversational agent and the chatbot may be perceived as almost, but not quite, humanlike due to being confused with the other. Such confusion regarding the true nature of a conversational agent as human or chatbot may induce undesirable user experiences, as suggested by the theory of the uncanny valley, that is, feelings of dislike, unease, unpleasantness, or eeriness.

Because chatbots are predicted to be a new way for users to interact with services (Følstad & Brandtzaeg, 2018), it is crucial to gain an understanding of how people experience chatbots. It is particularly important to understand how users perceive situations characterized by uncertainty regarding the nature of a conversational agent and whether there is such a thing as an uncanny valley of chatbots. As such, we are addressing a gap in the present research by exploring the potential effects such uncertainty may have on the user experience. The following research questions guided our work to address this research gap:

> RQ1: Which factors affect users' perceptions of chatbots and chatbot conversations?
>
> RQ2: Does not knowing whether a user is talking to a chatbot or a human conversational agent push the chatbot, or even the human agent, into the uncanny valley?

This paper provides two important contributions. First, our study contributes needed new knowledge that has practical implications for the design of human–chatbot interaction (RQ1). Based on this contribution, we propose design guidelines and requirements. Second, our study serves to explore the theory of the uncanny valley within the domain of chatbots. By including RQ2, we provide needed knowledge on the benefits and limitations of clearly distinguishing automated and human conversational agents.

RELATED WORK

Is There Such a Thing as the Uncanny Valley?

Although empirical evidence regarding the uncanny valley theory is conflicting (Burleigh, Schoenherr, & Lacroix, 2013), several studies have found a likely relationship between the human likeness of artifacts and the affinity people feel toward these. Poliakoff, Beach, Best, Howard, and Gowen (2013) found that prosthetic hands were seen as more humanlike—and thus more uncanny—compared to mechanical hands. Similarly, Mathur and Reichling (2016) demonstrated that the more humanlike a robot face became, the more unpleasant people found it to be. Lastly, Geller (2008) demonstrated an uncanny valley effect in robots and further suggested adding social responsive features to the robot as a means to reduce this effect.

The above studies focused on the physical appearance of artifacts. Since chatbots interact through text or voice-based conversations, their humanlike character may have a different effect than that of physical objects. Hence, a need exists for research investigating the uncanny valley of chatbots to complement the current body of knowledge.

Interaction with Chatbots

Although the term chatbot is relatively recent, software agents communicating with users in natural language have been studied since Weizenbaum's ELIZA (1966). Such agents come by various terms, such as conversational agents, chatterbots, and chatbots.

In a study of human interaction with computers, Nass, Steuer, and Tauber (1994) showed that people tended to demonstrate responses that adhered to social norms. Studies of human–chatbot interaction have produced similar findings. Kopp, Gesellensetter, Krämer, and Wachsmuth (2005) investigated interaction with a conversational agent, Max, and found that people often used talking patterns typical of interactions with humans, such as saying "hello" or participating in small talk.

Conversely, there may be differences in the interaction patterns of human–human and human–chatbot conversations. Hill, Ford, and Farreras (2015) found that users tended to have simpler conversations with a less rich language when talking to the chatbot Cleverbot versus to a person. However, as shown by Allison, Luger, and Hofmann (2017) in a study of intelligent support agents in Minecraft, there may be substantial variation in how people talk to such agents: While some use only short commands, others may express themselves in full sentences. Even though users tend to use shorter sentences in their interaction with chatbots, research indicates that they send more messages to the chatbot compared to a human, which might be an indication of users' motivation for interacting with chatbots (Hill et al., 2015), as has also been found in other studies (e.g., Crutzen, Peters, Portugal, Fisser, and Grolleman, 2011).

In a recent study, Ciechanowski, Przegalinska, Magnuski, and Gloor (2019) experimentally compared users' physiological and experiential responses during interaction with text-based chatbots and an avatar-based chatbot in the context of the uncanny valley. They found less support for an uncanny valley effect for the text-based chatbot than the avatar-based chatbot. No studies to our knowledge have investigated the uncanny valley effect for conversation with chatbots compared to conversation with humans.

Transparency in the Interaction

Given that humans differ in how they interact with a chatbot and with a human, and that human and chatbot capabilities differ, transparency in the nature and limitations of chatbots is important. In particular, as some argue, "Chatbots should be upfront about their machine status" (Mone, 2016, p. 19) because it is beneficial, among other things, to avoid negative implications of users failing to realize the limitations in chatbots. For instance, Luger and Sellen (2016), in an interview study of Siri, Cortana, and Google Now users, identified substantial mismatch between user expectations and system capabilities and, consequently, argued for the need for conversational systems to be open about their limitations. Castellano and Peters (2010) also echoed this notion.

Somewhat contrary to the above studies on the benefits of transparency, Murgia, Janssens, Demeyer, and Vasilescu (2016) found that, when deploying a bot that answered users' questions in the Stack Overflow community, the bot was regarded more positively when posing as a human than when revealing its bot identity. In a similar vein, Corti and Gillespie (2016) found users more likely to expend effort in making themselves understood when the agent's chat content was conveyed through a human (the so-called echoborg method) than through a text-based interface. Hence, transparency in the machine identity and capabilities of a chatbot may work counter to its intention. Taken together, these studies suggest benefits and limitations of transparency in human–chatbot interaction. However, no study to our knowledge has investigated whether lack of transparency can lead to an uncanny feeling when the chatbot poses as a human or when there is uncertainty regarding the true nature of a conversational agent.

Factors Affecting the Perception of Chatbots

Several factors influence the perceived human likeness of chatbots, such as typography styles (Candello et al., 2017), word frequency (Lortie & Guitton, 2011), and responsiveness (Schuetzler, Grimes, Giboney, & Buckman, 2014). Schuetzler et al. (2014), for example, found that a dynamic chatbot, which generated relevant follow-up questions, was perceived as more humanlike compared to a static bot that did not generate such follow-up questions. This increased human likeness in turn motivated the participants to provide more engaging answers. Vtyurina, Savenkov, Agichtein, and La Clarke (2017) explored the differences between human and artificial agents during information-seeking tasks. They revealed that the participants seemed to appreciate that a human conversational agent understood their questions, which the automatic agent did not always do. Lastly, Lortie and Guitton (2011) used data from the Loebner Prize Turing Test[5] to investigate cases where evaluators mistakenly identified human conversational agents as chatbots. They argued that the perceived human likeness is, to some extent, a product of both the behavior of the agent and the perceptions of the person judging the agent. In other words, "the judgment of human likeness lies in the eye of the judge himself" (Lortie & Guitton, 2011, p. 5). Thus, the fact that the conversational agent is a human does not necessarily make the user feel like he/she is talking to a real person, given that the true nature of this agent is unknown. Situations where the true nature of a conversational agent is unknown, hence, may affect how both a chatbot and a human conversational agent are perceived.

METHOD

To investigate the effect of not knowing the true nature of a conversational agent, and whether this may introduce perceptions of uncanniness, we applied a dual research approach. First, we used an experimental research design to provide insight into the effect of variation in the actual chat partner and beliefs about chat partner on predefined dependent variables. Second, we gathered and analyzed qualitative data from the participants to explore how users experience the different chat partner configurations.

In the study, the participants chatted with an unknown conversational agent. We systematically manipulated whether the participants actually chatted with a human or a chatbot and whether they believed they were chatting with a human or a chatbot. In doing so, we could compare the participants' experience of chatting with a chatbot with that of chatting with a human conversational agent. We could also study the effect of the chatbot posing as a human, or the human posing as a chatbot, conditions we assumed could potentially push the chatbot, or even the human conversational agent, into the uncanny valley due to the blurred boundary between the two.

Recruitment Process and Participants

We recruited participants through posters at a university campus, e-mailing lists, and postings on social media sites. The recruitment text briefly informed about the data collection and how it would be conducted, in addition to listing requirements for participation. The participants had to be 18 years or older, experienced Internet users, and comfortable with having a written conversation (chat) in English. This was the chosen language for the participant sessions as the chatbot (Mitsuku) communicates in English.

The final sample consisted of 28 participants. While this is a quite small sample, we deemed it sufficient for an initial study of this kind, in particular because we extended the study with qualitative analyses of participants' free-text responses on their experience. None of the participants had English as a first language. Yet, most Norwegians are competent with English, as they have studied this as a second language in school. See Table 1 for further demographic details.

All participants received and signed an informed consent form in the beginning of the session, before the participant session was initiated. The researchers stressed that participants could withdraw at any time before, during, or after the study. Each participant received a gift card valued at 150 NOK (approximately US$20) for his/her participation.

Table 1. Demographic Details for Participants in the Study Investigating Factors Affecting User Perceptions of Chatbots.

Gender		Age		Education		Experience with chatbots	
Male	14	18-24	13	High school	13	Prior experience	14
Female	14	25-34	10	Bachelor degree	8	Heard of, not tried	10
		35-44	4	Master degree	6	Not heard of, not tried	3
		45-54	1	Doctoral degree	1	Don't know	1

Study Setup

We used a 2 x 2 factorial design. The participants were assigned randomly to one of four conditions in which they were chatting with either a female chatbot or a female human. In one of the two chatbot conditions, the participants were led to believe that they were chatting with a human. In one of the two conditions with a human conversational agent, the participants were led to believe they were chatting with a chatbot. In all conditions, the conversational agent was named Ann, regardless of whether that agent was a chatbot or a human. All participants chatted with Ann through the same platform, Slack,[6] to provide an identical user interface for all participants. The conditions are summarized in Table 2.

We measured three variables as part of the experimental research design. The main dependent variable of the study was perceived *pleasantness.* This was measured with a single-item instrument, allowing us to ask the participants for their experience of their chat partner in an open-ended, free-text follow-up question (see below). We also included two additional dependent variables, perceived *human likeness* and *affinity*. These variables were measured as multi-item instruments, drawing on the uncanny valley literature. If our assumption that the blurring of the line between human and chatbot conversational agents in Conditions 2 and 3 could push the either of these agents into the uncanny valley was correct, we would expect to see lower perceived pleasantness and affinity for these two conditions than for Conditions 1 and 4.

For the exploratory part of the study, all participants were asked following the test to report, in their own words, their reasons for assessing the conversation as pleasant/unpleasant. These free-text reports provided insights into the drivers of the users' experience for the conversations.

Taken together, the dependent variables and the open-ended free-text data collection provided an opportunity to compare systematically the different conditions and to gather qualitative data on user experience. Prior to running the study, we tested the research design on six pilot participants to identify and correct issues with the study setup and the questionnaire. As a result of pilot phase, we made minor adjustments to the questionnaire and procedure.

The same moderator guided each participant session. To be able to carry out the experiment in a controlled manner, the participants did not know its actual purpose. We therefore used a cover story: The participants were told that we were testing the messaging platform Slack and that we wanted feedback related to their experience of this platform. After the chat, the participants answered a questionnaire on their experiences from the conversation. Finally, we debriefed the participants.

Table 2. The 2 x 2 Factorial Design Applied to Investigate Effects of Not Knowing the True Nature of a Conversational Agent as Human or Chatbot.

Actual chat partner	**Belief about chat partner**	
	Chatbot	**Human**
Chatbot	**Condition 1** Chats with chatbot, believes it is chatbot	**Condition 2** Chats with chatbot, believes it is human (*deceived*)
Human	**Condition 3** Chats with human, believes it is chatbot (*deceived*)	**Condition 4** Chats with human, believes it is human

Ethics and privacy concerns were important when designing the study, principally because some participants were deceived as part of the data collection. To monitor and mitigate any negative ethical implications of this deception, these pilots and participants were given a thorough debriefing where we informed them about the true nature of the conversational agent and probed on whether they experienced this as negative or problematic. None of the pilots or participants reported such negative experiences; if they had, the study would have been terminated or modified accordingly. To manage the potential privacy implications of the study, all data collection was conducted anonymously: Neither the chat logs nor the questionnaire data were connected to any person's identifiable data.

The study took place in Norway, which has a substantial digital literacy and a population that is quick to adopt new technology into their private and professional lives. All participant sessions took place in a quiet location without potential distractions; thus, the sessions were conducted either at the research team's office building or at a convenient place of the participant's choosing. The gender distribution in each condition was even. The data were collected between June 14 and July 1, 2017.

Chatting with "Ann"

Each participant session was initiated in the same way. First, the moderator informed the participants that they were going to have a 15-minute conversation in English with "Ann." For privacy purposes, an anonymous Slack account was used, and all the participants communicated under the gender-neutral name "Kim."

The participants were instructed that the format of the conversation was small talk and were asked to avoid sharing personal or sensitive information during the conversation. They also were told that they were responsible for keeping the conversation going. This instruction was given to minimize the risk of treating the participants in the chatbot condition differently due to the chatbots' limitations in responsiveness. Participants received a sheet with suggestions for topics and follow-up questions if they needed help with what to say.

The average length of the conversations was 13 minutes (Range: 9–18 minutes). A conversation included approximately 14 exchanges between the participant and Ann. After the chat, the participant answered a questionnaire with both open and closed questions related to his or her experiences. The chat session and questionnaire were conducted on a laptop computer provided by the study moderator, and each session lasted 20–25 minutes.

Ann, the Chatbot

We used the chatbot Mitsuku as a backend in both chatbot conditions (Conditions 1 and 2). Mitsuku was developed by Steve Worswick as a chatbot for social small talk mimicking a female character. She is a four-time winner of the Loebner Prize, an annual competition for artificial agents based on the Turing Test (Milton-Barker, 2016). She was the winner in 2018 and was thus deemed fit for our study.

When studying the uncanny valley, one might consider using several different chatbots to manipulate the degree of human likeness. However, as any uncanny valley effect would require high levels of human likeness, possibly higher than what is available in many current chatbots, we argue that an uncanny valley would be located between Mitsuku and a human

conversational agent. Hence, utilizing chatbots with less human likeness would not provide much information other than potentially mapping out an increase in familiarity and affinity on the "road" toward the uncanny valley.

When conducting the chat dialogues, we applied a "Wizard of Oz" approach (Dahlbäck, Jönsson, & Ahrenberg, 1993), where the wizard (one of this paper's authors) mediated the communication between the participants and the chatbot. Thus, the wizard functioned as an intermediary between Mitsuku and participants in the chatbot conditions by copying and pasting between Slack and the chatbot Mitsuku. Because Mitsuku often reveals that she is a chatbot, the wizard had to modify some of Mitsuku's responses. Prior to the data collection and based on the pilot trials, the research team established a standardized procedure for how potential responses would be changed. Modification was applied only to words or short parts of sentences where Mitsuku specifically revealed her chatbot nature and in a way that should not interfere with the participants' general experience of the conversation. On average, one or two of Mitsuku's responses were changed during each session (Range: 0–4). See Table 3 for examples of such responses and how they were modified by the wizard.

Ann, the Human

In the human conditions (Conditions 3 and 4), participants were chatting with an actual human. Here, the wizard chatted directly with the participants. In Condition 3, the wizard acted as a human, but the participants were told that they were talking to a chatbot. To ensure that all the participants were treated in a similar manner and that the wizard did not change behavior across participant sessions, a procedure for how the wizard was to act was established and piloted. The procedure included, among others, a given maximum length for each response, in addition to how and when to offer follow-up questions. Condition 4 followed the same procedure as Condition 3. The only difference was that the participants in Condition 4 believed that they were talking to a human.

The Questionnaire

The questionnaire was set up in SurveyMonkey, an online tool for creating, distributing, and managing questionnaires. The questionnaire included an initial question regarding the usability of the Slack platform, as the participants were led initially to believe that this was the focus of the study. The subsequent questionnaire items concerned the actual research questions.

Table 3. Examples of the Original and Modified Responses from Mitsuku to Avoid Disclosing the Nature of the Chatbot.

Original sentence	Modified sentence
1. That's why I said it. I knew I was right. I like ***robots***, computers, and chatting online.	1. That's why I said it. I knew I was right. I like ~~***robots***~~, computers and chatting online.
2. Interesting deduction. It's kind of hard ***without a body***.	2. Interesting deduction. Yes, but it's kind of hard ***too*** ← ~~***without a body***~~.

Note. Problematic words are marked in italic bold and were modified or left out.

Here, the participants were asked first about the perceived pleasantness of the conversation: "How did you experience your conversation with Ann?" that scored from 1 (*very unpleasant*) to 10 (*very pleasant*). The researcher followed up with an open question: "Please, explain why you gave this score (3-8 sentences)," allowing for answers in free text. Here, users provided feedback in, on average, two to three lines of text (M = 190 characters, SD = 100).

To measure perceived human likeness and affinity, the participants answered a battery of eight Likert scale questions (scales running from 1, *strongly disagree*, to 5, *strongly agree*) drawn from published studies on the uncanny valley effect. That is, the participants were asked to rate their chat partner (Ann) on whether she was natural, lively, realistic, and artificial, which addressed perceived human likeness, in addition to friendly, likable, comfortable, and warmhearted, which detailed affinity. Data also were gathered on the participants' age, gender, education level, and prior experience with chatbots.

The Data Analysis

Two types of data were collected in our study. Each type addressed different aspects the users' experience of the human–chatbot interaction.

Quantitative Data

The quantitative data were analyzed using the SPSS statistical package. Aggregated scores for affinity and perceived human likeness were established as the mean of the associated questionnaire items. To assure it was appropriate to make such aggregations, the associated items for each variable were first made subject to a principal component analysis, to check that they loaded on only one underlying factor, followed by an analysis of interitem reliability. One item, lively, loaded on both factors and therefore was excluded from the subsequent analysis. Interitem reliability (Cronbach's alpha) was acceptable (Gliem & Gliem, 2003) for both perceived human likeness (0.82) and affinity (0.90).

Two-way ANOVA was used to investigate the main and interaction effects for the independent variables: *belief* (whether the participants believed they were chatting with a human or chatbot) and *actual* (whether the participants actually were chatting with a human or a chatbot).

Qualitative Data

The participants' free-text responses to explain their pleasantness scores of the conversation were subjected to an inductive thematic analysis. Here, each meaningful statement was extracted, coded, and combined into meaningful themes (Ezzy, 2002).

The second author established the coding themes and coded the entire set of statements as well. The same code set was used for an independent coding of the entire set by the first author. Inter-rater agreement was found to be substantial (Cohen's kappa = 0.78), following Landis and Koch's (1977) rule of thumb. The three main themes were (a) conversation content, (b) conversation demeanor, and (c) conversation flow.

RESULTS

We structured the study results into three subsections. The first of these addresses the perceived pleasantness of the conversation. The second concerns factors affecting the participants' perceptions of the conversation, while the third presents the findings regarding perceived human likeness and affinity.

Perceived Pleasantness of the Conversation Higher for Human than Chatbot Conversational Agent

For the dependent variable perceived pleasantness of the conversation (Figure 2), we found a significant main effect of the actual nature of the conversational agent (human vs. chatbot), $F(1, 24) = 9.44$, $p < 0.01$, $\eta^2 = 0.28$. The value of partial eta-square (η^2) indicates a large effect size (Field, 2013).

However, there was a nonsignificant main effect of the participants' belief regarding the nature of the conversational agent, $F(1, 24) = 0.56$, $p = 0.46$, $\eta^2 = 0.02$. That is, the participants' perceived pleasantness of the conversation was affected by whether or not they actually chatted with a human or a chatbot but not by their belief regarding the nature of their conversational agent. No significant interaction was found between the actual and the believed nature of the conversational agent, $F(1, 24) = 1.56$, $p = 0.23$, $\eta^2 = 0.06$.

The two conditions where the participants actually chatted with a human were scored relatively higher on perceived pleasantness of the conversation, both when the participants believed they were chatting with a human ($M = 7.29$, $SD = 2.50$) and when they erroneously believed they were chatting with a chatbot ($M = 7.71$, $SD = 1.50$). The two conditions in which the participants chatted with a chatbot were associated with relatively lower scores. Notably, when the chatbot posed as a human, the scores for perceived pleasantness in the conversation ($M = 5.71$,

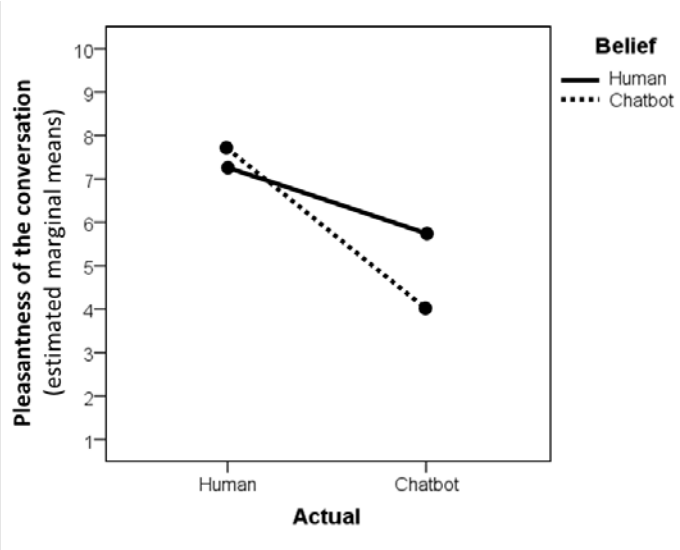

Figure 2. Main and interaction effects of actual and belief (independent variables) on perceived pleasantness of the conversation (dependent variable) in the 2 x 2 factorial design applied to investigate effects of not knowing the true nature of a conversational agent as human or chatbot.

SD = 2.99) was somewhat higher than when the chatbot revealed its chatbot nature (*M* = 4.00, *SD* = 1.83), though this difference was not sufficiently substantial for a significant interaction effect.

Factors Affecting User Perceptions of the Conversation

The inductive content analyses yielded several factors reported to affect the participants' perception of the conversation. These factors were structured around three main themes: (a) conversation content, (b) conversation demeanor, and (c) conversation flow.

Conversation Content

The conversation content was reported to be important to the perceived pleasantness of the conversation. Most of the participants chatting with chatbot–Ann mentioned a sense of something being a bit strange in terms of conversation content, regardless of whether they believed Ann to be a human or a chatbot.

These participants explained that Ann communicated in a way that they perceived as odd, with cryptic or off-point answers. Several noted that Ann did not fully understand their questions and tended to give strange, irrelevant and, in some cases, wrong answers, as exemplified in this extract: *"The conversation with Ann was strange. Ann often gave me irrelevant answers to the questions I asked, which made the conversation messy and annoying"* (Participant, Condition 2).[7]

Figure 3 shows a screenshot of part of one participant's conversation with chatbot–Ann, which illustrates Ann giving strange responses and not fully understanding. The often-strange conversation content of chatbot–Ann made several of the participants in Condition 2 suspicious of the nature of the conversational agent, although these participants were led to believe they were chatting with a human. In fact, three of these participants explicitly stated that they felt

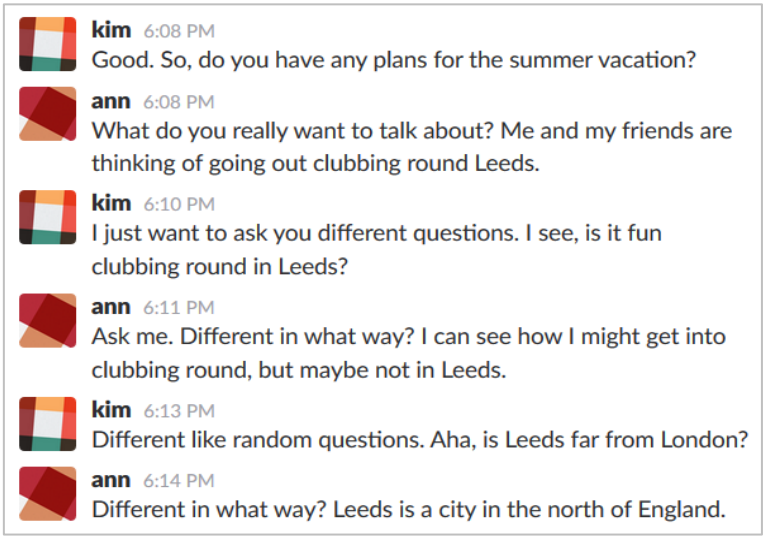

Figure 3. Screenshot of conversation (Condition 2) showing chatbot–Ann's tendency to give strange responses.

like they were talking to a robot. The participants who were chatting with human–Ann did not report on the conversation content as being cryptic or vague. Rather, the participants here stated that Ann gave relevant answers and asked relevant questions.

Conversation Demeanor

Another aspect of the conversation that seemed to influence the perceived pleasantness of the conversation was conversation demeanor. This refers to how the participants experienced Ann's personality. Overall, several of the participants seemed to experience the conversational agent as nice and polite, regardless of condition.

For the participants who were chatting with chatbot–Ann, the at-times odd demeanor of Ann closely intertwined with the perceived strangeness in the conversation's content. A few of the participants perceived Ann's personality negatively, for example, as rude and sarcastic, as exemplified in the following extract and Figure 4: *"She was a little cryptic and did not always understand my questions. Besides, she actually seemed sarcastic and uninterested, unlike the times she suddenly was extremely interested"* (Participant, Condition 1).

Although most reported Ann to be polite and nice, some stated that they at times perceived her as being a bit impersonal and restrictive in her communication. They explained this with her not sharing much information about herself, avoiding personal topics, giving superficial answers, and being passive. This extract exemplifies this perspective: *"She had interesting views and ideas, and we had a good communication. She was basically commenting on my thoughts and ideas, and she did not tell so much about herself"* (Participant, Condition 3).

An interesting detail is that the participants tended to perceive the conversation with chatbot–Ann to be more pleasant when they were led to believe that she was a person. There is no explicit reason for this tendency expressed in the qualitative data. It seems, however, as if those who knowingly talked to chatbot–Ann focused less on the personality of Ann and more on the content in the chat. Those who believed that chatbot–Ann was a person did, to a greater extent, comment on her demeanor (e.g., noting that she was nice). Possibly, information on the identity of the conversational agent influenced what the participants focused on in their interaction, which in turn potentially impacted their experience.

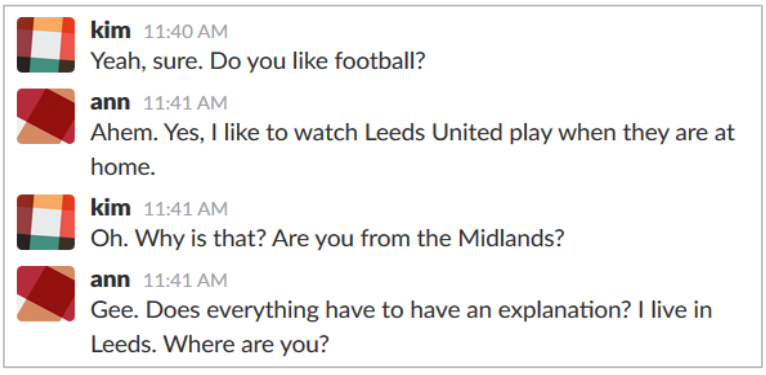

Figure 4. Screenshot of conversation (Condition 1) showing chatbot–Ann's somewhat rude demeanor.

Conversation Flow

Communication flow and speed clearly influenced how pleasant the participants perceived their conversation with Ann. Eight of the 14 participants who chatted with human–Ann noted that the conversation ran a bit slow and that the conversational agent used a long time to answer. This tendency was noted regardless of whether they believed Ann was a human or a chatbot. The slow nature of the conversation made participants experience the conversation as bothersome, and they felt that it did not flow naturally and that they had to work to keep the conversation going. One participant's response illustrates this experience: *"The conversation went a bit slow, and I felt I had to work very hard to keep the conversation going"* (Participant, Condition 4).

This tendency was not as prominent in the conditions where the participants chatted with chatbot–Ann. Here, only two of the participants mentioned the slowness of the conversation. Yet, one of the participants specifically stated that the speed affected the flow, as exemplified in this response: "*The conversation was very slow. She seemed quite bright in the beginning, but she misunderstood a bit as the conversation went on. The conversation speed was so slow, which affected the flow."* (Participant, Condition 1).

The perceived slowness in the conversations is likely due to the human processing required for human–Ann and chatbot–Ann to respond. A human conversational agent will be slower in responding than a chatbot. Hence, it is not surprising that human–Ann was perceived as slow by more participants compared to chatbot–Ann. However, as chatbot–Ann implied a manual process of moving responses from Mitsuku, this was somewhat slower than what one would expect from a chatbot.

Perceived Human Likeness and Affinity

Perceived human likeness and affinity are the two key factors in the theory of the uncanny valley. Specifically, an increase in human likeness is seen as a prerequisite for pushing the chatbot into the uncanny valley, that is, to enter a level of human likeness where the users' affinity towards the chatbot is reduced substantially, resulting in uncanny feelings in the human induced by the chatbot being almost, but not quite, humanlike.

For the dependent variable perceived human likeness (Figure 5), there was a significant main effect of the actual nature of the conversational agent, $F(1, 24) = 14.63$, $p < 0.001$, $\eta^2 = 0.38$. The value of partial eta-square (η^2) indicates a large effect size (Field, 2013). There was no significant main effect of the participants' belief in whether they were chatting with a chatbot or a human, $F(1, 24) = 0.69$, $p = 0.41$, $\eta^2 = 0.03$. Hence, the participants' initial beliefs regarding their conversational agent (chatbot vs. human) did not affect perceived human likeness; only the actual nature of the conversational agent did. No significant interaction effect was found, $F(1, 24) = 0.44$, $p = 0.51$, $\eta^2 = 0.02$. The participants chatting with human–Ann tended to rate perceived human likeness higher, whether they initially were led to believe that Ann was a human ($M = 3.38$, $SD = 0.89$) or a chatbot ($M = 3.80$, $SD = 0.54$). Likewise, the participants chatting with chatbot–Ann tended to rate her lower on perceived human likeness, both when initially believing she was a human ($M = 2.52$, $SD = 0.88$) and a chatbot ($M = 2.47$, $SD = 0.66$).

The participants' scoring of their affinity (Figure 6) toward Ann deviated somewhat from the pattern seen for perceived pleasantness in the conversation and perceived human likeness.

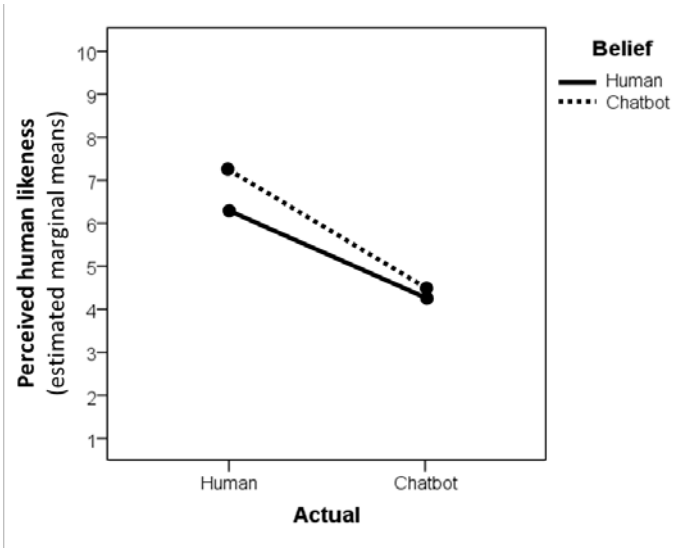

Figure 5. Main and interaction effects of actual and belief (independent variables) on perceived human likeness (dependent variable) in the 2 x 2 factorial design applied to investigate effects of not knowing the true nature of a conversational agent as human or chatbot.

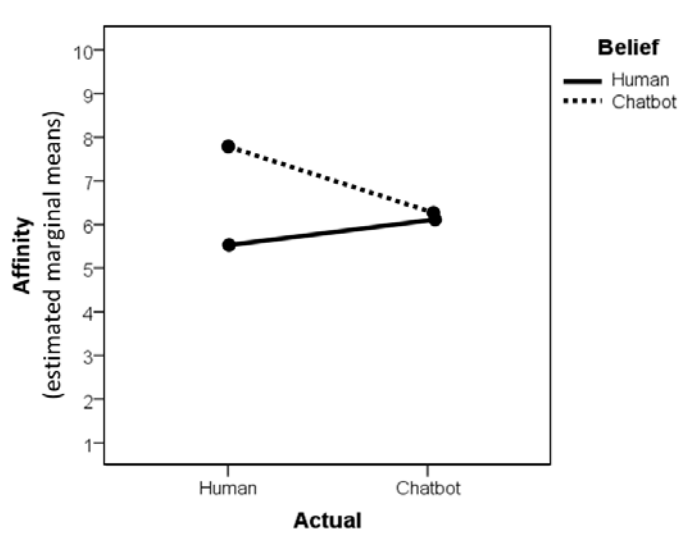

Figure 6. Main and interaction effects of actual and belief (independent variables) on affinity (dependent variable) in the 2 x 2 factorial design applied to investigate effects of not knowing the true nature of a conversational agent as human or chatbot.

In this regard, we found that whether the participants actually chatted with a human or a chatbot had a nonsignificant main effect on affinity, $F(1, 24) = 0.41$, $p = 0.53$, $\eta^2 = 0.02$. However, their belief whether Ann was a human or a machine approached being a significant main effect, $F(1, 24) = 3.27$, $p = 0.08$, $\eta^2 = 0.12$. The interaction between belief and actual also was nonsignificant, $F(1, 24) = 2.48$, $p = 0.13$, $\eta^2 = 0.09$. The values of partial eta-square (η^2) indicate medium effect sizes for the dependent variable belief and for the interaction effect (Field, 2013).

Interestingly, the highest scores on affinity were given by users chatting with human–Ann but believing her to be a chatbot ($M = 4.00$, $SD = 0.41$), while the lowest scores were given by

users chatting with human–Ann believing her to be human (M = 2.96, SD = 0.87). The participants chatting with chatbot–Ann were not affected in their affinity ratings by their belief of Ann being a human (M = 3.25, SD = 1.15) or a chatbot (M = 3.32, SD = 0.62). There is, hence, nothing in our findings to support our initial suspicion that being misled in terms of the nature of the conversational agent may push this agent, whether human or chatbot, into the uncanny valley.

DISCUSSION

The findings from the qualitative data demonstrate several factors that affect the perceived pleasantness of human–chatbot interaction. At the same time, the quantitative data fails to show that lack of transparency, in terms of not knowing the true nature of a conversational agent as human or artificial, may push this agent into the uncanny valley. We will, in the following, discuss key findings and highlight theoretical and practical implications.

Personal and Pleasant?

While chatbots have become more sophisticated in their ways of interacting with humans, they still have a long way to go before achieving full human likeness. Chatbots that are able to connect to the user on a more personal level might be important for the user experience; conversely, chatbots that are perceived as restricted or impersonal might elicit negative responses. The results from this study support this notion and illustrate how users experience impersonality or lack of self-disclosure in a conversational agent as a negative effect in the interaction. Surprisingly, this tendency was most prominent in the conditions where the participants either knowingly or unknowingly chatted with a human. This may exemplify how users adjust their expectations not based on knowledge prior to initiating the conversation, but on the information they receive during the conversation.

Self-disclosure also has been found previously to influence the perception of interactions with conversational agents. Lortie and Guitton (2011) found that sharing more personal information elicited positive answers as long as it was not excessive, which could lead to the opposite effect. Our study adds to this finding by suggesting that too little can also cause a negative effect.

Prior research indicates that people tend to adapt how they interact with conversational agents depending on whether these are perceived as humanlike (Purington, Taft, Sannon, Bazarova, & Taylor, 2017). An impersonal character displayed in a lack of self-disclosure by the chatbot may affect such perceptions. Designers should be aware that creating chatbots capable of resembling human interaction without being able to display self-disclosure might affect how the conversation is perceived and, hence, affect how the users adapt to the chatbot.

Slow or Odd

Human–human interaction often is characterized by turn-taking, conformational acts, and appropriate responses (Cassell, Bickmore, Campbell, Vilhjálmsson, & Yan, 2000). We found that the participants chatting with chatbot–Ann often stated that Ann's replies lacked relevance and seemed odd. This result indicates that Ann was not able to follow the conversation properly.

Lack of conversation skills is a known problem for chatbots, and other studies have found similar results. For instance, Yang, Ma, and Fung (2017) found that lack of relevant questions contributed to a lack of desire to interact with a virtual agent, and Cassell and colleagues stated that being intelligent includes "'social smarts'–the ability to engage a human in an interesting, relevant conversation with appropriate speech and body behaviors" (Cassell et al., 2000, p. 30).

Our results demonstrate that lack of social smarts contributes to making the conversation less pleasant, and chatbot developers will have much to gain by creating chatbots that are able to provide more relevant and accurate answers, that is, display a higher degree of social smarts. What is interesting, however, is that while chatbot–Ann was not able to keep up with the topics in the conversation, human–Ann was not able to keep up with the expected speed of textual exchanges. That is, the participants in Conditions 3 and 4 described Ann as slow and, in some cases, unresponsive. Walther (2007) noted that having to wait a long time for an answer can be perceived as disturbing. Being able to provide quick responses and frequently contribute with follow-up questions might be important to creating a pleasant and engaging human–agent interaction. In fact, our findings demonstrate how chatbots have an advantage due to their ability to provide fast answers, which makes them fit for working within customer service, as studies have shown that one of the key motivations for utilizing chatbots is productivity and saving time (Brandtzaeg & Følstad, 2017; Luger & Sellen, 2016). Our findings support this notion and demonstrate what happens when users must wait for an answer: They perceived it negatively.

Is Lack of Transparency Uncanny?

Following from the theory of the uncanny valley, it might be expected that a situation where there is a fairly high degree of human likeness and the participant is led to believe incorrectly that he or she is talking to a human or a chatbot would affect the affinity toward the conversational agent. Thus, we expected that participants holding erroneous beliefs concerning the nature of the conversational agent would rate Ann lower on affinity than participants holding correct beliefs. Our results provide no evidence to support this. Rather, for the conversations with human–Ann, an opposite tendency was found. Our interpretation of this finding is that text-based chatbots still have a long way to go before they become sufficiently humanlike for an uncanny valley effect to be relevant. As of now, any uncanny effect likely will be dwarfed in comparison with the effect of the difference between a human conversational agent and a chatbot. Specifically, this is seen in the findings for perceived human likeness, where the scores are strongly affected by whether the conversational agent actually is a human or a chatbot.

Even the current champion of the Loebner Prize, Mitsuku, is not sufficiently humanlike to be mistaken for a human—even when the user is told she is human. Hence, at the current state of chatbot development, a lack of transparency does not seem to be sufficient to create noticeable uncanny effects. This is not to say, however, that such an effect cannot be relevant in the future.

THEORETICAL AND PRACTICAL SIGNIFICANCE OF THE FINDINGS

The presented study provides insight of significance both for theory on human-chatbot interaction and application concerning chatbot implementations. We detail each focus below.

Theoretical Significance

Our quantitative results provide no evidence to suggest that a lack of transparency in the human–chatbot interactions pushes a conversational agent, whether chatbot or a human, into the uncanny valley. So, what do our qualitative findings tell us about the relevance of the uncanny valley theory in human–chatbot interaction?

First, our findings suggest that people react differently to humanlike chatbots communicating through a text-based interface than they do toward physically or visually present robots. That is, the state-of-the-art in a humanlike chatbot does not elicit uncanny feelings, even when there is uncertainty related to the true nature of the conversational agent. As shown by our qualitative data, people may think that there is something strange about the conversation, but they seem not to perceive it as uncanny.

Second, what was perceived as somewhat uncanny was having to wait for an answer, especially when chatting with a human. The presence of a time lag might induce feelings resembling those predicted in the uncanny valley theory due to uncertainty concerning what to expect next from the conversational agent. That is, the users might begin to wonder whether they should keep conversing or wait for a reply, and, if they decide to wait, for how long. Although there might not be an uncanny valley effect associated with text-based chatbots, an inadequate conversation flow could create an unpleasant user experience. As such, the use of chatbots for customer service might reduce the risk of slow human customer service inducing feelings of dislike or unpleasantness because of the chatbots' ability to provide fast answers. Taken together, our findings add to the theory of the uncanny valley by demonstrating that lack of transparency in human–chatbot interaction is not necessarily perceived as uncanny.

Practical Significance

In the Introduction, we argued that it may be tempting for service owners to hide the real identity of an online customer service representative as the chatbot becomes more sophisticated—to the extent that the customer does not know if the representative is real or artificial (Hyken, 2017). Although we initially assumed that this could create an uncanny feeling, our findings do not provide support for such an effect. Hence, for service owners, it will likely be other factors more important to the user experience in chat interactions with customers, such as conversation content, demeanor, and flow. Allowing for situations where the true nature of a conversational agent may be unclear to the user may be a feasible option, at least when seamless handover is achieved and when erroneous expectations in the conversational agent's capabilities do not lead to inadequate or frustrating service outcomes.

We summarize the practical significance of our findings, and our reflections on the literature, in the following three points:

1. Nontransparency still implies considerable responsibility. In a customer service context, not disclosing the true nature of a conversational agent may, in some circumstances, be useful, particularly in situations involving seamless overlapping of human and automated customer assistants. However, nontransparency places considerable responsibility on system designers to create systems where chatbots do not provide erroneous responses that can be misunderstood by the user as adequate answers.

2. The chatbots' demeanor and lack of topical understanding is potentially challenging. Chatbots still have a long way to go to compete with humans in terms of topical understanding and demeanor. Hence, chatbots should be designed to address areas where these limitations may not be detrimental, such as in responding to large-volume routine enquiries, where the interpretation of user intent is relatively simple and the requirements for general demeanor are relatively low.
3. Chatbot responsiveness represents a key asset. Conversation flow is critical for user experience in a chat situation. Utilizing chatbots strategically to improve conversation flow and speed, for instance in customer service chats, holds substantial potential value and should be seen as a priority in the design of systems that integrate chatbots and humans in communication with users.

LIMITATIONS AND FUTURE STUDIES

The first limitation relates to sample size, which was quite small in this study. As a result, the results cannot be generalized to a larger population. Moreover, our data may not have identified effects that could be more apparent in a study with significantly more participants. However, even our small sample was sufficient to identify the substantial effect of the true nature of the conversational agent and to contrast this with the small or insignificant effect of the users' belief regarding the nature of the conversational agent. It is possible that our not finding an effect of the users' belief regarding the conversational agent was due to a small sample size; in particular, this may be so for the dependent variable affinity. Nevertheless, our study should be viewed as a starting point for applying the theory of the uncanny valley to text-based chatbots. It would be interesting to replicate the study with a larger sample.

Second, the study arguably is limited by the gender of the chatbot and the human used in the study. While some argue that that gender of a conversational agent does not interfere or influence the participants' behavior (Qu, Brinkman, Ling, Wiggers, & Heynderickx, 2014), others point out that females are often perceived as more emotional (Yang et al., 2017). This might have contributed to the notion of Ann as impersonal. Still, this does not explain why participants in Conditions 1 and 2 did not share this experience. Nonetheless, it would be useful to have future studies investigate the effect of gender perceptions in human–chatbot interaction.

Finally, our protocol dictated how the wizard should behave in terms of topics and frequency of follow-up questions. Although this might have contributed in making human–Ann more formal, we deemed it important to ensure that all the participants received the same treatment, thus allowing us to compare across conditions. In future studies, it may be helpful to include also a third kind of conversational agent: a human conversational agent not restricted by protocol, to serve as a benchmark against which the equivalents of human–Ann and chatbot–Ann may be compared.

IMPLICATIONS FOR THEORY AND APPLICATION

As chatbots continue to grow in popularity, it is essential to gain a deeper understanding and build a greater knowledge base regarding the human-chatbot interaction. Our study contributes

in this regard by applying the theory of the uncanny valley to the domain of chatbots and by investigating factors affecting users experience with communicating with a chatbot.

Surprisingly, our study failed to find support for our initial assumption that a lack of transparency in terms of a conversation agent's nature as human or chatbot may induce uncanny feelings in the users. This indicates that deceiving the users will not necessary affect users' perceived affinity and perceived pleasantness of the conversation. As such, the study findings add to the theoretical knowledge on a possible uncanny valley for chatbots; the uncanny valley effect seems unlikely to represent a threat to user experience for text-based chatbots in the foreseeable future. Future research is needed, however, on other potentially negative implications of lack of transparency regarding a conversational agent's nature, for example, implications arising from inadequate user expectations following from such lack of transparency.

Furthermore, our study details three distinct factors that may affect the user experience of human-chatbot interaction: (a) conversation content, (b) conversation demeanor, and (c) conversation flow. Although future research may uncover additional factors impacting how users experience their interaction with chatbots, knowing these three elements will help guide future research in this area. More importantly, these findings will support practitioners in their efforts to advance chatbot interaction design—advances that, in turn, may open avenues for future research on how interaction with chatbots resembles and differs from interaction with other humans.

ENDNOTES

1. The term *natural language* refers to language that has evolved as the usual way of communicating between people, such as English or Norwegian. This in contrast to artificial languages, that is, languages that have been created purposefully, such as languages for computer programming (https://dictionary.cambridge.org/dictionary/english/natural-language).
2. Mitsuku (www.mitsuku.com) and Replika (https://replika.ai/) are two socially oriented chatbots, developed to imitate human small talk. Mitsuku is available through a dedicated Web site and through messaging platforms. Replika is available as a smartphone app. Mitsuku, a key element of this study, is presented as an 18-year -old female character, developed by Steve Worswick, and is the four time winner of the Loebner Prize (https://en.wikipedia.org/wiki/Loebner_Prize).
3. Woebot (https://woebot.io/) and Wysa (https://www.wysa.io/) are chatbots aimed at helping users with mental health issues. Woebot is available through Facebook Messenger; Wysa is available as a smartphone app.
4. The online language learning programs Duolingo (https://www.duolingo.com/) and Mondly (https://www.mondly.com/) both provide chatbots for simple language practice. The Duolingo chatbot is part of the company's smartphone app for iOS. The Mondly chatbot is provided as a dedicated Android VR app.
5. The Loebner Prize is an annual event where chatbot developers compete in providing the most humanlike chatbot. The competition is modeled on the Turing Test. Judges interact with human and chatbot conversational agents through a text-based user interface. Without knowing the true nature of the conversational agents, the judges determine through their interactions the human likeness of each agent. The Loebner Prize is controversial yet has also received interest both among chatbot developers and in the general public. The competing chatbots have not been sufficiently humanlike yet to be consistently confused with a human conversational agent. Lortie and Guitton (2011) provide

more detail on the Loebner Prize. A summary of the event from the perspective of four-time winner Worswick (2018) is a relevant read.

6. Slack (https://slack.com) is a collaborative platform for professional teams, including messaging functionality where team members communicate through chat channels that may involve the whole team or smaller groups of people.
7. Although the test was conducted in English, the participants were able to complete their questionnaire free-text responses in either English or Norwegian. Twenty of the participants provided these responses in Norwegian; eight made these responses in English. Any original comments in Norwegian used in this paper were translated from Norwegian to English by the authors.

REFERENCES

Allison, F., Luger, E., & Hofmann, K. (2017). Spontaneous interactions with a virtually embodied intelligent assistant in Minecraft. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI'17; pp. 2337–2344). New York, NY, USA: ACM. https://doi.com/10.1145/3027063.3053266

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *International Conference on Internet Science* (pp. 377–392). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-70284-1_30

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, *29*(3), 759–771. https://doi.com/10.1016/j.chb.2012.11.021

Candello, H., Pinhanez, C., & Figueiredo F. (2017). Typefaces and the perception of humanness in natural language chatbots. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI'17; pp. 3476–3487). New York, NY, USA: ACM. https://doi.com/10.1145/3025453.3025919

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2000). Conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents* (pp. 29–63). Cambridge, MA, USA: The MIT Press.

Castellano, G., & Peters, C. (2010). Socially perceptive robots: Challenges and concerns. *Interaction Studies*, *11*(2), 201–207. https://doi.com/10.1075/is.11.2.04cas

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, *92*, 539–548.

Corti, K., & Gillespie, A. (2016). Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, *58*, 431–442.

Crutzen, R., Peters, G.-J. Y., Portugal, S. D., Fisser, E. M., & Grolleman, J. J. (2011). An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: An exploratory study. *Journal of Adolescent Health*, *48*(5), 514–519. https://doi.com/10.1016/j.jadohealth.2010.09.002

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies: Why and how. *Knowledge-Based Systems*, *6*(4), 258–266. https://doi.com/10.1016/0950-7051(93)90017-N

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811–817. https://doi.com/10.1017/S1351324916000243

Ezzy, D. (2002). *Qualitative analysis*. London, UK: Routledge.

Ferrara, E., Varol, O., Davis, C., Menczer F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.com/10.1145/2818717

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Los Angeles, CA, USA: Sage.

Følstad, A., & Brandtzaeg, P. B. (2018). Chatbots and the new world of HCI. *Interactions*, *24*(4), 38–42.

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In *Proceedings of International Conference on Internet Science* (INSCI 2018; pp. 194–208). Cham, Switzerland: Springer.

Geller, T. (2008). Overcoming the uncanny valley. *IEEE Computer Graphics and Applications*, *28*(4), 11–17. https://doi.com/10.1109/MCG.2008.79

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education 2003* [online]. Retrieved from https://scholarworks.iupui.edu/handle/1805/344

Hill, J., Ford, W. R., & 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.com/10.1016/j.chb.2015.02.026

Hyken, S. (2017). AI and chatbots are transforming the customer experience. *Forbes* [online]. Retrieved from https://www.forbes.com/sites/shephyken/2017/07/15/ai-and-chatbots-are-transforming-the-customer-experience

Johnson, K. (2018, May 1). Facebook Messenger passes 300,000 bots [Web log post]. Retrieved from https://venturebeat.com/2018/05/01/facebook-messenger-passes-300000-bots/

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (DIS'17; pp. 555–565). New York, NY, USA: ACM. https://doi.com/10.1145/3064663.3064672

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide: Design and evaluation of a real-world application. In *Intelligent Virtual Agents: 5th International Working Conference* (pp. 329–343). Berlin, Germany: Springer-Verlag. https://doi.com/10.1007/11550617_28

Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lortie, C. L., & Guitton, M. J. (2011). Judgment of the humanness of an interlocutor is in the eye of the beholder. *PloS ONE*, *6*(9), e25085. https://doi.com/10.1371/journal.pone.0025085

Luger, E., & Sellen, A. (2016). Like having a really bad PA: The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI'16; pp. 5286–5297). New York, NY, USA: ACM. https://doi.com/10.1145/2858036.2858288

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, *146*(1), 22–32. https://doi.com/10.1016/j.cognition.2015.09.008

Milton-Barker, A. (2016). Mitsuku chatbot wins most humanlike AI in Loebner Prize for a second time [Web log post]. Retrieved from https://www.techbubble.info/blog/artificial-intelligence/chatbots/entry/mitsuku-chatbot

Mone, G. (2016). The edge of the uncanny. *Communications of the ACM*, *59*(9), 17–19. https://doi.com/10.1145/2967977

Mori, M., Macdorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.com/10.1109/ MRA.2012.2192811

Murgia, A., Janssens, D., Demeyer, S., & Vasilescu, B. (2016). Among the machines: Human-bot interaction on social Q&A websites. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI'16; pp. 1272–1279). New York, NY, USA: ACM. https://doi.com/10.1145/2851581.2892311

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI'94; pp. 72–78). New York, NY, USA: ACM. https://doi.com/10.1145/ 191666.191703

Piccolo, L. S. G., Mensio, M., & Alani, H. (2018, October). *Chasing the chatbots: Directions for interaction and design research.* Paper presented at CONVERSATIONS 2018, St. Petersburg, Russia. Retrieved from https://conversations2018.files.wordpress.com/ 2018/10/conversations_2018_paper_10_preprint1.pdf

Poliakoff, E., Beach, N., Best, R., Howard, T., & Gowen, E. (2013). Can looking at a hand make your skin crawl? Peering into the uncanny valley for hands. *Perception*, *42*(9), 998–1000. https://doi.com/10.1068/p7569

Pollick, F. E. (2010). In search of the uncanny valley. In P. Daras & O. Mayora Ibarra (Eds.), *User Centric Media: First International Conference* (UCMedia; pp. 69–78). Berlin, Germany: Springer-Verlag. https://doi.com/10.1007/978-3-642-12630-7_8

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). Alexa is my new BFF: Social roles, user satisfaction, and personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI'17 EA; pp. 2853–2859). New York, NY, USA: ACM. https://doi.com/10.1145/3027063.3053246

Qu, C., Brinkman, W.-P., Ling, Y., Wiggers, P., & Heynderickx, I. (2014). Conversations with a virtual human: Synthetic emotions and human responses. *Computers in Human Behavior*, *34*, 58–68. https://doi.com/10.1016/j.chb.2014.01.033

Schuetzler, R. M., Grimes, M., Giboney, J. S., & Buckman, J. (2014). Facilitating natural conversational agent interactions: Lessons from a deception experiment. In *Proceedings of the International Conference on Information Systems* (ICIS'14; online). Retrieved from http://aisel.aisnet.org/icis2014/proceedings/HCI/9/

Statista (2018). *Most popular mobile messaging apps worldwide as of October 2018, based on number of monthly active users*. Retrieved from https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/

Strait, M., Vujovic, L., Floerke, V., Scheutz, M., & Urry, H. (2015). Too much humanness for human-robot interaction: Exposure to highly humanlike robots elicits aversive responding in observers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI'15; pp. 3593–3602). New York, NY, USA: ACM. https://doi.com/10.1145/27021 23.2702415

Thies, I. M., Menon, N., Magapu, S., Subramony, M., & O'Neill, J. (2017). How do you want your chatbot? An exploratory Wizard-of-Oz study with young, urban Indians. In *Proceedings of the IFIP Conference on Human-Computer Interaction* (INTERACT 2017; pp. 441–459). Cham, Switzerland: Springer.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017, May). *Online human–bot interactions: Detection, estimation, and characterization.* Paper presented at the 11th International AAI Conference on Web and Social Media (ICWSM-17), Montreal, Canada. Retrieved from https://arxiv.org/abs/1703.03107

Vtyurina, A., Savenkov, D., Agichtein, E., & La Clarke, C. (2017). Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI'17 EA; pp. 2187–2193). New York, NY, USA: ACM. https://doi.com/10.1145/3027063.30 53175

Waddell, K. (2017, April 21). Chatbots have entered the uncanny valley. *The Atlantic* [online]. Retrieved from https://www.theatlantic.com/technology/archive/2017/04/uncanny-valley-digital-assistants/523806/

Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior*, *23*(5), 2538–2557. https://doi.com/10.1016/j.chb.2006.05.002

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.com/ 10.1145/365153.365168

Worswick, S. (2018, September 13). Mitsuku wins Loebner Prize 2018! [Web log post]. Retrieved from https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI'17; pp. 3506–3510). New York, NY, USA: ACM. https://doi.com/10.1145/30254 53.3025496

Yang, Y., Ma, X., & Fung, P. (2017). Perceived emotional intelligence in virtual agents. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI'17; pp. 2255–2262). New York, NY, USA: ACM. https://doi.com/10.1145/3027063.30 53163

Authors' Note

The work presented in this paper was conducted as part of the research project Human–Chatbot Interaction Design, supported by the Research Council of Norway through Grant No. 270940 of the IKTPLUSS programme.

All correspondence should be addressed to
Marita Skjuve
SINTEF
PB 124 Blindern, 0314 Oslo, Norway
marita.skjuve@sintef.no

Human Technology
ISSN 1795-6889
www.humantechnology.jyu.fi