IADIS International Journal on WWW/Internet
Vol. 19, No. 1, pp. 17-36
ISSN: 1645-7641

UNDERSTANDING HOW CHATBOTS WORK: AN EXPLORATORY STUDY OF MENTAL MODELS IN CUSTOMER SERVICE CHATBOTS

Stine Ordemann[1], Marita Skjuve[2], Asbjørn Følstad[2] and Cato Alexander Bjørkli[1]
[1]University of Oslo, Norway
[2]SINTEF, Norway

ABSTRACT

Chatbots are changing customer service interactions, enabling higher reliance on self-serving behavior. There seems to be an emphasis on designing these chatbots to be as humanlike as possible. One drawback is that such humanlike design might lead users to apply the same mental models when interacting with chatbots as they do when interacting with human customer service personnel. Arguably, this may cause issues in the chatbot interaction because the technology may not be capable of handling interactions at the same level of sophistication as human personnel. Thus, the mental models that users apply are important for successful system interactions, but little research have been dedicated towards understanding these mental models. To close this research gap, we asked 16 users to interact with two customer service chatbots to explore the mental models guiding their interactions. Based on qualitative interviews and screen-captured videos of the participants' dialogues, an exploratory analysis indicated that the participants drew on two types of mental models to understand, predict, and interact with chatbots: a human-oriented model and a technology-oriented model. We discuss our findings and their theoretical and practical implications.

KEYWORDS

Chatbot, Mental Models, Customer Service, Social Cues

1. INTRODUCTION

Did you chat with a human or a machine the last time you sought customer service online? It is very likely that you interacted with a chatbot, a software program that utilizes natural language to answer inquiries (Dale, 2016). Humans are growing more accustomed to interacting with such agents, and this shift is likely to increase even more in the years to come (Moore, 2018). Powered by artificial intelligence (AI) and machine learning, chatbots have emerged in various fields supporting people with tasks ranging from banking inquiries (Følstad and Skjuve, 2019) to health advice (Skjuve and Brandtzæg, 2018). Combined with their cost-effectiveness and ability to operate 24/7 in unlimited parallel conversations, there is substantial incentives for businesses and organizations to implement chatbots for customer service on a broader scale (Adam et al., 2020).

As customer service shifts towards increased reliance on self-service, new roles and strains on users have emerged (Larivière et al., 2017). The adoption of any self-service technology will require new skills for the users, and new systems for self-service need to be user friendly and intuitive (Meuter et al., 2005; Parasuraman and Colby, 2015). Chatbots are often viewed as intuitive user interfaces due to their resemblance to interactions on well-known chat platforms, such as Facebook Messenger (Følstad and Brandtzæg, 2017; Jain et al., 2018). Their human likeness and design for social interaction may further contribute positively to relationship building with organizations (Araujo, 2018; Sheehan et al., 2020).

While it is argued that conversational interaction with chatbots is intuitive, it might also create unforeseeable problems. The human brain has evolved over eons to interact with other human brains, but it may not have adapted to communicate well with artificial entities (K.M. Lee, 2009). When humans communicate with other humans, they adapt reciprocally and in accordance with assumptions regarding the other's expertise in order to enhance cooperation (Johnson-Laird, 1980). For example, research has shown how children try to understand the chatbot as a human being and they assume that chatbots have an intellect that is similar to theirs (Druga et al., 2017). Adults have also been found to hold high expectations towards chatbots (Luger and Sellen, 2016) and to neglect the fact that machines have certain limitations (K.M. Lee, 2009). This can cause conversations between humans and chatbots to go astray, resulting in fallback answers from chatbots, such as: "Sorry I don't understand that question" (Druga et al., 2017, p. 598). Therefore, it is important to ask: Why does such overconfidence in chatbot capabilities occur? It might be easy to assume that the technology is not sufficiently sophisticated. While this might be the case, user behavior indicating inflated expectations concerning chatbot capabilities could be caused by the chatbot triggering the "wrong" mental model.

Mental models are understood as "the mechanism whereby humans are able to generate a description of system purpose and form, explanation of system states, and prediction of future states" (Rouse and Morris, 1986, p. 351). Mental models support users in understanding and predicting what the software will do next and modifying their own behavior accordingly. Discrepancy may occur between the designer's conceptual model of a target system, such as the interface of chatbots and its underlying features, and the users' mental model (Norman, 1983). When the user's mental model does not correspond to the target system, errors are more likely to occur or the user may use the system ineffectively (Preece et al., 2015). Therefore, a functional mental model is essential when problems arise (Staggers and Norcio, 1993).

While the importance of designing a system that corresponds to the users' mental model is well established in the literature (Endsley et al., 2003), to the best of our knowledge, no study has tried to understand the mental models that the user relies on when interacting with text-based customer service chatbots. As chatbots are increasingly implemented for self-service purposes, it is essential to understand how chatbot design may elicit appropriate mental models in users.

2. BACKGROUND

2.1 The Social Design of Chatbots

Ever since the initial conversational computer systems were developed there has been a tendency towards designing chatbots to appear as humanlike as possible. When reflecting on his pioneering program for conversational computing, ELIZA, Weizenbaum (1976) reported that he was surprised by the willingness displayed by users and the interested public in describing and considering this system in terms typically applied only to humans. Another example of the aim for human likeness in chatbots is the Loebner Prize tournament in which judges attempted to predict whether conversational partners were of human or chatbot origin (Coniam, 2014). Similarly, in the Alexa Prize competition, teams compete to design Alexa skills that enable Alexa to interact socially with users for a prolonged period of time (Ram et al., 2018).

Chatbots for customer service purposes are also designed with different attributes to enhance their human likeness and facilitate rudimentary social interaction (Nordhem et al., 2019). Such attributes may include the use of informal language, names, avatars (Araujo, 2018), and gender (McDonnell and Baxter, 2019)—all of which seem to have several positive effects. For instance, the use of avatars has been found to heighten humans' trust resilience in chatbots when conversational flows are abrupted (De Visser et al., 2016), and the use of gendered avatars has been shown to increase user satisfaction (McDonnell and Baxter, 2019). However, human likeness may also contribute to the deception of users in the sense that they engender imprecise perceptions of the nature and capabilities of the chatbots (Luger and Sellen, 2016).

The implementation of social cues influences the users' perception of chatbots and facilitates the phenomenon of anthropomorphism (Araujo, 2018), that is, attribution of "humanlike properties, characteristics, or mental states to real or imagined nonhuman agents and objects" (Epley et al., 2007, p. 865). Anthropomorphism has been found to be positively related to users' adoption of chatbots, especially for customers who seek social interaction (Sheehan et al., 2020). Likewise, humanlike cues in the chatbot language and interface can generate a feeling of social presence, which is defined by K.M. Lee (2004, p. 37) as "a psychological state in which virtual (para-authentic or artificial) social actors are experienced as actual social actors in either sensory or non-sensory ways." The participants in Araujo's (2018) study reported higher emotional connection to organizations when chatbots were able to induce such states. People can experience a sense of connection within their automated discourse; thus, chatbots might be construed by humans as something more than a mindless software system.

Nass and Moon (2000) asserted that users are explicitly aware that computers do not warrant social treatment because they are non-living entities, nor do users believe that they are communicating with the programmer. Nevertheless, users tend to engage in social actions under various conditions, and the explanation for this behavior may be grounded in social scripts that are specialized for human-human interactions.

As such, the preceding elaboration exhibits a well-grounded knowledge base about how diverse social design in chatbots can generate positive outcomes for both the users and the organizations that implement them. It also seems that chatbots are designed to generate psychological reactions in users that are similar to those that occur in human-human interactions. Arguably, this can lead to users adopting a humanlike mental model and associated social scripts for interaction guidance and applying these to understand the software. However, this may be problematic since chatbots are often not capable of handling complex interaction similar to what occurs between humans.

Even with diverse implementation of chatbots, vast technological development, and high grammatical precision, they still have issues with understanding the deeper meaning of words, which can lead to conversational breakdowns (Asktorab et al., 2019). For example, Myers et al. (2018) found that language processing errors were the most frequent obstacle when the user interacted with a voice-based calendar manager. It has also been observed that users need to ask concrete questions and have vocabulary that is congruent with the chatbots' textual content for successful dialogue (Kvale et al. 2019). Breakdowns in conversation are a common problem between humans and chatbots (Ashktorab et al., 2019) and these may be caused by the complexity of human language. Human-human dialogues are "multi-threaded, hop back and forth, and circle around" (Grudin and Jacques, 2019, p. 6). To compensate for chatbots' inability to engage in such dynamic dialogue, effective repairs can be incorporated, such as presenting a set of pre-programed alternatives where users can choose from specific options with content labels (Ashktorab et al., 2019). Thus, there seems to be a conflict where chatbots are designed to trigger humanlike mental models without having the capabilities to manage such interaction.

2.2 Users' Mental Models in Chatbots

Some researchers have reported on what may be considered to be a "gulf" between user expectations and the realities of conversational user interfaces. Luger and Sellen (2016) interviewed users of conversational agents (e.g., Siri and Google Now) and found that they adopted what seemed to be a humanlike mental model for interactional purposes that guided their communication when interacting with the system. Specially, the participants exhibited unrealistically high expectations regarding the agents' intelligence. Luger and Sellen (2016) attributed this behavior to the agents' use of social cues and natural language. Similar high expectations towards customer service chatbots (Kvale et al., 2019) and calendar managers with voice user interface (Myers et al., 2018) have also been found. In Myers et al.'s (2018) study, the users either communicated with the system in a way that the software could not understand, or they tried to execute an operation that was out of the scope of the software. The users often resorted to guessing tactics to determine what "language" the software could support, such as hyper-articulation, simplifying information, or giving too much information. This behavior was attributed to an incomplete mental model. Interestingly, feedback from the system seemed to build a more appropriate model. In contrast, Følstad and Skjuve (2019) found that user expectations were reasonably accurate for text-based customer service chatbots, and the users in their study did not expect the chatbots to have human expertise.

Individual differences may also influence the users' mental model when interacting with text-based chatbots (Liao et al., 2016). Liao et al. (2016) found that users with a high desire for social interactions viewed chatbots as being more humanlike and they interacted in a more sociably way with the chatbots, while users with a lower desire for social interactions were less likely to understand chatbots as being humanlike and more likely to view them as being software systems for gathering information. It was also found that subjects with higher technical knowledge seemed less persuaded by the social cues in conversational agents and have more suitable mental models (Luger and Sellen, 2016). This is further supported by Chen and Wang (2018) who found that technically knowledgeable subjects had a higher understanding of how the chatbots work and adapted their behavior more appropriately.

However, at this point in time, customers who use chatbots are presumably a population in which technical knowledge varies as much as other attributes, such as age, cognitive abilities, and personality. Moreover, current literature dedicated to understanding mental models in relation to using chatbots have mainly addressed personal assistants and voice-based chatbots (Chen and Wang, 2018; Luger and Sellen, 2016; Myers et al., 2018), but users have exhibited what seems to be incomplete models towards both voice-based and text-based chatbots (Kvale et al., 2019; Liao et al., 2016). However, these studies have often not directly investigated mental models.

3. THEORETICAL FRAMEWORK

3.1 The Attributes of Mental Models

There are several attributes of mental models that can facilitate problems during interaction with chatbots. Norman (1983) has argued that internal mental models can lack cohesiveness, which may contribute to interaction difficulties. Mental models may also be incomplete, prone to memory loss, and generate superstitious behavior. Nevertheless, Norman (1983) noted that users may lack in-depth technological knowledge as long as the models are functional and lead to desired outcomes. Users' mental models will help them understand and predict what the software will do next and modulate their own behavior accordingly (Rouse and Morris, 1986). For example, mental models have been found to affect performance in navigation on websites (Wagner et al., 2014) and the ability to handle novel problems in calculators (Halasz and Moran, 1983).

Mental models may not be fully accessible for explicit introspection (Rouse and Morris, 1986). Consequently, they can generate behavior that contradicts what is found through the users' explicit reasoning (Knaeuper and Rouse, 1985). Additionally, individuals can depend on several models when encountering a problem (Staggers and Norcio, 1993). For example, subjects have been found to exhibit such tendencies when reasoning about specific technical equipment (Williams et al., 1983), and different mental models may be used interchangeably. The literature on mental models also suggests that novices and experts differ in their mental models; experts may have a more accurate model of a system (Rouse and Morris, 1986). Mental models are developed by relevant experience over time (Endsley, 1995; Rouse and Morris, 1986), but naive theories may still persist even as more accurate models are available (Rouse and Morris, 1986).

When interacting with a new system, users' mental models can draw on analogies with a similar and familiar system (Staggers and Norcio, 1993). Such assumptions are utilized in graphical interface design by designing systems that support mental model development (Wickens et al., 2013), and they may be the reason for designing chatbots similar to social media applications (Jain et al., 2018).

4. THE PRESENT STUDY

Chatbots are implemented at a high rate, especially in customer service. There seems to be an emphasis on designing chatbots that are as humanlike as possible. Studies have shown how implementing social cues, such as avatars, names, and informal language in chatbots, can contribute to positive human-chatbot interaction and user adoption. Presumably, these positive outcomes will increase the human likeness of the chatbot, both intentionally or non-intentionally, and induce mental models in the user where the chatbot is likened to a human in terms of its cognitive and interactional capabilities. We refer to such mental models as being "human-oriented". However, a human-oriented mental model may lead to inflated expectations regarding the chatbot's capabilities and the user's assumptions that social interaction similar to that of human-human interaction is appropriate. Therefore, problems arise when chatbots have difficulties in understanding users' requests; this can cause dialogues to break down, which, arguable, will contribute to negative user experiences.

Based on these observations, there is a need for a deeper understanding of the mental models that are triggered when users interact with a chatbot. Such knowledge is essential if we seek to design chatbots that facilitate good user experience. However, few studies have had this as their primary research objective, so this knowledge is lacking. Therefore, the current study aimed to contribute to closing this knowledge gap by investigating the mental models that are triggered when users interact with a customer service chatbot. This paper seeks to address the following research question:

Research question: What characterizes the mental models that individuals apply during customer service chatbot interactions?

5. METHOD

5.1 Recruitment and Sample

The sample consisted of 16 participants, nine women (56%) and seven men (44%). They were recruited through a students' group on social media and posters at different institutes at the University of Oslo. Individuals with computer science education were excluded from sampling, as technical knowledge may significantly influence users' mental models in comparison to the general population (Chen and Wang, 2018).

The participants ranged in age from 21 to 47 (mean age: 27). Five participants (31%) had achieved a master's degree as their highest educational degree, seven (43%) had a bachelor's degree, and three (18%) had finished one year of higher education. The participants' educational background was diverse and some had some prior use and knowledge of chatbots. The participants' native language was Norwegian, which corresponded to the language used in the chatbots applied for the study.

5.2 Data Collection

We conducted a laboratory study in which the participants were asked to carry out a task that was developed to facilitate their interactions with two different customer service chatbots.

5.2.1 The Chatbots

Two chatbots, A and B, were used in the study to facilitate more extensive interactions during the sessions. Both chatbots were operative for customer service in retail banking. Both chatbots welcomed the users and provided a short introduction for interaction guidance. Users' requests to the chatbots were to be provided in natural language, but the chatbot interaction also included predefined answer alternatives (presented as buttons). When the chatbots were unable to answer a question correctly or interpret the user's intent, both presented the opportunity of talking to a human customer service agent, although only one of the chatbots provided this option within the same chat interface. Both chatbots were presented visually by an avatar image: Chatbot A by a female avatar and Chatbot B by a gender-neutral robot avatar.

5.2.2 Task

We developed a task to guide the interaction, with the objective of creating a realistic user experience. This has been done successfully in previous chatbot research (Chen and Wang, 2018). In the task, the participants were instructed to find general information about mortgages, an area for which both chatbots had extensive information. If the chatbots suggested that the participants visit a website for further information, they were asked to continue with the task of conversing with the chatbots. The task was fairly open-ended, with few requirements to make the participants choose an interaction style that came naturally to them.

5.2.3 Procedure and Interviews

The participants were given the task and provided a computer with access to the two chatbots. The sequence of the presentation of the chatbots was alternated between the participants to control for order effects. The participants' interactions with the chatbots were video recoded using screen-capture software. After completing their chatbot interaction, each participant was asked to report on how realistic they perceived the task to be (scores ranged from 1 to 5 on a numbered scale).

We performed a think-aloud-interview while the participants conducted the task to gain insight into their perception and thinking (Koro-Ljungberg et al., 2013). Each participant was asked at random time-intervals to verbalize their thoughts (e.g., "What do you think now?"). Statements were followed-up with either paraphrasing, e.g., "You mentioned ... can you elaborate?" or general elaboration, e.g., "Why do you think that?" (Whiting, 2008). After answering the questions, the participant continued the task until the next prompting interval. This cycle continued until the participants were out of questions to ask the chatbots. Prompting by the researcher occurred during both successful (if the chatbot was able to answer appropriately) and non-successful communication.

We conducted a semi-structured interview once the participants were finished with their task. All interviews were audio recorded. Questions in the semi-structured interviews were based on literature related to mental models. The main topics covered where the participants' perceptions of the chatbots (Endsley et al., 2003) and their associations with and metaphors to similar systems (Wickens et al., 2013). The participants were also asked about their system understanding, their prediction of the chatbots behavior (Rouse and Morris, 1986), how to behave with the chatbots (Wickens et al., 2013), and if experience had changed their understanding (Endsley, 1995). The topic generated questions such as:

1. How do you think chatbots are constructed?
2. How should you communicate with chatbots to get a desired answer?

5.3 Analysis

5.3.1 Interview analysis

All the interview recordings were transcribed. A thematic analysis was adopted to identify common themes across the transcripts. We used an inductive approach where themes were generated from the participants' statements without placing them in a specific theoretical category. We followed Braun and Clarke's (2006) steps for thematic analysis (see Braun and Clarke, 2006, for an in-depth description). We started by coding the entire dataset using NVivo version 11 (a data program for qualitative data manipulation) to code the interviews. The codes were then collapsed into broader common themes. The codes and themes were evaluated using a two-step procedure. In the first step, each code was re-read to examine the internal validity of the units placed under the code. Some of the codes were revised. We then reviewed the initial themes and the codes against the original transcription. Small adjustments were made in this second phase.

5.3.2 Chatbot Dialogue Analysis

The chatbot dialogues were transcribed and analyzed to gain insight into the participants' tendency to engage in social interaction, that is, their tendency to phrase their interaction in ways resembling what would be expected if they were interacting with a humanlike conversational partner. The first step in the analysis was to develop an appropriate coding procedure. The procedure included definitions to guide the analysis process.

When interacting with digital technology, social interaction can include a variety of different behaviors, ranging from small talk to the use of gender stereotypes. There is no clear consensus on what constitutes social interaction with text-based chatbots, but we looked for the following aspects of the dialogues when evaluating the presence of social interaction.

- **Application of social rules:** The application of social rules or markers is a useful indicator of the users' social interaction with chatbots. Therefore, we found the following to be particularly relevant: the use of first- and second-person pronouns and polite remarks. This is in line with Brennan and Ohaeri (1994), who defined an anthropomorphic sentence towards a computer agent as consisting of a first-person pronoun. They also found a higher use of second-person pronouns and polite remarks towards computers, indicating social interaction with the computer.
- **Use of complete sentences or keyword tactics:** Social interaction with chatbots can be characterized by the use of longer or more complete sentences. While it is challenging to establish rules for what constitutes a complete or longer sentence, we discriminated between sentences and keyword tactics as follows: A sentence was defined as the use of four or more words in one message, while keyword tactics was defined as the use of three or fewer single words in one message.

Once the coding procedure was established, the first author analyzed the participants'dialogues. The coding was conducted in Excel, and this process was repeated three times to check for omissions in classification. A few omissions were found in the first analysis and corrected.

6. RESULTS

The study aimed to explore the participants' mental models while using customer service chatbots. The research question was investigated by conducting interviews with the participants during and after their chatbot interactions, and through an analysis of the written dialogues between the participants and the chatbots. In this section, we will first present the results from the interviews, followed by analysis of the dialogues. The participant quotes and example sentences included in the results presentation were translated from Norwegian by the first author.

6.1 Interviews

A cluster of six themes emerged from the analysis: (1) human-oriented and technology-oriented cues, (2) user-chatbot dialogue as a source of understanding, (3) adaptations of user communication, (4) perceptions of chatbot functionality, (5) factors affecting the users' trust and (6) perceived utility.

Human-oriented and technology-oriented cues. In the interviews, the participants detailed what they perceived as cues to a humanlike character in the chatbot. Most of the participant reports were directed towards Chatbot A, often referring to the chatbot as "she" and commenting on the humanlike avatar. A few participants pointed out that they almost forgot that Chatbot A was not a human being. The participant reports clearly indicate that the use of social cues in a chatbot dialogue impact the users' understanding of a chatbot's character. Specifically, human-oriented cues were found to contribute to anthropomorphism and to the experience of social presence in the chatbot dialogues. For Chatbot A, both the humanlike avatar and humanlike language contributed to the human-oriented cues.

"She appears more as a human. Also, the way she opens (the conversation) with an emoji and 'hi'." (P 1)

The participants mentioned that their most dominant associations related to the design of the chatbots were messaging platforms, mainly Facebook Messenger. For many of them, the general theme was linking the ease of use of the chatbot to the chat interaction design metaphor.

"Most (people) are probably using Facebook chat once a day, and it is a familiar format that it is easy to use." (P 11)

While for short periods of time the participants perceived the chatbots as somewhat human, they were able to revise this understanding and they also had access to an alternative understanding: The understanding of the chatbot as a technological support system. All the participants knew that both chatbots were software systems, and they could readily generate descriptions of a technological solution enabling the chatbots—in addition to reflecting on their humanlike character. Cues to support such a technology-oriented understanding were the rapid text responses from the chatbots, their inflexible replies, and the provision of predefined alternatives. The visual presentation of the chatbot could also contribute to a technology-oriented understanding of it. Specifically, the robot avatar of Chatbot B was reported to have this effect, as well as the buttons with predefined answer alternatives that were sometimes provided in the dialogues for efficient interaction.

"When you chat with a person, they would answer with text. Not alternatives." (P 11)

User-chatbot dialogue as a source of understanding. The dialogue between the participants and the chatbots was found to provide an important foundation for the participants' understanding of the chatbot, as well as which operations the chatbot could perform. When the chatbots were able to understand the users' requests or provide relevant information that the participants enjoyed, this could strengthen the human-oriented understanding of the chatbot. Such interactions also were taken to indicate that the participants were able to interact in accordance with the system capabilities.

"Even if you don't get the correct answer, you get a lot of useful information that you may not have been aware of." (P 6)

While the participants were pleased when the chatbot was able to provide detailed answers, they also obtained irrelevant answers and experienced unhelpful repetition of information that they had already received earlier in the conversation. When this occurred, the participants became confused or assumed that their own behavior was incompatible with the system capabilities.

"I don't understand what it (the chatbot) did. I asked how much mortgage I could get with an annual income of 600 000 (KR). Then it (the chatbot) asked if this concerned a new or existing loan. Then I wrote new, and it asked if I was renewing or moving my mortgage." (P 5)

The chatbot dialogues provided an important feedback loop for the participants to evolve their comprehensions of the chatbots' capabilities. If the system output was incongruent with their expectations, revisions and adaptations of their behavior were seen as necessary.

"I need to rephrase in a way that the chatbot understands (...), and it is almost more complicated in a bot (chatbot) than in a search engine." (P 10)

Adaptations of user communication. The participants seemed to start off with an assumption and understanding that natural human language would be appropriate when interacting with the chatbots. During the interaction that shifted. The participants stated the importance of using short and concise sentences or keywords. Some had a pre-study experience of using keywords as a tactic. Thus, they comprehended the need to adjust their own specific tactics and the language requirements of the context to facilitate successful communication.

"I tried at one point to write: requirements for interest rent, and it (the chatbot) did not understand. Then I wrote: interest rate and I got it (the answer)." (P 16)

The participants also generated an understanding of more specific language demands. Spelling errors, wrong declension, compounding words, and the use of dialect would be too difficult for the chatbots to handle. Additionally, many stated they had a problem finding the accurate semantic words to access the system content. Therefore, the use of alternatives was preferred for navigation, due to memory strains or the lack of visibility of the chatbot content.

"I don't have much familiarity with this (topic), and it's difficult to ask questions. It is very helpful that it (chatbot) asks on my behalf (with use of alternatives)." (P 3)

Perceptions of chatbot functionality. The participants initially had a general implicit assumption of the chatbots being able to understand context, something that was clearly shown in the way they formed their requests to the chatbots. During the chatbot interactions, they moved towards describing the chatbot as non-adaptive and lacking in creativity, as well as having the ability to think and to engage in reciprocal communication. When asked about the chatbots' ability to learn, their assumptions were mixed. Some assumed that human involvement was needed for actual improvement to occur; others did not have an assumption about or explanation for the matter. As Chatbot A stated an ability to learn, it caused uncertainty among the participants about its capabilities for self-improvement.

"If I previously pressed (the alternative) that I am between 18 and 34, then she (Chatbot A) should know. I did not celebrate my birthday in the meantime. I am still in the same age group, and it (Chatbot A) should remember that." (P 14)

Every participant also developed a keyword hypothesis, where they assumed that the chatbots had pre-programmed answers, which had been manually entered by humans. That is, they assumed that when they sent a message to the chatbots, the keywords in the message would be detected and based on those keywords the appropriate information would be provided.

“They (chatbots) pull out keywords, and do not look at the sentence. But at the same time, that's a bit weird, because Chatbot B was specific about writing concretely. But maybe it is like that, so it is easier for it to see what's relevant.” (P 8)

Different cues in the interaction generated an expectation that the chatbots were depleted of content or had reached their “end” of the conversation thread. Furthermore, the participants had experiences that gave them negative predictions concerning satisfactory answers.

“I get a bit blind to the answers. Because I assume to get the same (information) as previously given, and I forget to read properly.” (P 10)

The participants expected that the chatbots could assist them by answering concrete and straightforward questions, providing definitions, and guiding them to the correct information site. Everything that deviated from such simple operations was anticipated to be too complicated, including more complex questions, personal inquiries, or discretionary decisions. There was a consensus that such operations would be preferred in the future.

“Right now, I think it (chatbot) can deal with the absolute simplest things and bank related questions.” (P 6)

Factors affecting the users’ trust. Many of the participants stated the possible risk of the chatbot not being able to assist them properly. This was due to communication difficulties and the lack of transparency in the chatbot.

“I fear to miss out on information (…), when I don’t get the full informational picture.” (P 12)

If the participants experienced a general lack of expertise and human qualities in the chatbots, this made them uncertain of the chatbot’s ability to assist them. In such cases, the participants assumed that humans would better handle ambiguities in communication, thereby providing better aid than chatbots. However, a few of the participants assumed that the machines were more reliable than humans, which facilitated a higher level of trust.

“They (humans) will listen to your intonation and your demeanor, what your question really is. This (chatbot) would not; they (the chatbots) will only look at what you wrote.” (P 16).

Perceived utility. The participants experienced the chatbots to be time-efficient as they provided information quickly. The chatbots were seen as being faster than human customer service personnel because they reply instantly and do not require users to wait for their turn. At the same time, the participants sometimes experienced the chatbots to be more time-consuming due to the lack of appropriate answers.

“Because I need to start searching around (for information), and then the benefit of the chatbot is reduced. Because it (the chatbot) is supposed to be quick access to information.” (P 2)

Many participants were positive about trying the chatbots again. At the same time, contact with a human operator was mentioned as being necessary either due to the complexity of the inquiries or because it was the participant’s preferred way to obtain information. Some of the participants also preferred to search the internet by themselves. Nevertheless, the participants understood the potential benefits of using chatbots for customer service; it is both economically beneficial and time-efficient for the customer support system.

“I am able to use Google. The need to then go to the home page to use a somewhat advanced search engine seems meaningless.” (P 12)

6.2 Chatbot Dialogue Analysis

The purpose of the chatbot dialogue analysis was to investigate whether and how the participants interacted socially with the chatbots in the study. The participants sent a total of 229 messages to the two chatbots, and the messages were analyzed according to the following categories: (1) use of keyword vs. complete sentences, (2) addressing the chatbot with personal pronouns, and (3) polite remarks. Table 1 presents an overview of sample statements written in the interactions with the chatbots, which are representative of the overall sample in each category.

6.2.1 Descriptive Statistics

The 16 participants spent more time with Chatbot A (M = 2.62 min, SD = 0.98 min) than Chatbot B (M = 1.74 min, SD = 0.51 min), and a little over 4 minutes collectively with both chatbots (M = 4.25 min, SD = 1.20 min). Most participants perceived the task as being representative of their natural usage with such systems (M = 4.25, SD = 1.00).

6.2.2 Prevalence of Keywords or Sentences

The participants seemed to prefer using complete sentences, such as “I want to apply for a mortgage. Can you give me information about this?”, rather than just writing one word, such as “loan” (see Table 1).

Of the 229 messages, 171 were comprised of complete sentences (four words or more). The remaining 58 reflected a keyword tactic (defined as three words or fewer). Only two participants used a keyword tactic in their interaction in more than half of their messages.

6.2.3 Prevalence of Social Rules

The use of personal pronouns was prevalent in the dataset. The participants typically referred to themselves using first-person pronouns when asking a question, such as “Do I need a permanent job to get my first mortgage?” (P 8). Of the 229 messages, 104 contained first- and second-person pronouns, and seven participants used first- and second-person pronouns in half of their messages or more.

Politeness was also used by almost all the participants. This typically included saying hello or goodbye to the chatbot and using phrases such as “Have a nice day :)” (P 13). Twenty-seven of the 229 messages contained politeness, and 11 participants displayed some form of politeness towards the chatbot.

Table 1. Written Examples of Different Language Categories

Sentences	*“What do you need to know when applying for a mortgage?”* (P 5) *“If taking out a mortgage for a residence worth 5 million, how much equity is needed”* (P 6)
Keywords	*“Information on mail”* (P 9) *“Loan”* (P 16)
Personal Pronouns	*“Do I need a permanent job to get my first mortgage?”* (P 8) *“Do I have any benefits as a first-time buyer with regard to mortgage?”* (P 2)
Polite Remarks	*“Hi, I have questions about mortgage.”* (P 11) *“Have a nice day :)”* (P 13)

Note. The example sentences were translated from Norwegian by the first author. The numbers in the parenthesis reflect the participant’s ID.

7. DISCUSSION

This study aimed to gain insight into users' mental models during interactions with chatbots for customer service. The analysis of the interview data showed that six distinct themes related to the participants' mental models emerged. Furthermore, the analysis of the chatbot dialogues showed that most of the participants used socially-oriented language in their conversation with the chatbots. We interpret these two sets of results to show that the participants may have shifted between two distinctive types of mental models: a human-oriented model and a technology-oriented model. While the participants appeared to apply both types of mental models, we also note that, in reality, the mental models that guide users' understanding and interactions with chatbots are more continuous and complex than a strict dichotomy of two models. Therefore, these two distinct models may be considered to be a theoretical simplification of a messier set of constructs applied by users.

7.1 Indications of a Human-Oriented Mental Model

Our results show how the participants made anthropomorphic statements about the chatbots. Social robots and software seem to affect human social cognition, which in turn may lead to ambiguity with regard to whether or not the object should be considered to be a social partner (van der Woerdt & Haselager, 2019). Thus, in the current study, evidence of anthropomorphism towards chatbots and the participants' perceptions of social presence is not surprising; it is also in line with existing research findings (Araujo, 2018). In chatbots, even disembodied social cues can lead to anthropomorphic interpretations. Araujo (2018) found that the use of natural language and endowing the chatbot with a name is sufficient for users to perceive and evaluate it as being humanlike. Moreover, the similarity of chatbot dialogue and user interface to that of messaging platforms may also help induce social presence (Araujo, 2018). Hence, human perception is affected, and the familiarity may ultimately prime the subject into using a mental model that is more appropriate in human-human interaction in the digital world than for interactions with automated software systems. Transferring such models are well reasoned, as they are functional in a similar situation online.

Johnson-Laird (1980) assumed that a mental model consists of knowledge about others, which will contribute to communicational adaptation towards the receiver (Brennan and Ohaeri, 1994; Johnson-Laird, 1980). Our participants seemed to expect, at first, the pragmatics and norms of natural human language in both the communication and the recipient. They also held expectations of interacting with a "smart" system. Such expectations are reasonable, as chatbot language has become so seemingly sophisticated that it can pass the Turing Test under some conditions (Warwick and Shah, 2016). Furthermore, Heyselaar and Bosse (2019) found that subjects adapt their behavior when completing a task with a text-based chatbot in comparison to conducting a task individually. Specifically, they argued that such findings indicate that users may have an implicit understanding of text-based chatbots as having mental states. Nevertheless, applying a mental model with human content when interacting with chatbots can lead to overly inflated expectations towards the chatbots (Go and Sundar, 2019; Luger and Sellen, 2016). A human-oriented mental model may consequently create a "gulf" between the users' assumptions about the system and the actual ways to efficiently interact with the system (Norman, 2013).

We also found that, in terms of obtaining the desired answers, trial-and-error was a recurring theme. It has also been shown that users of voice interfaces often communicate in a way that the software system cannot interpret or they try to execute an operation that the system cannot support (Myers et al., 2018). This has previously been attributed to incomplete mental models, and Myers et al. (2018) found tactic changes, such as guessing, simplifying, quitting, or restarting the operation. Such findings generally point towards the lack of a proper understanding and predicting how a system will respond to the users' input.

7.2 Indications of a Technology-Oriented Mental Model

When automated artifacts encounter a situation that they are not programmed to grasp, or the artifact fails to conduct a task, it becomes essential that the user understand why complications occur (Endsley et al., 2003). Our results show how several interactions and the chatbots' responses did not meet the participants' expectations, and in response to challenges, they updated their comprehension of the chatbots. The participants seemed to generate a keyword hypothesis about the underlying chatbot technology and how to operate them more as a search engine with keywords. Therefore, we argue that a different mental model was also triggered—one that contains information about how to operate the chatbot as a non-living entity that does not warrant polite or social responses. We call this a technology-oriented mental model.

The chatbots' answers were found to be an important trigger of a technology-oriented mental model. Even though there has been significant technological development in producing natural language, chatbots may still fail to correctly interpret the meaning of a message and generate an appropriate answer (Kvale et al., 2019; Myers et al., 2018; Skjuve et al., 2019). Such interactions may be an important informational cue for the user to understand and predict how the chatbot will behave. The participants also explained that information cues, such as providing alternatives, and fast answers seem to highlight the distinction between human and non-human interactions and reduced the feeling of communicating with a person. Previous research has also found that fast answers from the chatbot make them seem less humanlike (Gnewuch et al., 2018).

When this understanding was triggered, the participants seemed to switch and apply a mental model containing a more accurate understanding of the chatbots' capabilities; therefore, they lowered their expectations. We also found that they evaluated the chatbots as lacking relevant information and expertise and were not particularly useful. When users have negative experiences, this information can be integrated into a mental model. It can result in the system being considered as having low value, which can demotivate future use (Følstad and Skjuve, 2019). Luger and Sellen (2016) also pointed to the risk of insufficient use and abandoning functionalities that would otherwise be helpful for the user.

Previous research has found that perceived expertise is essential for trusting chatbots (Nordheim et al., 2019). Therefore, we believe that being primed into initially understanding the chatbots as humanlike comes with a price. When a human-oriented mental model is violated during the interaction, the limitations in the chatbots are demonstrated (J.D. Lee and See, 2004). Trust can then be affected, especially for novices without technical competencies (Luger and Sellen, 2016). The user may then perceive the system to be less predictable, as found in research on mental models and trust for other technological systems (Beggiato and Krems, 2013).

7.3 Language use in chatbot dialogues

The analysis of the chatbot dialogues showed that the participants generally preferred the use of natural language and engaged less in keyword tactics. This finding is in marked contrast to the findings from the interviews where the users were found to assume that the chatbots operated on keywords.

The analysis of chatbot dialogue also demonstrated the use of anthropomorphic communication by all the participants in many of their interactions. A few studies have previously investigated users' behavior in chatbot dialogue, all reporting similar results as the current study (Jenkins et al. 2007; Kopp et al. 2005; Liao et al., 2016). Research has found that chatbot users tend to apply polite phrases, personal pronouns, and anthropomorphic questions. Thus, our results contribute to the existing knowledge base, but they also show that such anthropomorphic behavior occurs even with the verbalized rejection of anthropomorphism towards chatbots. A participant in the current study explicated the following realization when communicating about transferring loans:

> Participants (10) writes: *"I am first interested in the conditions."*
> Chatbot B answers: *"Sorry, I do not understand your question."*
> Participants (10) articulates: "*Maybe the word 'first' is difficult for the chatbot to understand. Also, that I am 'interested' is something that chatbots don't think about."*

Nass and Moon (2000) called this *ethopoeia*, a "direct response to an entity as human while knowing that the entity does not warrant human treatment or attribution" (p. 94). Social scripts are assumed to be the underlying explanation of this phenomenon and this further indicates that the participants used a human-oriented mental model in their interaction.

The explanation can also be related to the concept of overlearned social behaviors, which are so ingrained and automatic that they occur without conscious attention (Nass and Moon, 2000). Chatbots may trigger overlearned responses due to digitalization of human communication. The users may have difficulties with suppressing habitual behavior in this new interaction, even when the situation requires the users to change an action (Endsley, 1995). The use of a human-oriented mental model may be the cause of many of the observed interactions between humans and chatbots that are suboptimal and breakdowns in conversations. Would framing the chatbots more as a search engine (and less socially) generate more congruent user behavior towards the chatbots?

Nevertheless, the current results should be viewed with some caution. Our findings do not mean that communicational acts with chatbots are identical to communication with humans for most of the adult population. This is supported by Hill et al. (2015), who has shown that human-chatbot communication is qualitatively (e.g., more pronouns, swear words) and quantitatively (e.g., more words and massages) different in comparison to human-human dialogues. Therefore, a nuanced and plausible explanation is that people are generally aware of these distinctions.

7.4 Implications

7.4.1 Theoretical Implications

While the findings from the interviews and chatbot dialogue analyses corroborate existing knowledge on human-chatbot interaction, this study also contributes new insights. First, it provides new knowledge on mental models in chatbot interactions by demonstrating how users seem to utilize two different mental models for understanding the chatbot and predicting its abilities: a human-oriented model that is similar to the mental models applied in human-human interaction and a technology-oriented model specific to the assumed technological underpinnings of human-chatbot interactions. The present study also identified a contradiction between the expressed communicational tactic of using keywords and the actual behavior, which consisted of using more natural language. This indicates that the human-oriented mental model, and associated scripts or overlearned behavior, become the dominant interaction style for users even when they have an explicit awareness of the chatbot's as a software system.

7.4.2 Practical Implications

The overall findings indicate that the current design of chatbots for customer service may underutilize perceptual information (e.g., how to communicate) or emphasize suboptimal informational pieces (e.g., social cues). The findings from the present study suggest that users should be made aware of limitations in chatbot capabilities before a message is sent to the chatbots, rather than getting feedback after a message is sent, which is the current strategy in many text-based chatbot. Such up-front information may modulate the negative user experience that can follow from breakdowns in dialogue. It will also give users an indication and a comprehension that their behavior is a (mis)match with system functionality.

On the basis of our findings, we note specific opportunities for improvement, in particular with regard to issues in chatbot prediction capabilities. If a user's text does not match the specific textual content in the chatbot or has a high probability of generating breakdowns in dialogue, a notification could be provided. For example, this could be provided visually in the user's typing window where perceptual attention is fixated. Perhaps a red line under a specific word, as many individuals are familiar with such system feedback in spelling corrections applications (e.g., Microsoft Word, SMS, or Facebook Messenger). This type of notification could also be provided in expressions of uncertainty in the textual response of the chatbot, as suggested by Ashktorab et al. (2019). Presumably, this feedback will modulate the users' understanding of and expectations towards the chatbots, as suggestions will indicate how the users should interact. This feedback could possibly be more effective than initial introductions of communication guidance typically provided in current chatbots for customer service because such guidance is not immediately available throughout the chatbot dialogue.

7.5 Limitations and Future Studies

This study has several limitations. First, the sample consisted of participants with a higher educational background. Variations in cognitive abilities, age, technical skills, and relevant knowledge may affect the user's mental model. Therefore, there is uncertainty regarding the representativeness of the sample, which might affect the generalizability of the findings beyond the study context. Future research should investigate the impact of such individual variations.

Second, the sample size can also be criticized for being small. However, it is reported that nine participants are enough to generate coding saturation, and 16 to 25 subjects are necessary to generate meaning saturation, where higher in-depth information from codes is achieved (Hennink et al., 2017). Therefore, the current sample size of 16 participants was deemed sufficient, but a larger sample size may have contributed to additional meaning saturation in several sub-themes.

Lastly, Chatbot A and Chatbot B had different levels of social cues. This may have affected the overall results by making the experience of one chatbot affect how the other is interpreted, where Chatbot B could be evaluated as being more humanlike due to priming from Chatbot A. To reduce the risk of skewing the results based on priming, the two chatbots were presented in an alternated order. In principle, this should correct for any systematic order effects that potentially could have biased the participants' mental models. Future studies should experimentally examine how different levels of social cues, chatbots' general language (anthropomorphic or keywords), introduction text, and framing of the chatbots effect a user's mental model.

8. CONCLUSION

The present study identified six themes of relevance for users' mental models when interacting with a chatbot. The findings indicated the use of human-oriented and technology-oriented mental models to describe, explain, and predict customer service chatbots. A dialogue analysis of the dialogue between the participants and the chatbots showed that conversational behavior was guided by a human-oriented mental model in many of their interactions. This occurred even if the participants explicitly rejected the chatbot as a social partner.

REFERENCES

Adam, M. et al. 2020. AI-based Chatbots in Customer Service and their Effects on User Compliance. *Electronic Markets*, 1-19. doi:10.1007/s12525-020-00414-7

Araujo, T. 2018. Living Up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions. *Computers in Human Behavior, 85*, 183-189. doi:10.1016/j.chb.2018.03.051

Ashktorab, Z. et al. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, SCT, pp. 1-12.

Beggiato, M., and Krems, J. F. 2013. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour*, Vol 18, pp. 47-57. doi:10.1016/j.trf.2012.12.006

Braun, V. and Clarke, V. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology,* Vol. 3, No. 2, pp. 77-101. doi:10.1191/1478088706qp063oa

Brennan, S.E. and Ohaeri, J. O. 1994. Effects of Message Style on Users' Attributions toward Agents. *Conference Companion on Human factors in Computing Systems.* Boston, MA, pp. 281-282.

Chen, M.L. and Wang, H.-C. 2018. How Personal Experience and Technical Knowledge Affect Using Conversational Agents. *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. Tokyo, Japan.

Coniam, D. 2014. The Linguistic Accuracy of Chatbots: Usability from an ESL Perspective. *Text & Talk,* Vol. 34, No. 5, pp. 545-567. doi:10.1515/text-2014-001

Dale, R. 2016. The return of the chatbots. *Natural Language Engineering,* Vol. 22, No. 5, pp. 811-817. doi:10.1017/S1351324916000243

De Visser, E.J. et al. 2016. Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *Journal of Experimental Psychology: Applied,* Vol. 22, No. 3, p. 331. doi:10.1037/xap0000092

Druga, S. et al. 2017. "Hey Google Is It OK if I Eat You?" Initial Explorations in Child-Agent Interaction. *Proceedings of the 2017 Conference on Interaction Design and Children*. Stanford, CA, USA, pp. 595-600.

Endsley, M.R. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors,* Vol. 37, No. 1, pp. 32-64. doi:10.1518/001872095779049543

Endsley, M.R. et al. 2003. *Designing for Situation Awareness: An Approach to User-centered Design.* CRC Oress, Boca Raton, FL.

Epley, N. et al. 2007. On Seeing Human: A Three-factor Theory of Anthropomorphism. *Psychological Review,* Vol. 114, No. 4, p. 864. doi:10.1037/0033-295X.114.4.864

Følstad, A. and Brandtzæg, P.B. 2017. Chatbots and the New World of HCI. *Interactions*, Vol. 24, No. 4, pp. 38-42. doi:10.1145/3085558

Følstad, A. and Skjuve, M. 2019. Chatbots for Customer Service: User Experience and Motivation. *Proceedings of the 1st International Conference on Conversational User Interfaces*. Dublin, Ireland.

Gnewuch, U. et al. 2018. Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-chatbot Interaction. *Research Papers*, Vol. 113, pp. 1-17.

Go, E. and Sundar, S.S. 2019. Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions. *Computers in Human Behavior*, Vol. 97, pp. 304-316.

Grudin, J. and Jacques, R. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland, pp. 1-11.

Halasz, F.G. and Moran, T.P. 1983, December. Mental Models and Problem Solving in Using a Calculator. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston, MA, pp. 212-216.

Hennink, M.M. et al.. 2017. Code Saturation versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health research*, Vol. 27, No. 4, pp. 591-608. doi:10.1177/1049732316665344

Heyselaar, E., and Bosse, T. 2019. Using Theory of Mind to Assess Users' Sense of Agency in Social Chatbots. Paper presented at the *International Workshop on Chatbot Research and Design*. Amsterdam, the Netherlands, pp. 158-169.

Hill, J. et al. 2015. Real Conversations with Artificial Intelligence: A Comparison between Human–human Online Conversations and Human–chatbot Conversations. *Computers in Human Behavior*, Vol. 49, pp. 245-250. doi:10.1016/j.chb.2015.02.026

Jain, M. et al. 2018. Convey: Exploring the Use of a Context View for Chatbots. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal, Canada, pp. 1-6.

Jenkins, M.C. et al. 2007. Analysis of User Interaction with Service Oriented Chatbot Systems. Paper presented at the *International Conference on Human-Computer Interaction*, Beijing, China, pp. 76-83.

Johnson-Laird, P.N. 1980. Mental Models in Cognitive Science. *Cognitive Science*, Vol. 4, No. 1, pp. 71-115. doi:10.1016/S0364-0213(81)80005-5

Knaeuper, A. and Rouse, W.B. 1985. A Rule-based Model of Human Problem-solving Behavior in Dynamic Environments. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 15, No. 6, pp. 708-719. doi:10.1109/Tsmc.1985.6313454

Kopp, S. et al. 2005. A Conversational Agent as Museum Guide–design and Evaluation of a Real-world Application. Paper presented at the *International Workshop on Intelligent Virtual Agents.* Kos, Greece, pp. 329-343.

Koro-Ljungberg, M. et al. 2013. Reconceptualizing and Decentering Think-aloud Methodology in Qualitative Research. *Qualitative Research*, Vol. 13, No. 6, pp. 735-753. doi:10.1177/1468794112455040

Kvale, K. et al. 2019. Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues. Paper presented at the *International Workshop on Chatbot Research and Design*.Amsterdam, the Netherlands, pp. 187-200.

Larivière, B. Et al. 2017. "Service Encounter 2.0": An investigation into the roles of technology, employees and customers. Journal of Business Research, Vol 79, pp. 238-246. doi:10.1016/j.jbusres.2017.03.008

Lee, J.D. and See, K.A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors,* Vol. 46, No. 1, pp. 50-80. doi:10.1518/hfes.46.1.50_30392

Lee, K.M. 2004. Presence, Explicated. *Communication Theory*, Vol. 14, No. 1, pp. 27-50. doi:10.1111/j.1468-2885.2004.tb00302.x

Lee, K.M. 2009. Presence Theory. In S. W. Littlejohn & K. Foss, A (Eds.), *Encyclopedia of Communication Theory* (pp. 794-796). Retrieved from https://sk-sagepub-com.ezproxy.uio.no/reference/communicationtheory/n301.xml

Liao, Q. V. et al. 2016. What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems.* Brisbane, Australia, pp. 264-275.

Luger, E. and Sellen, A. 2016. " Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. San Jose, CA, pp. 5286-5297.

McDonnell, M. and Baxter, D. 2019. Chatbots and Gender Stereotyping. *Interacting with Computers*, Vol. 31, No. 2, pp. 116-121. doi:10.1093/iwc/iwz007

Meuter, M.L. et al. 2005. Choosing Among Alternative Service Delivery Modes: An Investigation of Customer Trial of Self-service Technologies. *Journal of Marketing,* Vol. 69, No. 2, pp. 61-83. doi:10.1509/jmkg.69.2.61.60759

Moore, S. 2018. Gartner says 25 percent of customer service operations will use virtual customer assistants by 2020. Article on Gartner. com. Available online at https://www. gartner. com/en/newsroom/press-releases/2018-02-19-gartner-says-25-percent-of-customer-service-operations-will-use-virtualcustomer-assistants-by-2020.

Myers, C. et al. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montréal, Quebec, pp. 1-6.

Nass, C. and Moon, Y. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues,* Vol. 56, No. 1, pp. 81-103. doi:10.1111/0022-4537.00153

Nordheim, C.B. et al. 2019. An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study. *Interacting with Computers,* Vol. 31, No. 3, pp. 317-335. doi:10.1093/iwc/iwz022

Norman, D.A. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, New York, NY.

Norman, D.A. 1983. Some Observations on Mental Models. In D. Gentner & A. L. Stevens (Eds.), *Mental Models* (pp. 7-14). Retrieved from https://www-taylorfrancis-com.ezproxy.uio.no/books/e/9781315802725

Parasuraman, A. and Colby, C.L. 2015. An Updated and Streamlined Technology Readiness Index: TRI 2.0. *Journal of Service Research*, Vol. 18, No. 1, pp. 59-74. doi:10.1177/1094670514539730

Preece, J. et al. 2015. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons Inc., United Kingdom.

Ram, A. et al. 2018. "Conversational ai: The science behind the alexa prize."arXiv preprint arXiv:1801.03604-

Rouse, W.B. and Morris, N.M. 1986. On Looking into the Black Box: Prospects and Limits in the Search for Mental Models. *Psychological Bulletin,* Vol. 100, No. 3, p. 349. doi:10.1037/0033-2909.100.3.349

Sheehan, B. et al. 2020. Customer Service Chatbots: Anthropomorphism and Adoption. *Journal of Business Research*, Vol. 115, pp. 14-24. doi:10.1016/j.jbusres.2020.04.030

Skjuve, M. and Brandtzæg, P.B. 2018. Chatbots as a New User Interface for Providing Health Information to Young People. In Y. Andersson, U. Dahlquist, & J. Ohlsson (Eds.), *Youth and News in a Digital Media Environment–Nordic-Baltic Perspectives*. Retrieved from https://www.nordicom.gu.se/sv/system/tdf/publikationer-hela-pdf/youth_and_news_in_a_digital_media_environment.pdf?file=1&type=node&id=39917&force=0

Skjuve, M. et al. 2019. Help! Is My Chatbot Falling into the Uncanny Valley? An Empirical Study of User Experience in Human-Chatbot Interaction. *Human Technology,* Vol. 15, No. 1. doi:10.17011/ht/urn.201902201607

Staggers, N. and Norcio, A.F. 1993. Mental Models: Concepts for Human-computer Interaction Research. *International Journal of Man-Machine Studies*, Vol. 38, No. 4, pp. 587-605. doi:10.1006/imms.1993.1028

van der Woerdt, S. and Haselager, P. 2019. When Robots Appear to Have a Mind: The Human Perception of Machine Agency and Responsibility. *New Ideas in Psychology*, Vol. 54, pp. 93-100.

Wagner, N. et al. 2014. The Impact of Age on Website Usability. *Computers in Human Behavior*, Vol. 37, pp. 270-282. doi:10.1016/j.chb.2014.05.003

Warwick, K. and Shah, H. 2016. Can Machines Think? A Report on Turing Test Experiments at the Royal Society. *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 28, No. 6, pp. 989-1007. doi:10.1080/0952813X.2015.1055826

Weizenbaum, J. 1976. *Computing Power and Human Reason: From Judgemnent to Calculation.* W. H. Freeman and Company., New York, NY.

Whiting, L.S. 2008. Semi-structured Interviews: Guidance for Novice Researchers. *Nursing Standard*, Vol. 22, No. 23, pp. 35-41. doi:10.7748/ns2008.02.22.23.35.c6420

Wickens, C.D. et al. 2013. *Introduction to Human Factors Engineering: Pearson New International Edition*. Pearson Higher Ed.

Williams, M. D. et al. 1983. Human Reasoning About a Simple Physical System. In D. Gentner & A. L. Stevens (Eds.), *Mental Models* (pp. 131-153). Retrieved from https://www-taylorfrancis-com.ezproxy.uio.no/books/e/9781315802725