"Understanding How Chatbots Work"

An Exploratory Study of Mental Models in Customer Service Chatbots

Stine Ordemann

Master Thesis at the Department of Psychology

UNIVERSITY OF OSLO

June 2020

2020

"Understanding How Chatbots Work":
An Exploratory Study of Mental Models in Customer Service Chatbots

Stine Ordemann

http://www.duo.uio.no/

Abstract

Chatbots are changing customer service interactions, generating a higher reliance on self-serving behavior. User's mental model is acknowledged as important for successful system interactions but have received limited attention in chatbot research. We asked an adult population (n = 16) of students and non-students to interact with two customer service chatbots to explore their mental models. Based on qualitative interviews and screen-captured videos of participant's dialogues, the exploratory analysis indicated the use of two separate mental models to understand, predict, and interact with chatbots. One humanlike model and one model containing chatbot knowledge. We further wanted to explore if the situational awareness framework could provide additional understanding of the eight emerging themes due to theoretical reciprocity between mental models and situational awareness. Six out of eight main themes from the exploratory analysis were relevant in the framework, indicating that Situational Awareness can be used in a chatbot context for practical design purposes.

Keywords: *Chatbots, Human-Chatbot Interaction, User Mental Model, Situational Awareness*

Acknowledgment

I want to thank my supervisors and SINTEF for letting me explore the topic of psychology and technology. I also want to thank the participants for dedicating time and providing insight into their thinking.

More specifically, I want to thank my supervisor Cato Bjørkli for advice and guidance when my "situational awareness" for this thesis was astray. You have also given me great interest and motivation to learn about the Human Factors field in the last two years, which I am ever grateful for. Asbjørn Følstad, thank you for all of your assistance. Some parts of this thesis would not have been achievable without your insight. Most importantly, I want to thank my supervisor Marita Skjuve, for all the help and advice you have given me the last year (even outside regular work hours). I will be forever grateful for your involvement, especially in the final weeks when everything seemed so distorted.

To my family and friends, thank you for being patient with me as my time has been absorbed in this work. To Bjørn H., this thesis would not have been possible without your support and continuous input. I will return the favor on your coming adventures.

Oslo, June 2020
Stine Ordemann

Table of Content

"Understanding How Chatbots Work"

Did you chat with a human or a machine the last time you sought customer support online? It is very likely that you interacted with a chatbot, a software program that utilizes natural language to answer your inquiries. Humans are growing more accustomed to interacting with such agents, and this shift is likely to increase even more in the years to come (Gartner, 2018). Powered by Artificial Intelligence (AI) and machine learning, these chatbots has emerged in various field supporting people with tasks ranging from banking inquiries (Følstad & Skjuve, 2019) to health advice (Skjuve & Brandtzæg, 2018). Combined with their cost-effectiveness and ability to operate 24/7 in unlimited parallel conversations, this creates an incentive for businesses and organizations to implement them on a broader scale (Adam, Wessel, & Benlian, 2020).

As customer service changes towards a more self-serving model, it generates new roles and strains on the users (Larivière et al., 2017). The adoption of such technology will require new skills for the costumer, and the systems need to be user friendly and intuitive (Meuter, Bitner, Ostrom, & Brown, 2005; Parasuraman & Colby, 2015). However, chatbots can be viewed as intuitive due to their resemblance to human communication that are digitalized through social media (Følstad & Brandtzæg, 2017; Jain, Kota, Kumar, & Patel, 2018). Their human likeness and social design further contribute positively to relationship building with organizations rather than the experience of impersonal encounters (Araujo, 2018; Sheehan, Jin, & Gottlieb, 2020).

While human brains have evolved over eons to interact with other human brains, it may not have adapted to communicate well with artificial entities (Lee, 2009). When humans communicate with other humans, they adapt in reciprocal ways and have theories of the others' expertise in order to enhance cooperation (Johnson-Laird, 1980). Children have been observed in trying to understand the chatbot as a human being, and assumed that chatbots had similar intellect as themselves (Druga, Williams, Breazeal, & Resnick, 2017). Adults have also been found to have high expectations towards chatbots (Luger & Sellen, 2016) and neglecting the fact that machines have certain limitations (Lee, 2009). Therefore, dialogues often go astray and users receive uncomprehending answers form the chatbots such as, "Sorry I don't understand that question" (Druga et al., 2017, p. 598).

When chatbots use social cues and natural language, it may cause an individual to use mental models of human beings (Luger & Sellen, 2016). In the Human Factor and Human-Computer Community, mental models can be understood as cognitive constructs that guide

the user's understanding of the system, and how to operate it (Preece, Rogers, & Sharp, 2015; Wickens, Lee, Liu, & Gordon-Becker, 2013). Discrepancy often occurs between the designer's conceptual model of a target system, such as the interface of chatbots and its underlying features, and users' internal mental model (Norman, 1983). When the user's mental model is not corresponding with the target system, errors are more likely to occur, or the user may use the system ineffectively (Preece et al., 2015). A functional internal model is therefore especially important when problems transpire (Staggers & Norcio, 1993). While the importance of designing a system that corresponds to the users' mental model is well established within the literature (Endsley, Bolte & Jones, 2003), no study to our knowledge 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-serving purposes and handle costumer requests, it is essential that we understand how to design for an appropriate mental model in users. We want to contribute to bridge such knowledge gaps by taking a direct approach to investigate user's mental models.

Based on this motivation, relevant background knowledge will (1) first explore chatbots and their corresponding interface design, before elaborating on the findings that exist regarding user's mental models and chatbots. (2) Thereafter, relevant background knowledge will elaborate on the attributes of mental models and its relationship to Situational Awareness (SA). Both concepts may contribute to a deeper understanding of the user in a chatbot context.

The Emergence of Chatbots

In the 1950s, an academic revolution within fields such as cognitive psychology, linguistics, and computer science transpired and laid the foundation for the emergence of the current chatbot technology (Miller, 2003). The development of artificial intelligence (AI) in particular generated the possibility of a software system to exhibit intelligent behavior, such as perceiving, reasoning, and behaving competently (Tørresen, 2013). Not long after, the first chatbot were built. In 1966, Weizenbaum created ELIZA, an agent that imitated a therapeutic conversation with a human. Since ELIZA, chatbots have emerged in various domains with diverse attributes. Some chatbots communicate with voice or text-based dialogues (Dale, 2016), and some text-based chatbots also provide options or alternatives for navigational purposes (Ashktorab, Jain, Liao, & Weisz, 2019) Chatbots are also referred to by many different names such as, cognitive agents (De Visser et al., 2016), chatterbots and conversational agents (Skjuve, Haugstveit, Følstad, & Brandtzaeg, 2019). We will refer to

them as chatbots, a software system "where users speak and listen to an interface" (Preece et al., 2015, p. 194).

Chatbots are additional made for different purposes (Følstad, Skjuve & Brandtzaeg, 2019). Personal assistants provide daily assistance in turning on different devices or finding information. Coaches help the user with a specific task such as learning a new language or provide psychological therapy. Customer service chatbots are made to replace human-human services online and provide automated help to costumers. In content duration, chatbots are used to generate information like weather forecasts, news or tweets. Due to diverse attributes and purposes, Følstad et al. (2019) argue that chatbots can be classified along the dimensions: *duration of relationship* and *locus of control*. Along the first dimension, *duration of relationship*, chatbots can differ based on the temporal extent of their relationship. Personal assistants and coaches are developed to engage users in an ongoing relationship over time, whereas customer service chatbots and content duration are designed for short-term relationships. In the latter users are treated as strangers in every interaction. Along the second dimension, chatbots differ in their *locus of control*. Personal assistants and customer service chatbots let the user have a high locus, meaning that users drive the conversation. On the other hand, coaches and content duration exhibit higher control over the dialogue, not giving the user freedom to go outside a prefixed script (Følstad et al., 2019).

Even with diverse implementation of chatbots, vast technological development and high grammatic precision, they still have issues with understanding the deeper meaning of words which can contribute to meaningless answers (Coniam, 2014). For instance, a related finding regarding voice-based calendar manager, was that natural language processing errors were the most frequent obstacle (52%) in user interaction (Myers, Furqan, Nebolsky, Caro, & Zhu, 2018). Also, it is observed that users needed to ask concrete questions and have congruent vocabulary to chatbots textual content for successful dialogue (Kvale, Sell, Hodnebrog, & Følstad, 2019). Breakdowns in conversation are a common problem between humans and chatbots (Ashktorab et al., 2019) and may be caused by the complexity of human language. Human-human dialogues are "multi-threaded, hop back and forth, and circle around" (Grudin & 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).

However, the chatbot community have increasingly focused on the incorporation of social cues to ensure successful implementation and user adoption of chatbots (Sheehan et al., 2020). The next section will highlight the positive effect of such designs.

The Social Design of Chatbots

There is a general trend in the chatbot community to strive towards creating the most humanlike agent as possible. In an annual tournament, Loebner Prize Turing Test, chatbots are evaluated on how they have evolved since ELIZA (Coniam, 2014), where a judge decide if their conversational partner is of human or chatbot origin. An interesting observation from this tournament is that both humans and chatbots have been mistakenly evaluated to be the opposite (Lortie & Guitton, 2011).

Chatbots are designed with many different attributes to enhance their human likeness and this may contribute to the deception of humans. They use informal language, names, avatars (Araujo, 2018) and gender (McDonnell & Baxter, 2019). Research has discovered several positive effects from the use of humanlike cues in chatbots. Avatars have been found to heighten trust resilience in chatbots when conversational flows were abrupted (De Visser et al., 2016). Some chatbots also have gender in their avatars, and the inclusion of gender has been shown to increase user satisfaction (McDonnell & Baxter, 2019). Beattie, Edwards, and Edwards (2020) found that the use of emojis in the dialogue generated higher ratings of message credibility, social attraction, and chatbot competency in performing a task. Humorous output have similarly been shown to contribute to increase users motivation for exploring chatbots functionalities (Luger & Sellen, 2016).

Implementations of social cues influence user's thinking towards chatbots, such as the phenomena of anthropomorphism (Araujo, 2018). Anthropomorphism refers to "attributing humanlike properties, characteristics, or mental states to real or imagined nonhuman agents and objects" (Epley, Waytz, & Cacioppo, 2007, p. 865). Anthropomorphism is found to positively relate to user 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. Social presence is defined 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" (Lee, 2004, p. 37). 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, and chatbots are thereby construed by humans as something more than mindless software system.

Interestingly, Beattie et al. (2020) found that impressions of chatbots and humans were rated comparably with regard to message credibility, social attraction and chatbot competency in performing a task. It was argued that "people seemed to evaluate chatbots like other people" (p. 12). In Computers as Social Agents (CASA)-paradigm, small manipulations such as framing the computer as a teammate, produced an ingroup effect of higher reciprocal cooperation towards computers (Nass, Fogg, & Moon, 1996). It has also been shown to cause a tendency for gender stereotyping based on gender cues in the computer (Nass, Moon, & Green, 1997), and induce polite behavior towards computers (Nass, Moon, & Carney, 1999). Nass and Moon (2000) assert that users are explicitly aware that computers do not warrant social treatment because they are non-living entities. Nevertheless, users tend to engage in such social acts under various conditions and may be grounded in scripts that are specialized for human-human interactions.

The preceding elaboration exhibits a well-grounded knowledgebase about how diverse social design in chatbots can generate positive outcomes for both the users and organizations who implements them. It also seems that chatbots are designed to generate similar psychological reactions in users as those that occur when humans interact with other humans. It can lead to users adopting a humanlike mental model and associated social scripts for interaction guidance and applying these to understand the software.

Users Mental Models in Chatbots

The literature describes a "gulf" between the user and personal assistant chatbots. Luger and Sellen (2016) interviewed regular chatbot users and found that they adopted a humanlike mental model for interactional purposes, which guided their communication. They also exhibited unrealistically high expectations regarding the chatbots intelligence. Over time, users changed their perception of chatbots into viewing them as less intelligent and this led to abandoning functionalities in the chatbot and only using them for simple purposes. E.g., setting on a timer rather than use chatbots to make a call during multitasking activities. There is also found similar high expectations towards customer service chatbots (Kvale et al., 2019) and calendar managers with voice user interface (Myers et al., 2018). In Myers et al. (2018) study, users either communicated with the system in a way that the software could not interpret or tried to execute an operation that were out of the software scope. Individuals often resorted to guessing tactics to figure out what "language" the software could support such as

hyper-articulation, simplifying- or giving too much information. Such behavior was attributed to an incomplete mental model. Yet, 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 users in their study did not expect the chatbots to have human expertise.

It is also found that users may differ in their mental model content regarding text-based chatbots. In a field study spanning 17 days, subjects were instructed to use chatbots and measured afterwards with regards to social-agent orientation (desire for social interactions). Their dialogues were also examined. Eight of these users were further interviewed post-interaction concerning their mental models (Liao, Davis, Geyer, Muller, & Shami, 2016). If users scored high on social-agent orientation, chatbots were viewed in a more humanlike lens and users interacted more socially towards the chatbots, such as using polite phrases. If users had a lower social-agent orientation, chatbots were understood as less humanlike and more concretely as software systems for gathering information. These individuals tended to lack conceptualization of which operations the chatbots could perform, indicating the need for affordance in the design (Liao et al., 2016). Affordance in design is to visually exhibit and signal the possibilities of actions that users can perform (Norman, 1999). Such lack of conceptualization were not found among users with high social-agent orientation (Liao et al., 2016).

It is also found that subjects with higher technical knowledge seemed less persuaded by the social cues in chatbots and have more suitable mental models (Luger & Sellen, 2016). This is further supported by Chen and Wang (2018) who found that technically knowledgeable subjects had higher understanding of how the chatbots worked and ability to adapt their behavior accordingly. The amount of use was also positively associated with the perceived usefulness of the chatbots. However, customers who use chatbots at this point in time are presumably a population in which technical knowledge varies as much as other attributes like age, cognitive abilities and personality.

In summary, the presented studies on chatbots and mental models have primarily been directed towards personal assistants and voice-based chatbots (Chen & Wang, 2018; Luger & Sellen, 2016; Myers et al., 2018). Users have exhibited incomplete models towards both voice and text-based chatbot (Kvale et al., 2019; Liao et al., 2016), with the exception of the Følstad and Skjuve (2019) study. There has also been shown different models towards text-based chatbots. However, mental models' interviews are often conducted after actual use with a considerable time-lag or assessed as secondary findings to the researcher's primary aims. In

some instances, relevant findings are not discussed in light of mental models at all. The present study will therefore take a more direct approach by focusing specifically on the nature of mental models. The next section will elaborate on the attributes of mental models and how to study them.

The Attributes of Mental Models

Norman (1983) argue 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 (Norman, 1983). Nevertheless, Norman (1983) argue that users may well lack in-depth technological knowledge as long as the models are functional and lead to desired outcomes. In the human factor community, the most cited definition of mental models is given by Rouse and Morris (1986). They define mental models 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 & Morris, 1986, p. 351). Such models will help the individual understand and predict what the software will do next and modulate their own behavior accordingly. It is found to affect older adults' performance in navigation on websites (Wagner, Hassanein, & Head, 2014), and the ability to handle novel problems in calculators have also been shown to be affected my mental models (Halasz & Moran, 1983).

Mental models are described with a range of different characteristics. They can be implicit, multiple, differ between experts-novices and rely on analogies. Due to the lack of full accessibility to the model from explicit introspection (Rouse & Morris, 1986), it can generate behavior that contradict what is found in explicit reasoning (Knaeuper & Rouse, 1985). Individual can additionally depend on several models when encountering a problem (Staggers & Norcio, 1993). For example, subjects exhibited such tendencies when reasoning about heat exchangers (Williams, Hollan, & Stevens, 1983), and it seemed like models were constantly switched and used interchangeably. As with recent findings on chatbots and technical knowledge, the literature on mental models also assumes that novices and experts differ. Experts may have more accurate mental models of the system (Rouse & Morris, 1986). Their models are developed by relevant experience over time (Endsley, 1995b; Rouse & Morris, 1986), but naive theories may still persist even as more accurate models are available (Rouse & Morris, 1986).

When interacting with a new system, mental models can draw on analogies with a similar and familiar system (Staggers & Norcio, 1993). Such assumptions are utilized in

graphical interface design by designing systems that support mental model development (Wickens et al., 2013), and may be the reason for designing chatbots similar to social media applications (Jain et al., 2018). However, metaphors in mental models can generate unexpected outcomes. A classical study by Kempton (1986) illustrates this nicely. It was found that subjects drew on gas burner models when operating residential heaters. The subjects assumed that setting the heater at a higher level would accelerate heat flow. Nevertheless, the actual system action will cause a longer period of operating before reaching desired state.

The ultimate goal that could be attained by having insight into the subject's mental models is the enhancement of system design (Rouse & Morris, 1986). However, models are argued to be challenging to measure. Nonetheless, insight can be gathered trough indirect and direct methods. (1) In indirect methods, interference methods can be used, by manipulating variables to see their effect on behavior and assume that models cause a difference in behavior. (2) In direct measures, think-aloud protocol, interviews, or questionnaire methods can be used. They can provide knowledge beyond what is found from interference by providing knowledge of subjects actual thinking. Nevertheless, each method has several drawbacks, and it may be impossible to have a full overview of the content of a given mental model. Yet, generating knowledge about the content of mental models is still highly important, despite these methodological issues (Rouse & Morris, 1986). Endsley (2000) and Zhang, Kaber, and Hsiang (2010) propose a different strategy to generate data about mental model content. As SA and mental models are thought to interact and mutually influence each other, SA methods can be used to generate further insight.

Theory of Situational Awareness

The most cited definition of SA is coined by Endsley (1995b), where she describes SA as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (p 36). According to the model, SA can be separated into three different levels: perception (level 1), comprehension (level 2) and projection (level 3). Perception focus on the subjects lower-level cognition and the perceived elements in a given situation. Comprehension is a higher-order state where elements are interpreted together to form a coherent understanding. In the third level, mental simulation of future states of the system or situation are predicted.

Important elements to form a cohesive SA in the current situation will vary based on the situation that the subject are operating in (Endsley, 1995b). Information to generate SA

would therefore be different in distinctive domains like aviation, military or operating power grids. Endsley (1995b) also points out that SA operates in the present but is affected by the past and the future. A situation can develop in dynamic ways, in which it is necessary for the operators to continuously update their awareness (Endsley et al., 2003). SA is therefore revised according to such changes and feedback from the environment.

SA is a concept that is mostly adopted in highly complex socio-technological systems, such as the occupations mentioned above (Endsley, 1995b). Endsley (1995b) stresses the need to incorporate the construct in system design to support operators SA under such conditions. Severe consequences can occur when operators fail to achieve an appropriate SA, which could be mediated by interface design or other factors (Jones & Endsley, 1996). The emphasis on high risk environments are presumably the reason for not adopting SA in less complex technology. It may even be inappropriate to adopt a construct that is so specialized for dynamic and straining tasks. It stands the risk of generating the same criticism as directed towards mental models. Payne (2003) argue that the concept of mental models is used in so many different fields that it risks meaningful explanatory value and becomes a generic idea about users' knowledge of systems they use.

Nevertheless, it is argued that an individual conducting less cognitively straining tasks also needs SA for optimal functioning. As technology are developing towards increased complexity, designers need new methods to enhance interface design (Endsley, 2008). As the previously cited literature demonstrate, interactions with chatbots is not a passive-response technology and also seem to be affected by user's comprehension and prediction of the system. We therefore adopt SA in the current study for system design purposes, as SA can be used as a tool for user-centered design (Endsley et al., 2003). SA has previously been used in chatbot contexts as well (Luria, Hoffman, & Zuckerman, 2017; Robb et al., 2018). For example, participants had higher SA when using chatbots in combination with the standard control interface of autonomous underwater vehicles. This was in contrast to participants who relied more heavily on the standard control interface (Robb et al., 2018). Luria et al. (2017) however, found that users of voice-based chatbots had lower SA when controlling smart homes. Higher SA was found for physical robots, wall-mounted screens, and apps. Luria et al. (2017) attributed the low SA in chatbot interaction to the lack of transparency and psychical interaction.

Mental models and Situational Awareness

SA is affected by many external and internal characteristics, such as workload, training and system design. The cognitive informational system will also contribute to construct user's SA, trough perception, attention, working- and long-term memory (Endsley, 1995b). As the present study has a main focus on mental models, which presumably reside in long-term memory (Endsley, 1995b), the next section will elaborate on their dynamic (see figure 1).

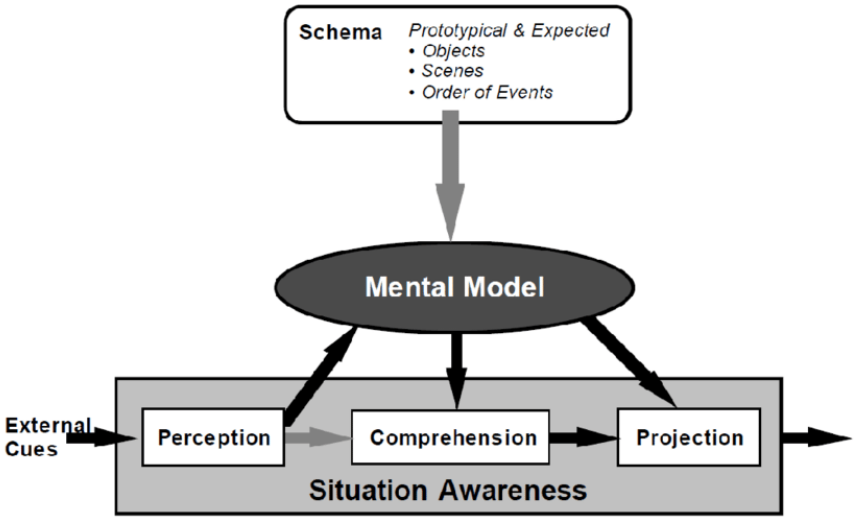

Figure 1. Relationship between the mental model and the situation model (situational awareness). From, Endsley, M, R., (2000). *Situation Models: An Avenue to the Modeling of Mental models*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting 44*(1),* p. 62. Copyright 2000, Human Factors and Ergonomics Society.

Endsley et al. (2003) argue that "a person would be very poor at understanding what is happening" (p. 23) without a mental model that guides them. This is also the case with prediction of future events. Mental models can contribute to SA in two main directions. On one hand, information that is perceived in the environment will affect the adoption of a specific mental model (Endsley et al., 2003), due to matching of cues to the model (Endsley, 2008). This will then affect the subject's comprehension and projection of the current situation. On the other hand, prior experiences or goals will also affect the adoption of a specific model. This will additionally affect which information a person attends to. It can contribute to more efficient processing and cognitive economy but can also drive attention away from important elements (Endsley et al., 2003).

Rouse and Morris (1986) also stress the importance of utilizing cues from the environment for adopting a mental model. Mental models and SA may therefore seem to conceptually overlap (Sarter & Woods, 1991). Sarter and Woods (1991) nevertheless state that the two constructs are distinct. An aspect that highlight their distinction, is the fact that mental models are developed and revised over time due the recurring exposure to situations and information (Endsley, 1995b). SA on the other hand, is a situational understanding in a specific event, and evolve with the dynamic of the environment in matter of seconds to a few hours at most (Sarter & Woods, 1991).

Several errors in SA are attributed to the selection of an errant or incomplete model (Jones & Endsley, 1996). Such errant models may be formed based on a limited set of informational cues, which underscore the need for appropriate system design (Endsley, 1995b). For instance, in framing experiments it is found that how information is presented in a task profoundly affects problem-solving strategies (Tversky & Kahneman, 1981). Endsley (1995b), propose that mental models can contribute to such effects. Jones (1997) also showed how resistant an adopted model can be. Errant mental models were induced and in 65% of the cases the controller did not detect that their mental models consisted of errant information. This occurred even when "bizzare" informational cues were presented to indicate that an errant model was held. Additionally, novices are especially prone to complications with gaining proper comprehension and projection due to the lack of a well-developed mental models (Endsley, 1995b).

In Endsley (1995b) theory, schemata and scripts will further affect SA (see figure 1). Schemata are understood as a cognitive concept that provides mental models with prototypic information. Additionally, schemata can operate in comprehension and projection in a single step when an individual are in well-known situations. As schemata has associated scripts that generate behavioral guidance, it causes highly effective interactions for experts (Endsley et al., 2003). Scripts can additionally generate errors in SA when a current situation require other behavioral scrips than a habitual propensity (Endsley, 1995b). These notions apply well with the findings from the CASA-paradigm and chatbots. SA may therefore contribute to a deeper understanding on some of the findings regarding the interaction between humans and chatbots, which the present study will explore.

Summary

In summary, two main topics from the preceding elaboration of relevant literature should be emphasized. (1) Chatbots are implemented at a high rate and social cues in chatbot

design can contribute to positive human-chatbot interaction and user adoption. The positive outcomes will presumably heighten chatbot designers desire to implement humanlike cues. Social cues may also contribute to affordance, in signaling what chatbots can do and how to use it. Designers may therefore intentionally or non-intentionally design chatbots to induce humanlike mental models in the user. (2) Research on mental models and chatbots confirm the use of such models, creating high expectations and the assumption that human natural language is appropriate. Nevertheless, chatbots have difficulties in understanding users' requests and dialogue can break down. Chatbots may not be as straight-forward to use as the affordance in design signal, and this ultimately cause frustration for users.

Based on these observations there is a need for a deeper understanding of the content of user's mental models, but few studies have had this as their primary research objective and have not investigated content in user mental models while interacting with chatbots. Therefore, the aim of the current study is to contribute in bridging such knowledge gaps as self-serving customer interactions with chatbots are increasingly being implemented on a wide scale. The existing literature on mental models shows that the construct is somewhat of an enigma. Mental models are incomplete and complex cognitive structures, which may be challenging to access and elucidate for researchers. Endsley (2000) propose that SA can be used as a lens into the construct. A review of Endsley (1995b) theory shows that SA and mental models are highly interconnected from a theoretical perspective, we therefore adopt both concepts in order to more thoroughly explore and understand user experiences in the current context.

The Present Study

The present study had two main objectives. The first and primary objective was to investigate the content of the participant's mental models in customer service chatbots. The second objective was to investigate if SA could be used as a framework to understand the themes that emerged when analyzing participants mental models. If the SA construct is applicable in a chatbot context, SA can be used as a design tool to enhance human-chatbot interactions. To investigate these research objectives, an exploratory qualitative approach was adopted. The present study set out to answer the following questions:

1. What characterizes the mental models that individuals apply in customer service chatbot interactions?
2. Is Situational Awareness a suitable framework for understanding interactions with costumer service chatbots?

These research questions were studied by exposing participants to two customer service chatbots in a laboratory setting. A task-based interaction guided communication between the participants and chatbots. The primary data consisted of qualitative information. Data were collected by think-aloud interviews, semi-structured interviews, and screen capture videos to capture the conversation between the subject and chatbots. Some quantitative data was additionally collected, where descriptive information about prior use, chatbot knowledge, task realism, and time spent with the two chatbots were obtained.

Method

The Project

This thesis was carried out in collaboration with the research project, "Chatbots for loyalty." Chatbots for Loyalty is managed by SINTEF Digital and aims to increase knowledge about user-centered design in the context of chatbots. The study presented in this thesis was done by the author (Master student at the Department of Psychology, University of Oslo). The author had responsibility for the choice of method, recruitment process, data collection, analysis and reporting. Feedback and guidance were mainly given in analysis and reporting by Cato Bjørkli (Associate Professor at the University of Oslo), Marita Skjuve (PhD fellow at SINTEF Digital), and Asbjørn Følstad (Senior Research Scientist at SINTEF Digital).

The following section will describe the recruitment process and sample, preparations and setup, data collection procedure, the quantitative and qualitative analysis, as well as ethical considerations.

Recruitment and Sample

The study was conducted between November 11 to December 6, 2019. The participants were recruited through a convenience sample, which entails sampling available and easily accessed individuals (Robinson, 2014). Individuals with computer science education were, however, excluded from the study. This was seen as necessary due to their presumed knowledge of chatbot technology, which may influence their mental models substantially from the general population (Chen & Wang, 2018).

Information about the research was posted at various student-groups (Facebook) or Institutes (University of Oslo). While 18 participants signed up for the study, the first two participants were removed from the finale sample and are considered as pilot participants. This was due to changes in the procedure after the data collection.

The final sample ($N = 16$) consists of nine women (56%) and seven men (44%). The mean age was 27, ranging from 21 to 47. Five participants (31%) had achieved a master's degree as their highest educational degree, seven participants (43%) had a bachelor's degree, and three (18%) participants had finished one year of higher education. Educational background was heterogeneous (e.g., law, biology, finance and other social studies).

Preparations and Setup

Pilot testing. Three separate pilot tests were conducted, with six pilot participants to find an appropriate study protocol.

Two pilot participants, prior to recruitment, were asked to conduct a think-aloud procedure (Ericsson & Simon, 1984). They were asked to find information about mortgage using chatbots from one Norwegian bank, and verbalize their thinking. Participants had difficulties with such multitasking and short interactions (less than a minute).
Four pilot participants (the former pilot participants and the two recruited participants) went through a revised study protocol. The protocol consisted of two similar chatbots, and a task with several topics to increase interaction time. Post-task semi-structured interviews based on mental model literature was conducted (Rouse & Morris, 1986). Answers exhibited more global evaluations and not the mental models that appear in the perception and comprehension during exposure to a stimulus (Klein & Hoffman, 2008). Changes in the procedure were therefore tested with two new pilot participants. A think-aloud interview strategy was implemented in the protocol (Wickens et al., 2013).

The latter approach became the final protocol to generate as much in-depth information as possible. The questions that were asked and the task underwent small changes to make them more understandable.

Choice of chatbots. Two chatbots were used in the present study to create a lengthier interaction. The chatbots were chosen based on similarity. Chatbot A and B are developed by the same company and consist of the same technology. They use AI, machine learning, natural language- processing and understanding to engage in dialogue (Thakur, 2018). They are also used in prior user's research (Følstad & Skjuve, 2019).

The interface of Chatbot A and B are similar in several ways. They welcome the users and generate a short introduction text for interaction guidance. The first interaction is based on natural language, afterwards both natural language and choice buttons (alternatives) can guide the conversation. When chatbots are unable to answer correctly or understand the user's intent, both present the opportunity of talking to a human customer service agent. The main

visual differences between the two chatbots are gender, avatar, and communication style. Chatbot A uses a female avatar, emoticons, and a more humanlike communication style. Chatbot B has a genderless robot avatar with more formal communication (Følstad & Skjuve, 2019).

Task development. A task was developed to give the participants a context for their interaction, which is done in previous chatbot and mental model research (Chen & Wang, 2018). In the task participants were instructed to find general information about mortgages (see Appendix A). Some examples were given (e.g., equity requirements) to ensure lengthier interactions. If chatbots suggested that the participants enter a website for further information, they were asked to continue with the task of conversing with the chatbots. The task was relatively open-ended with few requirements to make participants choose an interaction style that came naturally to them.

Data Collection Procedure

Data collection occurred in the same room at SINTEF Digital in Oslo, Norway. The author completed the study alone in one-on-one interviews. Each participant was (1) given formal consent and (2) a short introduction of the study purpose and collection procedure. The present study used audiotape and screen capture videos to conserve data.

Participants were asked demographic questions regarding age, gender identity, and educational background. The participants were also asked to generate a self-rating score ranging from one to five, on two different variables: level of prior use of chatbots and general knowledge of chatbots. The task was then provided for them. Two chatbots and screen capture videos were pre-prepared on the computer. The sequence of presentation of the two chatbots were alternated to control for order effects. During each chatbot interaction and after task completion, participants were asked questions in line with the protocol (think-aloud and semi-structured interviews). After the interviews, each participant was asked to generate a self-rating score from one to five on how realistic the task was perceived.

The mean length of data collection (time spent completing task and answering questions) was 53 minutes, with the shortest time to complete all steps being lasting 31 minutes and the most extended being 67 minutes.

Think-Aloud interviews. During chatbot interactions, a think-aloud-interview approach was implemented to gain insight into their thinking while interacting with the chatbots (Koro-Ljungberg, Douglas, Therriault, Malcolm, & McNeill, 2013). Each participant was asked at random time-intervals to verbalize their thoughts (e.g., "what do you think

now?"). The participant stopped the chatbot interaction and answered the question. Verbalizations from the participants were followed-up by the researcher with 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 with the task until the next prompting interval. This cycle continued until the participants were out of inquiries to ask the chatbots. Prompting by the researcher occurred during both successful and non-successful communication (if the chatbot can answer appropriately).

Semi-Structured interviews. Questions in the semi-structured interview were based on literature related to mental models (see Appendix B). The procedure were used to allow for a more flexible interview structure and follow-up questions with paraphrasing (Whiting, 2008). Some questions were asked immediately after one chatbot interaction, but most questions were asked after task completion (see Appendix B).

Data Analysis

Quantitative and qualitative data were collected during the current study. The quantitative data consisted of descriptive information and the time spent with the two chatbots. The qualitative data consist of information from both interviews and screen capture video. The following section will describe the method and data manipulation used in the present study and was conducted by author.

Quantitative Analysis

Statistical analysis was conducted by the use of IBM SPSS (Statistical Package for the Social Sciences statistics), version 25. The descriptive information (demographics, prior use, chatbot knowledge and task realism) and estimated time spent with the two chatbots were coded and descriptive analysis were done. Time spent with the chatbots was coded from the screenshot videos and estimated from the first typed letter to the response given from the chatbot. As participants were asked questions during their interaction, this estimation was conducted for each message and the total time spent interacting with the chatbots were calculated for every participant.

Qualitative Analysis

Transcriptions

Interviews. Each audio recording was transcribed verbatim to preserve the meaning and context for all statements. Less rigorous transcription was, however, conducted due to the use of thematic analysis (Braun & Clarke, 2006). Meaning, morphemes (the smallest

meaningful speech unit, (Warren, 2013) were written, but other speech sounds and positive reinforcement from the interviewer (e.g., yes, mm) was left out. Such speech sounds were thinking pauses (e.g., hm), laughter, or coughing. A question mark was put to signal ambiguities or unclear verbalizations.

Three interviews were randomly selected for a transcription check, by re-listening to the audio recording while examining the corresponding transcriptions. Only small morphemes were found to be left out form the audiotape transcriptions (e.g., repetition of the same word). The semantic meaning and general message of all statements were preserved, and transcription quality was deemed adequate.

Chatbot dialogue. Each message written by the participants were transcribed from the screen capture videos. Each statement was written identically to the data, meaning that capital or small letters, punctuations, and use of emoticons was preserved.

Thematic Analysis

Inductive analysis. A thematic analysis was adopted to identify common themes across the data (Braun & Clarke, 2006). An inductive analysis was first applied to identify the content of the participant's mental models. It entails constructing themes based on statements that occur in the transcriptions, without placing them in a specific theoretical category. However, a purely inductive approach can be challenging to achieve, as prior knowledge from both the participants and researcher can affect the analysis. The inductive analysis focused on the semantic meaning (Clarke, Braun, & Hayfield, 2015). According to Braun and Clarke (2006) six steps are necessary for a valid analysis to emerge. The steps described by Braun and Clarke (2006) will be elaborated on in the following section.

In the *familiarization phase*, the researcher should get a comprehensive understanding of the data content and to generate ideas of potential themes (Braun & Clarke, 2006). The author conducted all the interviews and transcriptions, as well as preliminary ideas for codes during transcription. Three of 16 transcripts were additionally re-read to generate a potential structure for coding.

In the *unitizing phase*, NVivo version 11 (data program for qualitative data manipulation) was applied in the coding procedure. Codes are generated by units in the transcribed data (Braun & Clarke, 2006). A unit can be defined as "the smallest meaningful unit that reflects the informant's experience and understanding of the topic of interest" (Hoff et al., 2009, p. 8). It can also consist of "a sentence, a whole sentence or several sentences " (Hoff et al., 2009, p. 8). Throughout the process, initial codes were continually revised to

capture similar units in the transcripts. Some additional information was sometimes attached to avoid losing the overall meaning. A residual code was also created where units were placed due to little relevance.

In the *generating theme phase*, codes are collapsed into broader common themes (Braun & Clarke, 2006). Using NVivo, codes were placed under a new category, making it possible to conduct future revisions in the structure. Each category was given an initial name for the presentation.

In the *reviewing phase*, the codes and themes are evaluated in a two-step procedure. In the first step, each code is re-read to examine the internal validity of the units placed under the code (Braun & Clarke, 2006). Some codes were revised, developed, or left in the initial code. The residual code was analyzed, where some statements were assigned to a different code or left in the category. The remaining content in the residual category was coded out of the final analysis. The second step in the reviewing stage is to evaluate the initial themes and code against the original transcription. It is done by re-reading the transcripts and evaluating the codes and themes from a final meta-perspective on to ensure that overall meaning is captured (Braun & Clarke, 2006). Small adjustments were made in this second phase.

In the *naming phase*, the overall thematic structure is reviewed to ensure that labels and structure create a meaningful story for the reader. During this phase, a summary of each theme and sub-theme is written (Braun & Clarke, 2006). As all interviews were conducted in Norwegian, the content was translated to English before the *reporting phase*.

Deductive analysis. Themes from the inductive analysis were placed in a theoretical framework (Clarke et al., 2015). SA were used to evaluate if the themes from the inductive analysis could fit within the three levels this model proposes. A strict placement of themes was conducted. Meaning, each theme could only be placed within one level and semantic meaning were considered to be most relevant to decisions of assigning particular themes to a corresponding level. Final placement of themes was approved by two independent researchers, a PhD fellow at SINTEF and an Associate Professor at the University of Oslo.

Language Analysis of Chatbot Dialogue

A discrepancy was observed during data collection between the statements from participants (use of keywords) and their actual interaction with the customer service chatbots. Additionally, mindless application of social rules towards the chatbots were observed in the transcribed dialogues by the author and a Senior Researcher at SINTEF.

To evaluate the data, the author and a Senior Researcher at SINTEF developed an appropriate coding procedure. A message was defined as a single inquiry sent to the chatbot by the participants and could include a single word or a string of words (Hill, Ford, & Farreras, 2015). To examine the discrepancies between actual behavior (messages to chatbots) and assumptions (verbal statements to researcher), the author sat a cut-off point at three single words in one message to indicate a keyword tactic. Keyword tactics were often in contrast to more complete Norwegian sentences in the remaining dataset (Simonsen, 2019).

Additionally, there are no agreed upon definition that characterizes social behavior towards text-based chatbots. To evaluate the presence of mindless application of social rules, the author decides to look at the use of first- and second-person pronoun and use of polite remarks. This is in line with Brennan and Ohaeri (1994) who defined an anthropomorphic sentence towards a computer agent as consisting of first-person pronoun. They also found a higher use of second-person pronoun and polite remarks towards computers, indicting a social act towards the computer.

The analysis was conducted three times to check for omissions in classification. A few omissions were found in the first analysis and corrected.

Ethical Considerations

The current study was approved (ref. 284040) by the Norwegian Center for research data (NSD), and the ethical research guidelines by University of Oslo and NSD was followed. Each participant was assigned an ID-code that were stored in a secure storage service provided by the University of Oslo, and the formal informed consent procedure was approved by NSD. The formal consent (Appendix C) consists of an introduction that describes the current study, a description of the purpose of the study, statements that underscore the possibility to withdraw from the study at any given time, and information about how the collected data will be stored and handled. All participants consented to participate based on this information.

To ensure that no personal data were given to the companies from the chatbot interactions (e.g., personal IP-addresses), the author's own computer was used. Each participant was also clearly instructed to not write any personal information, and the author ensured that the instructions was followed by monitoring the task completion. Audio was recorded with "Nettskjema" Dictaphone, which sends audio files directly to a secure storage service provided by the University of Oslo. The screen capture videos only record the screen,

and no personal information was collected. All data analysis was done through secure networks of the University of Oslo.

To reduce the potential stress of being evaluated or tested, each participant was given an introduction before the task. Furthermore, each participant was informed that the questions asked by the author had no right or wrong answer. Their only assignment during the study were to elaborate on information that came to mind. After the data collection phase, each participant was then given a debrief that elaborated on the study purpose. Every participant was asked about partaking, and the author is confident that each individual left the study well informed and without any negative affect. Each participant was thanked for their contribution and given a gift card of 300 KR as compensation. SINTEF Digital provided the vouchers.

Results

The aim of the current study was to explore the participants mental models while using customer service chatbots, and to investigate if the SA-framework could be used to elucidate the themes that emerged from this exploration. Information about the participant's mental models was gathered trough two interview strategies (think-aloud and semi-structured interviews) and screen capture videos. The analysis generated three sets of main data: inductive analysis, deductive analysis, and language analysis. The following section will present the results from the three analyses, as well as the collected descriptive information.

Descriptive Statistics

Each participant (n = 16) was asked to generate a self-rating score from one to five, on three different variables: prior use, chatbot knowledge and task representativeness. Descriptive statistic shows that most participants had some prior use of chatbots (M = 2.56, SD = 0.96), and some knowledge of chatbots (M = 2.25, SD = 0.77). Most participants perceived the task as representative of their natural usage (M = 4.25, SD = 1.00).

A calculation of time spent with the two chatbots were also conducted, to give an overview of interaction time during the data collection. Descriptive statistics show that participants spent more minutes with chatbot A (M = 2.62 min SD = 0.98 min) than Chatbot B (M = 1.74 min, SD = 0.51 min), and little over four minutes collectively with both chatbots (M = 4.25 min, SD = 1.20 min).

Inductive Thematic Analysis

A cluster of eight main themes and 19 sub-themes emerged from the inductive analysis, and table 1 provides an overview of the results. As table 1 shows, all main themes contained relevant statements by each participant, except for the main theme *User Trust.* In the sub-theme section, 12 out of 19 themes contained relevant statements by each participant, *Trust in Machines* and *Organizational Viewpoint* had least included participants. The following table 2 contains a summary describing each main theme and sub-theme.

Table 1: Overview of the Inductive Thematic Analysis

Themes	**Sub-Themes**
Human-Chatbot Cues (16)	Human-Like Traits (16) Chatbot-Like Traits (16)
Design Attributes (16)	Positive Characteristics (16) Negative Characteristics (14) Affective response (13)
User-Chatbot Dialogue (16)	Successful Conversational Exchange (16) Problematic Conversational Exchange (16)
User Communication (16)	Communications Strategy (16) Communication Demands (16)
Chatbot Functionality (16)	Keyword Hypothesis (16) Chatbots General Intelligence (16)
Chatbot Assumptions (16)	Current Expectancy (14) Area of Use (16) Future Expectancy (14)
User Trust (15)	Risk Perception (14) Trust in Machines (10)
Perceived Utility (16)	Perceived Efficiency (16) Perceived Usefulness (16) Organizational Viewpoint (10)

Note. Numbers in the parenthesis reflect the number of participants that made relevant statements in each theme.

Table 2: A Summary of the Inductive Thematic Analysis

Themes	Description
Human-Chatbot Cues	In the sub-theme, *Human-Like Traits*, participants describe chatbot attributes which resemble human features and human-human conversation. Most comments were directed towards Chatbot A, e.g., referring to the chatbot as "she" and commenting on the humanlike avatar. A few participants also pointed out that, for a brief moment, they almost forgot that chatbot A was not a human. On the other hand, every participant was explicitly aware that the chatbot was a machine, and some rejected the social cues. In the sub-theme, *Chatbot-Like Traits*, cues in the design (e.g., fast answers and providing alternatives), and Chatbot B's visual impact facilitate this understanding due to the robotic avatar. Some participants did not attend to Chatbot B's avatar, and most participants were generally pleased that the avatar looked like a robot.
Design Attributes	The Utterances were diverse, from the colors to the ease of use. Most associations towards the software were related to social media (like Facebook Messenger) as well as the Google search engine. In the sub-theme, *Positive Characteristics*, a general topic for many participants was linking the ease of use to the metaphoric design of social media. They were also pleased about the ability to access human communication if necessary. In the sub-theme, *Negative Characteristics*, too much text was provided in the interface, or negative to the delay of reaching human contact was verbalized. In the sub-theme, *Affective Responses*, some participants described positive affect towards the social cues, like amusement towards humanlike answers. A few participants also perceived the use as more relaxing than human communication. Also, when the chatbot interaction was difficult, most participants in this sub-theme experienced frustration.
User-Chatbot Dialogue	In the sub-theme, *Successful Conversational Exchanges*, participants experienced useful and relevant answers to their inquiries. They were also pleased when the chatbot was able to answer in a more detailed fashion. In the sub-theme, *Problematic Conversational Exchanges*, participants experienced irrelevant answers and non-useful repetition of information that they had already received earlier in the conversation. Some participants also got information on a previous question in a later conversational exchange, which ended in low reliability regarding when the relevant information was received.
User Communication	In the sub-theme *Communication Strategy*, participants experienced that the design made a human-human communication style seem natural. On the other hand, when they reflected on which behaviors that efficiently triggered the needed information, participants mentioned short and concise messages or use of keywords. Some participants also mentioned that polite phrases were unnecessary. All participants stated a favorable attitude towards being provided with alternatives, except for a few occasions where the alternatives did not fit their informational need. In the sub-theme, *Communication Demands,* specific language barriers were mentioned. Such as users spelling errors, wrong declension, compounding words and the use of dialect. Semantic knowledge created difficulties for users due to the lack of expertise in banking terminology. A general problem in the interactions was the need to concretize their questions, and many participants went through a trail- end error process in order to be understood by the chatbots.
Chatbot Functionality	In the sub-theme, C*hatbots General Intelligence*, each participant commented on a lack of intelligent behaviors. This was a contradiction to their expectations of a smarter and more sophisticated system. Such missing behaviors were lack of memory, understanding of context, and ability to learn and engage in reciprocal interaction. Participants were unsure of the chatbots' ability to learn, but most assumed that humans needed to upgrade the system for actual change to occur. In the sub-theme, *Keyword Hypothesis*, which is a term used by participants, most participants assumed that the chatbots consist of pre-programmed and automated answers, which were plotted in by humans. Based on keywords from the participant's message, the chatbots provided the corresponding information in the conversation.

Chatbot Assumptions	In the sub-theme, *Current Expectancy*, cues in the interaction generated the assumption that the chatbots had no future information to provide. Such cues were: repetitions of the same answer or lack of additional alternatives. Prior or current experience also gave most participants negative expectations of the chatbots ability to help. In the sub-theme, *Area of Use*, participants elaborated on what the chatbots could assist with, like answering simple inquiries (e.g., give definitions) or guide them to the correct information site. Everything that deviates from this was reasoned to be too difficult for the chatbots to help with (e.g., help with personal inquiries, discretionary decisions). In the sub theme, *Future Expectations*, they hoped chatbots would evolve and conduct operations that are experienced as too difficult for the current technology, and act more like a human advisor. E.g., give more personal and specific information, handle follow-up questions.
User Trust	In the sub-theme, *Perceived Risk*, some participants assumed that there is a risk of not receiving all the relevant information. Some also articulated risk regarding misunderstandings due to lack of intelligent behavior from the chatbots. In the sub-theme, *Trust in Machines*, participants also discussed a more global distrust towards machines. They assumed that humans would better handle ambiguities in the communication, and therefore provide better assistance than a chatbots. On the other hand, a few participants assumed that machines were more reliable than humans, which facilitated higher level of trust in the chatbots.
Perceived Utility	In the sub-theme, *Perceived Efficiency*, participants perceived the chatbots to be both time-efficient by providing information at a rapid pace and time-consuming due to the lack of appropriate answers. In the sub-theme, *Perceived Usefulness*, many participants were positive with regards to using the chatbots again. They also discuss how search engines or navigating banking websites on their own would be more productive, and most participants felt they needed human communication to get the desired information. In the sub-theme, *Organizational Viewpoint*, most comments were directed towards economic benefits and time efficiency for the customer support system.

Deductive Thematic Analysis: Situational Awareness

SA (Endsley, 1995b) were used to see if the isolated themes could be placed within the different levels of the framework. Placement of themes within this framework would potentially also generate a more coherent and meaningful understanding of the inductive results. Table 3 presents an overview of the deductive analysis. Table 3 shows that six out of eight main themes were placed under level 2 and 3. No main theme was placed under level 1.

Table 3: Overview of The Deductive Thematic Analysis

Situation Awareness	**Themes**
Perception (Level 1)	
Comprehension (Level 2)	Human-Chatbot Cues (16) User-Chatbot Dialogue (16) User Communication (16) Chatbot Functionality (16)
Projection (Level 3)	Chatbot Assumptions (16) Users Trust (15)

Note. No main theme was placed under level 1 due to a strict placement of themes, see discussion for further elaboration. Numbers in the parenthesis reflect the number of participants that made relevant statements in each theme.

Perception (Level 1). Perception (Level 1). The relevant elements to perceive in a current context or task will depend on the specific environment the subject(s) are in (Endsley, 1995b). In a chatbot context, such information can be the design features or the conversational output. Many participants commented on the dialogue and few participants spontaneously commented on the design:

"*wow, wow, wow (...) that was a nice avatar*" (P 3)

However, such statements did not form a cohesive and meaningful main theme in the current study. Such comments were either few or more appropriate to consider under Level 2 and 3. Considerations regarding this finding will be elaborated further in the discussion section.

Comprehension (Level 2). After perceiving the relevant elements in the current environment, a subject can form a cohesive understanding of the situation (Endsley, 1995b). Additionally, the mental models that are in use can also contribute to level 2 understanding when the environment is lacking important level 1 values to perceive (Endsley, 2015). This is especially relevant whenever available information in a situation is incomplete or ambiguous. As chatbots are generally lacking transparency with regard to their underlying dynamics, themes in level 2 seem affected by both bottom-up and top-down processing (mental models is an example of the latter). Four themes were appropriate to designate to this level, *Human-Chatbot Cues*, *User-Chatbot Dialogue*, *User Communication,* and *Chatbot Functionality.*

In the *Human-Chatbot Cues*, participants expressed an understanding of the system as having humanlike qualities. They also revised this construal or had access to an alternative understanding, which consisted of a complete awareness of conversing with an artificial software system. Statements in *User-Chatbot Dialogue* also exhibited that the ongoing dialogue provided an important feedback loop to their comprehension of system capabilities. Additionally, answers from the chatbot could elicit reactions ranging from confusion to satisfaction with the system output they received. If the system output were incongruent with their current goal, revisions and adaptations of their behavior were considered to be necessary. Instances of this is reflected in *User Communication.* Participants thus comprehended the need to adjust their own specific tactics and the language requirements of the particular context for successful communication to occur. In the last theme, *Chatbot Functionality*, participants generated explanations of how the system works. Initial

assumptions were generally of a smart and adaptive system. During interactions or by remembering pre-study experiences, their explanations shifted towards describing a rigid system operating on basic keyword recognition. See table 4 for in-depth information and relevant statements.

Table 4: Overview of Relevant Themes for Comprehension (Level 2)

Themes	Content of Relevance for Comprehension	Relevant Statements
Human-Chatbot Cues	As chatbots use social cues, it seems to generate an understanding of the chatbots ontology. This understanding contributed to anthropomorphism towards both chatbots and the experience of social presence. Chatbot A facilitated this understanding considerably, possibility due to the embedded humanlike avatar and language.	*"She appears more as a human. Also, the way she opens (the conversation) with an emoji and "hi""* (P 1) *"You also get a slight feeling of talking to a person"* (P 6)
	All participants knew that both chatbots were software systems, and also generated descriptions of a qualitatively different system understanding than that of a human. Cues to support such system knowledge were quick responses, inflexible replies and presentation of alternatives. Chatbot B visual avatar also had an impact on participant's understanding, with the use of less social cues.	*"When you chat with a person, they would answer with text. Not alternatives"* (P 11) *"I like that it is a cute robot, it might be more encouraging to treat it as such. More as a machine and less as a person"* (P 7)
User-Chatbot Dialogue	Overall, dialogue served as an important foundation for system understanding, and about which operations the chatbot could perform. In some interactions, the chatbots were able to understand the user's message or provide relevant information. Participants were pleased in such instances. These interactions also indicate that users were able to interact in accordance with the system capabilities	*"You start to see a pattern of what it (chatbot) can answer"* (P 13) *"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)
	Both chatbots answered out of context, repeated previous answers or stated a lack of comprehension of users' messages. In such incidents, participants became confused or assumed that their own behavior was incompatible with the system capabilities.	*"Know I don't understand what it (chatbot) did. I asked how much mortgage I could get with an annual income of 600 000 (KR). Then it (chatbot) asked if it was regarding a new or existing loan. Then I wrote new, and it asked if I was renovating or moving my mortgage"* (P 5) *"I need to rephrase in a way that the chatbot understand (...), and it is almost more complicated in a bot (chatbot) than in a search engine"* (P 10)
User Communication	Participants started with an assumption and understanding that natural human language would be appropriate. During the interaction, such comprehension shifted. Participants stated the	*"Such chat format creates the expectation that I can formulate myself as I would do in a chat (with humans)"* (P 4)

	importance of being short, concise, or the use of keywords. Some had pre-study experience of using keywords as a tactic. They understood that multiple question and long sentences were difficult for the chatbot to handle. Polite phrases were unnecessary as chatbots were understood as a software system.	*"I tried at one point to write: requirements for interest rent, and it did note understand. Then I wrote: interest rate and I got it (the answer)"* (P 16)
	Participants generated an understanding of not only tactical needs but also more specific language demands. Participants therefore assumed and experienced that spelling errors, wrong declension, compounding words, and dialect use would be too difficult for the chatbots to handle. Additionally, many stated a problem with finding the accurate semantic words to access system content. The use of alternatives was therefore preferred for navigation, due to memory strains or lack of visibility of chatbot content.	*"This is the reason why google is such a good search engine. You can write utterly wrong, and it still find articles"* (P 9) *"I feel I need to have a lot of insight (to communicate), and write the correct keyword"* (P 4) *"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)
Chatbot Functionality	Participants had a general implicit assumption of the chatbot being able to understand context. When disproven, such assumptions came to light. Their system understanding was revised and described the chatbot as non-adaptive and lacking in creativity, ability to think and to engage in reciprocal communication. When asked about the chatbot's ability to learn, their assumptions were mixed. Some assumed that human involvement was needed for actual change to occur; others did not have an assumption or explanation on the matter. As chatbot A stated an ability to learn, it caused uncertainty among participants about its abilities	*"If I previously pressed (the alternative) that I am between 18 and 34, then she (Chatbot A) should think that. 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) *"I was a bit disappointed, I thought it was smarter"* (P 13)
	In the keyword hypothesis, participants exhibited a system understanding of the chatbots. They assumed that chatbots had pre-programmed answers, which had been manually plotted in by humans. When participants sent a message to the chatbots, keywords in the message would be detected and based on those keywords the attendant 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)

Note. Relevant Statements are translated from Norwegian to English, se appendix D for original version. Numbers in the parenthesis reflect participants ID.

Projection (Level 3). At the highest level of Endsley (1995b) model, individuals will engage in mental simulation based on a holistic understanding of the perceived elements and their meaning. As with perception and comprehension, the mental models that are adopted can contribute to projection (Endsley, 2015). Two main themes from the inductive analysis were considered appropriate to place under this level, *Chatbot Assumption* and *Users Trust*.

In the present study, participants generated assumptions about which operations the chatbot could perform. In *Chatbot Assumption*, low expectations of chatbot capabilities were verbalized due to both pre-study experiences and the current dialogues. They did not predict that they would gain information beyond the banking domain and hoped that chatbot technology would evolve to support more complex interactions in the future. In the main theme, *Users Trust*, participants elaborated on the chatbots ability to assist them. Participants anticipated potential risks with using the chatbots. Participants also predicted that humans, in general, would better sort out their inquiries due to the chatbots lack of expertise. See table 5 for in-depth information and relevant statements.

Table 5. Overview of Relevant Themes for Projection (Level 3)

Themes	Content of Relevance for Comprehension	Relevant Statements
Chatbot Assumptions	Different cues in the interaction generated an expectation that the chatbots were depleted of content or reached its "end". Furthermore, participants had experiences that gave them negative prospects 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)
	Participants expected that the chatbots could assist 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. Such as 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) "*I would expect that regular opening hours would be inside, and I would ask about that. But if I have a more advanced question (...). It is difficult to come up with an example, but I would not ask if the beer sales had regular sales hours on May 17 (Constitution Day)"* (P 14)
Users Trust	Many participants stated the prospective risk of the chatbot not being able to assist them properly. This was due to communicational difficulties, or lack of transparency in the chatbot.	*"I think that, maybe not dangerous, but if you are uncertain of what you are looking for (...), then you really need to hunt and find what you need to ask about"* (P 15) *"I feel frightful of losing information (...), when I don't get the full informational picture"* (P 12)
	A general lack of expertise and human qualities made participants uncertain of the chatbot's ability to assist. They assumed that humans would better handle ambiguities in communication, thereby providing better aid than chatbots. On the other hand, a few of the participants assumed that the machines were more	*"They (humans) will listen to your intonation and your demeanor, what your question really is. This (chatbot) would not, they will only look at what you wrote"* (P 16).

	reliable than humans, which facilitated a higher level of trust.	

Note. Relevant Statements are translated from Norwegian to English, se appendix D for original version. Numbers in the parenthesis reflect participants ID.

Others. Two out of eight main themes were not placed in the SA-framework as these themes seemed to be a more general evaluation of the two chatbots (*Perceived Utility*) as well as a subject satisfaction and affective responses to the interface (*Design Attributes*). See table 6 for in-depth information and relevant statements

Table 6. Overview of Themes not Relevant for SA

Themes	Content of Relevance for Comprehension	Relevant Statements
Perceived Utility	Participants experienced the chatbots to be time efficient as it provided information quickly. They are faster than human customer service by not requiring the user to wait for their turn and replying instantly. At the same time, they experienced the chatbot to be more time consuming by being a detour to gather information due to lack of appropriate answers.	*"Because I need to start searching around (for information), and then the purpose of the chatbot disappears. Because it's (chatbots) supposed to be quick access to information"* (P 2)
	Many participants were positive about trying the chatbot again. At the same time, contact with a human operator was mentioned as necessary due to the complexity of their inquiry or preferred way to get information. Also, searching the internet by themselves was described as preferable. Nevertheless, they understood the reasons for their existence, by being both economically beneficial and time-efficient for the customer support system.	*"I am able to use Google. The need to then go to the homepage to use a somewhat advanced search engine seems meaningless"* (P 12)
Design Attributes	Participants mentioned that their most dominant associations related to the chatbots design was with social media platforms, mainly Facebook Messenger. A general theme for many participants was linking the ease of use to the metaphoric design. Others had an association with search engines. Participants also mentioned that too much text was provided in the interface or were uncertain of the meaning behind some of the graphical interface design.	*"Most (people) are probably using Facebook chat once a day, and it is a familiar format that it is easy to use"* (P 11)
	Participants described a positive affect towards the social cues. Some also perceived the use of chatbots as more relaxing than talking to a human regarding such matters. Also, when the chatbot interaction was difficult, most participants experienced frustration.	*"I feel that it (chatbot interaction) can be a bit frustrating. It is like talking to a person that don't understand"* (P 8)

Note. Relevant Statements are translated from Norwegian to English, see appendix D for original version. Numbers in the parenthesis reflect participants ID.

Language Analysis of the Chatbot Dialogue

A language analysis was conducted for two purposes. Frist to look at the discrepancies between the participants statements on how they assumed the customer service chatbots to work and their actual interaction with the chatbots. Secondly, to evaluate mindless application of social rules towards the chatbots, as post-evaluation video exhibited such behavior.

The participants sent a total of 229 messages to the two chatbots. Out of the 229 messages a total of 58 messages consisted of keyword tactic (defined as three words or less). Such messages were in contrast to the more complete Norwegian sentences that most participants primarily used with both chatbots. The results indicate that users prefer using natural language when interacting with the chatbots.

Out of the 229 messages, a total of 104 messages contained Norwegian first- and second-person pronoun, and 27 messages contained polite remarks. This indicates a mindless application of social rules towards the chatbots. Table 7 gives an overview of examples that are written to the chatbots and are representative for the overall sample in each category.

As the total written messages (M = 14.31, SD = 4.92), and the use of linguistic characteristics varied in the sample, the author decided to summarize the three categories per participant in table 8. The table shows that two out of 16 participants used a keyword tactic in their interaction in more than half of their messages. The remaining participants used more complete sentences as their most dominant messaging tactic. Table 8 also show that seven out of 16 participants used first- and second-person pronouns in half of their messages or more. Polite remarks on the other hand, were used less by the overall sample.

Table 7. 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) *"My partner bought an apartment last year"* (P 4) *"do bsu account as the same as other equity"* (P 14)
Keywords	*"Information on mail"* (P 9) *"Loan"* (P 16) *"equity"* (P 13) *"will refinance mortgage."* (P 10)
Personal Pronouns	*"How much equity do you demand with a mortgage?"* (P 7) *"Do I need a permanent job to get my first mortgage?"* (P 8) *"Do I get more loan if I have a guarantor?"* (P 1) *"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)

"Hi!" (P 15)
"thanks for the help Chatbot A" (P 12)

Note. The sentences are translated from Norwegian to English, see appendix E for original version. Numbers in the parenthesis reflect participants ID.

Table 8. Prevalence of Different Language Categories for Each Participant's Messages

Participants ID	Use of Keywords	Use of Personal Pronoun	Use of Polite Remarks
1	12%	50%	25%
2	25%	61%	
3	11%	77%	
4	7%	66%	
5	12%	50%	25%
6	7%	30%	15%
7	15%	61%	15%
8		44 %	
9	31%	57%	5%
10	4%	41%	
11		41%	16%
12	43%	43%	21%
13	36%	42%	31%
14	20%	40%	13%
15	62%	25%	6%
16	68%	18%	12%
Total	25%	45%	11%

Note. The percentage is calculated based on the total written messages per participant with both chatbots. Empty spaces represent a lack of applied category.

Discussion

The purpose of this study was to gain insight into the participant's mental models in customer service chatbots interactions, and to evaluate the utility of implementing the SA-framework to understand these mental models. In order to get insight into the research questions, two interview strategies were implemented, as well as screen capture videos to assess the user's chatbot dialogue. The overall results generated three streams of main findings that will be discussed in the following order. (1) Eight distinct aspects of the participant's mental models were discovered. Participants also seemed to shift between two distinctive models, one consisting of humanlike content and the other consisting of chatbot specific content. The findings from the inductive analysis is discussed in light of relevant literature.

(2) Most participants preferred use of natural language and applied social rules towards the chatbots. These results indicate that social scripts in a human mental model guided their behavior. The findings will be discussed in light of the inductive results and existing research and theory. (3) Six out of eight emergent themes could be placed within the SA-framework. Assignment of themes proved to be challenging and produced somewhat strange results, as no themes were considered to properly represent the level of perception. These results will be discussed in light of Endsley (1995b) theory. The last section will discuss the limitations and implications of the current findings, and present recommendations for future research

Frist Stream: The Characteristics of User's Mental Models

The first stream of results was generated from the inductive analysis. The results revealed some interesting findings regarding the user's mental models, as similar content in the models were evident across the study sample. Additionally, participants seemed to experience two qualitatively different models in their chatbot interactions. One mental model that consisted of humanlike knowledge, and one that had more chatbot-like knowledge. This is in-line with previous assumptions about the quality of mental model in the sense that humans can use multiple models in reasoning tasks instead of relying on a single unchanging model (Williams et al., 1983).

Eight main themes were found in the analysis: *Human-Chatbot Cues*, *Design Attributes, User Communication*, *User-Chatbot Dialogue*, *Chatbot Functionality*, *Chatbot Assumptions*, *Users Trust* and *Perceived Utility*. Seven themes where discussed by all participants (n = 16), and 15 participants articulated relevant statements in the main theme *Users Trust*. In the main theme, *Human-Chatbot Cues* and *Design Attributes*, participants made comments about the chatbots interface. Two main themes, *User-Chatbot Dialogue* and *User Communication* reflected communicational characteristics from the user and chatbots perspective. In *Chatbot Functionality*, participants verbalized assumptions about their understanding of the system, and in *Chatbot Assumptions*, they made predictions of what operations the system could perform. In the two last themes, *Users Trust* and *Perceived Utility*, participants elaborated on their level of trust towards the chatbots and the degree of perceived benefits provided by the chatbots as an artifact.

Human-Chatbot Cues. Anthropomorphism is the mental operation of attributing human traits to non-living objects (Nass & Moon, 2000), and previous research has found that such attributions can be directed towards chatbots (Araujo, 2018). It even occurs for chatbots

with disembodied social cues, such as lack of avatar and name. Therefore, it comes as no surprise that participants in the current study engaged in such attribution. Chatbot A was called "she" and all participants made comments about the humanlike appearance of both chatbots. A part of the theoretical explanation for anthropomorphism is the assumption that the content of mental models may contribute to this phenomenon (Breazeal, 2003; Culley & Madhavan, 2013; Epley et al., 2007). Mental models may consist of human knowledge schemata that guides an understanding of the system, or an egocentric knowledge of the self that is projected towards the artifact (Epley et al., 2007). Such findings and theoretical assumptions indicate that subjects in the current study where applying a human mental model towards the customer service chatbots. Mental models are also described as the cognitive process behind the concept of social presence, where humans engage in mental simulation of another mind (Biocca, 1997). Lee's (2009) definition of social presence reflects what participants in our study articulated, that they had the experience of talking to another human. This is also consistent with pervious findings in chatbot studies (Araujo, 2018; Følstad & Skjuve, 2019).

However, participants never lost awareness of the chatbot origin as a non-human artifact, which is comparable to what has been found in similar studies (Følstad & Skjuve, 2019). Cues in the chatbot, such as providing alternatives and fast answers seem to highlight this distinction between human and non-human and reduced the feeling of communicating with a person. Former research has also found that use of fast answers makes the chatbot seem less humanlike (Gnewuch, Morana, Adam, & Maedche, 2018). Even the use of typography in the chatbots influence social perception (Candello, Pinhanez, & Figueiredo, 2017). It can therefore be said that participants rejected the anthropomorphism, at least explicitly insofar as they never truly confused the bots with a person. Similarly, in several studies anthropomorphism towards machines has been denied by participants as they always report an explicit awareness of the system ontology. However, they still behaved socially towards machine (Nass & Moon, 2000).

Design Attributes. In the present study, participants reported that their main associations towards the chatbots were Facebook Messenger and other related human digital interaction mediums. The association with social media was also descried as a positive attribute of the system by the participants, making it user friendly and familiar. Metaphoric design can reduce cognitive load when interacting with a system (Gambino, Fox, & Ratan, 2020). Araujo (2018) also found that feeling of presence can be induced in a chatbot context without social cues as well, due to dialogue and interface resemblance to mediums that the

user is familiar with, such as Messenger. Therefore, it seems as the participants in the current study were relying on knowledge of human communication trough social media.

The consequence of using such knowledge as guidance for how to interact with automated chatbots, may be the experience of frustration as they often fail to respond in an appropriate manner according to such expectations (Luger & Sellen, 2016). Frustration were the most frequently described affective state in the present study. However, the social cues also induced some positive affective states. Human likeness in robotics has have been found to be generally preferable (Walters, Syrdal, Dautenhahn, te Boekhorst, & Koay, 2008), and use of humor and personality in chatbots have been found to generate positive states (Jain, Kumar, Kota, & Patel, 2018). The subjective effects of including social cues are therefore somewhat ambiguous as the implementation has both cost and benefits and may relate to individual differences (Walters et al., 2008).

User-Chatbot Dialogue. The chatbots ability to correctly interpret the intended meaning of a message and answer appropriately with regards to the users goal, is reported to be mixed (Kvale et al., 2019; Myers et al., 2018; Skjuve et al., 2019). One can argue that this is a purely technical issue, but the current thesis also argues otherwise. As the present discussion has pointed out, it seems that as users adopt a mental model with human content that contribute to difficulties. Mental models also consist of scripts for behavioral guidance (Endsley et al., 2003). The mental model that is adopted may be errant (Jones & Endsley, 1996), which generate a "gulf" between the users' assumptions about the system and the actual ways to efficiently interact with the system (Norman, 2013). Evidence of such becomes clear in statements covered in the next theme.

User Communication. Johnson-Laird (1980) assume that a mental model consists of the knowledge about others. This will contribute communicational adaptation towards the receiver (Brennan & Ohaeri, 1994; Johnson-Laird, 1980). It can be argued that participants in the current study engage in such mentalization, where participants expect that the pragmatics and norms of human natural language could be applied. This is consistent with previous findings about early interactions with personal assistance chatbots, in which less experienced users tend to rely more heavily on natural language (Luger & Sellen, 2016). Such expectations are reasonable, as chatbot language has become so seemingly sophisticated that they can pass the Turing Test in some conditions (Warwick & Shah, 2016).

When interacting with robotics and machines, it seems that human apply the cognitive process of constructing a Theory of Mind (ToM). ToM is the process of making inferences about another person's knowledge and beliefs (Premack & Woodruff, 1978). There has even

been found that some of the same brain regions are active when interacting with machines with anthropomorphism cues as those that are involved in ToM directed at humans (Krach et al., 2008). However, engagement in ToM is also done by adapting our own behavior towards the receiver. A ToM study with chatbots has found that subjects adapt their behavior when completing a task with text-based chatbot, in comparisons to how they behave when conducting a task individually (Heyselaar & Bosse, 2019). Heyselaar and Bosse (2019) argue that such findings indicate that users have an implicit understanding of the text-based chatbot as having a mental state, further supporting our notion of a humanlike mental model in this study.

This may have contributed to the experiences reported in *User-Chatbot Dialogue,* where trial and error with regards to getting desired answers were a recurring theme in our study. It is shown that users of voice user interface often communicated in a way that the software system cannot interpret, or tries to execute an operation that the system are unable to support. (Myers et al., 2018). This was attributed to an incomplete mental model, and Myers (2018) found tactic change such as guessing, simplifying, quitting or restarting the operation. This indicate that our participants had similar difficulties and lacked a complete system understanding.

Participants in the present study also seemed to change tactics, by proceeding from initially using natural language to a tactic of using only a few keywords. Luger and Sellen (2016) also found that their subjects used keywords and less complex language with more chatbot experience. Such findings indicate that users' mental model is revised by experience, and increasingly containing less humanlike content. Hill et al. (2015) also found that human-chatbot communication is qualitatively (e.g., more pronouns, swearwords) and quantitatively (e.g., more words and massages) different than human-human dialogues. Such results indicate a chatbot mental model, with the assumption of interacting with an entity that lack a mind.

Chatbot Functionality. Participants also had pre-existing expectations of interacting with a "smart" system. For instance, expectations about the chatbots ability to understand context were exhibited. It has been argued that the chatbot should explicitly state its limited intelligence, as novel users have too high expectations towards chatbots capabilities (Luger & Sellen, 2016). During the interactions and after task completion, participants described the chatbots as not sufficiently adaptive and intelligent and the probable need for human intervention if actual changes in the system were to be realized.

Mental models can be updated with experience (Endsley, 1995b; Rouse & Morris, 1986). Epley et al. (2007) argue that an updated understanding of the system consisting of

non- anthropomorphic descriptions can be acquired. Epley et al. (2007) consider findings about autism as evidence of such models being a relevant alternative, as this population often uses other semantic categories than anthropomorphic accounts to describe moving objects (Castelli, Frith, Happé, & Frith, 2002). Our study seems to find a keyword hypothesis among participants as a tactic that fits how they presume the system to work. This mental model also influences the participants rejection of polite phrases and their use of keywords in *User Communication*. However, Epley et al. (2007) also points out that system mental models may not be applied, even if it is available for the user.

Chatbot Assumptions. Previous research has found that a user's mental model seems to affect expectations of what information the chatbot are able to provide. Such expectations were based on limited background knowledge such as in which country the chatbot was produced (Lee, Lau, Kiesler, & Chiu, 2005). Participants in Lee et al. (2005) study assumed that a conversational robot's made in New York would have the same knowledge of local landmarks as people from New York. This also seem to be the case in the current study, where topics beyond banking were not mentioned. Participants assumptions about the chatbots abilities are also similar to what is found in previous research, and are described as realistic (Følstad & Skjuve, 2019). However, low expectations regarding the chatbots ability were also evident. A consequence of this is the insufficient use of chatbots features that would provide the user with useful information (Luger & Sellen, 2016). A pessimistic chatbot mental model with low understanding of the software operational features can be equally undesirable as the unrealistic expectations that follows with a human mental model (Luger & Sellen, 2016).

User Trust. Anthropomorphism is found to have a positive effect on trust in chatbots (De Visser et al., 2016). However, the current study found mistrust towards the chatbots. As the system reveals difficulties with answering appropriately to inquiries, the chatbot will demonstrate limitations that are in violation of the humanlike mental model (Lee & See, 2004). A recent study of trust in chatbots has shown that expertise and providing appropriate answers are the most important factors with regards to chatbot trust (Nordheim, Følstad, & Bjørkli, 2019). The current study also seems to find similar results, as lack of both relevant information and expertise is perceived as negative.

Research on mental models, trust and Adaptive cruise control (ACC) systems show that incomplete mental models has a negative effect on trust (Beggiato & Krems, 2013). However, individuals who have appropriate mental models of ACC systems can predict problems, and trust is thus not negatively affected. Such findings illustrate the risk of

incomplete mental models and may be generalized to the current study. Research has found that individuals with higher technical knowledge seem to be more forgiving of such issues than novices with regards to trusting the system (Luger & Sellen, 2016). The risk of denigrating the system and not trusting it, which was found in in the current study, may be caused by a violation of an errant mental model which cause the system to be perceived as less predictable.

Perceived Utility. The findings in the last theme corroborate earlier findings in the field of chatbot research, where previous studies have reported that time efficiency is an important motivational factor for engagement in chatbot communication (Følstad & Skjuve, 2019; Luger & Sellen, 2016). The present study is in line with previous research with regards to chatbot usefulness (Følstad & Skjuve, 2019). The mental model definition also pinpoints the need for individuals to understand the rationale for why a system exists (Rouse & Morris, 1986). However, the negative perceptions about usefulness among our participants indicates that users are integrating negative experiences in their mental models which can result in the system being considered to have low value and thus demotivate future use (Følstad & Skjuve, 2019).

In summary, the results of the present study indicate that both humanlike and chatbot mental models were available for participants during and post-interaction. The results are largely in line with previous findings, which suggest that themes that emerged in our research and other studies are important factors for understanding the mental models involved in chatbot interactions. Nevertheless, the author encourage a cautious interpretation of what has been uncovered so far, as many attributes of mental models may be implicit and unavailable for human articulation (Staggers & Norcio, 1993).

Second Stream: User's Language in Costumer Service Chatbots

Participants in the current study explicitly states distinct conversational tactics, of using human natural language and a keyword hypothesis. The analysis of the actual chatbot dialogue show that participants generally preferred the use of natural language and engaged less in keyword tactics. Demonstrating some contradictions between explicit thinking and actual behavior, as Knaeuper and Rouse (1985) found when studying problem-solving in human-machine interaction. The analysis also demonstrates the use of anthropomorphic communication by all participants in some or most of their interaction. The finding indicates what Nass and Moon (2000) call mindless application of social rules towards machines.

A few studies have previously looked at behavior in chatbot dialogue, all finding similar results as the current study (Allison, Luger, & Hofmann, 2017; Jenkins, Churchill, Cox, & Smith, 2007 ; Kopp, Gesellensetter, Krämer, & Wachsmuth, 2005; Liao et al., 2016). Research find the use of polite phrases, personal pronouns and anthropomorphic questions. Our results therefore contribute to the existing knowledge base, but also show that such anthropomorphic behavior occur 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 interested in the condition first"*
> 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 chatbot don't thinks about"*

Nass and Moon (2000) calls 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 gives further support for the notion that humanlike mental models were applied in the current study. Lee (2004) suggests that human cognition is "tricked" by evolutionary based dispositions and social scripts are therefore applied in this novel context. The explanation can also be related to the concept of overlearned social behavior, which are so ingrained and automatic that they occur without conscious attention (Nass & Moon, 2000). Use of natural language in chatbots may trigger such overlearned responses due to digitalization of human communication.

However, this do not mean that communicational acts with chatbots are identical to communication with humans for most of the adult population. One should be cautious when interpreting the present results, as it seems unlikely that the participants would have had identical communication with the customer service chatbots as with customer service chats operated by humans. This is supported by Hill et al. (2015) findings on human communication with chatbots and humans, in which the respective interactions exhibited a significant difference in many attributes. Therefore, a nuanced and plausible explanation is that people are generally aware of these distinctions and adapt accordingly without fully abandoning a humanlike mental model.

Third Stream: Situational Awareness in Costumer Service Chatbots

The third stream of results was generated from the deductive analysis, where themes found in the inductive analysis were placed in a SA-framework. SA was used in the present research to see if the distinct themes could be placed in the SA framework, and to consider if the framework could be used in a chatbot context for future design purposes.

SA has primarily been adopted in relation to complex socio-technological systems such as aviation, power plants and tactical systems. Nevertheless, Endsley (2008) argue that SA occurs even in less straining tasks. On one hand, chatbots can be viewed as a relatively simple system to use. There may not be a need to fully understand how chatbots works in order to use them. On the other hand, our study demonstrated that use of chatbots are not always straight-forward. The ability to interpret the intended meaning of messages were mixed, which is consistent with previous research (Kvale et al., 2019; Myers et al., 2018). Participants also expressed some confusion about how to operate the chatbot, as a trial and error approach were considered necessary. Endsley et al. (2003) argue that even simple systems and automation can impose operational complexity, especially when there is a need for joint cooperation to complete a task. When automated artifacts encounter a situation that they are not programmed to grasp, or the artifact fails to conduct a task, it becomes important that the user understand why complications occur.

Adoption of SA in a chatbot context were mixed. Two out of eight themes were deemed as not appropriate in the framework, and no main themes were placed under level 1 (Perception). This finding does not imply that users did not engage in perceptual processing but rather that all identified themes rely heavily on other levels. Endsley (2015) argue that SA is not a divided process of three distinct levels, but a mental operation that is highly interconnected. This is especially so for experts who have rich mental models to rely on. This makes a clear-cut separation into the three levels impossible at times (Endsley, 2015). Participants in the current study are not experts on chatbots per se. However, the participants have expertise in operating similar systems, as chatbots draw on similarities to social media communication between humans (Følstad & Brandtzæg, 2017). The remaining six themes were evaluated as appropriate. The SA model contributed to a more cohesive understanding of the current results by placing the themes in the framework. However, the placement of themes was challenging and will be further elaborated on in the section of theoretical implications.

User's Situational Awareness. Overall, the current study found two separate mental models. The current thesis proposes an overall dynamic where social cues in chatbots interface induce a mental model of humanlike content, almost in a habitual way. This model guide both comprehension of the software and projection of the chatbots abilities. During their interactions, participants revised and updated their comprehension of the system, mainly due to interaction difficulties which indicated that their mental model may have been errant. Chatbot knowledge became available for altered and more realistic understanding and projection. Nevertheless, social scripts associated with humanlike mental models seemed to guide their actual behavior. This was the case even if their system understanding changed, indicating a discrepancy between conscious analysis and concrete actions.

As such, it is highly probable that the proposed dynamic was non-linear. Meaning that humanlike and chatbot mental models were simultaneously available, generating a continuous shift and feedback loop at different levels in a dynamic interplay (Endsley, 2015). However, a linear model will be presented to make the results understandable. It should therefore be remembered that this is a somewhat simplified version of underlying cognitive dynamics. The next sections will describe the results in more depth.

Perception (Level 1). Chatbots are not a complex system to process perceptually, and participants had no problem detecting information presented in the interface. This was exhibited in the theme, *Design Attributes,* where participants evaluated the interface's overall usability. Also, content in one sub-theme in *Design Attributes* seemed relevant for constructing SA. Users had associations to Facebook Messenger, which may have been critical to induce a habitual schema. Themes placed under comprehension and projection further indicated which information that participants were attentive too when forming their understanding of the chatbots. The inductive results showed that participants commented on avatar, gender, and use of emoticons, which covered in *Human-Chatbot Cues*. Also, some attentional narrowing seemed to occur (Endsley et al., 2003). E.g., participants did not attend to chatbot B avatar. When asked about their perception, they reported that they noticed the avatar and were pleased that the avatar had an informational character that indicated a machine-like origin. Therefore, some informational cues in the interface may have been lacking salience to facilitate an initial comprehension (Endsley et al., 2003).

Additionally, participants articulated that other informational cues were important for their understanding. Fast answers and the use of alternatives were revealing with regards to the chatbot ontology. Other dialogue characteristics also received attention, especially if the answers were incongruent with participants prior messages. This were exhibited in *User-*

Chatbot Dialogue. Such information seemed to be important for their revised mental model and understanding of the chatbot, as well as what information they predicted that the system could provide them with. According to Endsley (1995b) model, new information will contribute in the formation of a more cohesive situational understanding.

Comprehension (Level 2). Endsley (1995b) argue that bottom-up stimuli are matched with the preliminary content of a mental model in an automatic fashion. The updated mental model will then guide system comprehension. As mental models contribute to the attribution of anthropomorphism and the feeling of social presence towards objects (Biocca, 1997; Epley et al., 2007) finding of such processes (see, *Chatbot Functionality* and *User Communication*) indicated which type of information the participants were using to understand the chatbots. However, as the dialogue progress from the first greeting, complications transpire for all participants. A lack of awareness may occur for "what the system is doing, and why it is doing it" (Endsley et al., 2003, p. 175). This is indicated in *User-Chatbot Dialogue*, where a loss of appropriate SA seemed to cause confusion about the chatbots answers. Previous research in aviation has shown that 18% of errors in SA can be traced back to mental models that are errant, incomplete or too heavily reliant on default values in subjects' mental model (Jones & Endsley, 1996). Such findings may extend to our study, as the participants humanlike model contribute to their SA and generated a mismatch between user action and chatbots capabilities.

Chatbots answers also facilitated understanding among participants that a tactic change was needed, and specifically that the use of keywords was useful (see, *User Communication*). Endsley (1995b) argue that such tactic change occurs when a subject becomes aware that their current understanding is incongruent with the situation. This were also exhibited in *User-Chatbot Dialogue,* as participants assumed that their own behaviour was incompatible with the system capabilities. Endsley (1995b) argues that new information in the environment causes the operator to form a more appropriate SA of the system, and in our study such information seem highly dependent on the chatbots behaviour.

A qualitative change in participant SA became more apparent as the dialogue progressed. Participants rejected anthropomorphism (see, *Human-Chatbot Cues* and *User Communication*). They also moved towards a system understanding that operated on keywords and pre-programmed answers (see, *Chatbot Functionality*). Even though their situational understanding changes, their actual behaviour seemed to be more consistent throughout the interaction. The language analysis indicates that participants were still relying on a humanlike mental model and social script. Endsley (2008) argue, however, that the

concepts of SA, decision making, and performance are not the same theoretical concepts. They operate in continuous interaction with each other, where SA often correlates with performance (Endsley, 1995a). Endsley (1995a) also found that errors in performance can occur even if the subject has an acceptable SA. Therefore, it is up to the user to understand when behavioural scripts are appropriate (Endsley et al., 2003). Nevertheless, subjects can have difficulties with suppressing such habitual behaviour even when the situation requires the subject to change action (Endsley, 1995b).

Projection (Level 3). According to Endsley (1995b) comprehension and mental models have the critical function of creating expectations, which in turn will support the projection of system behaviour. As with comprehension, inadequate mental models will also affect the accuracy of projection (Endsley et al., 2003). The initial high expectations towards the chatbots indicate that this is the case in the current study, where participants assumed that the chatbots were able to remember prior dialogue (see, *Chatbot Functionality*).

As their understanding changed, participants observed which massages the chatbot was able and not able to answer. Such observations often violated their more anthropomorphic initial understanding. Over time, the projection of the system's ability declined (see, *Chatbot Assumption*). Nevertheless, expectations outside banking terminology were never mentioned. Accurate projections are highly challenging to perform (Jones & Endsley, 1996) and the author presumes that the chatbots act of providing specified alternatives with labels of content contributed to their expectations. However, a lack of overall level 1 information may affect their ability to simulate the chatbots capability beside the current dialogue output (Endsley, 1995b).

Trust is not part of the original model of Endsley (1995b) and should be understood as both an affective and cognitive state in automated systems (Lee & See, 2004). A recent model has, however, integrated trust in a SA framework (Morita & Burns, 2014), supporting the placement of the theme under SA in the current evaluation. The content in *User Trust* also seem relevant to Endsley (1995b) description of the mechanism in projection. The participants had a system understanding (e.g., communication difficulties, chatbots lack expertise and intelligent behaviour), which generated a prediction of not receiving information or proper assistance. Endsley et al. (2003) argue that accurate mental models and SA of system behavior can decrease uncertainties, which indicate a need for future SA support in interface to overcome the lack of trust in chatbots.

In summary, the preceding discussion indicate that the SA-framework displays a moderate ability to account for our findings. Several themes could be placed the framework

and explained in light of SA theory. Even if the SA theory is complex and integrated with a broad range of research on cognition, the theory itself is simple and easy to understand. It consists of three levels structured in ascending progression (Endsley, 2015). This means that adequate perception at level 1 will support level 2 comprehension, which in turn will support better predictions as level 3. Attention to the needs at different levels could, therefore, support a more global SA and contribute to better interactions and build a more appropriate mental model of the system over time (Endsley, 2008).

Limitations

The next section will elaborate on relevant limitations of the present study.

Sample. 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 generalizability beyond the study context. Children, for example, are found to engage in higher anthropomorphic behavior towards voice-based chatbots than what was found in the current study (Druga et al., 2017). However, the present study needed some restrictions for practical feasibility and our research should be viewed as a foundation for future hypothesis testing on mental models and use of SA in chatbots.

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, Kaiser, & Marconi, 2017). The current sample size of 16 participants was therefore deemed sufficient, but a larger sample size may have contributed to additional meaning saturation in several sub-themes.

Procedure. There are several limitations related to how the study was carried out. First, the current study was conducted in a laboratory setting which may have influenced the results due to the environment being artificial and thereby reducing ecological validity. Second, there is also a risk that the task framed the participant's mental model and their problem-solving strategy Endsley (1995b), which could in turn effect their language output or assumptions regarding chatbot content. However, the task-wording (Appendix A) were constructed to not influence a specific interaction strategy. Also, previous research finds similar tendencies in language output (Luger & Sellen, 2016) and expectations towards

chatbots in more realistic environments (Følstad & Skjuve, 2019). Task were also evaluated as realistic by participants.

Lastly, chatbot A and 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 more humanlike due to priming from chatbot A. To reduce the risk of skewing the results based on priming, chatbots were presented in an alternated order. This should in principle correct for any systematic order effects that potentially could have biased the participants mental models.

Interviews. First, the use of interviews in general imposes a risk that participants engage in demand characteristics, providing answers they presume the researchers wish to hear (Orne, 1962; Rouse & Morris, 1986). However, interviews consisted of open-ended questions without leading the respondent, which tend to reduce this effect (Powell, Hughes-Scholes, & Sharman, 2012). Nonetheless, in a qualitative and explorative study of this kind there is still a risk that the author's behavior affected the results (Kvale & Brinkmann, 2015; Yardley, 2015).

It should be noted that Situation Awareness Global Assessment Technique (SAGAT) may yield more comprehensive insight into SA and mental models than the current procedure (Zhang et al., 2010). However, SAGAT were found to be difficult to conduct in customer service chatbots studies. As participants in the current study used an average of four minutes collectively with both chatbots interactions may me be too short for SAGAT methods, in which prompting SA questions are prompted after three to five minutes (Endsley et al., 2003).

Analysis. Researches bring their knowledge and subjectivity into coding, which may affect the overall results (Braun & Clarke, 2006). Additionally, no inter-reliability checks were conducted in the current thesis. This may affect the inductive, deductive and language analysis. For example, as it were difficult to place themes in the SA framework, other researches might have decided to consider SA inappropriate altogether or place themes differently within the framework. Nevertheless, inter-rater reliability is not a requirement in the thematic analysis approach. This approach emphasize that codes and themes should be continuously revised as new insights come to light and that research "subjectivity" can generate valuable insight (Braun & Clarke, 2013).

Implications and Future Research

Theoretical implications. While the findings from the inductive and language analyses corroborate existing knowledge on human-chatbot interaction, the study also contributes with some new insights. First, the study provides new knowledge on mental models in chatbot interaction by demonstrating how the participants seems to utilize two different mental models for understanding the chatbots and predicting its abilities: one that is similar to mental models used in human-human interaction, and one that is specific to human-chatbot interaction. The present study also identified a contradiction between expressed communicational tactic of using keywords and actual behavior which consisted of more natural language.

Second, no previous studies have used the SA-framework in the context of chatbots to such an extent. The current study provides indications that the SA-framework could be applied, as it was feasible for most emergent themes at level 2 and 3. This suggests that SA can add theoretical value to understand a user's cognition in a more cohesive manner. Nevertheless, the placement was not without challenges, indicating that SA theory may need future conceptualization for chatbot purposes. For SA to be appropriate, future research can conceptualize which elements that have relevance for building appropriate SA, such as social cues or dialogue. Correspondingly, to investigate and hypothesize a "correct" comprehension and projection, which correspond with chatbot functionality. Design solutions for appropriate SA and mental models may then become more evident.

Practical implications. The overall findings indicate that current design may underutilize perceptual information (e.g., how to communicate), or place salience at the wrong informational pieces (e.g., social cues). The present study propose that users should be made aware of potential issues before a message is sent to the chatbots, rather than getting feedback after a message is sent which is the current strategy by many text-based chatbot. Such feedback 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.

The following changes could be made. (1) If user's text does not match with specific textual content in the chatbot or has a statistically high probability of generating breakdowns in dialogue, a warning should be provided. Preferably visual, such as red sign in the user's typing window where perceptual attention is fixated. This could be a red line under specific word, as many individuals already are familiar with such system feedback in spelling corrections applications (E.g., Microsoft Word, SMS or Facebook Messenger). (2) When

visual signs (E.g., red line under words) are provided, the users could be given word substitutions. An alternative solution is to provide users with reframing example or a textual rationale for why their current message is not supported by the software system. Such information could be automatically presented visually in the interface when incongruent wording is written, rather than in the initial introduction of communication guidance that disappears in the dialogue feed. Such feedback will presumably modulate users' expectations to a more realistic level. Over time, such strategies may be less needed in the interface as users learn how to use the chatbots, or the software develops and supports more complex natural language.

Future research can test the proposed solution against the current chatbots in experimental research design. It should test (1) which visual signs that have enough salience to attract user's attention. (2) To see if the visual cues that are presented has an effect on behavior, e.g., if users attend to the information and modify their conversational tactic. (3) It should also be tested how information should be presented and explained, e.g., if word substitution is enough or if more elaborate explanations are necessary.

Final Conclusion

The present study found eight themes that exhibited an aspect of a user's mental model when interacting with a chatbot. The results indicated that a humanlike and chatbot mental model was used by the participants to describe, explain, and predict customer service chatbots. A language analysis of the dialogue between subjects and chatbots showed that conversational behavior was guided by a humanlike mental model, in some or most of their interactions. This occurred even if the participants expressed explicit rejection of anthropomorphism. Implementation of the SA-framework created a more cohesive understanding of the isolated themes and were found to be a feasible framework in a chatbot context. SA can therefore support future theoretical and practical guidance of chatbot development and may become particularly important if chatbots are implemented in more critical tasks or complex user-technology.

References

Adam, M., Wessel, M., & Benlian, A. (2020). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 1-19. doi:10.1007/s12525-020-00414-7

Allison, F., Luger, E., & Hofmann, K. (2017). *Spontaneous interactions with a virtually embodied intelligent assistant in Minecraft.* Paper presented at the Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver, CO.

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., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). *Resilient chatbots: Repair strategy preferences for conversational breakdowns.* Paper presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland.

Beattie, A., Edwards, A. P., & Edwards, C. (2020). A Bot and a Smile: Interpersonal Impressions of Chatbots and Humans Using Emoji in Computer-mediated Communication. *Communication Studies*, 1-19. doi:10.1080/10510974.2020.1725082

Beggiato, M., & 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, 18*, 47-57. doi:10.1016/j.trf.2012.12.006

Biocca, F. (1997). The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of computer-mediated communication, 3*(2), JCMC324. doi:10.1111/j.1083-6101.1997.tb00070.x

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101. doi:10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. London: Sage.

Breazeal, C. (2003). Toward sociable robots. *Robotics and autonomous systems, 42*(3-4), 167-175. doi:10.1016/S0921-8890(02)00373-1

Brennan, S. E., & Ohaeri, J. O. (1994). *Effects of message style on users' attributions toward agents.* Paper presented at the Conference companion on Human factors in computing systems, Bosten, MA.

Candello, H., Pinhanez, C., & Figueiredo, F. (2017). *Typefaces and the perception of humanness in natural language chatbots.* Paper presented at the Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*(8), 1839-1849. doi:10.1093/brain/awf189

Chen, M.-L., & Wang, H.-C. (2018). *How personal experience and technical knowledge affect using conversational agents.* Paper presented at the Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, Tokyo.

Clarke, V., Braun, V., & Hayfield, N. (2015). Thematic analysis. In J. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 222-248). Thousand Oaks: SAGE Publications Inc.

Coniam, D. (2014). The linguistic accuracy of chatbots: usability from an ESL perspective. *Text & Talk, 34*(5), 545-567. doi:10.1515/text-2014-0018

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in human behavior, 29*(3), 577-579. doi:10.1016/j.chb.2012.11.023

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

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331. doi:10.1037/xap0000092

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). *" Hey Google is it OK if I eat you?" Initial Explorations in Child-Agent Interaction.* Paper presented at the Proceedings of the 2017 Conference on Interaction Design and Children, Stanford, CA.

Endsley, M., R. (1995a). A taxonomy of situation awareness errors. In R. Fuller, N. Johnston, & N. McDonald (Eds.), *Human factors in aviation operations* (pp. 287-292). Aldershot: Avebury Aviation, Ashgate Publishing Ltd.

Endsley, M., R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human factors, 37*(1), 32-64. doi:10.1518/001872095779049543

Endsley, M., R. (2000). Situation models: An avenue to the modeling of mental models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 44*(1), 61-64. doi:10.1177/154193120004400117

Endsley, M., R. (2008). Theoretical underpinnings of situation awareness: A critical review. In M. Endsley, R & D. Garland, J (Eds.), *Situation awareness analysis and measurement* (Vol. 1, pp. 3-32). New York: CRC Press.

Endsley, M., R. (2015). Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making, 9*(1), 4-32. doi:10.1177/1555343415572631

Endsley, M., R, Bolte, B., & Jones, D., G. (2003). Designing for situation awareness: an approach to user-centered design. *Boca Raton, FL: CRC Oress*.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review, 114*(4), 864. doi:10.1037/0033-295X.114.4.864

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*: the MIT Press.

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *interactions, 24*(4), 38-42. doi:10.1145/3085558

Følstad, A., & Skjuve, M. (2019). *Chatbots for customer service: user experience and motivation.* Paper presented at the Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, Scotland.

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2019). *Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Understand Interaction Design.* Paper presented at the International Conference on Internet Science, St. Petersburg, Russia.

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication, 1*(1), 5. doi:10.30658/hmc.1.5

Gartner. (2018). Gartner Says 25 Percent of Customer Service Operations Will Use Virtual Customer Assistants by 2020 [Press release]. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2018-02-19-gartner-says-25-percent-of-customer-service-operations-will-use-virtual-customer-assistants-by-2020

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction. *Research Papers, 113*.

Grudin, J., & Jacques, R. (2019). *Chatbots, humbots, and the quest for artificial general intelligence.* Paper presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland.

Halasz, F. G., & Moran, T. P. (1983). *Mental models and problem solving in using a calculator.* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Boston, MA.

Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research, 27*(4), 591-608. doi:10.1177/1049732316665344

Heyselaar, E., & 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, Netherlands.

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. doi:10.1016/j.chb.2015.02.026

Hoff, T., Flakke, E., Larsen, A.-K., Lone, J. A., Bjørkli, C. A., & Bjørklund, R. A. (2009). On the validity of M-SWOT for innovation climate development. *Scandinavian Journal of Organizational Psychology, 1*(1), 3-11.

Jain, M., Kota, R., Kumar, P., & Patel, S. N. (2018). *Convey: Exploring the use of a context view for chatbots.* Paper presented at the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, QC

Jenkins, M.-C., Churchill, R., Cox, S., & Smith, D. (2007). *Analysis of user interaction with service oriented chatbot systems.* Paper presented at the International Conference on Human-Computer Interaction, Beijing, China

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive science, 4*(1), 71-115. doi:10.1016/S0364-0213(81)80005-5

Jones, D. G. (1997). Reducing situation awareness errors in air traffic control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 41*(1), 230-233. doi:10.1177/107118139704100152

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, space, and environmental medicine, 67*(6), 507-512.

Kempton, W. (1986). Two theories of home heat control. *Cognitive science, 10*(1), 75-90. doi:10.1207/s15516709cog1001_3

Klein, G., & Hoffman, R. R. (2008). Macrocognition, mental models, and cognitive task analysis methodology. In M. J. Scraagen, L. G. Militello, T. Ormerod, & R. Lipshitz (Eds.), *Naturalistic decision making and macrocognition* (pp. 57-80). Hampshire: Ashgate Publishing.

Knaeuper, A., & Rouse, W. B. (1985). A rule-based model of human problem-solving behavior in dynamic environments. *IEEE transactions on systems, man, and cybernetics, 15*(6), 708-719. doi:10.1109/Tsmc.1985.6313454

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (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.

Koro-Ljungberg, M., Douglas, E. P., Therriault, D., Malcolm, Z., & McNeill, N. (2013). Reconceptualizing and decentering think-aloud methodology in qualitative research. *Qualitative Research, 13*(6), 735-753. doi:10.1177/1468794112455040

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS one, 3*(7), e2597. doi:10.1371/journal.pone.0002597

Kvale, K., Sell, O. A., Hodnebrog, S., & Følstad, A. (2019). *Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues.* Paper presented at the International Workshop on Chatbot Research and Design, Amsterdam, Netherlands.

Kvale, S., & Brinkmann, S. (Ed.) (2015). *Det kvalitative forskningsintervju* (T. M. Anderssen., J. Rygge. Johan, Trans.). Oslo: Gyldendal akademisk.

Larivière, B., Bowen, D., Andreassen, T. W., Kunz, W., Sirianni, N. J., Voss, C., . . . De Keyser, A. (2017). "Service Encounter 2.0": An investigation into the roles of technology, employees and customers. *Journal of Business Research, 79*, 238-246. doi:10.1016/j.jbusres.2017.03.008

Lee, Lau, I. Y.-m., Kiesler, S., & Chiu, C.-Y. (2005). *Human mental models of humanoid robots.* Paper presented at the Proceedings of the 2005 IEEE international conference on robotics and automation, Barcelona, Spain.

Lee, J., D, & See, K., A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50-80. doi:10.1518/hfes.46.1.50_30392

Lee, K. M. (2004). Presence, explicated. *Communication theory, 14*(1), 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., Davis, M., Geyer, W., Muller, M., & Shami, N. S. (2016). *What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees.* Paper presented at the Proceedings of the 2016 ACM Conference on Designing Interactive Systems, Brisbane, QLD.

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. doi: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.* Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA.

Luria, M., Hoffman, G., & Zuckerman, O. (2017). *Comparing social robot, screen and voice interfaces for smart-home control.* Paper presented at the Proceedings of the 2017 CHI conference on human factors in computing systems, Denver, CO.

McDonnell, M., & Baxter, D. (2019). Chatbots and gender stereotyping. *Interacting with Computers, 31*(2), 116-121. doi:10.1093/iwc/iwz007

Meuter, M. L., Bitner, M. J., Ostrom, A. L., & Brown, S. W. (2005). Choosing among alternative service delivery modes: An investigation of customer trial of self-service technologies. *Journal of marketing, 69*(2), 61-83. doi:10.1509/jmkg.69.2.61.60759

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in cognitive sciences, 7*(3), 141-144. doi:10.1016/s1364-6613(03)00029-9

Morita, P. P., & Burns, C. M. (2014). Understanding 'interpersonal trust' from a human factors perspective: insights from situation awareness and the lens model. *Theoretical Issues in Ergonomics Science, 15*(1), 88-110. doi:10.1080/1463922X.2012.691184

Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018). *Patterns for how users overcome obstacles in voice user interfaces.* Paper presented at the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, QC.

Nass, C., Fogg, B., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45*(6), 669-678. doi:10.1006/ijhc.1996.0073

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues, 56*(1), 81-103. doi:10.1111/0022-4537.00153

Nass, C., Moon, Y., & Carney, P. (1999). Are respondents polite to computers? Social desirability and direct responses to computers. *Journal of applied social psychology, 29*(5), 1093-1110. doi:10.1111/j.1559-1816.1999.tb00142.x

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology, 27*(10), 864-876. doi:doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study. *Interacting with Computers, 31*(3), 317-335. doi:10.1093/iwc/iwz022

Norman, D. A (2013). *The design of everyday things: Revised and expanded edition*. New York: Basic books.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 7-14). New York: Psychology Press

Norman, D. A. (1999). Affordance, conventions, and design. *interactions, 6*(3), 38-43. doi:10.1145/301153.301168

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist, 17*(11), 776–783. doi:10.1037/h0043424

Parasuraman, A., & Colby, C. L. (2015). An updated and streamlined technology readiness index: TRI 2.0. *Journal of service research, 18*(1), 59-74. doi:10.1177/1094670514539730

Powell, M. B., Hughes-Scholes, C. H., & Sharman, S. J. (2012). Skill in interviewing reduces confirmation bias. *Journal of Investigative Psychology and Offender Profiling, 9*(2), 126-134. doi:10.1002/jip.1357

Preece, J., Rogers, Y., & Sharp, H. (2015). *Interaction Design: Beyond Human-Computer Interaction*. United Kingdom: John Wiley & Sons Inc.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences, 1*(4), 515-526. doi:10.1017/S0140525X00076512

Robb, D. A., Chiyah Garcia, F. J., Laskov, A., Liu, X., Patron, P., & Hastie, H. (2018). *Keep me in the loop: Increasing operator situation awareness through a conversational multimodal interface.* Paper presented at the Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO.

Robinson, O. C. (2014). Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative research in psychology, 11*(1), 25-41. doi:10.1080/14780887.2013.801543

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin, 100*(3), 349. doi:10.1037/0033-2909.100.3.349

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology, 1*(1), 45-57. doi:10.1207/s15327108ijap0101_4

Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research, 115*, 14-24. doi:10.1016/j.jbusres.2020.04.030

Simonsen, H. G. (2019, November 28). Setning. Retrieved from https://snl.no/setning

Skjuve, M., & 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., Haugstveit, I. M., Følstad, A., & Brandtzaeg, P. B. (2019). HELP! IS MY CHATBOT FALLING INTO THE UNCANNY VALLEY? AN EMPIRICAL STUDY OF USER EXPERIENCE IN HUMAN-CHATBOT INTERACTION. *Human Technology, 15*(1). doi:10.17011/ht/urn.201902201607

Staggers, N., & Norcio, A. F. (1993). Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies, 38*(4), 587-605. doi:10.1006/imms.1993.1028

Thakur, A. (2018, November 14). How virtual agents work and why you should care. Retrieved from https://www.boost.ai/articles/how-chatbots-work-and-why-you-should-care

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science, 211*(4481), 453-458. doi:10.1126/science.7455683

Tørresen, J. (2013). *Hva er kunstig intelligens*. Oslo: Universitetsforlaget.

Wagner, N., Hassanein, K., & Head, M. (2014). The impact of age on website usability. *Computers in human behavior, 37*, 270-282. doi:10.1016/j.chb.2014.05.003

Walters, M. L., Syrdal, D. S., Dautenhahn, K., te Boekhorst, R., & Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots, 24*(2), 159-178. doi:10.1007/s10514-007-9058-3

Warren, P. (2013). *Introducing psycholinguistics*. Cambridge: Cambridge University Press.

Warwick, K., & Shah, H. (2016). Can machines think? A report on Turing test experiments at the Royal Society. *Journal of experimental & Theoretical artificial Intelligence, 28*(6), 989-1007. doi:10.1080/0952813X.2015.1055826

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

Whiting, L. S. (2008). Semi-structured interviews: guidance for novice researchers. *Nursing standard, 22*(23), 35-41. doi:10.7748/ns2008.02.22.23.35.c6420

Wickens, C. D., Lee, J., Liu, Y. D., & Gordon-Becker, S. (2013). *Introduction to Human Factors Engineering: Pearson New International Edition*. Harlow: Pearson Higher Ed.

Williams, M., D,, Hollan, J., D, , & Stevens, A., L. (1983). Human Reasoning About a Simple Physical System. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 131-153). New York: Psychology Press

Yardley, L. (2015). Demonstrating validity in qualitative psychology. In J. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 258-272). Thousand Oaks: SAGE Publications Inc.

Zhang, T., Kaber, D., & Hsiang, S. (2010). Characterisation of mental models in a virtual reality-based multitasking scenario using measures of situation awareness. *Theoretical Issues in Ergonomics Science, 11*(1-2), 99-118. doi:10.1080/14639220903010027

Appendix A. Task provided for interaction purposes

Tenk deg at du skal kjøpe en ny leilighet eller bolig og ønsker å innhente informasjon rundt temaet. Eksempelvis informasjon om: boliglån, renter, kausjonistordning, og/eller egenkapital. Du velger å innhente informasjon ved bruk av to ulike kundeservicer Chatbots fra banker som du vurderer å søke lån hos.

[redacted] – Chatboten til [redacted]

[redacted] – Chatboten til [redacted]

Forestill deg at dette er ekte og at det derfor betyr noe for deg.

Generelle retningslinjer du må følge:
Ikke skriv personlig informasjon som navn og personnummer.
Gå tilbake til chatten om den sender deg til en informasjonsside.
Bruk Chatboten til å bli kjent med hva den kan hjelpe deg med angående teamet, og avslutt oppgaven når du føler at du har fått tilstrekkelig med informasjon.

Appendix B. Interview guide with theoretical explanation

Mental Model Attribute	Description of Relevance	Question Asked
Perception	Mental models are suggested to both guide which perceptual information that is attended to as well as affect which mental model that is activated in a current situation due to perceptual information (Endsley et al., 2003).	Hva er det med utseende til Chatbot [redacted] som du legger merke til?[1]
Associations	In the human-computer interaction field, it is generally recommended to design systems based on metaphors. This is done to supporting the development of a mental model for the current system (Wickens et al., 2013).	Hvilke andre systemer forbinder du med Chatbot [redacted]?
System understanding	System understanding is an important part of mental models, as the model help explain and understand the underlying structure of the artifact (Rouse & Morris, 1986; Wickens et al., 2013).	Hvordan tror du Chatbot [redacted] er konstruert? Tror du chatboten lærer?[1]
System prediction	System prediction is generated by the use of mental models. Mental model forms expectations of the system behavior where the user can predict what the system will do next. In the chatbot context, what the artifact can assist with (Rouse & Morris, 1986; Wickens et al., 2013).	Hva forventer du at Chatbot [redacted] kan hjelpe deg med? Noe Chatboten kunne gjort annerledes?[1]

Own behavior	System understanding and predictions give general guidance to the users' behavior, and the mental model will inform the user's for which action to take for task achievement (Wickens et al., 2013).	Hvordan må du kommunisere med Chatbot [redacted] for å få best mulig svar? Kan du beskrive hvordan det er å forholde seg til Chatbots sammenlignet med et menneske?
Expertise	Experts and novices have been described to have qualitative different mental models of a system. They can develop and change through experience (Endsley, 1995; Rouse & Morris, 1986).	Har forståelsen din endret seg gjennom bruken av Chatbot [redacted]? Og hvordan?
General experience	This question is not specifically related to mental models but allow the participant to give a general evaluation after task completion.	Hvordan opplevde du bruken av Chatbot [redacted]?[1]

Note. [1]Questions are asked between each chatbot interaction, and the unmarked questions are asked after task completion. Original question are provided in Question Asked.

Appendix C. Formal consent

Vil du delta i forskningsprosjektet «*Chatbots og mentale modeller*»?

Dette er et spørsmål til deg om å delta i et forskningsprosjekt om Chatbots og mentale modeller. I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formålet med prosjektet er å undersøke hvordan personer oppfatter og forstår Chatbots og hvilke forventninger man har til systemet. I dag har man lite kunnskap om slike mentale modeller i møte med teknologien. Datamaterialet vil danne grunnlag for Stine Ordemann sin masteroppgave ved Psykologisk Institutt, Universitetet i Oslo. Masterprosjektet har også inngått et samarbeid med SINTEF sin avdeling "Software and Service Innovation" som forsker på Chabots. Du har fått spørsmål om deltagelse da du representerer en vanlig forbruker som kan møte på Chatbots innen kundeservice. Hvis du velger å delta i prosjektet vil du bli spurt om å utføre en enkle oppgaver ved bruk av to ulike Chatbots. Før, under og etter samhandlingen vil det bli gjennomført intervju som vil bli tatt opp med båndopptaker, hvor du vil bli spurt om dine opplevelser i møte med Chatboten. Deltagelse vil ta ca. 60 minutter.

Det er frivillig å delta og du kan når som helst trekke deg fra undersøkelsen uten å oppgi informasjon om hvorfor.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger. Vi vil behandle informasjon og opplysninger om deg konfidensielt og i samsvar med personvernregleverket. Dialogen du har med Chatbotene og svarene du gir under intervjuet vil bli anonymisert og du vil ikke kunne gjenkjennes. Lydfilen som tas opp vil automatisk bli lagret på eget sikkert dataområdet hos universitetet i Oslo. Ingen andre enn mastergradsstudenten og prosjektansvarlig vil ha tilgang til lydopptak og intervjunotater. Skjermbildet av dialogen blir også tatt opp og slettet ved prosjektslutt. Prosjektet skal etter planen avsluttes juni 2020.

Hvem er ansvarlig for forskningsprosjektet og vil ha tilgang til datamaterialet? Følgende personer vil ha prosjektansvar og tilgang til datamaterialet: Stine Ordemann, masterstudent ved Arbeids- og Organisasjonspsykologi, Universitetet i Oslo. Cato Bjørkli, Førsteamanuensis ved Psykologisk Institutt, Universitet i Oslo. Marita Bjaaland Skjuve, Doktorgradsstipendiat, SINTEF, avdeling "Software and Service Innovation".

Dine rettigheter. Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke personopplysninger som er registrert om deg,
- å få rettet personopplysninger om deg,
- få slettet personopplysninger om deg,
- få utlevert en kopi av dine personopplysninger (dataportabilitet), og
- å sende klage til personvernombudet eller Datatilsynet om behandlingen av dine personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg? Vi behandler opplysninger om deg basert på ditt samtykke. På oppdrag fra Universitet i Oslo, psykologisk institutt har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Hvor kan jeg finne ut mer?
Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med:

- Student: Stine Ordemann. Telefonnummer: [redacted]. Epost: [redacted]
- Veileder: Cato Bjørkli. Telefonnummer: [redacted]. Epost: [redacted]
- Biveileder: Marita Bjaaland Skjuve. Telefonnummer: [redacted]. Epost: [redacted]
- Vårt personvernombud: [redacted] på e-post [redacted]
- NSD – Norsk senter for forskningsdata AS. Telefonnummer: [redacted]. Epost: [redacted]

Med vennlig hilsen

Cato Alexander Bjørkli
(Prosjektansvarlig og hovedansvarlig)

Stine Ordemann
(Mastergradsstudent)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet «Chatbots og mentale modeller» og har fått anledning til å stille spørsmål. Jeg samtykker til:

☐ å delta i intervju om mentale modeller og Chatbots.

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet, ca. Juni 2020.

(Signert av prosjektdeltaker, dato)

Appendix D. Translated statements from the thematic analysis

Theme	Translated Statement	Original statement
Human-Chatbot Cues	She appears more as a human. Also, the way she opens (the conversation) with an emoji and "hi" (P 1)	Men hun fremstår som mer menneske. Også litt måten hun åpner på med smileys og "hei på deg" (P 1)
	You also get a slight feeling of talking to a person (P 6)	Du føler jo også litt at du snakker med en person på en måte (P 6)
	When you chat with a person, they would answer with text. Not alternatives (P 11)	Når man chatter med en person så ville jo de svart med tekst Ikke med spørsmål med alternativer (P 11)
	I like that it is a cute robot, it might be more encouraging to treat it as such. More as a machine and less as a person (P 7)	Likte at det var en sånn søt robot og sånn. Men det blir kanskje mer oppfordret da til å behandle det som en slags, mer som et dataprogram og mindre som en person (P 7)
User-Chatbot Dialogue	You start to see a pattern of what it (chatbot) can answer (P 13)	Nei det er jo at man begynner å se mønster i hva den kan svare på og når den sender meg til denne lånesiden (P 13)
	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)	Selv om man kanskje ikke får svar så får man jo mye nyttig informasjon som man kanskje ikke har tenkt på tidligere (P 6)
	Know I don't understand what it (chatbot) did. I asked how much mortgage I could get with an annual income of 600 000 (KR). Then it (chatbot) asked if it was regarding a new or existing loan. Then I wrote new, and it asked if I was renovating or moving my mortgage (P 5)	Nei nå skjønte jeg ikke helt hva den gjorde her. Jeg bare spurte hvor mye boliglån jeg kunne fått på en årsinntekt på rundt 600000 og så spør den om det gjelder et nytt eller eksisterende lån. Og så skrev jeg nytt og så begynte den å spørre om jeg skal pusse opp eller flytte boliglånet (P 5)
	I need to rephrase in a way that the chatbot understand (...), and it is almost more complicated in a bot (chatbot) than in a search engine (P 10)	Må omformulere meg på en måte som boten skjønner (...) og det er nesten mere komplisert i en bot sammenheng enn i et søkefelt (P 10)
User Communication	Such chat format creates the expectation that I can formulate myself as I would do in a chat (with humans) (P 4)	Altså det at det er en sånn chatt format skaper en slags forventing fra min side om at her kan jeg formulere meg sånn som jeg ville gjort i en chatt (P 4)
	I tried at one point to write: requirements for interest rent, and it did note understand. Then I wrote: interest rate and I got it (the answer) (P 16)	Jeg prøvde det på et tidspunkt og skrive krav til hva slags rentenivå. Da fikk jeg beskjed at "jeg forstår ikke spørsmålet ditt". Og da skrev jeg bare rentenivå og da fikk jeg (P 16)

	This is the reason why google is such a good search engine. You can write utterly wrong, and it still find articles (P 9)	Fordi dette er grunnen til at google er så bra søkemotor for du kan skrive ting helt. Veldig feil og fortsatt så finner den ting (P 9)
	I feel I need to have a lot of insight (to communicate), and write the correct keyword (P 4)	for å få best mulig svar så føler jeg at når det gjelder bolig lån og sånn så føle jeg egentlig må ha endel innblikk selv fra før og så må du nesten rett og slett skrive det mer sentrale nøkkelordet (P 4)
	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)	Jeg vet jo ikke så mye om det her og da vet jeg kanskje heller ikke hva jeg skal spørre om, og da er det veldig hjelpsomt at den spør for meg på en måte for å fiske etter riktig svar (P 3)
Chatbot Functionality	If I previously pressed (the alternative) that I am between 18 and 34, then she (Chatbot A) should think that. 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)	Vist jeg trykket tidligere at jeg er mellom 18 og 34, så bør hun jo tenke at. Jeg har jo ikke feiret bursdagen min i mellomtiden. Jeg er jo fortsatt i samme aldersgruppe og det burde den jo huske da (P 14)
	I was a bit disappointed, I thought it was smarter (P 13)	Jeg ble litt skuffet. Tenkte den var litt smartere (P 13)
	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)	Ja at de trekker ut stikkord og at de ikke ser så mye hvordan setningen er. Hvordan man stille spørsmål på og at, men samtidig er det litt rart. For den ene den der [redacted] var veldig sånn at den spesifiserte at man skal være konkret. Men kanskje det bare er for at det skal være lettere for den å se hva som er relevant da (P 8)
Chatbot Assumptions	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)	Jeg blir litt sånn blind på svaret for jeg regner med å få det samme som tidligere så jeg glemte å lese ordentlig (P 10)
	I would expect that regular opening hours would be inside, and I would ask about that. But if I have a more advanced question (...). It is difficult to come up with an example, but I would not ask if the beer sales had regular sales hours on May 17 (Constitution Day) (P 14)	Vanlig åpningstider ville jeg forventet at ligger inne og det ville jeg spurt om. Men har jeg et mer sånn avansert spørsmål (...). Nå kommer jeg ikke på noen eksempler da men er ølsalget åpent til vanligtid 16 mai dagen før eller 17 mai (P 14)
	Right now, I think it (chatbot) can deal with the absolute simplest things and bank related questions (P 6)	Akkurat nå så tror jeg at den bare forholder seg til det aller enkleste og kun bank relaterte spørsmål egentlig (P 6)
Users Trust	I think that, maybe not dangerous, but if you are uncertain of what you are looking for (…), then you really need to hunt and find what you need to ask about (P 15)	Jeg tror at det eneste som er, ikke farlig, men litt sånn med chatbot er at når man selv er usikker etter hva man ser (...), da må du virkelig jakte og prøve å finne ut av hva du selv spør etter (P 15)

	I feel frightful of losing information (…), when I don't get the full informational picture (P 12)	jeg føler at jeg blir redd for å gå glipp av informasjon (...), når jeg ikke får fult informasjonsbildet (P 12)
	"They (humans) will listen to your intonation and your demeanor, what your question really is. This (chatbot) would not, they will only look at what you wrote" (P 16)	For de vil gjerne høre på ditt tonefall og din væremåte hva du egentlig lurer på. Det vil jo ikke den her. Den vil jo bare se på hva du skrev på og hvordan du formulerte deg (P 16)
Perceived Utility	Because I need to start searching around (for information), and then the purpose of the chatbot disappears. Because it's (chatbots) supposed to be quick access to information (P 2)	For da må jeg jo begynne å lete rundt og da forsvinner jo hensikten med en chatbot. For det skal jo være en rask tilgang på spørsmål (P 2)
	I am able to use Google. The need to then go to the homepage to use a somewhat advanced search engine seems meaningless" (P 12)	Jeg klarer å bruke google og liksom det å måtte gå inn på hjemmesiden bare for å bruke en, hva skal man si, litt små avansert søkefunksjon. Det virker meningsløst (P 12)
Design Attributes	Most (people) are probably using Facebook chat once a day, and it is a familiar format that it is easy to use (P 11)	De fleste er jo sikkert inne på den Facebook chatten iallfall engang om dagen. Og at det er et kjent format som er lettere å bruke da (P 11)
	I feel that it (chatbot interaction) can be a bit frustrating. It is like talking to a person that don't understand (P 8)	Så jeg føler at det kan være litt sånn frustrerende. Eller det er litt som å prate med en person som ikke skjønner (P 8)

Ex. in text	**Translated Statement**	**Original statement**
Level 1	wow, wow, wow (…) that was a nice avatar (P 3)	Oi oi oi (…) Var fin avtar (P 3)
Language in Costumer Service Chatbots	I am interested in the condition first. "Sorry, I do not understand your question". Maybe the word "first" is difficult for the chatbot to understand. Also, that I am "interested" is something that chatbot don't thinks about. (P 10)	Jeg er interessert i betingelsene først. "Jeg forstår ikke spørsmålet ditt". Det kanskje er dette ordet «først» som er vanskelig for en bot å forstå. Altså at jeg er «interessert i» er kanskje ikke noe en bot tenker på heller (P 10)

Appendix E. Translated statements from the language analysis

Language Categories	**Translated Statement**	**Original statement**
Sentences	what do you need to know when applying for a mortgage? (P 5)	hva trenger man å vite når man skal søke boliglån? (P 5)
	If taking out a mortgage for a residence worth 5 million, how much equity is needed? (P 6)	Dersom man tar opp boliglån for et bolig på 5 millioner, hvor mye egenkapital trenger man? (P 6)

	My partner bought an apartment last year (P 4)	Samboeren min kjøpte leilighet i fjor (P 4)
	do bsu account as the same as other equity (P 14)	teller bsu likt som annen egenkapital (P 14)
Keywords Tactic	Information on mail (P 9)	Informasjon på mail (P 9)
	Loan (P 16)	Lån (P 16)
	equity (P 13)	egenkapital (P 13)
	will refinance mortgage. (P 10)	vil refinansiere boliglån. (P 10)
Personal Pronouns	How much equity do you demand with a mortgage? (P 7)	Hvor mye egenkapital krever dere ved boliglån? (P 7)
	Do I need a permanent job to get my first mortgage? (P 8)	Må jeg ha fast jobb for å få førstehjemslån? (P 8)
	Do I get more loan if I have a guarantor? (P 1)	Får jeg mer lån om jeg har kausjonist? (P 1)
	do I have any benefits as a first-time buyer with regard to mortgage? (P 2)	har jeg noen fordeler som førstegangskjøper med tanke på boliglån? (P 2)
Polite Remarks	hi, I have questions about mortgage (P 11)	hei, jeg har spørsmål om boliglån (P 11)
	Have a nice day :) (P 13)	Ha en fin dag :) (P 13)
	Hi! (P 15)	Hei! (P 15)
	thanks for the help Chatbot A (P 12)	takk for hjelpen [redacted] (P 12)