Humanlike Customer Service Chatbots:

The effect of humanness on user experience and user behavior

Alexander Jakobsen

Master thesis

University of Oslo

Department of Psychology

May, 2021

Author: Alexander Jakobsen

Title: Humanlike Customer Service Chatbots: The Effect of Humanness on User Experience and User Behavior.

Supervisors: Knut-Inge Fostervold (UiO) and Asbjørn Følstad (SINTEF)

Abstract

Utilizing **c**hatbots in customer service is becoming more prominent due to their potential to cut costs for companies while provide immediate assistance to customers. Understanding user experience associated with customer service chatbot interactions is essential in order to provide customers with productive and satisfying interactions. Designing chatbots with humanlike features is a common method for increasing user experience. However, there is currently a knowledge gap regarding how user experience and behavior is impacted by such human likeness in chatbots.

The current study involved an experiment (N = 120) to test the effect of humanness in a chatbot avatar and dialogue on user experience and user behavior. Participants interacted with a chatbot in a 2 (avatar humanness low/high) x 2(dialogue humanness low/high) randomized factorial design. Results revealed that participants interacting with the chatbot with high humanness dialogue had increased levels of mindless anthropomorphism and mindful anthropomorphism and also experienced the chatbot as more socially present. However, contrary to expectation, humanness had no effect on the hedonic quality of user experience. Furthermore, participants interacting with the high humanness dialogue were more inclined to behave socially towards the chatbot. These results suggest a greater focus on user behavior is needed. Limitations and suggestions for future research are discussed.

Keywords*: Chatbots, Humanness, Customer Service, User Experience, User Behavior, Experimental Study.*

Acknowledgements

The decision to study chatbots was based upon a previous collaborative project between SINTEF and UiO, in which I had the pleasure of working with a talented team of researchers which inspired me to pursue a study of chatbots myself. Being able to collaborate with SINTEF has been a truly valuable experience and has greatly increased my desire to conduct research in the future.

I would first like to extend a sincere gratitude towards Asbjørn Følstad from SINTEF. Your genuine enthusiasm, combined with your seemingly endless knowledge is truly inspiring. This thesis would not be possible had it not been for your continuous guidance, support, and never-ending patience. Thank you!

I would also extend my gratitude towards Knut-Inge Fostervold from UiO, a wizard of statistical knowledge. Your support through sharing knowledge and digressional pirate-stories have helped me more than you can imagine.

I would also like to thank Boost.ai for providing the chatbot used in this study.

A special thank you goes out to my family for supporting me through this period, especially during a pandemic.

Finally, I would like to thank Erika D. Simanjuntak, for being an endless source of emotional support.

Alexander Jakobsen

May, 2021

Table of Contents

List of Tables and Figures

Introduction

Recent years have seen a surge in businesses implementing chatbots in customer service. A chatbot can be defined as "software that accepts natural language as an input and generates natural language as an output, engaging in conversation with the user" (Griol, Carbó, & Molina, 2013). The use of chatbots compared to employees allows companies to address a larger number of customers without the costs associated with employees such as salary, office space or healthcare. For instance, using chatbots, Autodesk was able to cut response time by 99%, from 38 hours to 5.4 minutes for simple inquires (Reddy, 2017). In the banking and healthcare sector, an average of more than 4 minutes per chatbot inquiry for customer service was saved compared to human operators (Smartmessage, 2019). These successful business anecdotes illustrate some of the advantages of chatbots in customer service. As one of the main motivators for people to use chatbots is productivity, chatbots are able to provide fast and efficient assistance to customers, which in turn increases user experience (Brandtzaeg & Følstad, 2017).

As users have been found to typically respond more favorably to chatbots exhibiting human characteristics, much research has been dedicated to make chatbots feel as "human" as possible, (Chaves & Gerosa, 2019). Research shows that different manipulations of a chatbots humanness have found to impact users perceptions of the chatbot as measured through self-reported measures (Go & Sundar, 2019). Araujo (2018) found that increased humanness in a customer service chatbot increased users self-reported emotional connection to the chatbots company. Smestad and Volden (2018) demonstrated how chatbots with different personality traits influenced users self-reported perceptions of user experience. However, while previous work has addressed how humanness in a chatbot affect users self-reported user experience, there is a lack of knowledge as to how humanness in chatbots impact users' behavior. Analyzing conversational behavior is insufficiently studied in current research on chatbots. It is important to not only study what users report, but also how they behave to properly increase the user experience in chatbots. This is an important knowledge need that needs to be addressed to design the next generation of chatbots that are able to strengthen user experience. To guide the design of future chatbots for customer service, it is important to advance current knowledge about how humanness in chatbots impact user experience and, in particular, users' interactions with chatbots.

The aim of this paper was to address the effect of humanness in customer service chatbots on user experience and user behavior. To do this, a chatbot was created to be used as a customer service solution in a fictious company. The chatbot had four conditions, allowing for a 2 (avatar humanness low vs high) x 2 (conversational humanness low vs high) between-subjects design. Participants completed a questionnaire to measure their perception of the chatbot and user experience, while user behavior was investigated based on the user dialogue with the chatbot.

This study will add to the existing literature regarding the effect of humanness in chatbots on user experience. In particular, the study sheds new light on how user experience is affected by differences in humanness in chatbot avatar and dialogue. This insight will be valuable for future chatbot design and research. Furthermore, the study contributes new knowledge in the field of user behavior. Understanding how humanness in a chatbot affects the behavior of users enables the design of chatbots more suited to users' needs and preferences and paves the way for future research on user behavior in chatbot interactions.

The structure of this thesis is as follows. First, an overview of the background related to chatbots, user experience, user behavior, humanness, anthropomorphism, and social presence is provided, Followed by a presentation of research questions and hypotheses. Then, the methodology, procedure, materials used and information regarding the experiment and participants will be presented, followed by a description of the study findings. Lastly, the study findings are discussed along with the limitations and suggestions for future research.

Background information

Chatbots and customer service

ELIZA, a "program which makes natural language conversation with a computer possible" (Weizenbaum, 1966, p. 36) could be classified as one of the earliest chatbots developed. Since then, chatbots have been developed and implemented in several fields such as healthcare (Laranjo et al., 2018) and education, (Hobert & Meyer von Wolff, 2019; Krassmann, Paz, Silveira, Tarouco, & Bercht, 2018), private sector such as a personal assistant for new employees in a business, (Liao, Davis, Geyer, Muller, & Shami, 2016; Liao et al., 2018) or as customer service agents (Araujo, 2018; Go & Sundar, 2019; Koetter et al., 2018). In addition, different categories of chatbots has been developed such as voice activated dialogue systems (e.g. Siri or Alexa), (Cho, Lee, & Lee, 2019) and embodied conversational agents (Bergmann,

Eyssel, & Kopp, 2012; Kopp, Gesellensetter, Krämer, & Wachsmuth, 2005). Chatbots are well suited for the role as a customer service agent, in part due to their easy adaptation into already existing technology such as messaging apps installed in smartphones and tablets. More importantly, due to the ability to cut cost for companies, while at the same time provide immediate assistance to customer whenever needed has resulted in an increased interest in how to improve customer service chatbots.

Customer service is considered a task-oriented environment in which efficiency and productivity is highly valued (Goldstein, Johnston, Duffy, & Rao, 2002). Due to user's prioritization of goal-oriented behavior such as acquiring information or assistance, customer service chatbots are typically designed first and foremost with efficiency in mind (Følstad & Skjuve, 2019). Also, according to a study by Brandtzaeg and Følstad (2017) on peoples motivations for using chatbots, 68% of the participants cited productivity as the main motivator. For instance, Cui et al. (2017) demonstrated the usefulness of an e-commerce chatbot assisting customers during online shopping by utilizing the information from both the webpage the users visited and the users input, greatly increasing online shopping efficiency while reducing the workload of human customer service personnel.

Typically, customer service chatbots rely on users initiating the dialogue and try to interpret user intent and match the text by a predefined matching answer (Kvale, Sell, Hodnebrog, & Følstad, 2019). As such, some customer service chatbots utilize predefined options in the form of buttons, or hyperlinks for the user to *click*, instead of engaging in a regular conversation with the chatbot in order to save time (Jain, Kumar, Kota, & Patel, 2018). Although the use of buttons enables a more effortless interaction, it comes at the cost of reducing the natural conversational feeling provided by free-text conversations. Indeed, some users routinely prefer a more humanlike interaction associated with social cues presented in chatbots, and are more likely to engage in social chit-chat, or elicit polite behavior (e.g., "thank you" or greeting the chatbot) towards the chatbot (Følstad & Brandtzaeg, 2020; Følstad & Skjuve, 2019; Liao et al., 2016). In addition, as chatbots are designed to interact with users through natural language, substantial research has been done to design customer service chatbots capable of interacting more humanlike. Such as interacting using informal language (Araujo, 2018), message interactivity (Go & Sundar, 2019), social chatbots able to cooperate with the user (Gnewuch, Morana, & Maedche, 2017) or showing empathy (Xu, Liu, Guo, Sinha, & Akkiraju, 2017).

However, Følstad and Skjuve (2019) study of user experience in customer service chatbots suggests that while humanness in chatbots could have a positive effect on user experience, most participants reported features related to productivity as most important for the chatbot interaction. In sum, several strategies have been utilized to satisfy the needs of users associated with both productivity and need for human interaction in customer service chatbots.

User experience and user behavior

As shown by Brandtzaeg and Følstad (2017), people use chatbots for a variety of reasons such as entertainment, curiosity, or productivity. However, this multifaceted motivational drive for using chatbots creates a definition of user experience that is difficult to capture as the user experience depends to what degree the chatbot fulfil the motivations of the user (Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). For instance, someone interacting with a purely task-oriented customer service chatbot to solve inquiries regarding a product or service will likely have a vastly different experience compared to someone interacting with the same chatbot to be entertained. The International Organization for Standardization defines user experience as "…person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" (ISO, 2019). User experience is also dynamic and changing. As shown by Luger and Sellen (2016) in their study of user expectations of conversational agents in households, after initially testing and experimenting with its capabilities, user experience quickly changed. While users initially focused on playful interactions for entertainment purposes, it gradually changed to goal-oriented tasks focusing on efficiency after interacting with the conversational agent and discovering its limitations. A recent study by Følstad and Brandtzaeg (2020) explored user experience of chatbots based on the *hedonic* and *pragmatic* qualities of user experience, a holistic model of user experience by Hassenzahl (2018). By interviewing over 200 participants regarding their positive and negative experiences using chatbots in the past, Følstad and Brandtzaeg (2020) categorized participants answers based on the hedonic/pragmatic framework. Results indicated that users' appreciation of chatbot attributes varied greatly, with 42% reported positive experiences relating to providing help/assistance or finding information, while 36% of the participants reported positive attributes such as the chatbot being entertaining or inspiring. Hassenzahl's (2018) hedonic- pragmatic user experience addresses the broad differences in user experience by categorizing the experiences as either pragmatic or hedonic. According to Hassenzahl (2018), one of the primary uses of interactive technology is to

manipulate the surroundings. However, a product requires functionality and usability in order to achieve any manipulation. These aspects of user experience are called pragmatic qualities, and in the context of customer service chatbots would be how straightforward it is to use, how fast it provides answers to the user requests or how easy it is to communicate with. In addition, pragmatic qualities are mainly used to achieve behavioral goals of the user, either internally driven or provided to the user externally. If for instance a customer interacted with a banks chatbot and wanted to know the location of the nearest bank, the chatbot would be evaluated based on how efficiently the chatbot fulfilled the behavior goal of the customer.

Not all users interact with chatbots purely to fulfill a behavioral goal. For someone interacting with a chatbot due to it being *interesting* or *fun*, pragmatic attributes does not necessarily matter. Hedonic attributes are independent of pragmatic attributes, and emphasis individuals' psychological well-being. As such, users will be more likely to continue interacting with a chatbot the user find enjoyable, even if its behavioral goal is already satisfied. On the other hand, a chatbot strong on pragmatic attributes but weak on hedonic attributes have a high appeal to the user due to the behavioral goals of the user. For instance, in their study of chatbot user experience, Følstad and Brandtzaeg (2020) found that over one-third of the participants reported hedonic qualities in a chatbot important, with entertainment (29% of participants) being considered the most important hedonic quality. Although hedonic qualities are generally not considered important for behavioral goals, research suggests some users find hedonic qualities in task-oriented chatbots, such as customer service chatbots, important for their user experience (Følstad & Skjuve, 2019). A more humanlike customer service chatbot could increase familiarity making the chatbot more pleasant and trustworthy while at the same time be productive (Følstad, Nordheim, & Bjørkli, 2018; Go & Sundar, 2019). Indeed, as Hassenzahl (2018) argues, as pragmatic and hedonic qualities are independent and not mutually exclusive, a customer service chatbot can solve user inquires while at the same time be fun and interesting. Ideally, this would be considered a desired customer service chatbot, solving user inquiries in an efficient and productive manner, while at the same time interacting with the user similar to a human customer service agent, thus maximizing user experience. However, rarely is a perfect balance achieved.

Chatbot user experience depends on the interaction between the user and the chatbot, that is, the chatbot dialogue. Hence, it is important to understand how this dialogue impacts user experience. For example, Liebrecht and van Hooijdonk (2019) argues communication techniques

used in online web care conversations could be used in chatbot design as they share similarities such as being private and being online interactions. Conversational human voice techniques such as personal greeting, using second person pronouns and stimulating dialogue (e.g. "how may I help you?") has been found to increase perception of a more personal, natural and engaging conversation (Liebrecht, Tsaousi, & van Hooijdonk, 2021). Furthermore, using several techniques within the same conversation was found to have a larger positive effect on perceived personalization than only using one (Liebrecht et al., 2021).

User behavior

Research on chatbot user experience has largely focused on self-reported measures relating to users' perceptions of chatbots, often neglecting user's *responses.* However, to accurately understand the effect of chatbots dialogue on user experience, it is important to understand the users' responses, or behavior when interacting with chatbots. An important part of chatbot interaction is the sequential nature of conversations in which each participant of the conversation takes turns interacting. Similarly to a human conversation, conversations with chatbots follows a similar structure where an answer to an inquiry is expected, or a greeting following a greeting, referred to as action pair sequences (Moore, 2018). Based on observational science, Moore (2018) presents a natural conversational framework in which chatbot interactions are designed based on natural human conversations. The framework suggests that the smallest conversations are comprised of minimum one sequence. The end of a sequence is achieved when both the chatbot and the user reach a mutual understanding and close the sequence (e.g., saying "goodbye"). Longer sequences on the other hand, are comprised of either several completed small sequences, or sequences containing behaviors such as repeating utterances or paraphrasing (Moore, 2018). As the natural conversation framework is based on expandable sequences, its effectiveness and efficiency are often measured based on 1) The number of sequences that were initiated by the user or chatbot, where user-initiated sequences are an indicator of user engagement. 2) Chatbot success rate is measured by the percentage of sequences completed by the user or the chatbot, and 3) interactional efficiency, which is a measure of how much "work", such as paraphrasing or repeating, the user or chatbot must do to complete a sequence (Moore, 2018).

In addition to the natural conversation framework, other measures of user behavior have been proposed. For instance, in a comparison of human conversations and chatbot conversations

Hill, Ford, and Farreras (2015) noted that while conversations with chatbots was significantly longer that human conversations, they were shorter in terms of words per conversations, words per message and were considerably lacking in vocabulary depth. Indeed, similar results were reported by Lortie and Guitton (2011) as they explored why some humans in the Loebner Prize (a recent version of the Turing Test) were judged to be machines. They found that people judged to be machines used significantly less words per message, similarly to machines in the test. Hill et al. (2015) argues that this difference is at least in part due to humans mirroring the conversational style of their partner (as when talking to a child).

According to the theory of agent orientation, a user either view a chatbot as a sociable or utilitarian tool, and behave accordingly (Liao, Geyer, Muller, & Khazaen, 2020). For instance, social agent oriented users, or social behavioral oriented users, defined as "the preference for humanized social interactions with an agent interface, such as having natural conversations and social dialogues (Liao et al. (2016, p. 265) are more likely to engage in social chit-chat with the chatbot (Liao et al., 2018). Users engaging in social chit-chat typically ask the chatbot socializing questions (e.g., "what is your favorite color") which are unrelated to the tasks the chatbot was designed for, yet typical for regular human interactions (Liao et al., 2016). In addition, social behavior oriented users are more likely to use politeness (e.g., "thank you"), greetings (e.g., "good morning") and farewells (e.g., "goodbye") when interacting with a chatbot (Liao et al., 2016).

While some users prefer to interact with highly humanlike chatbots, capable of imitating a human interaction using proper turn taking and engage in social chit-chat. Some users consider humanlike features in a chatbot to be unnecessary, and instead prefer a design more similar to a traditional search system. Liao and colleagues (2020) refer to users on the opposite side of the social agent orientation scale as utilitarian agent oriented, or having a utilitarian *behavioral* orientation, favoring chatbot features such as ranked list of answers and query-like input. For instance, utilitarian behavioral oriented users are less likely to engage in casual testing of the chatbots abilities than social behavioral oriented users. Liao and colleagues (2020) argue this difference is due to less curiosity regarding the intelligence of the chatbot, as utilitarian behavioral oriented users think of the chatbot more as a regular information searching system.

By understanding how users behave during interactions with chatbots, it is possible to design chatbots more suitable for users. For instance, users more inclined to behave socially

would greatly benefit from interacting with chatbots incorporate greetings in the conversation and stimulate dialogue. On the other hand, utilitarian behavioral oriented users would instead prefer a chatbot devoted of social-chit chat and assist the user as efficient as possible. However, knowing the behavioral orientation of a user prior to interacting with a chatbot proves problematic. Liao and colleagues (2020) suggests users be able to customize the way the chatbots in the workplace interact, to align with the individual users' behavioral orientation. However, while user customizable chatbots are feasible in an environment where the user repeatedly interacts with the same chatbot over a long period of time. Chatbot interactions in customer service rarely lasts long and thus impractical for the user to spend time customizing the chatbot prior to a short interaction.

Humanness and anthropomorphism

The addition of humanlike features to the chatbot facilitate a more natural conversational feeling. Thus, chatbots are often designed with a plethora of humanlike features in order to increase its *humanness*. A chatbots humanness is defined as "the extent to which an agent is designed to act and appear human… encompassing the objectively established human capabilities such as having eyes, face, or the ability to respond politely" (Meyer, Miller, Hancock, de Visser, & Dorneich, 2016, p. 281). In other words, a chatbot can be designed to be more human either visually, or behaviorally. Although visual humanness in chatbots is limited to a small avatar, the effect of continuous exposure to the chatbot avatar during interactions should not be disregarded (Nowak & Rauh, 2005). For instance, Go and Sundar (2019) found that an avatar high on humanness (picture of a woman vs speech bubble) compensated for chatbots with impersonal conversations. In addition, (McDonnell & Baxter, 2019) reported a gender bias in chatbots deployed in gender stereotypical environment (banking vs mechanic), where male chatbots were rated higher on satisfaction that female chatbots in a mechanic domain. Although no effect for gender were found for non-gender stereotypical domain such as banking, participants preferred the non-gender chatbots in both conditions. However, research regarding visual humanness should be research further.

Similarly to visual humanness, research suggests people prefer to interact with chatbots that behave similarly to that of a human conversation. Researchers has been able to express conversational humanness in chatbots in several different ways, such as the ability to express humor (Niculescu & Banchs, 2019), empathy (Zhou, Gao, Li, & Shum, 2020), or different types

of personality (Mairesse & Walker, 2009). For instance, by increasing message interactivity, Go and Sundar (2019) found that participants interacting with a customer service chatbot felt more like they were interacting with a real person compared to a chatbot with low message interactivity. In addition, research indicates that informal customer service chatbots, interacting using informal language and having a human name are perceived as more friendly and likeable (Araujo, 2018). Informal chatbots utilizing a modest amount of emojis have also been found to be rated similarly in social attractiveness to that of human conversations (Beattie, Edwards, & Edwards, 2020). However, making the chatbot appear too humanlike could lead to an uncanny valley effect, an increased feeling of eeriness when interacting with technology that exhibiting high humanlike traits, though this is likely more relevant for embodied conversational agents than chatbots (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2019). Humanness has also been reported to increase user expectancy of interactivity, in which users overestimate the abilities of the chatbot and subsequently evaluate it worse than if interacting with a more machinelike chatbot (Go & Sundar, 2019).

Anthropomorphism

By increasing humanness in a chatbot, users perceive it as being more human, or elicit more *anthropomorphism,* which is defined as "the attribution of human personality or characteristics to something non-human, such as an animal, object, etc" (Smestad & Volden, 2018, p. 3). In other words, as humanness in the chatbot increases, it becomes easier for users to attribute human traits or characteristics to the chatbot, which in turn increases user experience. In addition, anthropomorphism can be *mindless* or *mindful.* As implied, mindless anthropomorphism occur automatically, where users treat a chatbot as if they were interacting with another human being due to humanlike attributes. Mindful anthropomorphism on the other hand is a conscious and sincere belief that the chatbot has human traits. For instance, Kim and Sundar (2012) demonstrated this difference in anthropomorphism by having participants interact with a webpage with or without an animated human agent. They found that participants interacting with the webpage with the human agent rated the webpage higher on human-like traits (mindless) perceived it as being less human compared to the web page without the human avatar when directly asked. Anthropomorphism has been found to be important to trust (Cassell et al., 1999; Følstad et al., 2018) as people tend to trust chatbots that behave and look more human. Anthropomorphism has also reported to be an important factor whether or not users

choose to cooperate with a chatbot (Laban & Araujo, 2019). In addition, anthropomorphism has been shown to lead to favorable product decisions in users by interacting with a chatbot with human qualities such as warmth or competence. (Roy & Naidoo, 2021). Indeed, several humanlike features both visual and conversational has been found to increase anthropomorphism. For instance, Araujo (2018) found that participants were more likely to engage in anthropomorphism and reported greater satisfaction when interacting with a chatbot using informal language and had a human name. In addition. Another humanlike feature found to elicit anthropomorphism is the use of first- and second-person pronouns. Although the effect is mostly research from a corporate-consumer online interaction perspective, pronouns are effectively used in web care conversations as a way to create personal and humanlike conversations (Liebrecht et al., 2021).

Social presence

People often apply the same interaction behaviors, or social rules, found in regular conversations such as politeness to conversations with chatbots (Nass, Steuer, & Tauber, 1994). As such, chatbots could be considered social actors, eliciting feeling of *social presence* in people interacting with them. Social presence can be defined as "a psychological state in which virtual (para-authentic or artificial) actors are experienced as actual social actors in either sensory or non-sensory ways." (Lee, 2004, p. 45), or the feeling of interacting with a "real person". Social presence has been shown to increase as users are exposed to more social cues (Oh, Bailenson, & Welch, 2018). Furthermore, humans automatically and effortlessly engage in social responses to computers when presented with social cues, or humanlike traits associated with social interactions (Nass & Moon, 2000). For instance, Nass, Steuer and Tauber (1994) reported in their study that participants who received praise from a male computer voice rated the praise as more assertive, affectionate, and sympathetic than compared to praise from a female computer voice. By increasing the politeness in the computer, Nass, Steuer and Tauber (1994) were able to manipulate the participants' perception of the computer, making it feel more friendly and competent. Furthermore, Bickmore and Picard (2005) demonstrated humans preference to embodied conversational agents which behavior resemblance that of another human. Participants interacting with such an agent designed to be more empathic, engage in more social small talk and use more non-verbal body language were perceived as more likable, were trusted more, and respected more by the participants than non-relational agents.

Measuring chatbot user experience

Several different measures have been utilized within the literature to capture the user experience of chatbot interactions. Typically, self-reported instruments are considered the most widely used method in which participants complete questionnaires designed to measure different aspects of user experience. For instance, Cameron et al. (2018) measured usability of a mental health chatbot using the System Usability Scale developed by Brooke (1996) designed to quickly measure the usability of a system. Yang, Aurisicchio, and Baxter (2019) used the Positive Affect, Negative Affect Scale (PANAS) (Watson, Clark, & Tellegen, 1988) measuring participants affective responses towards a chatbot in combination with critical incident method, consisting of users reporting on an experience they have had, followed by open-ended questions. One of the most holistic measures of user experience is the AttrakDiff developed by Hassenzahl, Burmester, and Koller (2003) as a tool to capture both pragmatic and hedonic qualities of user experience within interactive systems. Different instruments have been developed to measure the humanness on the chatbot and its effect on users, such as anthropomorphism (Powers & Kiesler, 2006) and social presence (Lee, Jung, Kim, & Kim, 2006). As behavior is considered important for user experience, it is useful to measure user behavior in chatbot interactions. Although some research have measured user behavior such as amount of messages, length of messages (Hill et al., 2015), or use of pronouns during interactions (Liao et al., 2020), considerably less research has been conducted on user behavior compared to user perceptions.

Research question and hypothesis

The section above provided a short summary of the literature regarding chatbots in customer service, and the effect of humanness on user experience and user behavior. There is a need to further understand the effect of humanness in chatbots on user experience, specifically there is a gap in the literature regarding the effect of humanness in chatbot's avatar and dialogue on user behavior. In order to fill this gap, the following research question was formulated:

"How is user experience and user behavior affected by humanness in a chatbots avatar and dialogue?"

The research question addresses the gap in the literature while simultaneously aims to replicate previous research' results on the self-reported effect of humanness in user experience. Based on previous research by Go and Sundar (2019), Hassenzahl (2018) and Araujo (2018), it is hypothesized that:

H1: Increase in human likeness in the chatbots avatar will increase perceptions of (a) mindful and mindless anthropomorphism, (b) social presence and (c) hedonic quality.

H2: Increase in human likeness in the chatbots dialogue will increase perceptions of (a) mindful and mindless anthropomorphism, (b) social presence and (c) hedonic quality.

Furthermore, as previous research on humanness suggests, people tend to mirror their conversational partner (Hill et al., 2015) . In addition, social cues increase anthropomorphism making the chatbot perceived as more human (Araujo 2018). It is therefore hypothesized that:

H3: Increase in human likeness in the chatbot (a) avatar and (b) dialogue will increase the likelihood of social orientation interaction in user behavior.

Likewise, research suggests that a combination of multiple social cues, or humanlike features in chatbots is more effective than few humanlike features, it is therefore hypothesized that:

H4: Increase in human likeness in the chatbot through visual appearance and conversational design will strengthen the effect of human likeness on (a) user behavior and (b) user perceptions compared to only increasing human likeness for one of these factors

Lastly, as humanness is considered to be less important in customer service chatbots it is postulated that:

H5: Increase in human likeness in the chatbot does not impact pragmatic qualities of user experience.

Method

The previous section provided a brief overview of the background of chatbots and the link between user experience, user behavior and humanness. In the following section, a description of the methods used along with a rationale for why it was conducted in such a way to test the hypothesis.

Research design

In order to measure the effect of humanness in a chatbots avatar and dialogue on user experience and user behavior, and to test the hypothesis stated previously, an explanatory hypothesis testing approach was used. A hypothesis testing approach was chosen given the existing knowledge base on chatbot user experience within the HCI field (Araujo, 2018; Go & Sundar, 2019; Kim & Sundar, 2012; Liao et al., 2020).

For the study, a between-subjects 2x2 factorial experimental design with randomized conditions was used. Randomizing the conditions allows to test for causality between the chatbot conditions. In addition, using a factorial design allows to test two different factors and their effect on user experience and user behavior, as well as a possible interaction effect. As each participant interacted with the chatbot which was either low or high on avatar humanness, and low or high on conversational humanness. A factorial randomized design allows to test for each of the main effects of the independent variables, as well as the interaction between them on both user experience and user behavior. The dependent variables can be divided into two parts, user experience was measured using a self-report questionnaire designed to capture social presence, anthropomorphism, and pragmatic/hedonic qualities of user experience. User behavior on the other hand, was measured by analyzing the dialogue of the participants. During participants interactions, their conversations was collected and saved to be recoded as quantitative data regarding user behavior.

About the project

The research was conducted in collaboration with SINTEF and boost.ai. The collaboration with SINTEF was conducted as part of a research project on chatbot interaction design, Human-Chatbot Interaction Design, and the study findings contributed to this project. boost.ai, a chatbot platform provider, collaborated by making available their platform for the experiment. The author led the research process, from the literature search, what independent variables to manipulate, dependent variables to measure, deciding the instruments necessary for

measuring the dependent variables, designing the study, and analyzing the data collected. Throughout the project, the author's supervisors provided assistance and feedback on a two and a half-week interval.

Participants and recruitment

Participants were recruited through Prolific, an online recruitment agency for people to sign up and participate in a variety of different studies. To limit any potential language barrier, all participants had to be fluent in English. It was also required for participants to complete the experiment on their desktops, this limitation was put in place to reduce any potential variance associated with completing the experiment in an unfamiliar environment and to ensure that the chatbot was presented similarly on the screens of the participants. Lastly, although some computer proficiency was needed due to the study being an unsupervised field experiment, having a registered account at Prolific was deemed sufficient for the experimental task. By accessing the link provided on the Prolific website, users would be presented with general information regarding the study, such as approximate duration, compensation, and purpose of the study. Participants received £2.50 as incentives for completing the study.

The sample size used in this study (N = 120) consisted of 73 females (60.8%) and 47 males (39.2%) with a mean age of 33.37 (SD = 11.95) ranging from 18 to 73. The participants resided in 13 different countries with the majority of participants reported their country of residence as United Kingdom (89 participants), North America (12 participants) and South Africa (9 participants). Prior to participating in the study, all participants were informed of the purpose of the study, what their roles would be as a participant and that they would be able to withdraw from the study at any point without providing any reason (see appendix A for information provided to the participants).

Materials

Chatbot

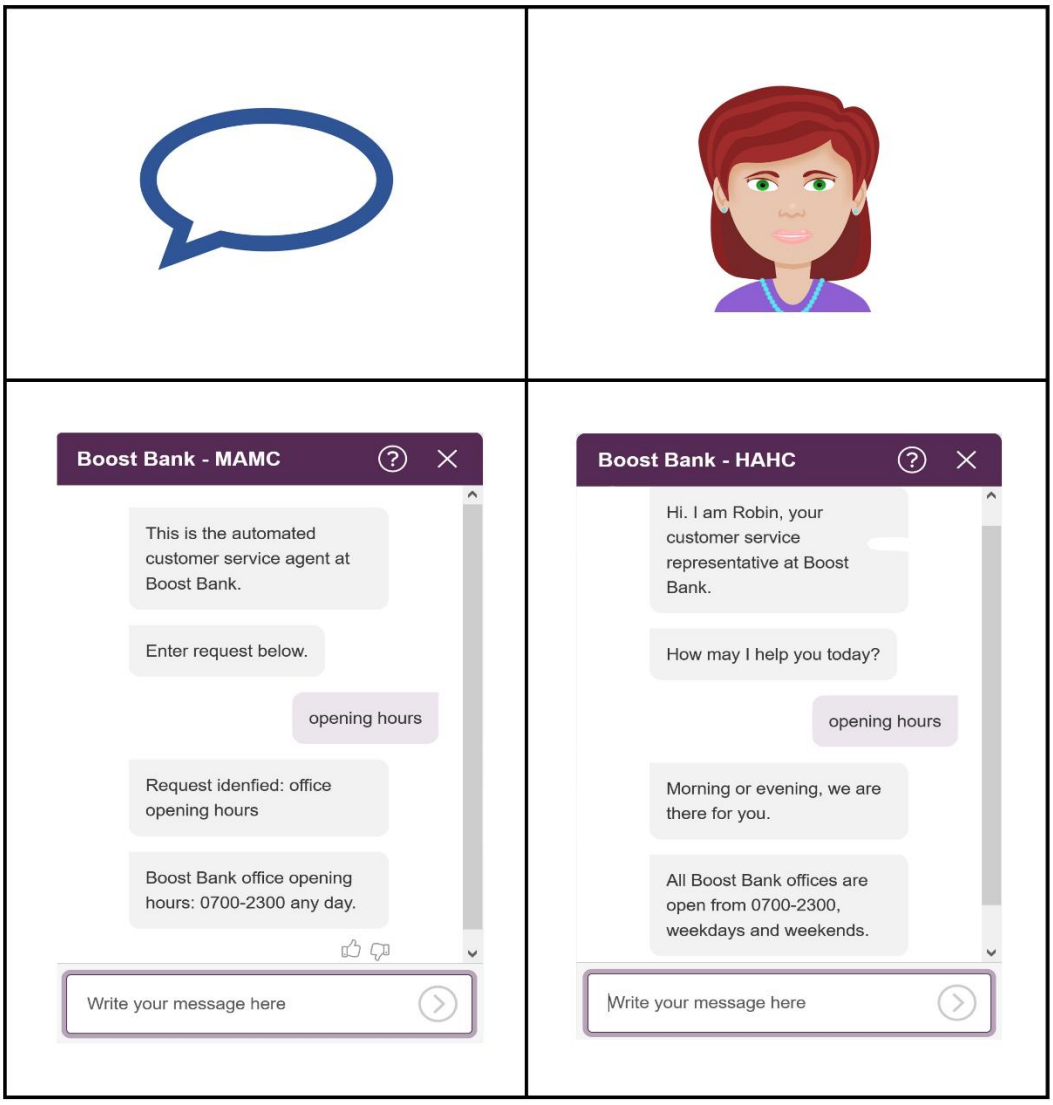
In order to answer the research question stated previously, a customer service chatbot provided by Boost.ai was used. The chatbot was designed as a regular customer service chatbot in a fictious bank called BoostBank, providing written answers to user inquiries. The chatbot responses were limited to topics surrounding opening hours, currency exchange and contact information for the purpose of the study. Limiting the chatbots topics allowed for a focus of data

and limiting unnecessary variation in the response data from the participants, thus increasing the success rate of the conversations.

The chatbot was designed with four conditions based on the level of humanness in either the avatar or the dialogue. The high humanness avatar consisted of an animated woman, while a simple "speech bubble" would be presented for participants interacting with the chatbot in the low avatar humanness condition. In the low humanness dialogue condition, the conversation was created to appear as impersonal and machinelike as possible, contrasting the more natural humanlike dialogue in the high humanness dialogue condition. See Figure 1. for a visual representation of the different chatbot humanness conditions.

Figure 1.

Four different manipulations of humanness in the chatbot.



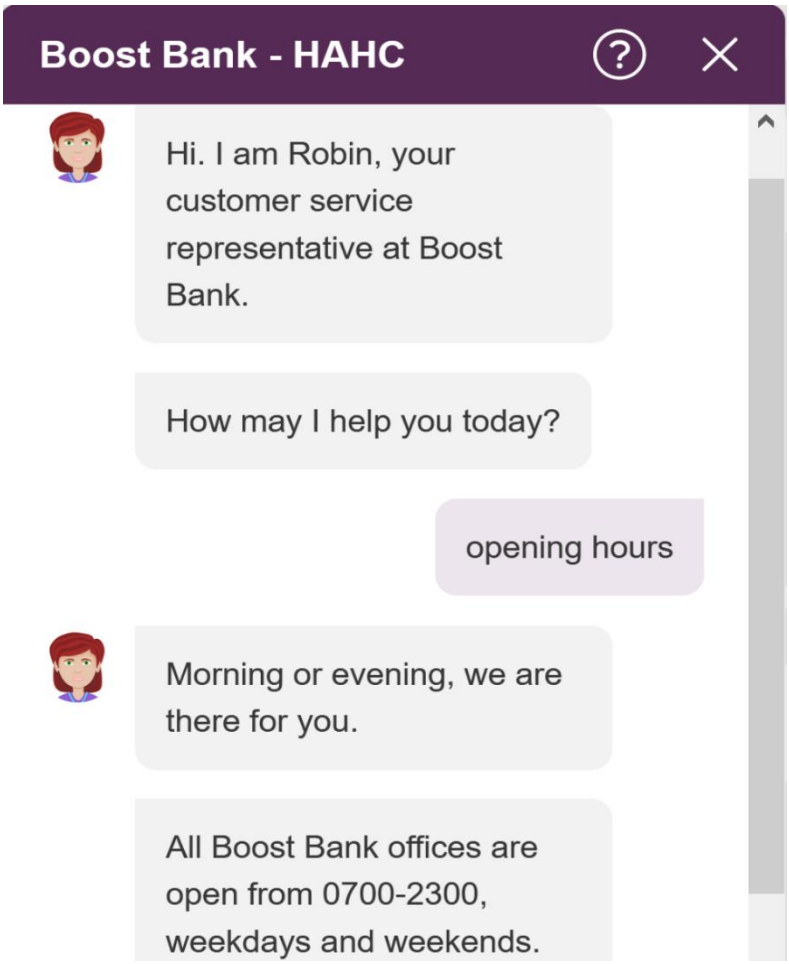
Note: Top left: Low humanness avatar. Top right: High humanness avatar. Bottom left: Low humanness dialogue. Bottom right: High humanness dialogue.

The experiment was conducted in a field setting, in which participants interacted with the chatbot in the comfort of their own home rather than in a laboratory experimental setting designed by researchers. Although the lack of laboratory experimental setting lends itself to unwanted variance that is difficult to control for. It does increase the external validity, which makes generalization to the general population easier as the study was more similar to how a natural chatbot interaction would be carried out (Cook, Campbell, & Shadish, 2002). In addition, an animated avatar instead of a photograph of a person was used as unnatural high humanness in technology lends itself to be vulnerable to the uncanny valley effect, or a feeling of eeriness associated interacting with technology that exhibiting humanlike traits to a varying degree (Ciechanowski et al., 2019). Users interacting with the high visual humanness and low conversation humanness condition, a photographically realistic avatar might have an adverse effect of increased humanness due to an increase in discomfort among participants.

The dialogue of the chatbot was designed to interact with the participants in either a high conversational humanness or low humanness condition. The high humanness condition interacted using informal language, which has been reported to increase perception of humanness in users (Araujo, 2018; Doyle, Edwards, Dumbleton, Clark, & Cowan, 2019). Before answering the participants inquiry, the chatbot would acknowledge the topic of the inquiry in an informal manner. Furthermore, the chatbot referred to itself as Robin and used first person pronouns when communicating with the participants. This was done in order to solidify an identity and increase anthropomorphism in participants, as identity and use of pronouns has found to increase interactivity and anthropomorphism in users (Go & Sundar, 2019; Hendriks, Ou, Amiri, & Bockting, 2020).

Figure 2.

Example of the high visual humanness and high conversational humanness chatbot condition.

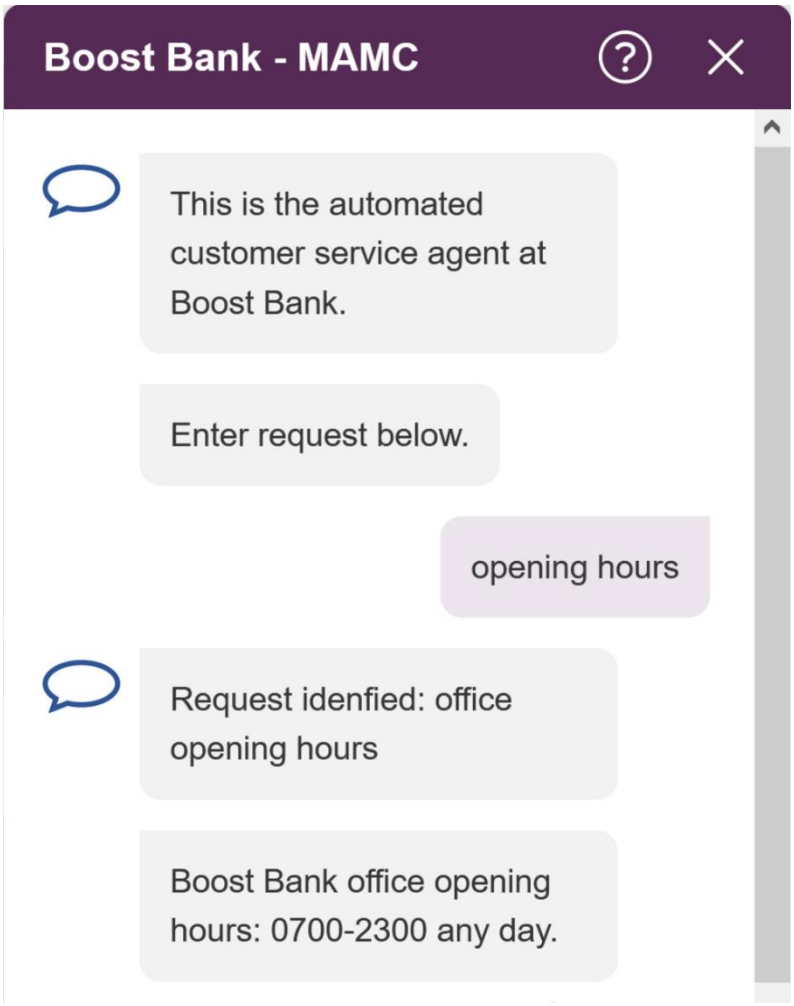


Lastly, in the high humanness conversation condition the chatbot would greet the participants before introducing itself to create a conversation similar to a human to human- interaction and enable a social oriented dialogue (Liao et al., 2020; Liebrecht et al., 2021). Following the interactions, all conversations were stored to be recoded later to test for levels of social/utilitarian oriented user behavior.

On the other hand, the low humanness conversational condition was designed to be as low on humanness as possible. Instead of a regular introduction the chatbot simply stated that it was a customer service agent. The conversation style was formal, machine-like, and was devote of any social chit-chat.

Figure 3.

Example of the low visual humanness and low conversational humanness chatbot condition.



Similarly to the high conversational condition, the low conversational condition acknowledged participants responses. However, emphasizing and reminding participants of the use of key words was used to create a feeling of communicating with a machine rather than a person (Liao et al., 2020).

Self-report measures

Currently there are no standardized measurement within the literature to comprehensively measure the effect of humanness in chatbots on user experience. As such, the set of measures used in this study has been gathered from different instruments from the literature. To accurately capture user experience, the self-report questionnaire in this study consisted of five sections measuring social presence, mindful/mindless anthropomorphism, and hedonic/pragmatic

qualities of user experience. Demographics such as age, gender, nationality, and educational background were also collected. In the following section, a brief description of the measures used in this study will be presented, while the questionnaire in its entirety will presented in Appendix B.

Social presence. Social presence was measured using a modified self-report questionnaire aimed at measuring participants feelings of interacting with another being (Araujo, 2018). Three items were adapted from Lee et al. (2006), and were measured on a 10-point semantic differential scale. Participants were to indicate how they felt about the chatbot using the word-pairs unsociable/sociable, machine-like/life-like and insensitive/sensitive. The second part of the social presence measure consisted of five items measured on a 10-point Likert scale, ranging from 1 (not at all) to 10 (extremely). The items were designed to measure participants feelings towards the chatbot regarding intelligence, sociability, communication, attention, and involvement. The items were combined to create an average social presence score of (α = .92).

Mindless and mindful anthropomorphism. Mindless anthropomorphism was measured based on the mindless anthropomorphic index adapted from Kim and Sundar (2012). It consists of four items measured on a 10-point Likert scale ranging from 1 (describes very poorly) to 10 (describes very well) in which participants were to rate the chatbot using the adjectives likeable, sociable, friendly and personal (α = .92). Mindful anthropomorphism was measured based on the mindful anthropomorphism index adapted from Kim & Sundar (2012). A mindful anthropomorphism index was created by directly asking participants to rate the chatbot using the word-pairs human-like/machine-like, life-like/artificial, natural/unnatural. The items were measured on a 10-point semantic differential scale (α = .93).

Hedonic and pragmatic qualities of user experience. In order to capture participants user experience interacting with the chatbot, a modified version of AttrakDiff, adapted from Hassenzahl et al. (2003), was used. The original questionnaire contains 28 items measured on a 7-point semantic differential scale items divided into three factors measuring pragmatic qualities, Hedonic qualities (identity) and hedonic qualities (stimulation). AttrakDiff has been successfully utilized in previous chatbot studies (Smestad & Volden, 2018). The pragmatic qualities measures aspects of user experience related to practicality and ease of use. While the hedonic qualities measures qualities such as creativity and captivation. In this study however, all items measuring identity was removed, as well as one item each from both the pragmatic quality and hedonic

quality (stimulation) factor, as these were considered irrelevant for a chatbot study. As a result, 6 items measuring pragmatic qualities of user experience (α = .782) and 6 items measuring hedonic qualities of user experience (α = .802) were included in the study.

User behavior

There is currently no standardized measurement available for the effect of user behavior. However, previous studies have measured the amount of social markers (e.g. “hello” and “thanks”) as an indicator of conversational orientation by (Liao et al., 2020). Therefore, in this study, user behavior will be measured using social makers in combination of the use of pronouns (Liao et al., 2020; Liebrecht et al., 2021). Conversational logs were inspected, and participants found to use both social makers and pronouns while interacting with the chatbot will be considered social-oriented. On the other hand, absence of one or more of these markers would categorize the participant as utilitarian oriented.

Open ended question

The questionnaire also included one item for qualitative data collection. The data from this item was not used in this study but collected for later future use for collaborating partner SINTEF. This item was placed so as not to interfere with the other participant reporting in the questionnaire.

Procedure

Prior to the data collection, a literature review was conducted to determine both the variables and measures typically used in the literature, as well as which sectors typically use chatbots. Several meetings were arranged, discussing what dependent variables should be included and how to measure them. The experimental tasks chosen for this study was based on what a typical banking customer might ask a customer service representative. The study was piloted by inviting 20 participants to respond. One participant did not answer most of the questionnaire, which resulted in a change in settings as to which questions were mandatory to proceed. However, no changes to the chatbot was needed. Following this, the remainder of the participants were recruited.

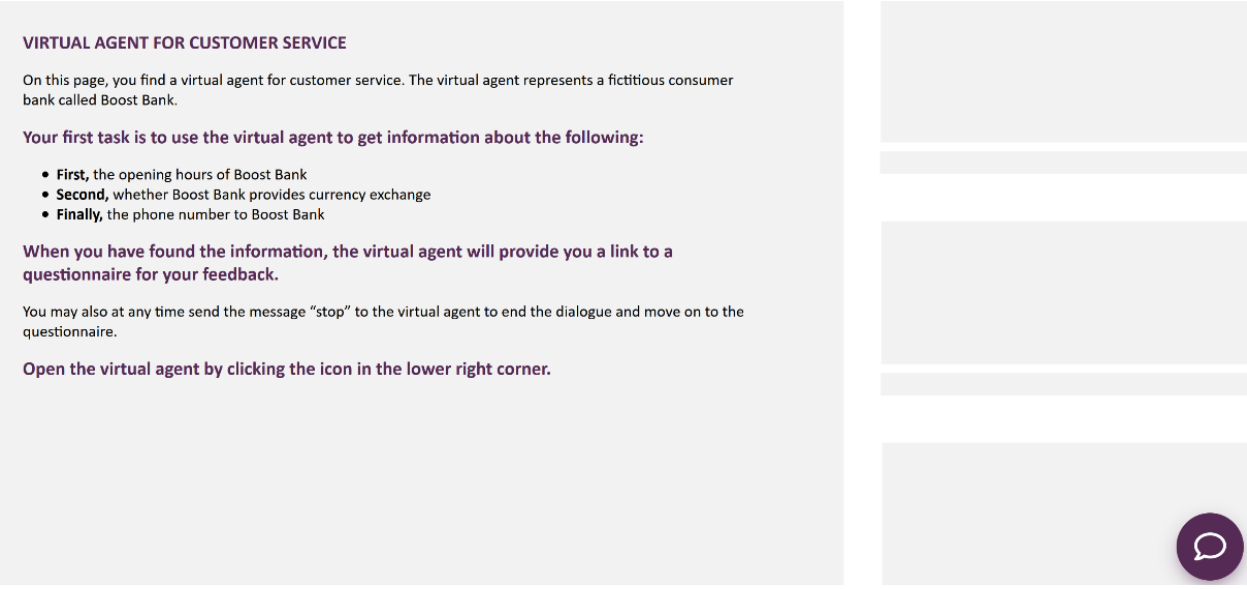
General procedure

Prior to being exposed to the chatbot, participants were first provided with information regarding the purpose of the study and what would be expected should they agree to participate.

After giving informed consent by clicking the “next” button, participants would be redirected to another webpage containing the chatbot as depicted in figure 4.

Figure 4.

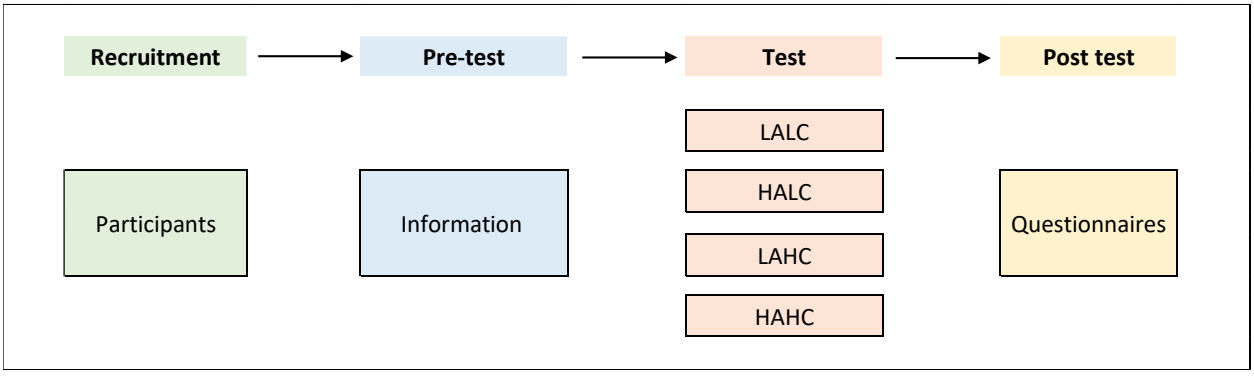
Webpage containing experimental instructions and the chatbot.



Participants were instructed to interact with a chatbot from a fictious bank called Boost Bank. Their task was to acquire information regarding opening hours, currency exchange and the phone number of Boost Bank by interacting with the chatbot. To standardize the test conditions, all participants were only able to interact with the chatbot using a desktop. Although the instructions were identical for all participants, after being redirected from the informed consent page, participants were randomly divided into four different groups as depicted in Figure 5.

Figure 5.

Participants study procedure.



Note: LALC = low (humanness) avatar, low (humanness) conversation. HALC = high avatar, low conversation. LAHC = low avatar, high conversation. HAHC = high avatar, high conversation.

Following the completion of the tasks, the chatbot would provide participants with a link to the questionnaire containing the self-report measures in the study. After completing the questionnaire, participants would again be provided with a link that would redirect them back to prolific, which would ensure that they completed the study and would be correctly reimbursed for their time spent.

Data collection

The data collection consisted of two steps: 1) during their interacting with the chatbot, participants dialogue was being recorded and stored for later analysis of user behavior. 2) following the interaction, participants completed a questionnaire consisting of a series of self-report measures designed to measure anthropomorphism, social presence, and pragmatic/hedonic qualities of user experience.

Analysis

Self-report measures

Raw data from the questionnaires were first exported from the online questionnaire to an excel file. From there, the data was imported to IBM SPSS (Statistical Package for Social Sciences) version 27, 64-bit Windows edition for further reconstruction and analysis. Composite scores were created for each dependent variable (mindful anthropomorphism, mindless anthropomorphism, social presence, pragmatic quality, hedonic quality), information regarding the data codes and setup is available in Appendix C. Prior to creating composite scores, all items that were negatively scored had been reversed so that the items were scored equally across the measures. Following the data reconstruction, an analysis of the descriptive statistics was carried out to investigate the distribution of age, gender, nationality, and education. Furthermore, the data was also tested for normality, which is an assumption needed to conduct a between-subjects ANOVA. Lastly, one 2-tailed 2x2 factorial between-subjects ANOVA for each self-report measure was conducted to provide the main effects of both avatar humanness and dialogue humanness, as well as the interaction effect between them.

User behavior

To test the hypothesis that humanness in the chatbot affect the conversational orientation (social vs utility oriented) of users, participants dialogue was first exported to excel. From there, user dialogue was binary coded so that the use of pronouns = 1, and the use of social makers = 1. To be considered socially oriented, participants had to use both pronouns and social makers. User behavior was then analyzed using logistic regression to determine if a chatbots humanness affected the user behavior of participants, causing them to behave more socially when interacting with a chatbot high on humanness.

Ethics

All participants provided informed consent prior to participating in this study. Participants could at any point withdraw from the study without providing any reason. As the study was anonymous, the data provided by the participants can in no way be traced back to the participants. Furthermore, no sensitive or identifying personal data was collected. The data will be stored on a password protected computer at UiO.

Results

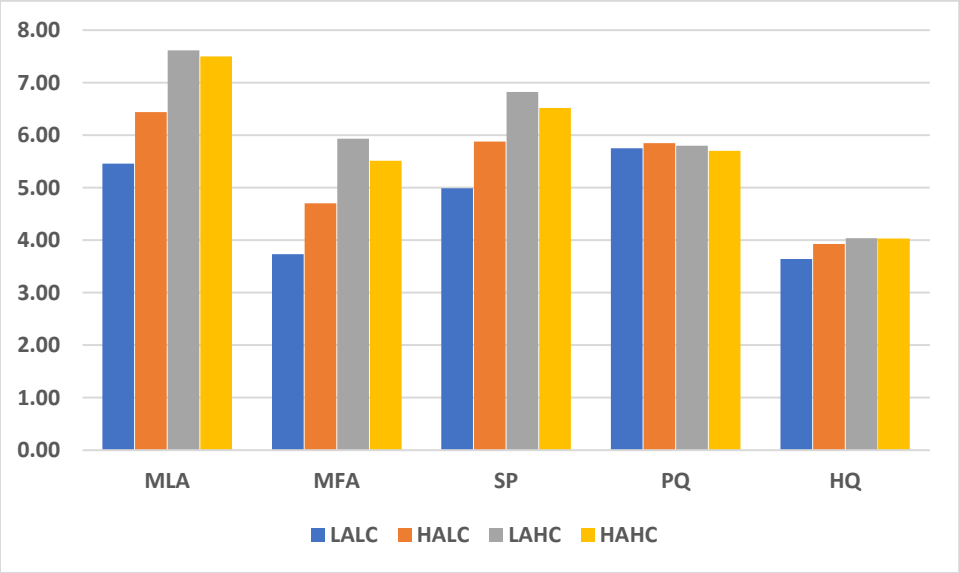
The purpose of this study was to measure the effect of humanness in chatbots on user experience and user behavior. In the following section, descriptive statistics regarding the participants score on the self-report measures will first be presented, followed by the result of the analysis of participants self-report scores. Lastly, an analysis of user behavior will be presented.

Descriptive statistics

The participants mean scores from the self-report measures are presented in Figure 6. and detailed below.

Figure 6.

Mean scores of chatbot condition grouped by self-report measures.



Note: MLA = mindless anthropomorphism. MFA = mindful anthropomorphism. SP = social presence. PQ = pragmatic qualities. HQ = hedonic qualities. LALC = low humanness avatar, low humanness conversation. HALC = high humanness avatar, low humanness conversation. LAHC = low humanness avatar, high humanness conversation. HAHC = high humanness avatar, high humanness conversation.

The participants reported the lowest mean values across mindless anthropomorphism (M = 6.75, SD = 2.05), mindful anthropomorphism (M = 4.96, SD = 2.31), and social presence (M = 6.05, SD = 1.88) when interacting with the chatbot low on both visual and conversational humanness. On the other hand, participants interacting with the chatbot low on avatar humanness

but high on dialogue humanness reported the highest mean values. However, it appears to be little meaningful difference in pragmatic (M = 5.77, SD = .85) and hedonic (M = 3.91, SD = .97) qualities of user experience across chatbot conditions. In order to measure the main effect of the humanness in the chatbots avatar and dialogue, and the interaction between them on the dependent variables. Chatbot conditions were split based on the levels of humanness in the avatar and dialogue as reported in Table 1.

Table 1.

Descriptive statistics grouped by levels of humanness in the chatbots avatar and dialogue.

		Low Humanness					High Humanness				
	Measures	N	M	SD	Skewness	Kurtosis	N	M	SD	Skewness	Kurtosis
Avatar	mindless	61	6.520	2.144	-0.480	-0.594	59	6.979	1.933	-0.613	0.128
	mindful	61	4.814	2.275	0.170	-1.013	59	5.113	2.349	0.233	-0.288
	social	61	5.891	1.805	-0.271	-0.261	59	6.203	1.965	-0.201	0.020
	pragmatic	61	5.773	0.851	-0.680	0.695	59	5.771	0.859	-0.686	-0.379
	hedonic	61	3.836	1.017	0.145	0.771	59	3.980	0.917	0.073	-0.052
Dialogue	mindless	60	5.933	2.139	-0.147	-0.568	60	7.558	1.592	-0.790	0.431
	mindful	60	4.200	2.129	0.666	0.476	60	5.722	2.242	-0.215	-0.652
	social	60	5.419	1.860	0.275	0.118	60	6.671	1.703	-0.710	1.243
	pragmatic	60	5.794	0.796	-0.537	-0.288	60	5.750	0.910	-0.756	0.287
	hedonic	60	3.777	0.996	0.127	-0.167	60	4.036	0.929	0.116	1.325

Note: composite score associated with the measures used in the study based on condition.

As can be seen, the data was considered normally distributed by having a skew of > |2| and kurtosis of > |9|. As predicted, participants that interacted with the chatbot low on visual and/or conversational humanness were associated with the lowest numerical mean values in mindless anthropomorphism, mindful anthropomorphism, and social presence. Interestingly, humanness appeared to have little effect on participants mean scores in pragmatic and hedonic qualities of user experience. However, to determine the nature of the numerical differences between the means, a series of 2x2 factorial between-group ANOVAs was carried out.

Mindless anthropomorphism

Differences in mindless anthropomorphism were investigated through a two-way ANOVA. A significant main effect was found for variations in humanness in conversation ($F(1,116) = 22.23$, $p > .001$ $\eta^2 = .161$), but not for variations in humanness in the avatar ($F(1,116) = 1.60$, $p = .21$). Furthermore, no significant interaction effect was found ($F(1,116) = 2.59$, $p = .11$). The corrected model accounted for 16.7 percent of the variance. The assumption of normality was tested and satisfied based on Levene's F test ($F(3,116) = 1.33$, $p = .27$).

Mindful anthropomorphism

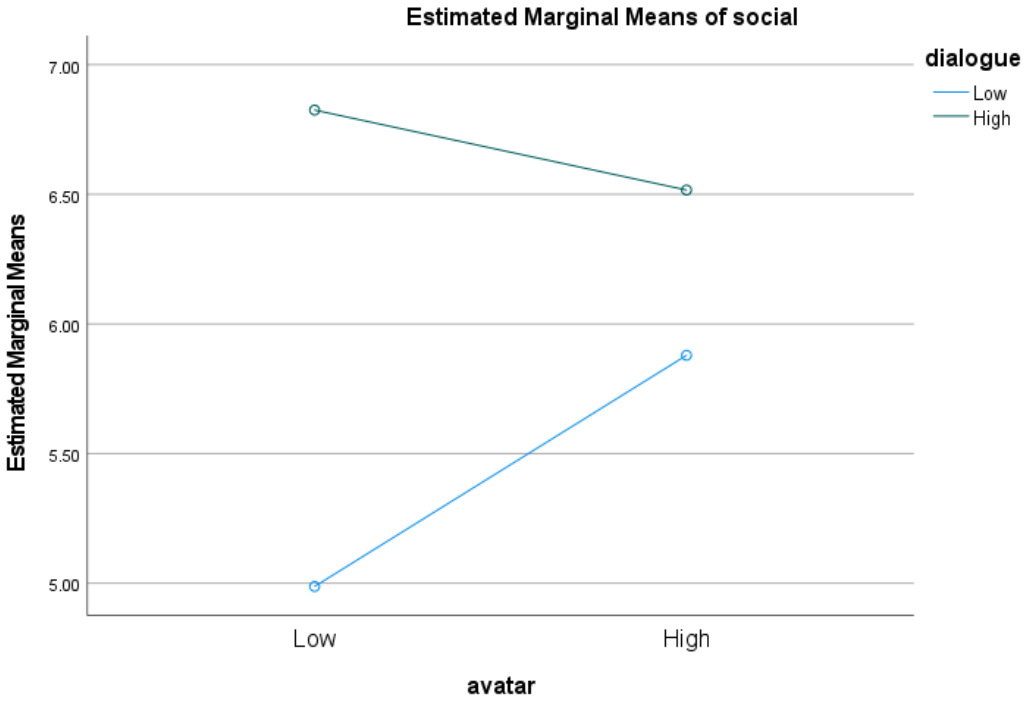
Differences in mindful anthropomorphism were investigated through a two-way ANOVA. A significant main effect was found for variations in humanness in the conversation ($F(1,116) = 14.420$, $p > .001$ $\eta^2 = .111$), but not for variations in humanness in the avatar ($F(1,116) = .477$, $p = .491$). Furthermore, no significant interaction effect was found ($F(1,116) = 3.080$, $p = .082$). The corrected model accounted for 11.1 percent of the variance. The assumption of normality was tested and satisfied based on Levene's F test ($F(3,116) = 2.18$, $p = .094$).

Social presence

Differences in social presence were investigated through a two-way ANOVA. A significant main effect was found for variations in humanness in the conversation ($F(1,116) = 14.716$, $p > .001$ $\eta^2 = .113$), but not for variations in humanness in the avatar ($F(1,116) = .817$, $p = .368$). As can be seen in Figure 7, although close to significance, the interaction effect between humanness in the chatbots avatar and dialogue was found not to be significant, ($F(1,116) = 3.459$, $p = .065$). The corrected model accounted for 11.3 percent of the variance. The assumption of normality was tested and satisfied based on Levene's F test ($F(3,116) = 2.550$, $p = .059$).

Figure 7.

Interaction plot between avatar humanness and dialogue humanness on social presence.



Pragmatic qualities of user experience

Differences in pragmatic qualities of user experience were investigated through a two-way ANOVA. No significant main effect was found for variations in humanness in conversation ($F(1,116) = .086$, $p = .770$), or for variations in humanness in the avatar ($F(1,116) = .000$, $p = .994$). Furthermore, no significant interaction effect was found ($F(1,116) = .395$, $p = .531$). The assumption of normality was tested and satisfied based on Levene's F test ($F(3,116) = .349$, $p = .790$).

Hedonic qualities of user experience

Differences in hedonic qualities of user experience were investigated through a two-way ANOVA. No significant main effect was found for variations in humanness in conversation ($F(1,116) = 2.076$, $p = .152$), or for variations in humanness in the avatar ($F(1,116) = .635$, $p = .427$). Furthermore, no significant interaction effect was found ($F(1,116) = .686$, $p = .409$). The assumption of normality was tested and satisfied based on Levene's F test ($F(3,116) = .203$, $p = .894$).

User behavior

To test the hypothesis that humanness in the chatbots avatar and dialogue influenced participants user behavior (social vs utilitarian behavioral orientation), user dialogue was coded so that the use of pronouns and social makers indicates social behavioral orientation. On the other hand, if participants do not use pronouns and/or social makers it would indicate a utilitarian behavioral orientation. The frequency of social oriented and utilitarian oriented participants is depicted in Table 2. Although there seems to be no difference in behavior when participants were interacting with the chatbot low on humanness in the avatar, there is a numerical increase in social oriented behavior in the high humanness avatar condition.

Table 2.

Behavioral orientation based on humanness in the chatbots avatar.

		Utility oriented	Social oriented	Total
Avatar humanness	Low	31	30	61
	High	21	38	59
	Total	52	68	120

Note: Frequency of participants behavior during chatbot interaction.

Furthermore, as can be seen in Table 3, there is a similar behavioral pattern among participants based on the chatbots dialogue conditions, in which a numerical increase in social oriented behavior is observed in the high humanness dialogue condition.

Table 3.

Behavioral orientation based on humanness in the chatbots dialogue.

		Utility oriented	Social oriented	Total
Dialogue humanness	Low	32	28	60
	High	20	40	60
	Total	52	68	120

Note: Frequency of participants behavior during chatbot interaction.

In order to analyze the nature of the difference in behavior across chatbot conditions, a logistic regression analysis was conducted to test the probability of correctly classifying the user behavior as either social behavior oriented or utilitarian behavior oriented based on the chatbot conditions. The analysis yielded a bordering significant result, $\chi^2(3) = 7.768$, $p = .51$. The model explained between 6% (Cox & Snell $r^2 = .063$) and 8% (Nagelkerke $r^2 = .084$) of the variance and predicted 62.5% of the user's behavioral orientation correctly. Due to the bordering significant *p*-

value, a further examination of the Wald-statistic was conducted, which revealed that humanness in the chatbots dialogue significantly contributed to the model (B= -0.417 (SE=0.191), Wald= 4.764(1), $p > .029$), with an odds ratio of 0.66 (95% CI [0.45, 0.96]). The avatar humanness (B= -0.318(SE=0.191), Wald= 2.773(1), $p = .096$) and the interaction between the avatar and dialogue on user behavior (B= -0.015(SE=0.191), Wald= 0.006 (1), $p = .936$) was not significant.

Discussion

The purpose of this study was to study the effect of humanness in a chatbots avatar and dialogue on user experience and user behavior. As the effect of humanness in chatbots on user experience through self-reported measures are commonly used measure within the research field. The interest of this study is particularly related to the effect of humanness on user behavior through behavioral measures. A summary of the findings in the study are presented in Table 4. Three findings were found to be particularly interesting; 1) participants perceived the chatbot as more human, and felt greater social presence interacting with a chatbot that behaved more humanlike. 2) increase in humanness in the chatbots dialogue increased the frequency of participants interacting socially with the chatbot. 3) humanness in the chatbot did not impact hedonic or pragmatic qualities of user experience.

Table 4.

Effect of chatbot humanness on user perception and user behavior.

Hypothesis	Findings
H1a. Avatar humanness → anthropomorphism	Not supported. Difference between low humanness chatbot and high humanness chatbot not significant.
H1b. Avatar humanness → Social presence	Not supported. Difference between low humanness chatbot and high humanness chatbot not significant.
H1c. Avatar humanness → Hedonic user experience	Not supported. Difference between low humanness chatbot and high humanness chatbot not significant.
H2a. Dialogue humanness → Anthropomorphism	Supported. High humanness chatbot associated with higher levels of mindless and mindful anthropomorphism than low humanness chatbot.
H2b. Dialogue humanness → Social presence	Supported. High humanness chatbot associated with higher levels of social presence than low humanness chatbot.
H2c. Dialogue humanness → Hedonic user experience	Not supported. Difference between high humanness chatbot and low humanness chatbot not significant.
H3a. Avatar humanness → User behavior	Not supported. Difference between high humanness chatbot and low humanness chatbot not significant.
H3b. Dialogue humanness → User behavior	Supported. High humanness chatbot associated with higher levels of social behavioral orientation than low humanness chatbot.

H4a. Avatar + dialogue humanness → User behavior	Not supported. Interaction effect of avatar and dialogue on user behavior not significant.
H4b. Avatar + dialogue humanness → user perceptions	Not supported. Interaction effect of avatar and dialogue on user perceptions not significant.
H5. Humanness → No increase in pragmatic user experience	Supported. No difference between high humanness chatbot and low humanness chatbot in pragmatic user experience.

The following section will discuss the hypotheses based on the results in relation to relevant existing literature, suggesting possible underlying mechanism for the observed causal relationship. In addition, implications of the results in the study will be presented, both practical and theoretical, along with limitations associated with the study. For the convenience of the reader, results will be presented first by the effect of humanness on anthropomorphism and social presence, followed by hedonic and pragmatic qualities of user experience. Lastly, measures of user behavior will be discussed.

Effect of humanness on self-report measures

It was hypothesized that an increase in the humanness of the chatbots avatar would increase H1a) mindful and mindless anthropomorphism, H1b) social presence, and H1c) hedonic qualities of user experience. However, no significant difference between chatbots with high avatar humanness and low avatar humanness was found. A possible explanation for the lack of significant effect could be that visual humanness found in the chatbots avatar does not appear sufficiently salient for users. As suggested by cognitive load theory, human are only capable of processing a limited amount of information at any given time (Sweller, 2011). As such, information considered irrelevant for the task at hand, such as the chatbots avatar, will not be processed during an interaction. Indeed, this might be particularly relevant for customer service chatbots as they operate in a highly task-oriented environment in which the primary motivation of users is productivity and efficiency (Følstad & Skjuve, 2019).

Prior research has found evidence for the effect of visual humanness on anthropomorphism in participants navigating through webpages by having a highly humanlike animated figured appear on the screen (Kim & Sundar, 2012). However, the visual cue featured a complete animated human body, and while the participants did interact with the webpage, they did not engage in a dialogue similar to a chatbot. In addition, Go and Sundar (2019) found an

interaction-effect for visual humanness and identity in their customer service chatbot on perceived homophily, or sense of perceived similarity. By interacting with a chatbot high on visual humanness and having the chatbot not reveal its identity as a chatbot, participants reported greater sense of homophily. However, unlike Go and Sundar (2019) which featured a highly salient visual avatar of a real person, the chatbot used in this study only featured a small, animated avatar. Possibly, a more salient visual stimuli, for example as in embodied conversational agents, might have produced a stronger effect. The result from this study suggests that users are not substantially impacted by visual expression of humanness in the avatar image of a customer service chatbot.

Humanness in the chatbot dialogue

As hypothesized, participants who interacted with a chatbot with a highly humanlike dialogue perceived it as being more humanlike. This study shows that using relatively few humanlike cues such as giving the chatbot a name (identity) and interacting using informal language was sufficient to facilitate both mindful and mindless anthropomorphism in participants. These results are consistent with past research on mindless anthropomorphism which suggests that humans subconsciously attribute human characteristics to computers when exposed to social cues (Kim & Sundar, 2012; Araujo, 2018). In addition, participants were found to consciously attribute human characteristics to the chatbot, consistent with research on the CASA paradigm (Naas & Moon, 2000; Araujo, 2018). It has been argued that humans engage in mindless anthropomorphism more easily than mindful anthropomorphism. For instance, Kim and Sundar (2012) argued that participants denied viewing the webpage in their study as humanlike and subsequently rated it lower than the same webpage not exhibiting humanlike traits. However, unlike a webpage, the conversational design of a chatbot might make it easier for people to consider the chatbot as a conversational partner similarly to that of another person, thus be more willing to engage in mindful anthropomorphism (Araujo, 2018).

Effect of humanness in the dialogue on social presence

As hypothesized, this study found a significant difference in social presence between participants interacting with the chatbot low on humanness in the dialogue compared to participants interacting with the chatbot high on humanness in the dialogue. The literature often discusses humanness and social presence as related constructs, in which increased humanness leads to increased social presence (Naas & Moon, 2000). Therefore, people engaging in

conversation tend to expect basic conversational rules to apply, such as returning a greeting, even if they know they are interacting with a chatbot (Jain et al., 2018). For instance, Go and Sundar (2019) reported that social presence was affected by message interactivity, in which their chatbot would acknowledge previous statements of the user and thus create a more engaging conversation similar to that of a regular human conversation. In addition, Araujo (2018) demonstrated how participants expectations interacted with their social presence. By informing participants they would interact with a highly advanced chatbot beforehand, participants reported increased social presence, compared to no difference in social presence without the positive priming. Although this suggests more research is needed on the effect of humanness in customer service chatbots on social presence. This study contributes to existing literature by demonstrating that the use of pronouns, informal language, and identity cues in a customer service chatbots dialogue is sufficient to affect social presence in users.

Humanness on hedonic and pragmatic qualities of user experience

As hypothesized, humanness in the chatbot dialogue did not affect pragmatic qualities of user experience. Participants did not perceive greater pragmatic user experience in either of the four different groups depending on the humanness in the chatbot. These findings are in line with previous research on user experience (Hassenzahl, 2018) and customer service chatbots in general (Følstad & Brandtzaeg, 2020). While some argue that increase of social characteristics in chatbots similar to those found in regular human conversations could reduce pragmatic qualities such as dissatisfaction or frustration (Chaves & Gerosa, 2019). However, customer service is a task-oriented field which favors pragmatic over hedonic qualities (Hassenzahl, 2018; Følstad & Skjuve, 2019). As such, hedonic features such as small talk could be seen as a hindrance to users effective interaction of the customer service chatbot (Svenningsson & Faraon, 2019). As Følstad and Brandtzaeg (2020) argues, due to the task-oriented field of customer service, humanness in a chatbot should only supplement user experience if it does not interfere with the chatbots perceived productivity and efficiency. However, results from this study shows that although humanness did not interfere with pragmatic qualities of the chatbot, humanness did not increase hedonic qualities either.

Hedonic qualities

Interestingly, no effect for humanness on hedonic qualities of user experience was found in any of the different groups. These results are interesting for a variety of different reasons.

Hassenzahl (2018) argues that a balance between pragmatic and hedonic qualities are desired in interactive systems as people use them for a variety of reasons, and motivations for use are often dynamic and changing. Increased humanness in chatbots through different personalities has been found to increase user experience in healthcare (Chaix et al., 2019), as an "general purpose chatbot" (Thies, Menon, Magapu, Subramony, & O'neill, 2017) or as a dinner planner (Smestad, & Volden, 2018). However, as argued by Følstad and Skjuve (2017), customer service is a highly task-oriented field in which users favor pragmatic qualities above all. Although increased humanness creates a more natural conversation between the chatbot and the user, creating feelings of familiarity (Diederich, Lembcke, Brendel, & Kolbe, 2021), it is not necessarily needed in simple task-oriented interactions. As shown in this study, participants acknowledged the humanness in the chatbot, and reported it as being more socially present, but it did not significantly affect hedonic qualities of user experience. Hassenzahl and Tractinsky (2006) argues user experience is a complex construct which depends not only on the individual user, but also the user's situation, motivations and needs. As such, humanness might not be sufficient to significantly impact user experience in customer service chatbots.

User behavior

It was hypothesized that an increase in humanness would alter participants behavior, causing them to behave more socially. Although the support for the hypothesis only bordered significance, it is speculated its mainly due to two reasons. 1) humanness in the avatar did not significantly predict behavioral orientation in users, only humanness in the chatbots dialogue. 2) behavioral orientation was found to be evenly split in the low humanness groups; thus, the model was only accurate in the high humanness group predicting social orientation. Early research on user interactions with computers postulate that users behave socially towards computers, especially when exposed to social cues (Nass et al., 1994). Although the majority of research rely on self-report measures, a select few studies have studied the user dialogue to assess people's behavior when interacting with chatbots. For instance, previous research on social chatbots shows that people behave differently when interacting with a chatbot compared to another human, using fewer sentence and less words per sentence (Hill et al., 2015). However, as the motivation from using social chatbots designed for *enjoyment* and *entertainment* is considerably different from a task-oriented customer service chatbot in which the driving motivation for use is productivity (Følstad & Skjuve, 2019). Extrapolating the results of studies using social chatbots

to studies using customer service chatbots should therefore exercise caution as people most likely behave differently interacting with a chatbot to complete a goal compared to fulfill their psychological well-being (Hassenzahl, 2018). While Moore (2018) provides convincing arguments for using conversation sequences as a behavioral measure of user experience in advanced chatbots utilizing natural conversation framework, no experiment was conducted to support the arguments presented.

Social behavioral orientation behavior has been observed in related studies (Liao et al., 2016; Liao et al., 2020). By observing employees in a company interacting with a human resource chatbot for up to six weeks, Liao et al. (2018) observed that a large portion of the participants interacted with the chatbot in a non-task related manner such as engaging in social chit-chat or asking agent related questions (e.g., "what is your favorite color"). Although the current study found similar evidence of increased humanness in a task-oriented chatbot leading to an increase in social behavioral orientation in users, no evidence of marked utilitarian behavioral orientation was observed. These results are in contrast with results by Liao et al. (2016) in their 17-days study of interactions with a human resource chatbot, which found that utilitarian oriented participants valued minimal user input, similar to a traditional information search system. A possible explanation for the discrepancy in results may be attributed to the duration of interactions in the different studies. As noted by Luger and Sellen (2016) in their study of conversational agents in households, users changed their user behavior over time. Although participants initially interacted with the conversational agents in a social manner, asking what it could do and engaging it in social chit-chat, user interaction gradually changed to be more task oriented. Luger and Sellen (2016) argued the change in user behavior was due to a learning process in which participants gradually became aware of the capabilities and limitations of the conversational agent, as well as how to interact with it in such a way that would optimally provide a suitable response. Indeed, as Liao et al. (2016) noted, some participants found it difficult to learn how to effectively communicate with the chatbot, suggesting utility-oriented behavior is an effortful learning process occurring over time. Due to the short interaction duration of customer service chatbots, the results from the current study indicate that people do not have sufficient time to adjust their behavior accordingly, and instead rely on previously learnt behavior of regular conversations.

Implications for practice

Research shows that while productivity is the driving motivator for customer service chatbots (Følstad & Skjuve, 2019), people still prefer to interact with a humanlike chatbot as long as it does not impact productivity (Følstad & Brandtzaeg, 2020; Thies et al., 2017). These findings echo Hassenzahl (2018) arguments for desired interactive products needing a balance between hedonic and pragmatic qualities. However, although participants reported no effect of humanness on self-report measures, people still were affected by humanness, behaving more socially when interacting to a humanlike chatbot. As such, designers should keep in mind the importance of humanness as it makes people behave more socially, which is easier than learning how to behave utilitarian (Liao et al., 2020; Luger & Sellen, 2016). For instance, this study shows that even though an increase in the chatbots humanness increased social orientation, people generally tend to behave socially towards customer service chatbots even if it is designed with few humanlike features. Thus, future design of customer service chatbots should consider users preference towards social oriented behavior as an increase in humanness facilitate habitual conversational behaviors in users.

In addition, this study demonstrates the lack of effect humanness has in a customer service chatbots avatar. Although previous research has reported an effect of humanness in chatbots avatar in task-oriented situations (Go & Sundar, 2019), this study failed to replicate similar results. In addition, this study demonstrates the relatively small effect, if any, humanness in the avatar poses on the perception of users when considered irrelevant for the task at hand. As such, designers of chatbots utilized in task-oriented fields would be wise to shift focus away from the chatbots avatar, and instead focus more on the chatbots dialogue to increase humanness.

This study supports the arguments presented by Følstad and Brandtzaeg (2020) in their study of user experience interacting with customer service chatbots, which found that the driving motivation for using chatbots is pragmatic qualities related to productivity and efficiency (Følstad & Skjuve, 2019). Although participants reported perceiving the chatbot with highly humanlike dialogue as more humanlike, it had no impact on either pragmatic or hedonic qualities of user experience. Hassenzahl (2018) argues the ideal interactive product is high on both pragmatic and hedonic qualities. However, as message interactivity (Go & Sundar, 2019), response time (Gnewuch, Morana, Adam, & Maedche, 2018) and message sequence (Moore, 2018) were identical in both conditions, it is possible that the chatbot was considered «human

enough» to complete the experimental tasks, and any additional humanlike features introduced in the high humanness condition was insufficient to contribute to increased user experience.

Implications for theory

The theoretical implications of this study contributes to the existing knowledge of the relationship between chatbot humanness and user experience. First, although an effect for visual humanness on perceived contingency has been reported by previous research on customer service chatbots (Go & Sundar, 2019). No effect for visual humanness was reported in the current study. Arguably the lack of effect is due to the lack of salient visual stimuli, causing users to ignore it and instead focus on the dialogue. Chatbots, unlike embodied conversational agents, do not use visual humanness as part of the conversation as non-verbal cues, forcing users to proccess the visual information (Wolfert, Robinson, & Belpaeme, 2021). In addition, this study contribute to existing knowledge by demonstrating that while users might percieve a chatbot as more human, and socially present, it does not nessecarily affect user experience. Although humanness was expected to not influence pragmatic qualities (Følstad & Brandtzaeg, 2020). Similar research on customer service chatbots found that personality influenced hedonic qualities of user experience (Smestad & Volden, 2018), suggesting humanness might affect hedonic qualities of user experience. However, as Hassenzahl (2018) argues, hedonic qualities of user experience consists of everything not associated with pragmatic qualities. As such, more than simple humanlike features might be needed to affect user experience.

In addition, this study contributes new knowledge by demonstrating the effect humanness in a chatbots dialogue has on users behavior. Unlike previous research on user behavior which showed evidence of social behavioral orientation after interacting with a chatbot for an extended period of time (Liao et al., 2020). Social behavioral orientation has been shown to occur after only a limited duration interaction, even in a highly task oriented customer service chatbot. In addition, this study demonstrate the importance of measuring user behavior, as even though humanness had no effect on user experience, participants still changed their behavior based on the level of humanness in the chatbots dialogue.

Limitations and future research

The current study has several limitations which should be discussed. The study utilized a field experiment instead of a laboratory experiment. Field experiments are ideal for generalization of the results of the study to the wider population, due to the experiment being

conducted in the participants regular environment, thus increasing the ecological validity of the study (Cook et al., 2002). However, field experiments entail less control for any potential confounding variables that might affect the results, such as distractions in the environment of the participants. Concerning the bordering significant results on user behavior, one opportunity for future research could be to measure the effect of humanness on user behavior in a more controlled laboratory setting. In addition, as no standardized measure of user behavior currently exists, user behavior was measured with a combination of pronouns and social markers. However, different measures of user behaviors such as message length (Hill et al., 2015), socializing questions (Liao et al., 2016), complaints (Liao et al., 2018) and message sequence (Moore, 2018) could be used instead, potentially providing different results. As such, future research may expand on the results from this study on the effect of humanness on user experience in customer service chatbots. Lastly, participants in this study were tasked with asking the chatbot three questions, which resulted in a relatively short interaction duration. Although evidence for increase social behavioral orientation was found by increasing humanness in the chatbots dialogue, no effect for utilitarian behavioral orientation was found. Luger and Sellen (2016) noted that users initially interacted more socially with chatbots, gradually changing to a more utilitarian oriented interaction over time as participants learned how to interact with the chatbots. Liao et al. (2018) found evidence for utilitarian behavioral orientation in their study in which participants interacted with a chatbot over the duration of six weeks. Future research should take this into consideration, designing longer duration interactions with customer service chatbots to observe the effects on user behavior.

Conclusion

This study contributes to existing knowledge by reducing the knowledge gap regarding the effect of a chatbot humanness on user experience and user behavior. As humanness is often utilized in chatbots to increase user experience, the focus in this study was the effect of humanness on user behavior. Findings in the study partially supported the hypothesis and were partially congruent with past research. However, three unexpected findings in this study were considered interesting. 1) Humanness in the chatbots avatar had no effect on any of the measures used in the study, this might suggest the chatbots avatar is insufficiently salient or important for the purposes of the task-oriented chatbot interaction. 2) even though participants acknowledged the chatbot as more human and more socially present with increased humanness, it had no effect

on hedonic qualities of user experience. These findings might suggest hedonic qualities of user experience consists of much more than humanness. 3) Although increased humanness increased the frequency of social behavioral orientation among participants, a lack of humanness in the chatbot did not increase the frequency of utilitarian behavioral orientation above social behavioral orientation. These findings suggest humans tend to behave socially towards chatbots, even in task-oriented situations. These findings can help future research on user behavior in task-oriented chatbots such as customer service chatbots. In addition, these findings can help chatbot designers and developers create chatbots better designed to interact with users.

References

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.

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.

Bergmann, K., Eyssel, F., & Kopp, S. (2012). *A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time.* Paper presented at the International conference on intelligent virtual agents.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI), 12*(2), 293-327.

Brandtzaeg, P. B., & Følstad, A. (2017). *Why people use chatbots.* Paper presented at the International conference on internet science.

Brooke, J. (1996). Sus: a "quick and dirty'usability. *Usability evaluation in industry, 189*, 189-194.

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., . . . McTear, M. (2018). *Assessing the usability of a chatbot for mental health care.* Paper presented at the International Conference on Internet Science.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., & Yan, H. (1999). *Embodiment in conversational interfaces: Rea.* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems.

Chaix, B., Bibault, J.-E., Pienkowski, A., Delamon, G., Guillemassé, A., Nectoux, P., & Brouard, B. (2019). When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR cancer, 5*(1), e12856.

Chaves, A. P., & Gerosa, M. A. (2019). How should my chatbot interact? A survey on human-chatbot interaction design. *arXiv preprint arXiv:1904.02743*.

Cho, M., Lee, S.-s., & Lee, K.-P. (2019). *Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten.* Paper presented at the Proceedings of the 2019 on Designing Interactive Systems Conference.

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

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*: Houghton Mifflin Boston, MA.

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2017). *Superagent: A customer service chatbot for e-commerce websites.* Paper presented at the Proceedings of ACL 2017, System Demonstrations.

Diederich, S., Lembcke, T.-B., Brendel, A. B., & Kolbe, L. M. (2021). Understanding the impact that response failure has on how users perceive anthropomorphic conversational service agents: Insights from an online experiment. *AIS Transactions on Human-Computer Interaction, 13*(1), 82-103.

Doyle, P. R., Edwards, J., Dumbleton, O., Clark, L., & Cowan, B. R. (2019). *Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction.* Paper presented at the Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services.

Følstad, A., & Brandtzaeg, P. B. (2020). Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience, 5*(1), 1-14.

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). *What makes users trust a chatbot for customer service? An exploratory interview study.* Paper presented at the International conference on internet science.

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.

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.

Gnewuch, U., Morana, S., & Maedche, A. (2017). *Towards Designing Cooperative and Social Conversational Agents for Customer Service.* Paper presented at the ICIS.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in human behavior, 97*, 304-316.

Goldstein, S. M., Johnston, R., Duffy, J., & Rao, J. (2002). The service concept: the missing link in service design research? *Journal of Operations management, 20*(2), 121-134.

Griol, D., Carbó, J., & Molina, J. M. (2013). An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence, 27*(9), 759-780.

Hassenzahl, M. (2018). The thing and I: understanding the relationship between user and product. In *Funology 2* (pp. 301-313): Springer.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003* (pp. 187-196): Springer.

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology, 25*(2), 91-97.

Hendriks, F., Ou, C. X., Amiri, A. K., & Bockting, S. (2020). The power of computer-mediated communication theories in explaining the effect of chatbot introduction on user experience. *interaction, 12*, 15.

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.

Hobert, S., & Meyer von Wolff, R. (2019). Say hello to your new automated tutor–a structured literature review on pedagogical conversational agents.

ISO (2019). ISO 9214-210:2019 (E). Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. International Organization for Standardization (ISO). Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). *Evaluating and informing the design of chatbots.* Paper presented at the Proceedings of the 2018 Designing Interactive Systems Conference.

Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in human behavior, 28*(1), 241-250.

Koetter, F., Blohm, M., Kochanowski, M., Goetzer, J., Graziotin, D., & Wagner, S. (2018). Motivations, classification and model trial of conversational agents for insurance companies. *arXiv preprint arXiv:1812.07339*.

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.

Krassmann, A. L., Paz, F. J., Silveira, C., Tarouco, L. M. R., & Bercht, M. (2018). Conversational agents in distance education: Comparing mood states with students' perception. *Creative Education, 9*(11), 1726.

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.

Laban, G., & Araujo, T. (2019). *Working together with conversational agents: the relationship of perceived cooperation with service performance evaluations.* Paper presented at the International Workshop on Chatbot Research and Design.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., . . . Lau, A. Y. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association, 25*(9), 1248-1258.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). *Understanding, scoping and defining user experience: a survey approach.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Lee, K. M. (2004). Presence, explicated. *Communication theory, 14*(1), 27-50.

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International journal of human-computer studies, 64*(10), 962-973.

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.

Liao, Q. V., Geyer, W., Muller, M., & Khazaen, Y. (2020). Conversational Interfaces for Information Search. In *Understanding and Improving Information Search* (pp. 267-287): Springer.

Liao, Q. V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M. P., . . . Geyer, W. (2018). *All work and no play?* Paper presented at the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

Liebrecht, C., Tsaousi, C., & van Hooijdonk, C. (2021). Linguistic elements of conversational human voice in online brand communication: Manipulations and perceptions. *Journal of Business Research, 132,* 124-135.

Liebrecht, C., & van Hooijdonk, C. (2019). *Creating humanlike chatbots: what chatbot developers could learn from webcare employees in adopting a conversational human voice.* Paper presented at the International Workshop on Chatbot Research and Design.

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.

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.

Mairesse, F., & Walker, M. A. (2009). Can conversational agents express big five personality traits through language?: Evaluating a psychologically-informed language generator. In: Cambridge & Sheffield, United Kingdom: Cambridge University Engineering ….

McDonnell, M., & Baxter, D. (2019). Chatbots and gender stereotyping. *Interacting with Computers, 31*(2), 116-121.

Meyer, J., Miller, C., Hancock, P., de Visser, E. J., & Dorneich, M. (2016). *Politeness in machine-human and human-human interaction.* Paper presented at the Proceedings of the human factors and ergonomics society annual meeting.

Moore, R. J. (2018). A natural conversation framework for conversational UX design. In *Studies in Conversational UX Design* (pp. 181-204): Springer.

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

Nass, C., Steuer, J., & Tauber, E. R. (1994). *Computers are social actors.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Niculescu, A. I., & Banchs, R. E. (2019). *Humor intelligence for virtual agents.* Paper presented at the 9th International Workshop on Spoken Dialogue System Technology.

Nowak, K. L., & Rauh, C. (2005). The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication, 11*(1), 153-178.

Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI, 5*, 114.

Powers, A., & Kiesler, S. (2006). *The advisor robot: tracing people's mental model from a robot's physical attributes.* Paper presented at the Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction.

Reddy, T. (2017, October 17). How chatbots can help reduce customer service costs by 30%. Retrieved from https://www.ibm.com/blogs/watson/2017/10/how-chatbots-reduce-customer-service-costs-by-30-percent/

Roy, R., & Naidoo, V. (2021). Enhancing chatbot effectiveness: The role of anthropomorphic conversational styles and time orientation. *Journal of Business Research, 126*, 23-34.

Smartmessage. (2019, July 17). Impressive Numbers and Stats From Chatbot Dominance So Far. Retrieved from https://www.smartmessage.com/impressive-numbers-stats-chatbot-dominance-far/

Smestad, T. L., & Volden, F. (2018). *Chatbot personalities matters.* Paper presented at the International Conference on Internet Science.

Svenningsson, N., & Faraon, M. (2019). *Artificial Intelligence in Conversational Agents: A Study of Factors Related to Perceived Humanness in Chatbots.* Paper presented at the Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37-76): Elsevier.

Thies, I. M., Menon, N., Magapu, S., Subramony, M., & O'neill, J. (2017). *How do you want your chatbot? An exploratory Wizard-of-Oz study with young, urban Indians.* Paper presented at the IFIP Conference on Human-Computer Interaction.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, *54*(6), 1063.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Wolfert, P., Robinson, N., & Belpaeme, T. (2021). A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *arXiv preprint arXiv:2101.03769*.

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). *A new chatbot for customer service on social media.* Paper presented at the Proceedings of the 2017 CHI conference on human factors in computing systems.

Yang, X., Aurisicchio, M., & Baxter, W. (2019). *Understanding affective experiences with conversational agents.* Paper presented at the proceedings of the 2019 CHI conference on human factors in computing systems.

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, *46*(1), 53-93.

Appendix A. Information regarding the study provided to the participants

Virtual agents for customer service – your feedback is wanted

Virtual agents are increasingly available on websites for customer service. When you need information or help, you can ask the virtual agent. You are invited to this study to help us get new knowledge on how virtual agents are experienced.

In this study, your first task is to use a virtual agent for customer service. Then you will give feedback on this experience in a questionnaire. After this, there will not be any other tasks. The duration of your participation will be approximately 15 minutes.

The study is conducted as a collaboration between two Norwegian research organizations: the University of Oslo and SINTEF. Data collection is fully voluntarily and you may withdraw from the study at any point in time without providing a reason for this.

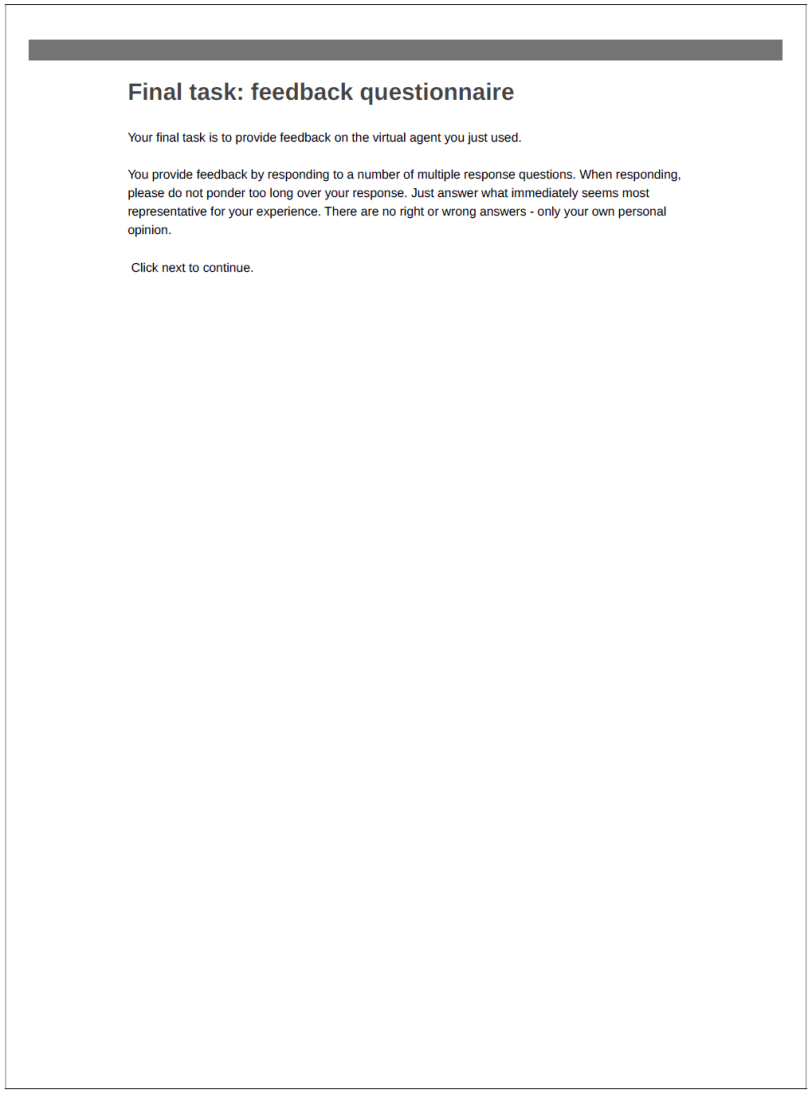
Data collection is fully anonymous. Data from the study will be used as basis for a master thesis and open research publications. The data may be made openly available to support open science.

By clicking next you agree to participate and enter the study.

Next

Responsible for the data collection:
Asbjørn Følstad | Senior researcher, SINTEF | epost: asf@sintef.no | phone: +47 93293013

Appendix B. Questionnaire

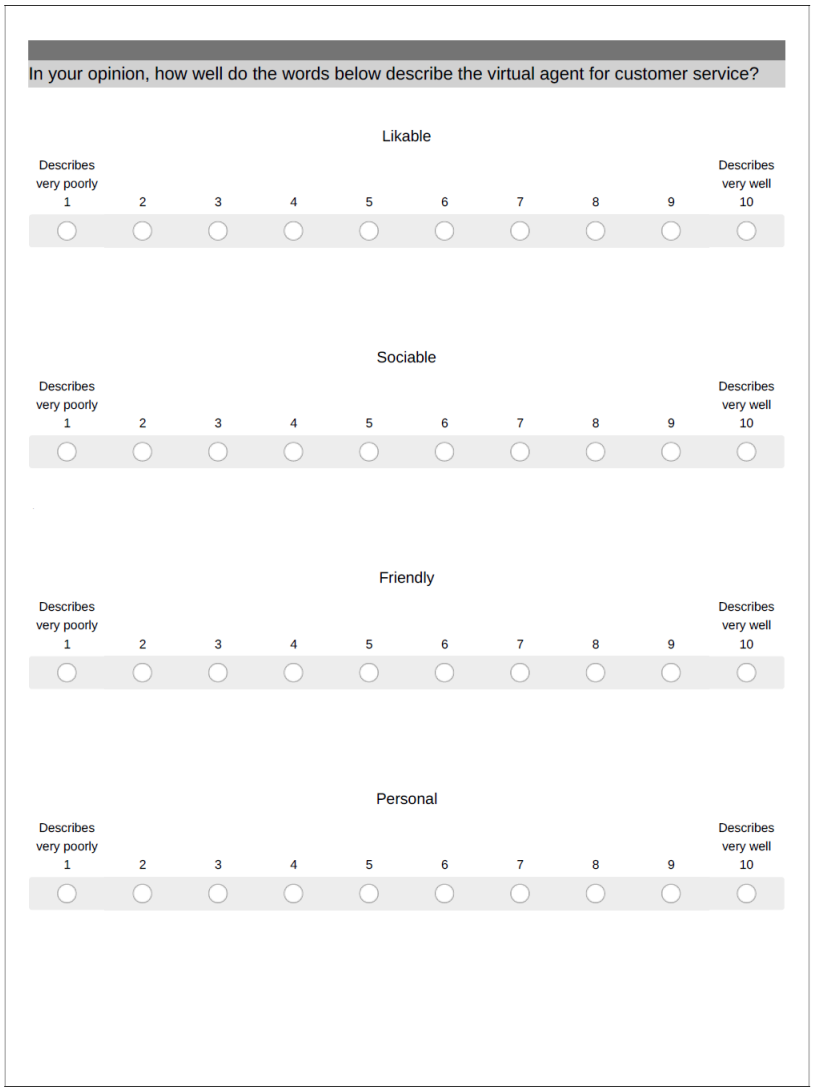


Final task: feedback questionnaire

Your final task is to provide feedback on the virtual agent you just used.

You provide feedback by responding to a number of multiple response questions. When responding, please do not ponder too long over your response. Just answer what immediately seems most representative for your experience. There are no right or wrong answers - only your own personal opinion.

Click next to continue.



In your opinion, how well do the words below describe the virtual agent for customer service?

Likable

Describes very poorly									Describes very well
1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

Sociable

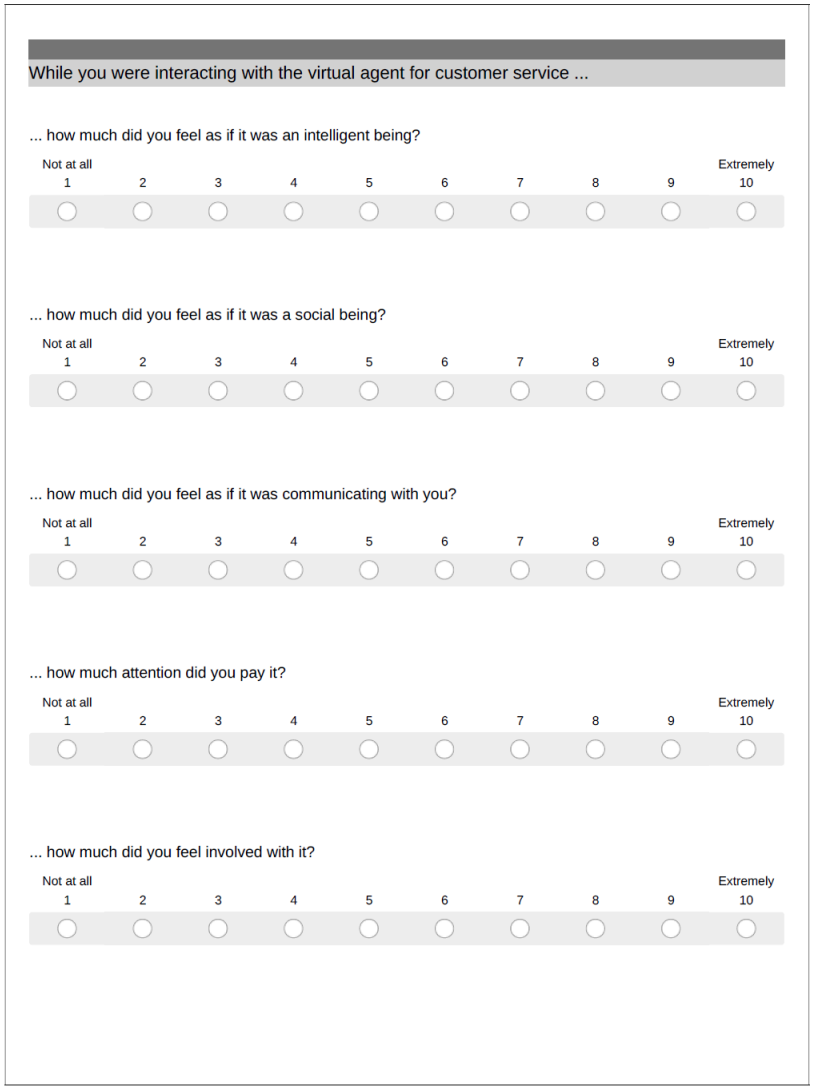
Describes very poorly									Describes very well
1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

Friendly

Describes very poorly									Describes very well
1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

Personal

Describes very poorly									Describes very well
1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○



While you were interacting with the virtual agent for customer service ...

... how much did you feel as if it was an intelligent being?

Not at all 1 2 3 4 5 6 7 8 9 Extremely 10

... how much did you feel as if it was a social being?

Not at all 1 2 3 4 5 6 7 8 9 Extremely 10

... how much did you feel as if it was communicating with you?

Not at all 1 2 3 4 5 6 7 8 9 Extremely 10

... how much attention did you pay it?

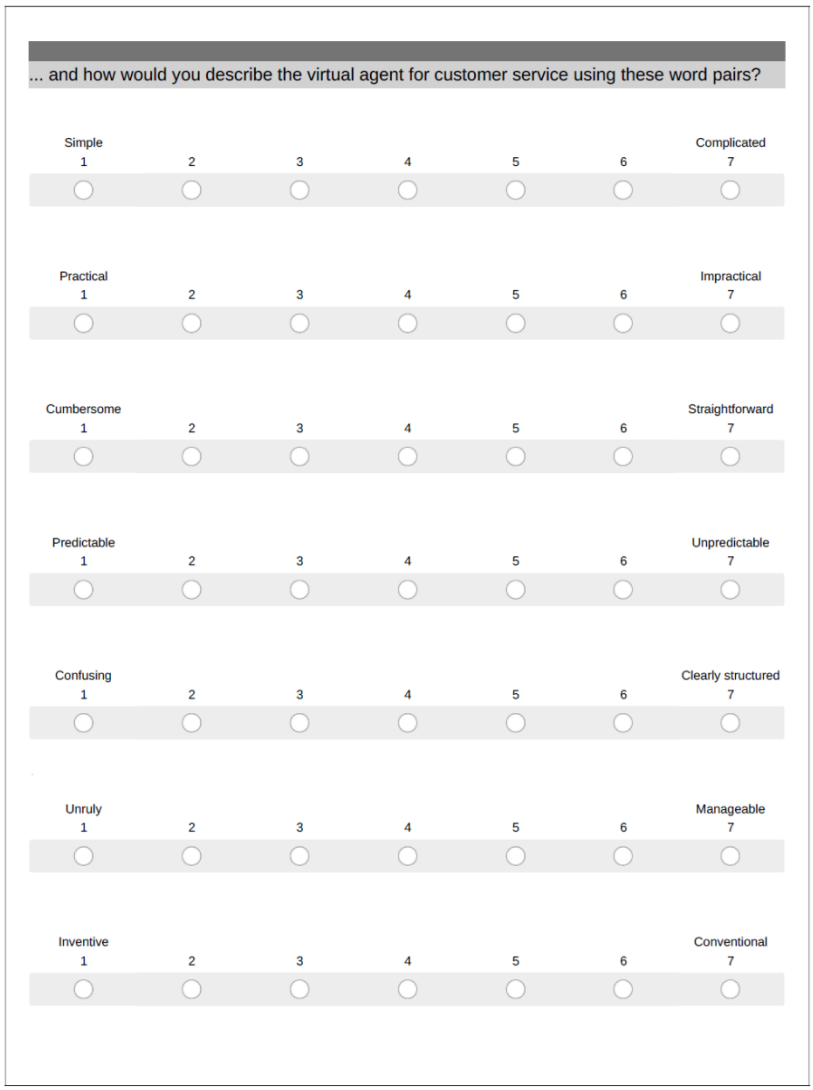
Not at all 1 2 3 4 5 6 7 8 9 Extremely 10

... how much did you feel involved with it?

Not at all 1 2 3 4 5 6 7 8 9 Extremely 10

Considering the word-pairs below, how would you describe the virtual agent for customer service?

	1	2	3	4	5	6	7	8	9	10	
Machine-like	○	○	○	○	○	○	○	○	○	○	Human-like
Unnatural	○	○	○	○	○	○	○	○	○	○	Natural
Artificial	○	○	○	○	○	○	○	○	○	○	Lifelike
Unsociable	○	○	○	○	○	○	○	○	○	○	Sociable
Machine-like	○	○	○	○	○	○	○	○	○	○	Life-like
Insensitive	○	○	○	○	○	○	○	○	○	○	Sensitive



... and how would you describe the virtual agent for customer service using these word pairs?

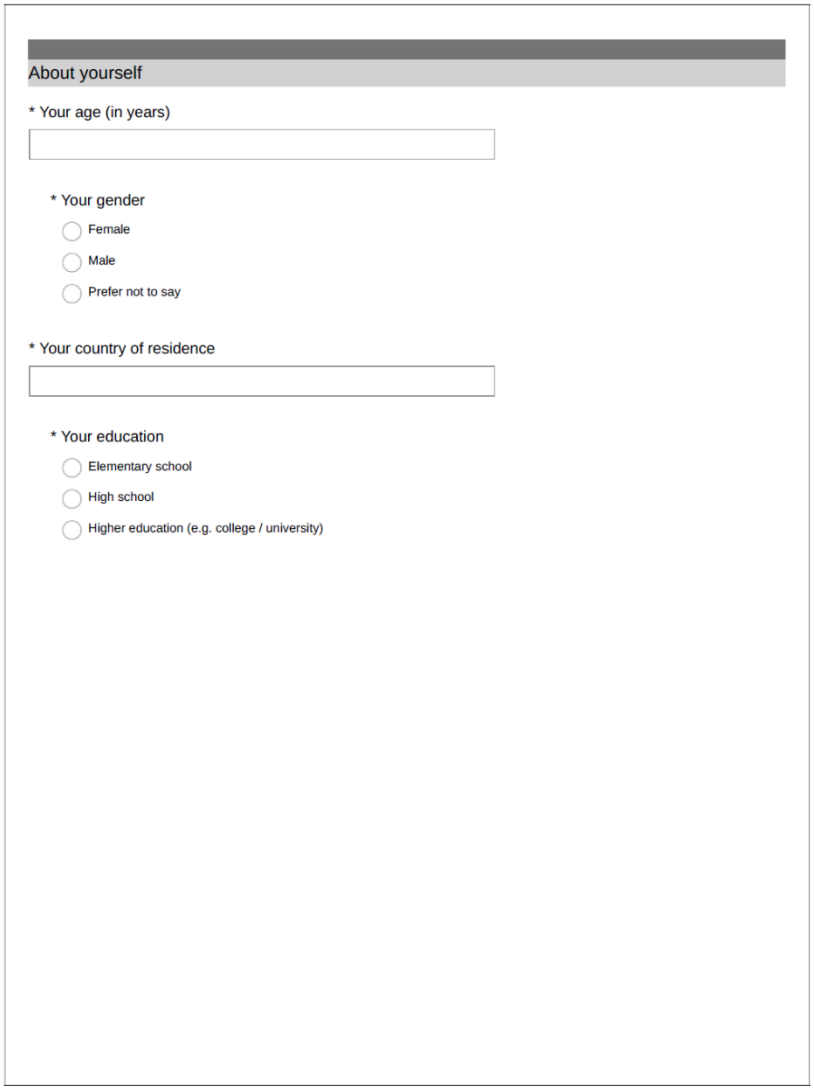
	1	2	3	4	5	6	7	
Simple	○	○	○	○	○	○	○	Complicated
Practical	○	○	○	○	○	○	○	Impractical
Cumbersome	○	○	○	○	○	○	○	Straightforward
Predictable	○	○	○	○	○	○	○	Unpredictable
Confusing	○	○	○	○	○	○	○	Clearly structured
Unruly	○	○	○	○	○	○	○	Manageable
Inventive	○	○	○	○	○	○	○	Conventional


Unimaginative
1
2
3
4
5
6
Creative
7
Bold
1
2
3
4
5
6
Cautious
7
Innovative
1
2
3
4
5
6
Conservative
7
Dull
1
2
3
4
5
6
Captivating
7
Ordinary
1
2
3
4
5
6
Novel
7

Your feedback - in your own words

Finally, we would love to hear your opinion about the virtual agent for customer service in your own words.

* Please provide a short description of your experience with the virtual agent (e.g. 15-25 words)



About yourself

* Your age (in years)

* Your gender

- Female
- Male
- Prefer not to say

* Your country of residence

* Your education

- Elementary school
- High school
- Higher education (e.g. college / university)

Appendix C. Variables and their SPSS input

Variable	SPSS Variable	Coding instruction
Interaction start	StartDate	Time participants started the interaction
Interaction end	EndDate	Time participants ended the interaction
Mindless anthropomorphism	MLA1, MLA2, MLA3, MLA4	10-point semantic differential scale
Items 1, 2, 3, 4		Low value = negative
		High value = positive
Social presence	SP1, SP2, SP3, SP4, SP5, SP6, SP7, SP8	SP1-SP5 = 10-point likert scale
Items 5,6,7,8,9, 13, 14, 15		SP6-SP8 = 10-point semantic differential scale
		Low value = negative
		High value = positive
Mindful anthropomorphism	MFA1, MFA2, MFA3	10-point semantic differential scale
Items 10, 11, 12		Low value = negative
		High value = positive
Pragmatic qualities of user experience	PQ1, PQ2, PQ3, PQ4, PQ5, PQ6	7-point semantic differential scale
Items 16, 17, 18, 18, 20, 21		Low value = negative
		High value = positive
Hedonic qualities of user experience	HQ1, HQ2, HQ3, HQ4, HQ5, HQ6	7-point semantic differential scale
Items 22, 23, 24, 25, 26, 27		Low value = negative
		High value = positive
Age	Age	Age in years
Gender	Gender	1 = male
		2 = female
		3 = other
Nationality	country	Enter the country received by the participant
Educational background	Education	1 = elementary school
		2= High school
		3= higher education
Participants ID	ID	Number assigned to each participant
chatbot condition	condition	Enter the condition of the participants
		1 = MAMC = low humanness avatar, low humanness dialogue
		2 = HAMC = high humanness avatar, low humanness dialogue
		3 = MAHC = low humanness avatar, high humanness dialogue
		4 = HAHC = high humanness avatar, high humanness dialogue