Why would I talk to you?

Investigating user perceptions of conversational agents

Anne Cathrine Saarem
Department of Product Design
Norwegian University of Science and Technology

ABSTRACT

This article looks into the use of Artificial Intelligence (AI) agents in Conversational User Interfaces (CUIs), and their effect on the user experience. A literature review was conducted, exploring how the human tendency to anthropomorphise computers that talk affects how we perceive and interact with them. Ethical issues such as trust, responsibility, and stereotyping of agents in relation to gender were also investigated. Based on this insight, the anthropomorphic traits of two commercial chatbots were evaluated. The results from this evaluations is presented, along with a set of proposed guidelines for anthropomorphic chatbot design and suggestions for further research. If chatbots are to be employed in commercial services, understanding their anthropomorphic qualities and their effect on user perceptions is crucial, as indiscriminate application of human-like traits can have unwanted effects. Consequently, a knowledge of the benefits and possible pitfalls of anthropomorphism is essential in the design of these types of systems.

KEYWORDS: Chatbots, Conversational Agents, Anthropomorphism, Human-Computer Interaction, User Experience

1. INTRODUCTION

The use of Artificial Intelligence (AI) agents in Conversational User Interfaces (CUIs) is becoming more and more common. In fact, 2016 has seen an upsurge of online articles, blog posts and discussions on the topic of chatbots and their business potential[1]. Many are predicting a future of conversation-based interactions where everything from ordering flowers to scheduling meetings will be in the form of requests made out to chatbots[2]. Simultaneously, technology companies such as Facebook and Google are launching digital assistants in their messaging applications, and voice user interfaces such as Apple's Siri, Microsoft's Cortana, and Amazon's Alexa are becoming household names.

As with many other technologies, academic research and technological advances have been the driving forces behind chatbot development. Most existing research on conversational agents reflects this by focusing on the technical capabilities and implications of the systems they are discussing. However, if chatbots are to be as useful and encompassing as some predict, there is also a need to establish principles for chatbot design in relation to user perceptions and experiences. What effects will this form of interaction have on the end user? How do we respond to and interact with computers that talk to us, and why would we want to talk back? This article aims to answer these questions by looking into existing research on user perceptions of conversational agents, and use this insight to evaluate two commercial chatbots. Hopefully, the findings from these studies can contribute to the ongoing discussion of chatbots – their potential, usefulness and limitations, and be a stepping stone towards a more established set of principles for user-centred chatbot design.

1.1 Artificial Intelligence today

Artificial Intelligence (AI) systems, has become as ubiquitous as the smartphone in the past decade.

This is all thanks to the advent of big data, and advances in AI technologies such as neural networks and machine learning[3]. The last couple of years have seen a renewed interest in AI; self-driving cars are now a real possibility and IBM's Watson[4], who recently won Jeopardy, is able to diagnose patients and recommend treatments. For most of us, the term Artificial Intelligence is associated with the likes of Siri and Cortana as well as the operating system 'Samantha' in Spike Jonze's film 'Her'. These are digital assistants, powered by AI, that are created to help us with our everyday tasks.

AI comes in many forms, and researchers have different views of the field. According to Duffy[5], supporters of what Searle once dubbed 'strong AI' have a mechanistic view of the human mind and believe it is possible to duplicate human intelligence in artificial systems. Proponents of 'weak AI' on the other hand, consider the term 'Artificial Intelligence' to be an oxymoron, as the expression implies a simulation of human intelligence. Still, they believe this simulation can be useful for studying the mind, as a computer is 'useful for studying the weather, economics, or molecular biology[6 p.26]'. The position of weak AI is that the only way for a system to become intelligent is to create an illusion of intelligence, primarily through projection on the part of the human observer. Duffy follows this stance and asks whether it matters how a system achieves its intelligence: 'Can the illusion of life and intelligence emerge through simply engaging people in social interaction? How much can this illusion emerge through people's tendency to project intelligence and anthropomorphise?[5]'.

The focus of this article is the user perception of the form of AI that requires direct interaction between user and software system. That is, through conversation. Following the stance of weak AI and Duffy, as long as the user *perceives* a system as intelligent, it is considered an AI system for the purpose of this article.

1.2 Chatbots and conversational agents

A chatbot system is defined as a 'software program that interacts with users using natural languages[7 p.29]', either through text, voice, or both. Chatbots are considered a form of weak AI[8], and they are usually powered by a set of scripted rules, AI, or a combination of AI and human assistance. According to Shawar and Atwell, the purpose of a chatbot is 'to simulate human conversation[7]'. There are many different terms used to describe chatbots, such as conversational agents, dialogue systems, chatterbots, machine conversation systems, virtual agents and so on. For the purpose of this article, the terms conversational agent and chatbot will be used interchangeably to describe these types of systems. Another variation of the chatbot is the Embodied Conversational Agent (ECA). Unlike chatbots, ECAs are often employed as service agents on commerce websites and employ a human-like appearance and behaviour (e.g. gestures and facial expressions) as a way to mimic human-human conversation[9]. Research into user perception of chatbots and ECAs often overlap and lessons learned from one area is likely to apply to the other.

1.3 Anthropomorphism

Anthropomorphism is the phenomenon of seeing nonhuman objects as having human-like qualities. The Oxford English Dictionary defines anthropomorphism as 'the attribution of human personality or characteristics to something non-human, as an animal, object, etc'. The term has its origins in the Greek words anthropos for man and morphe for form/structure, and different academic disciplines, such as psychology, Human-Computer Interaction (HCI), and biology, interpret anthropomorphism differently.

2. METHODS

This article aims to explore how humans perceive computers that talk, and how this knowledge can be utilised in chatbot design. It is based on a literature review of different approaches to

chatbot research, as well as interpretations of anthropomorphism and the use of this knowledge in chatbot development. The search covered the disciplines of HCI, psychology, Human-Robot Interaction (HRI), communication, and sociology, as well as keywords such as anthropomorphism, conversational agents, chatbots, social agents and digital assistants. Sources were gathered from digital information resources such as Association for Computing Machinery (ACM), ScienceDirect, Google Scholar, and Oria. Articles discussing the human implications of chatbot design in terms of perception, reception, and ethical implications, were selected based on their relevance to the topic and contribution to existing research.

In addition, a user evaluation of two commercial chatbots with focus on user perception and response was conducted. The evaluation consisted of a semi-structured interview, followed by a user test of the chatbots. This test included the task of becoming acquainted with the chatbots' capabilities, giving them a compliment, insulting them, and ending the conversation. Six students participated in the evaluation, all of which had limited experience with conversational agents. Notes were taken of the participants' reactions and responses to the chatbots actions, and supplementary questions were asked based on these observations. After the tests were completed, screenshots were saved to document the conversations, which were then analysed in combination with the notes from the interviews. The main findings from the evaluations can be found in section 4, followed by a list of proposed guidelines for anthropomorphic chatbot design.

3. LITERATURE REVIEW

3.1 The utility and limitations of chatbots

The earliest chatbots, such as Weizenbaum's ELIZA[10], were academic exercises with the purpose of imitating a human interlocutor. Current research, on the other hand, aims to establish practical and commercial applications for chatbots in fields such as education, health, and e-commerce. Improvements in data-mining and machine-learning techniques, decision-making capabilities and more have made today's chatbots more practical, and chatbot usage in digital services has become increasingly popular in the last couple of years[7]. The question is, how useful are the chatbots of today, and for what domains are they best suited? In addition, what are the benefits and limitations of conversational user interfaces (CUIs) compared to graphical user interfaces (GUIs)?

In a study examining the usefulness of chatbots, Shawar and Atwell assessed different chatbots used in practical domains such as language learning, information retrieval and e-commerce[7]. A case involving a prototype of a shopping assistant showed that the chatbot was perceived as easy to use and the users preferred it over a menu-driven search. This was because people liked being able to express their needs in their own language, they felt that the computer did most of the work for them (i.e. retrieving the relevant information), and thought the chatbot reduced the overall interaction time. Another case regarding a language learning chatbot presented many advantages to using these types of systems, such as availability, repetition of material, and allowing for self-analysis and evaluation[11]. Despite this, Shawar and Atwell agreed with Knill et al.'s conclusion that a chatbot could enrich the teaching process, but never replace the teacher's role in the classroom[12]. They concluded that the general aim for chatbot designers should not be to mimic human-human conversation or replace the human role entirely, but rather to utilize the strengths and capabilities of these systems to create helpful tools that can simplify people's work[7].

Zadrozny et al. saw the use chatbots as a compelling way to facilitate and personalize HCI, as it enables people to use their own words when interacting with a system, rather than having a small number of predetermined input options to choose from[13]. They envisioned different future scenarios where conversation-based interaction could enable smooth exchanges between customers and services in fields such as banking,

information retrieval, and investment advice. Still, they acknowledged the technological limitations of conversational agents, especially in terms of their Natural Language Processing (NLP) capabilities. Words are arbitrary symbols with different interpretations in different contexts, and AI researchers have yet to unlock the secret to understanding natural language completely[14]. This is seen as the greatest challenge in AI, as the systems of today can only follow simple commands and answer queries, but as of yet, they have no real understanding of the words they use. Due to these technical limitations, today's commercial chatbots work best in limited domains, such as giving information about a constrained topic (e.g. the weather).

Chatbots are also limited in terms of their interaction context. CUIs rely on spoken and/or written language, as opposed to GUIs, which are visual representations of possible actions in a system that allow for direct manipulation[15]. Because the brain processes language sequentially, we interpret verbal information more slowly than visual input, which is processed in parallel[16]. In addition, typing out queries on a keyboard can be a time-consuming process compared to pressing a button (although time spent finding the correct button must be considered in this estimation). Existing chatbot interfaces such as Google Allo' and Facebook Messenger attempt to solve this by utilising images, GIFs, and emoticons as well as quick reply buttons as a way to combine visual and verbal input and output, and thus simplify the conversation (see figure 1).

3.2 How we perceive computers that talk

Regardless of application, chatbots and digital assistants are systems based on conversational interaction between man and machine. As conversation is an essential component in social interaction, successful implementation of conversation-based interfaces will require more research into social interaction between computers and humans. How can we ensure that a conversation with a virtual agent is engaging and

worthwhile for the user? According to Duffy, social interaction is 'fundamentally observer-dependent[5 p.177]' and exploring the underlying mechanisms of anthropomorphism is essential in discovering the features required to make a machine socially engaging.

The term anthropomorphism is used differently in disciplines such as psychology, the natural sciences, and HCI. Duffy describes it as a phenomenon of 'ascribing human-like qualities to non-human entities[5 p.180]'. He sees anthropomorphism as a metaphor and a useful mechanism to facilitate social interaction between humans and machines. Waytz et al. define anthropomorphism as 'attributing capacities that people tend to think of as distinctly human to non-human agents[17 p.220]'. Their stance is that anthropomorphism is not a universal tendency but rather that the degree of anthropomorphism is dependent on the individual and his or her experience, culture, norms, and education. They stress the importance of understanding the determinants of the phenomenon, as this can help practitioners identify the conditions under which the effects of anthropomorphism will enhance HCI by increasing engagement and perceived intelligence.

Holtgraves et al. claim that 'one of the most uniquely human qualities is the ability to use language[18 p.2164]' and that chatbots exhibit this to some degree – despite the fact that they have yet to attain perfect conversational ability. Through experiments on human-chatbot interaction, they showed that people, while being aware of talking to a computer program, regarded a chatbot as having a personality and that communication factors such as level of politeness and response time affected their perceptions of it. Based on these findings, they suggest that conversational agents can work well as contact points on websites, but question whether the agent should be highly anthropomorphic or 'convey a distinctly non-human, robot-like persona[18 p.2172]'. They conclude, similarly to Waytz et.al, that the degree of humanness depends, in part, on the specific anthropomorphic

qualities chosen to include in the bot, as well as personal and cultural preferences.

Nass and Moon see anthropomorphism as a 'thoughtful, sincere belief that an object has human characteristics[19 p.93]', a literal interpretation that is in contrast to Duffy's metaphorical view of the phenomenon. They state that a 'computer is not a person and does not warrant human treatment or attribution[19 p.82]'. Working under the Computers Are Social Actors (CASA) paradigm, Nass and colleagues recreated studies from Human-Human Interaction (HHI) by replacing a human participant with a computer and thus demonstrating that principles from HHI apply to HCI. Their results showed people using polite behaviour towards computers, judging them based on their gender-like qualities, and feeling team spirit after being grouped with computers in specific tasks. Based on this, they state that people 'mindlessly apply social rules and expectations to computers[19 p.81]'. To them, the presence of contextual social cues such as interactivity, natural speech, and the filling of social roles, triggers automatic social responses in humans that are not actively processed. Their conclusion is that explanations such as anthropomorphism cannot justify the results of their studies, as all participants denied assigning human traits to computers, stating that this would be inappropriate. Kim and Sundar question this and ask whether anthropomorphism 'has to be necessarily mindful[20 p.241]'. They argue that the mindless social responses that Nass and Moon refer to as 'social cues' could very well be labelled anthropomorphic as they 'serve to remind individuals of human attributes (e.g. voice, language style, degree of introversion, etc.). on the interface[20 p.242]'.

3.3 Expecting too much

Whether consciously or not, people assert human-like qualities such as personality and sociability to conversational agents. This affects the way we see and interact with them in both positive and negative ways. Studies show that affective conversational agents can have a positive influence on users in certain domains such as education[11], therapy[21], as well as social interaction and companionship[22]. However, a high degree of anthropomorphism can also lead to disappointment if the system fails to live up to user expectations[5].

Shneiderman expressed concerns regarding the use of anthropomorphic traits in user interfaces. He believes that 'attributions of intelligence, autonomy, free will, or knowledge to computers can deceive, confuse, and mislead users[15 p.380]'. As of today, conversational agents such as Siri and Cortana have limited capabilities, especially when it comes to NLP; yet, their design and advertisement disguise this capability gap by appealing to our tendency to anthropomorphise. The reality however, is that these systems can only answer simple queries such as weather conditions, perform searches and set reminders, and we as users have to carefully adapt our language to something the machine can understand in order to achieve this. This discrepancy between user expectations and the agents' capabilities can cause a lot of frustration and discontent, as evidenced by the countless YouTube videos of people asking Siri, Cortana, and other digital assistant questions ranging from purely silly to impossibly complex in order to test their limits. Videos like these are usually made with the intention of amusing people, but at some level, they also reflect a general disappointment in systems that do not live up to their promises.

3.4 Trust and responsibility

Conversational agents are based on the premise of user cooperation, as a conversation requires collaboration and trust from both parties. If chatbots are to be employed in areas such as commerce and health, the latter becomes particularly crucial. Why make a chatbot that gives medical advice if people question its competence and credibility?

Cassel and Bickmore argue that conversation can demonstrate trustworthiness and that the conversational strategies in HHI, such as

greetings, small talk, and self-disclosure, can be used as way to increase user trust in conversational agents. By applying social (or anthropomorphic) cues in its design, the agent is seen as more reliable, knowledgeable, and competent[23]. Nass and Moon reported the same effect in their experiment on reciprocal self-disclosure[19]. Here, the participants that interacted with a computer that revealed information about itself before asking them personal questions responded with more intimate details than those who were asked the same questions in a straightforward manner.

Another way to look at the issue of trust is in relation to responsibility. Waytz et al. pose the question of whether an agent that is perceived as intelligent also can be held accountable for its actions, and if so, will the humans controlling those agents appear less responsible[17]? Gray et al. showed that people are more willing to penalise an agent they consider mindful[24] and Serenko demonstrated that people can treat chatbots as scapegoats when things go wrong[25]. The opposite might also occur, as shown in a study by Nass and Moon. People paired with a computer displaying similar personality traits to themselves were more likely to blame themselves and less likely to blame the computer for negative outcomes of collaborative tasks. When the outcome was positive, on the other hand, they were more likely to give credit to the computer instead of themselves[26]. These studies show that although perceived trustworthiness and competence can increase with the increase of anthropomorphic traits in chatbot design, it also poses the risk of making people feel less in control and less responsible for their own actions.

3.5 Enforcing stereotypes

Despite studies showing that utilizing anthropomorphism in the design of technological devices can enhance HCI by making them appear more understandable and predictable[17], it can also lead to unforeseen side-effects. For this reason, deciding which anthropomorphic cues to apply in the design of conversational agents is something that requires careful consideration. Over-expectation and confusion about a system's capabilities is one of the risks involved, as well as dislike of the agent (e.g. Microsoft Office's 'Clippy'). Another consequence of ill-applied anthropomorphism is the prevalence of enforced stereotypes. A stereotype is a fixed idea people have about what someone or something is like, which is closely linked with prejudice[27]. When it comes to chatbots and ECAs, considering stereotypes regarding gender and ethnicity becomes especially important. Rossen et al. showed that people with racist tendencies were more empathetic towards ECAs with a light skin tone than ones with a darker one[28] and Nass and colleagues demonstrated that people perceived a female-voiced computer as less friendly than a male-voiced one. This was despite the fact that they both provided the user with the same information[29]. In addition, the participants viewed the male-voiced computer as more competent and knowledgeable about technical topics, whereas the female-voiced computer was perceived as more informative about subjects concerning love and relationships. These and other studies has led to a consensus in the field of HCI that 'the standard of believability has become inextricably linked to gender personification, especially female personification [9 p.174]'.

Most of today's virtual assistants have female names (Siri, Cortana, Alexa) and voices, and the use of female representations is widespread in the field of ECAs. Brahnam and Weaver showed that female ECAs frequently employ the roles of assistants and service providers, while male ECAs are often portrayed as doctors and experts in technological fields[9]. According to Branham, 'most female ECAs are specifically scripted to conform to stereotypical specifications of what it means to communicate as a woman[9 p.175]', such as being compliant, helping and nurturing. De Angeli agrees that, whether intentionally or not, ECAs are designed with gender-specific intentions, which enforces and encourages stereotypes, leading to objectification and abuse that might affect real women[30]. She argues that

stereotypes create an 'us versus them' mentality which may 'induce and justify anti-social behaviour[30 p.55]' such as sexism and racism.

Studies have shown a prevalence of verbal abuse, such as swearing, name-calling, threats and ridicule in anonymous conversations with online chatbots[8]. As chatbots are software programs, they cannot feel pain and abuse the way humans do, and most are not programmed to respond to this sort of behaviour (e.g. Microsoft's Tay [31]). By appealing to our tendency to anthropomorphise, we are also creating software programs that could be loved, or possibly abused by someone. In light of this, designing conversational agents that enforce ethnic and gender stereotypes can affect how these groups are being treated in real life, and this is something that requires careful consideration. One approach could be to design out unsavoury behaviours and make conversational agents capable of handling abuse more appropriately[32]. In a recent article, Brahnam and Weaver propose a rhetorical approach to the design of conversational agents as way of increasing their credibility (as opposed to believability)[9]. Their recommendations include refraining from exaggerating gender representation and sexualisation in the design of ECAs, as well as randomly varying the form of embodiment applied to the agent (e.g. age, gender, ethnicity, height, and weight). They see this periodical altering of the agent's physical appearance as way of showing diversity, and avoiding the reinforcement of stereotypes[9].

Although stereotyping in relation to gender and ethnicity is mostly relevant in design of ECAs, where the agent's physical appearance and behaviour is perhaps the strongest anthropomorphic cue, the lessons from research into this topic are relevant for designers of all types of conversational agents. As soon as we give a chatbot a name, a language style, and (possibly) a human voice, we also give it human-like qualities that affect the way users perceive and interact with them. Being aware of the possible consequences of the design choices we make is thus extremely important.

4. USER PERCEPTIONS OF PONCHO AND THE GOOGLE ASSISTANT

The following section presents the results from six user evaluations of Poncho[33] and the Google Assistant[34]. These chatbots were selected based on their capabilities and anthropomorphic qualities. Although existing in different chat-applications, their input/output modalities are somewhat similar, allowing for little change in interaction when switching between them (see figure 1).

The purpose of this evaluation was to investigate the chatbots in relation to user perceptions and responses. Most existing evaluation methods focus on the utilitarian aspects of chatbots, such as measuring task achievement, believability, and general performance[35]. Although important, these are only part of a chatbot's overall capabilities. The findings from the literature review show that aspects such as perceived humanness, expectations, and gender associations, are important to consider as well, as these are all essential components that make up how people will perceive and respond to a chatbot. For this reason, the evaluation focused on answering the following questions: Do people respond to the chatbots as if they were human? Does this affect the way they perceive them, and would they want to interact with them again? Do people prefer interacting with highly anthropomorphic chatbot, or a more computer-like one?

4.1 Poncho and the Google Assistant

Poncho is a chatbot whose expertise is weather information. Although limited in language understanding, Poncho utilises many anthropomorphic traits, such as unprompted self-disclosure, and use of personal pronouns as a way to increase user engagement. The chatbot also claims to be a male cat. Poncho has built-in responses to off-topic behaviour such as expressions of affection and insults. If a user compliments Poncho, he or she is asked on a date. If they insult the chatbot however, Poncho will ask

them to apologise, or threaten to leave the conversation.

The Google Assistant is far more advanced than Poncho in terms of capabilities and language understanding. On the other hand, it uses fewer anthropomorphic traits. For instance, Poncho has a name and an iconic profile picture (a cartoon cat), whereas the Google Assistant has an impersonal name and an abstract icon as its profile picture. When asked about its gender, the Google Assistant's response is 'I try to stay neutral' and it reacts to compliments and insults with a deflective comment, deliberately steering the conversation back to task-related topics.

Figure 1: Screenshots from conversations with the Google Assistant (left) and Poncho (right).

4.4 Main findings

The user evaluation was conducted as described in section 2. The following is a summary of the most significant findings.

Perceived human likeness
All but one of the participants perceived Poncho and the Google Assistant as humanlike, and they attributed intentions and feelings to their messages. Poncho was described as nice, enthusiastic, and cute, but three out of five found it annoying and demanding when asked several questions in quick succession. The Google Assistant was perceived as less human-like. It was described as a service, a robot, and a computer, but also as helpful, more factual, and less annoying than Poncho.

Conflict about social responses
Several participants expressed conflicted feelings about talking to the chatbots. Half found Poncho's comments amusing, while the other half found them creepy. One said it was odd knowing that she was talking to a computer, but at the same time feeling as if she was talking to a friend.

Low expectations and feeling of control
None of the participants expected either of the chatbots to be very intelligent. Despite this, several expressed disappointment when the chatbots did not understand their requests, and sa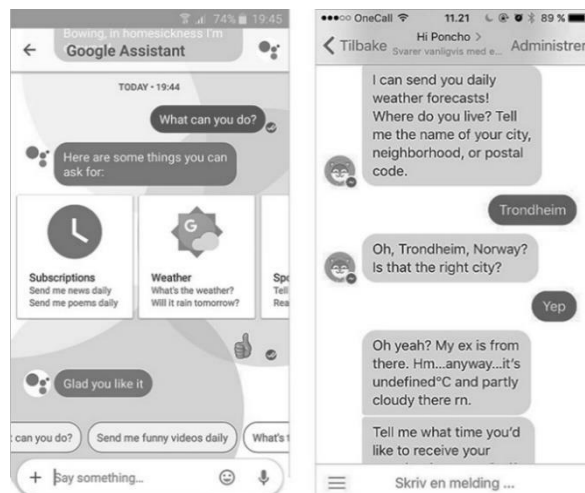id that this made them seem less intelligent. It was noted that the quick reply buttons helped simplify the interaction by suggesting answers for them, but they also made some of the participants feel less in control of the conversation, as they limited the available options.

Compliments and insults
Poncho asked all the participants on a date when they complimented it. Half found this funny, while the other half deemed it inappropriate and odd. After insulting Poncho, two out of six felt guilty, while the rest did not care. One stated that she was putting too much personality into the bot. Another said that he wanted to apologize as h felt uncomfortable being rude to Poncho, despite knowing that he was talking to a chatbot. There were mixed opinions concerning the Google Assistant's responses as well. Some preferred these, describing it as more objective and focused on providing assistance rather than being personal. Others thought its responses were boring, and one stated that it was easier being rude to the Google Assistant because it seemed less human.

Perceived gender
When asked, all participants, both male and female, thought Poncho was male. This was attributed to its name, profile picture, and general behaviour. The female participants also stated that the date invitation made them see it as male.

The Google Assistant on the other hand, was considered gender neutral by all the participants.

Preference
Five out of six participants preferred the Google Assistant. They described it as less personal and more to the point. One stated that 'If I am to use something like this, I'm interested in getting information, I'm not interested in talking to a machine (…) I would rather talk to people'.

5. DISCUSSION

5.1 Implications of the findings

The chatbots' anthropomorphic traits influenced the participants' impression of and interaction with them. They all responded to them in a social manner, and attributed intentions and feelings to both, though in varying degrees. This coincides with the findings from the literature review, especially Waytz et al.'s theory that anthropomorphism is individually dependent[17]. Whilst acknowledging that they were talking to a computer, all of the participants referred to Poncho as 'him', and said they felt as if they were talking to another person. In situations where the chatbot did not understand them, some expressed feelings of disappointment, saying that this made it clear that they were talking to a software program and not a person. This is in line with Shneiderman's argument that users might be confused and feel disappointment in systems that do not live up to their expectations.

While Poncho was seen as the most human-like of the two, the majority preferred interacting with the Google Assistant. It was perceived as more concise, factual, and less annoying than Poncho. One participant stated that the Google Assistant was more helpful, while Poncho kept trying to be her friend. Another said that she would probably only use Poncho for entertainment purposes such as showing it to her friends and testing its limits for fun. This implies that a applying many anthropomorphic traits is not necessarily the best option in chatbot design. While it can create excitement about the product, it also poses the risk of the chatbot being seen as annoying and tiresome once its novelty wears of.

Although the results from the evaluations correspond with findings from the literature review, it is important to keep in mind that this was a single case experiment, which might not be generalizable. The participants interacted with the chatbots in an unnatural setting where they were being observed and asked to think aloud. This might have affected how they interacted with the chatbots, what they said, and some information might have been missed in the note-taking process. In addition, the participants were asked to compliment and insult the chatbots in order to observe their reactions to the chatbots' responses. This is not something they might have done if they had not been asked to do so.

Another aspect to consider is that every conversation with a chatbot is different. Different commands and questions elicit different responses, creating new and unforeseen situations and experiences. For this reason, no two conversations in the user evaluations were alike (although they had similar traits). Poncho, for instance, has limited language understanding and there were times when the chatbot would 'break', and become stuck in a specific conversation pattern, regardless of user input. This might have affected the overall results. In addition, all participants talked to Poncho first, then the Google Assistant, which might have affected the participants' impressions of the two chatbots.

5.2 Proposed guidelines for chatbot design

This article has explored issues concerning the anthropomorphic aspects of chatbots, and their implications for the user perceptions of these types of systems. The following is a set of proposed guidelines for anthropomorphic chatbot design, which were developed based on the findings from this study:

Make it human-like, but only when necessary
Although basing a chatbot's behaviour on its human counterpart can be fruitful, there is no need to recreate the human mind in its entirety. A

chatbot is a software program, not a human, so why should it have a complex human-like personality? Rather than trying to make a chatbot as human-like as possible, it could be beneficial to focus on selecting the traits most relevant in a particular setting.

Design for the appropriate social context
The conversational setting of chatbots involves components of social interaction. Thus, it is important to consider what conversational strategies fit the situation at hand. For what purpose is the chatbot meant to be used, and when and how are people going to use it? For instance, a digital assistant designed to keep track of your appointments might use a more personal language than a shopping assistant chatbot, which might be more formal.

Acknowledge technological limitations
Achieving perfect language understanding is still a long way into the future and overselling a chatbot's capabilities will only lead to disappointment. Being upfront about what the chatbot can and cannot do can help alleviate the situation.

Avoid repetition
Talking to someone who does not remember what you are saying can be frustrating. The same goes for people who constantly repeat themselves. Memory retention, albeit technically difficult, as well as varying the chatbot's responses in a given situation, will be essential in ensuring a good conversation flow.

Account for flexible conversation flows
Because of the conversational context, user interactions with chatbots do not follow a set pattern. Allowing people to write anything they want instead of choosing between buttons means having to be prepared for anything. Creating fixed conversation flows for specific topics might work in certain situations, but it also limits the way people can interact with a chatbot and makes it seem less intelligent.

Design for long-term use
A highly anthropomorphic chatbot with funny remarks and jokes might be entertaining for a first time user, but it can also be tiresome once the novelty wears off. Gradually decreasing the degree of anthropomorphic traits, such as impromptu self-disclosure, as time passes could make the users more tolerant of the system.

Account for unsavoury behaviour
Situations where people make inappropriate advances, insult, and otherwise test the chatbot's limits are very likely to occur. Consequently, designing ways for chatbots to handle these types of situations will be important.

Consider gender representation
Gender is a sensitive subject and should thus be approached carefully. As a chatbot is a software program, assigning a gender-specific identity to it should not be a goal in itself. However, certain anthropomorphic traits can create certain gender-like associations. Therefore, considering how the implicit or explicit gender identity of a chatbot contributes to the overall perception of that system and the possible consequences of this is very important.

These guidelines do not aim to describe the definitive criteria for chatbot design, but rather to provide a list of important aspects to consider in the design of conversational agents. Collectively, they could work as a starting point towards a more established set of principles for user-centred chatbot design.

5.3 Further research

Further work in this area could be to test and refine these guidelines by looking more closely at the different variables that can affect people's opinions about conversational agents. Is it possible to uncover which anthropomorphic qualities will be appropriate in different situations and social contexts? This would require close collaboration with researchers from disciplines such as psychology, sociology, and communication theory. In addition, the conversational aspect of chatbot means that most

of a chatbot's anthropomorphic qualities will come from its language style. Because of this, more insight into theories from linguistics and how this can be incorporated in chatbot design will be essential.

6. CONCLUSION

In light of the findings from this study, it is possible to conclude that anthropomorphism, despite its different interpretations, plays a central role in our perception of chatbos. If applied thoughtfully, anthropomorphic traits can be utilised in chatbot design as a way to increase user engagement and preference of a commercial service, and make a chatbot appear more understandable and predictable to the user. Contrariwise, indiscriminate application of anthropomorphic cues can create expectations of capabilities that a system, due to technological shortcomings, is not yet able to meet, and possibly enforce stereotypes, which might in turn affect real people. Based on this, a set of proposed guidelines for chatbot design have been developed, and presented, along with suggestions for further research into this topic.

Chatbots are technologically limited, and as of yet they can do little beyond answering simple requests. However, if these systems are to be employed in commercial services in the future, it is important to understand their anthropomorphic qualities, and how this affects the end user's perception of them. Our anthropomorphic dispositions go hand in hand with our tendency to like and trust computers that talk to us and (appear to) understand us. Although an advantage for chatbot designers, this can also be exploited, as the blurred line between human and computers can make us easy targets for manipulation and propaganda (e.g. [36]). Transparency will thus be important in ensuring that users know they can trust the systems with which they are interacting. In sum, a thorough knowledge of the benefits and possible pitfalls of anthropomorphism is essential in the design of conversational agents. By considering these aspects, chatbot designers have the opportunity

to avoid unwanted side effects and create enjoyable services, benefiting both businesses and the end user.

7. REFERENCES

[1] M. Schlicht. (2016, April 20) The Complete Beginner's Guide To Chatbots - Everything you need to know. *Chatbots Magazine*. Available: https://chatbotsmagazine.com/the-complete-beginner-s-guide-to-chatbots-8280b7b906ca#.gf3tpfbjx [Accessed: 21-Oct-2016]

[2] C. Messina. (2016, January 19) 2016 Will Be the Year of Conversational Commerce *Medium*. Available: https://medium.com/chris-messina/2016-will-be-the-year-of-conversational-commerce-1586e85e3991#.4m7qfnscw [Accessed: 23-Aug-2016]

[3] K. Kelly. (2014, October 27) The Three Breakthroughs That Have Finally Unleashed AI on the World. *Wired*.Available:https://www.wired.com/2014/10/future-of-artificial-intelligence/ [Accessed: 10-Sep-2016]

[4] IBM Watson. *What is Watson?* Available: http://www.ibm.com/watson/what-is-watson.html [Accessed: 18-Oct-2016]

[5] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and autonomous systems,* vol. 42, pp. 177-190, 2003.

[6] J. R. Searle, "Is the brain's mind a computer program," *Scientific American,* vol. 262, pp. 26-31, 1990.

[7] B. A. Shawar and E. Atwell, "Chatbots: are they really useful?," in *LDV Forum*, 2007, pp. 29-49.

[8] A. De Angeli and S. Brahnam, "I hate you! Disinhibition with virtual partners," *Interacting with computers,* vol. 20, pp. 302-310, 2008.

[9] S. Brahnam and M. Weaver, "Re/Framing Virtual Conversational Partners: A Feminist Critique and Tentative Move Towards a New Design Paradigm," in *International Conference of Design, User Experience, and Usability*, 2015, pp. 172-183.

[10] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM,* vol. 9, pp. 36-45, 1966.

[11] L. Fryer and R. Carpenter, "Bots as language learning tools," *Language Learning & Technology,* 2006.

[12] O. Knill, J. Carlsson, A. Chi, and M. Lezama, "An artificial intelligence experiment in college math education," *Preprint available at http://www. math. harvard. edu/~ knill/preprints/sofia. pdf,* 2004.

[13] W. Zadrozny, M. Budzikowska, J. Chai, N. Kambhatla, S. Levesque, and N. Nicolov, "Natural Language Dialogue for personalized interaction," *Communications of the ACM,* vol. 43, p. 117, 2000.

[14] W. Knight. (2016, August 9) AI's Language Problem. *MIT Technology Review.* Available: https://www.technologyreview.com/s/602094/ais-language-problem/?set=602129 [Accessed: 12-Sep-2016]

[15] B. Shneiderman, *Designing the user interface : strategies for effective human-computer interaction*, 3rd ed. ed. Reading, Mass: Addison-Wesley, 1998.

[16] C. Ware, *Information visualization: perception for design*: Elsevier, 2012.

[17] A. Waytz, J. Cacioppo, and N. Epley, "Who sees human? The stability and importance of individual differences in anthropomorphism," *Perspectives on Psychological Science,* vol. 5, pp. 219-232, 2010.

[18] T. Holtgraves, S. J. Ross, C. Weywadt, and T. Han, "Perceiving artificial social agents," *Computers in Human Behavior,* vol. 23, pp. 2163-2174, 2007.

[19] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of social issues,* vol. 56, pp. 81-103, 2000.

[20] Y. Kim and S. S. Sundar, "Anthropomorphism of computers: Is it mindful or mindless?," *Computers in Human Behavior,* vol. 28, pp. 241-250, 2012.

[21] J. Grolleman, B. van Dijk, A. Nijholt, and A. van Emst, "Break the habit! designing an e-therapy intervention using a virtual coach in aid of smoking cessation," in *International Conference on Persuasive Technology*, 2006, pp. 133-141.

[22] T. W. Bickmore, L. Caruso, K. Clough-Gorr, and T. Heeren, "'It's just like you talk to a friend'relational agents for older adults," *Interacting with Computers,* vol. 17, pp. 711-735, 2005.

[23] J. Cassell and T. Bickmore, "External manifestations of trustworthiness in the interface," *Communications of the ACM,* vol. 43, pp. 50-56, 2000.

[24] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *Science,* vol. 315, pp. 619-619, 2007.

[25] A. Serenko, "Are interface agents scapegoats? Attributions of responsibility in human–agent interaction," *Interacting with computers,* vol. 19, pp. 293-303, 2007.

[26] Y. Moon and C. Nass, "Are computers scapegoats? Attributions of responsibility in human–computer interaction," *International Journal of Human-Computer Studies,* vol. 49, pp. 79-94, 1998.

[27] P. G. Devine, "Stereotypes and prejudice: Their automatic and controlled components," *Journal of personality and social psychology,* vol. 56, p. 5, 1989.

[28] B. Rossen, K. Johnsen, A. Deladisma, S. Lind, and B. Lok, "Virtual humans elicit skin-tone bias consistent with real-world skin-tone biases," in *International Workshop on Intelligent Virtual Agents*, 2008, pp. 237-244.

[29] C. Nass, Y. Moon, and N. Green, "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices," *Journal of applied social psychology,* vol. 27, pp. 864-876, 1997.

[30] A. De Angeli, "Ethical implications of verbal disinhibition with conversational agents," *PsychNology Journal,* vol. 7, pp. 49-57, 2009.

[31] G. Neff and P. Nagy, "Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay," *International Journal of Communication,* vol. 10, p. 17, 2016.

[32] S. Brahnam, M. Karanikas, and M. Weaver, "(Un) dressing the interface: Exposing the foundational HCI metaphor "computer is woman"," *Interacting with Computers,* vol. 23, pp. 401-412, 2011.

[33] Poncho. Available: http://poncho.is/

[34] Google Allo. *Say Allo*. Available: https://allo.google.com/

[35] B. A. Shawar and E. Atwell, "Different measurements metrics to evaluate a chatbot system," in *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, 2007, pp. 89-96.

[36] M. C. Forelle, P. N. Howard, A. Monroy-Hernández, and S. Savage, "Political bots and the manipulation of public opinion in Venezuela," *Available at SSRN 2635800,* 2015.