

Can we trust our health in the hands of chatbots?

An exploratory study investigating the effect of anthropomorphic design of e-Health chatbots on patients UX.

Name: Viktoriya Nikolaeva Petrova
number: 1974726

Email: v.petrova@student.utwente.nl

Study: Communication Science

Supervisor: Dr. J. Karreman (Joyce)

Date: 26.06.2020

Total number of words:

Abstract

The future of healthcare requires patients to become more autonomous and take actions into their own hands. Chatbots have been around for some time now, and since 2014 developers have been trying to integrate them as part of hospital e-Health systems. Medical chatbots have the potential to benefit both patients and doctors, by reducing workload and improving the chances of giving accurate diagnosis. Most empirical researches investigate the effect of chatbot design characteristics on the user experience (UX) in e-commerce, and not in a medical context. The present study uses a 2x2 research design to investigate how anthropomorphic visual and language cues affect patients' trust, perceived intelligence, satisfaction and willingness to use. To explore the extent to which these characteristics influence users' perceptions, each of the participants were presented with one of four videos that was designed to have either human avatar, or logo, and human language, or robot language. The study was conducted online and collected in total 120 responses. Due to the nature of the study there were no strict limitations regarding the target group, the age varied between 18 – 25 years old, and mostly included Bulgarian, Dutch and German participants. Based on the literature analysis conducted prior to the main study, it was expected that the chatbot, which integrated anthropomorphic characteristics would result in better overall UX. The results justify this assumption when it comes to implementing language cues. Furthermore it is confirmed that there is a positive effect of human avatar on the patient trust. Nonetheless, this study can serve a number of practical and theoretical implications in the field of medical chatbot design. It provides arguments and demonstrates how anthropomorphic design improves UX, and enables technological acceptance and adoption.

Keywords: Anthropomorphic design, chatbots, conversational agents, medical chatbot, social presence, social cues, user experience

Content table:

1. Introduction:	4
2. Theoretical framework	7
2.1 Chatbots and their development	7
2.2 Visual and Language characteristics of a chatbot:	8
2.3 Trust.....	11
2.4 Perceived intelligence.....	12
2.5 User satisfaction	13
2.6 Willingness to use.....	13
2.7 Hypotheses.....	14
2.8 Research model.....	15
3. Methods.....	16
3.1 Design:.....	16
3.2 Participants:	16
3.3 Materials:	17
3.4 Pre-test:	19
3.5 Procedure:	20
3.6 Measurements	20
4. Results:	26
4.1 Manipulation check	26
4.1.1 Manipulation for Visual cues	26
4.1.2 Manipulation for Visual cues	26
4.2 Multivariate Analysis of Variance:.....	27
4.2.1 Manipulation test for the independent variables	27
4.2.2 Trust	28
4.2.3 Satisfaction with the robot:	29
4.2.4 Willingness to use	29
5. Hypothesis overview:	31
6. Discussion:	32
6.1 Discussion of results:	32
6.1.1 Main effect of Visual cues: (exploring H1)	32
6.1.2 Main effect of language cues (exploring H2):	33

6.1.3	Interaction effect (exploring H3):.....	33
6.2	Limitations:.....	34
6.3	Implications	34
6.3.1	Theoretical implications:.....	34
6.3.2	Practical implications:	35
6.4	Conclusion:.....	36
7.	Acknowledgements:.....	37
	References:	38
	Appendix 1: Survey.....	45
	Appendix 2: Video conditions.....	48
	Appendix 3:Search log	50

1. Introduction:

The business and economy has faced rapid growth since automatization, as Artificial Intelligence (AI) has transformed the nature of work processes. According to an article from McKinsey (2018) these new opportunities will contribute not only to businesses, but also to major societal discourses concerning climate change and health. AI technology has the capacity to imitate human behaviour, it is able to seek patterns, learn and improve without being taught how to do so. Simply to say, nowadays AI has become the standard of every emerging technology product. Targeting smart work rather than hard work is the driving force behind corporate decisions. An example of this is the automatization of the online services, where it is estimated that by the end of 2020 approximately 80% of the business will rely on chatbots for their customer support.

Chatbot systems are not as ground breaking, as one may suspect. The first conversational agent enabling human-computer interaction dates back to 1966, and goes by the name Eliza. Nowadays, with the development of Natural Language Processing (NLP) these conversational agents have transformed into accessible and domesticated systems, such as Apples' Siri (launched in 2011) and Facebook messenger (launched in 2016).

As of now, the most common role chatbots take is that of virtual assistants in the customer service sector. Their ability to analyse customers input and provide the most accurate answers in the course of a dialogue has proved to be more efficient, compared to using real workers (Reshmi & Balakrishnan, 2018). It is highly profitable for companies to invest in the development of a conversational agent, as it cuts costs on human resources (Radziwill & Benton, 2017). Moreover, chatbots are able to develop relationships with multiple users at the same time, and their availability is not restricted by time zones, or working hours (Trivedi, 2019). According to Shum, He and Li (2018) the appeal of modern chatbots lies in their ability to establish an emotional connection with their users. To be able to achieve this, there should be cohesion between the social cues transmitted through the two main communication channels – visuals and textual.

The fundamental principle of a modern conversational agent, such as chatbot is to appear as human, as possible in their verbal and visual interaction (Cahn, 2017). Building a good conversational agent requires many technical, design and linguistic skills. In the process of technological acceptance of chatbots, applying the right combination of visual and textual information is expected to stimulate users' motivations and interests. This is crucial for achieving emotional appeal (Brandtzaeg & Følstad, 2017). Achieving this will result in a stronger trust bond, and longer commitment to the machine. As chatbots simulate human-to-human interaction, research has found that anthropomorphic features are essential for achieving this user experience (Qiu & Benbasat, 2009). Go and Sundar (2019) outline two easy ways to measure humanness of online conversational agents, this is through visual and language cues. For the purpose of this research visual cues would be limited to the chatbots' appearance (human avatar and logo). Moreover, language cues refer to the conversational style used by the chatbot (either human or robotic). The level of social presence conveyed by a conversational agent is found to be an important factor in building trust, satisfaction and perceived usefulness of the online agent (Etemad-Sajadi, 2016; Gefen & Straub, 2004).

E-Health is one of the industries that has benefited the most from the automation of human labour. Electronic health records, personal wearables and most recently portable communication systems such as chatbots have simplified the job of specialists (Bibault, Chaix, Nectoux, Pienkowski, Guillemasé, & Brouard, 2018). According to Pereira and Díaz (2019), this is a step towards improving healthcare and overall quality of life, by allowing patients to have more control, and be more autonomous in the process of taking care of their wellbeing (Bates, 2019). Some of the most common reasons why people do not consult with doctors when they have symptoms is the lack of personal time to do so, their inability to afford medical care, or the distance between their whereabouts and the hospital.

Since 2018 there have been many cases of experimentations with chatbots in health care. Denecke, Tschanz, Dorner and May (2019) consider that in the near future these chatbots will become a first contact point for primary care to those who doubt their mental or physical health. The idea is that medical chatbots will be able to use peoples input to keep track of their symptoms, providing recommendations and consultations. Moreover, they will take over administrative work, such as booking appointments, delivering and reading results. Thanks to NLP and AI they have the ability to personalize medical follow-ups and provide preliminary diagnosis. Chatbots do not display biased behaviour, or prejudice towards patients with certain demographics, or ethnicities (Palanica, Flaschner, Thommandram, Li & Fossat, 2019).

Another issue that medical bots aim to resolve is to substitute all the unreliable internet sources that patients might turn to check their symptoms. Self-diagnosing has proved to be harmful not only on a psychological level, but it may lead to people undergoing procedures that might be damaging to their physical health. In November 2018 Pfizer published statistics that explained how only 38% of patients find the data provided by the e-Health application to be credible. And 40% believed that it is as secure as the one on online forums.

Despite their recent popularity, there is scarce empirical research on the topic of visual design of medical chatbots. Most of the e-Health conversational agents incorporate minimalist design features, by simply using the logo of the organization/hospital (some examples are One Remission, Youper and Babylon Health). However, a few differ by integrating more realistic cues, for example, by using human avatars (such as GYANT). And, as of recently integrating an embodied avatar that moves and speaks when interacting with its users (Sensely).

The conversational style of the chatbot also plays a central role in establishing an emotional and intellectual relationship with the users (patients). Taking into account the delicacy of this interaction, there are certain emotions that are essential for a pleasant doctor-patient interaction. Chatbots should be able to adapt their answers throughout the interaction process, and based on the emotional and factual input. This is needed, in order to make the user more comfortable and have trust (Gennaro, Krumhuber & Lucas, 2020). Thus, it is important that chatbots are able to adapt their vocabulary and intonation in accordance to the situations (Müller, Mattke, Maier & Weitzel, 2020).

The aim of this research is to understand how different e-Health chatbot designs affect the users trust, satisfaction and willingness to adopt and use them. In the upcoming analysis the following research questions will be addressed:

‘To what extent do the visual and conversational style of the virtual assistant affect the users’ satisfaction, willingness to use and trust?’

To answer the research questions this paper is divided into multiple sections. As a starting point, an extensive literature review is performed. This is done so that the reader has a better understanding of the relationship between the chatbot characteristics (visual, identity and linguistic cues, which are the independent variables), and the UX (trust, perceived intelligence, user satisfaction and willingness to use - the dependent variables). Following this research a number of hypotheses are defined. These assumptions are measured using a 2x2 design research methodology. The results of this research are analysed and discussed in the final section of this report. Moreover limitations and practical implications for future research were discussed.

2. Theoretical framework

There is scarce scientific research exploring the effects of anthropomorphic design of chatbots designed to provide medical assistance. Hence, this framework will use relevant scientific studies from various fields (e-commerce and banking) where chatbots have been successfully implemented as part of the online customer services. More specifically, this section of the study will focus on the importance of human-like design of visual and language cues, which are the independent variables. And their effect on users' level of trust, perceived intelligence, user satisfaction and willingness to use an e-Health chatbot, which are the dependent variables.

2.1 Chatbots and their development

Modern day chatbots are reckoned as a leading tool that eases the interaction between organizations and their online visitors and clients. They allow these organizations to address the needs and issues of the users in a flexible and cost-efficient way (Trivedi, 2019). However, to those who are not acquainted with the history of chatbots, it may come as a surprise that they have existed for more than sixty years. The first conceptualization of a conversational agent dates back to the 1950s', when the researcher Alan Turing came up with the definition: 'online human-computer dialog system[s] with natural language' (as cited in Shum, He & Li, 2018). Despite the immense technological development following the 50s', modern scholars and developers have not deviated, but rather build upon this conceptualization. Hills, Ford and Farreras (2015) specify it as a software system that simulates human-human textual interaction, supported by AI, and expanded by the abilities of NLP. This interaction is mediated by the means of messaging applications and websites (Xu, Liu, Guo, Sinha & Akkiraju, 2017). The interaction process requires users textual input that the chatbot is able to understand and analyse, in order to answer in an appropriate manner.

The development of technology and user interfaces brought new opportunities to the human-computer interaction process. Sansonnet, Leray and Martin (2006) outlined an early framework with three basic functions expected to be fulfilled, in order to build a good conversation agent. First, the chatbot should be able to comprehend the input and generate appropriate responses. This means understanding the general definition of a word, while at the same time taking into account the fluidity of its cultural meanings in an everyday conversational language (Hill, Ford & Farreras, 2015). Second, the system should have access to external data and various sources that would help it acquire new knowledge independently, without the need of programming. According to Hussain and Athula (2018) this chatbot is called an open source chatbot, because it is able to independently grow its own knowledge base. And finally, chatbots should have a 'persona' in order to give an impression of a human agent. Both the linguistic style of interaction and the explicit visual cues should be coherent in order to portray a believable online persona (Qian, Huang, Zhao, Xu & Zhu, 2018).

With time this initial framework had to be expanded and new criteria were added to it. The reason for this revolves around the advancement of users' technological knowledge,

leading to higher degree of mistrust held over intelligent technology (Madhavan & Wiegmann, 2007). Most people feel uncomfortable interacting with AI that is as smart, if not more, than they are. On one hand, this could be explained using the uncanny valley theory. According to it, once technology becomes too human-like, it may lead to negative consequences, such as users feeling unease and discomfort (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2019). On the other hand, it may be explained as fear of substitution, as chatbots are already taking over jobs in the online customer service (Giard & Guitton, 2010).

To resolve this researchers have explored the effect of implicit design cues on the human-machine interaction experience. The goal is to find the right combination of cues that would optimize users' experience (UX) and emotional appeal of the chatbot (Go & Sundar, 2019). Thus, engaging users by taking into account both their intellectual and emotional quotient (Shum, He & Li, 2018). Information privacy is another criteria, which was recently added to the equation. Chatbots are expected to provide 'notice' information of how the system will utilize the data from the input. Hence enabling the users to handle this interaction consciously (Følstad, & Brandtzæg, 2017).

The success of automation in many sectors, including e-Health and medical assistance, requires better online customer service and communication. Having an excellent conversational agent, such as chatbot, is beneficial for both sides of the transaction process (Marcondes, Almeida & Novais, 2018). Chatbots are able to overcome boundaries that are physically, or psychologically impossible for a human worker. Such are the matter of availability, consistency and prejudice towards different races and ethnicities. Luo, Tong, Fang, Z and Qu (2019) found out that chatbots are able to consistently establish stronger positive emotional connection and higher trust rate, compared to a human worker. Moreover, chatbots not only work more efficiently, but are also cost-saving for the organization (Trivedi, 2019). Lastly, chatbots are able to communicate and serve multiple users at the same time. And their availability is not restricted by time zones and work hours (Shum, He & Le, 2018).

2.2 Visual and Language characteristics of a chatbot:

People possess two distinctive processing modes for information processing that support their process of making informed decisions. According to Korthagen (1993), one mode relies on the implicit signs that lay in the visual design of a technology. Whereas the second one focuses on the implicit signs communicated through the language used by the conversational system. Designing a good chatbot requires a combination of visual and language cues that are coherent with each other. This combination should serve the purpose and add value to the system, in order to enhance the UX (Cahn, 2017; Zhou, Gao, Li & Shum, 2020).

2.2.1 *Visual cues:*

Implementing the right visual cues when designing an interaction agent is crucial for the implicit behavioural impact of the system. These cues create the initial impression and expectations, therefore setting the tone of the dialogue that follows (Blascovich, 2001). It's not just about picking favourite colours and interesting images, but rather making an

embodied representation of the systems' purpose (Agrawala, Li & Berthouzoz, 2011). Moreover, in the context of e-services, they are perceived as more impactful than the text that follows. Thus influencing the willingness to use and develop customer loyalty (Appel, Pütten, Krämer, & Gratch, 2012; Brandtzaeg & Følstad, 2017). The visual interaction that can be conveyed through the conversational agents depends on their type, as well as the interfaces that enable the interaction process. Araujo (2018) refers to chatbots as examples of disembodied conversational agents that have limited capacity of visual cues that could be used for implicit communication: avatars and emoticons.

Chatbots are taking over jobs that were previously associated with human-human contact. In order to recreate the intimacy of a natural interaction designers aim to make the chatbot appear as human-like as possible (Kim & Sundar, 2012). This attribution of human characteristics to a non-human artefact is called anthropomorphic design (Kim & Sundar, 2012). Previous research on chatbot design has shown that anthropomorphic design tends to evoke the feeling of social presence (Hassanein & Head, 2007), as well as positively affect users satisfaction with the system (Brandtzaeg & Følstad, 2017; Radziwill & Benton, 2017). This is expected to prompt people to exhibit social behaviour towards the technology (Toader, Boca, Toader, Măcelaru, Toader, Ighian & Rădulescu., 2019). When talking about anthropomorphic visual architecture of a chatbot designers use contextual cues such as giving avatars a face and a name of a real person (Araujo, 2018).

Disembodied agents such as chatbots have limited visual cues to convey social presence among users. And the choice of profile image plays an important role in forming good UX (Gefen & Straub, 2004). Today most organizations and brands rely on using their own logos when designing their chatbots, despite the fact that people tend to have more meaningful responses when they see faces (Marino, 2014). Anthropomorphic visuals such as pictures of real people then to exhibit higher social cues compared to a logo. Ciechanowski, et al. (2019) performed study on artificial gaze and found out that avatar gaze has a significant effect on the speed, thoughtfulness and precision of human actions. Moreover, Gustavsson (2005), and McDonnell and Baxter (2019) suggest that gender biases should be taken into account when designing appropriate avatars, such as users preference towards female avatars.

This research paper will investigate the effect of anthropomorphic visual design of an e-health chatbot. More specifically, weather literature findings that have been mostly related to the e-commerce sector, can also be applicable for medical services.

2.2.2 *Language cues*

Conversational interfaces have transformed the human-machine interaction by giving users the privilege of communicating with the technology on human terms. The field of user experience (UX) has been trusted with the responsibility of making this interaction easier and understandable for the masses. The need to make technology accessible to the average person meant that the interaction process to be redesigned (Dale, 2016). Unlike IT professionals who understand how to enter syntax-specific commands and codes, the average person could only recognize, or use their native written and spoken language. This is why the interface design had to become more social by allowing users to contact, navigate independently and retrieve

data through the interaction.

One of the biggest achievements in the field of human-computer interaction is the development of natural language processing (NLP). This system allowed average consumers to have more tangible interaction with complex systems such as chatbots (Araujo, 2018). Duran, Hall, McCarthy and Mcnamara (2010) defines NLP as the ability of AI driven systems to understand everyday conversational language of a user, and answer them in an appropriate manner. Chatbots with NLP ability are able to constantly study and analyse how humans communicate and replicate it in a manner that makes them appear human-like (Hirschberg, & Manning, 2015). This simulation of human behaviour allows them to convey social presence by using language cues. Go and Sundar (2019) determine this as a key requirement for achieving favourable user behaviour and trust. Go and Sundar (2019) determine this as a key requirement for achieving favourable attitude and user behaviour. There are few empirical studies that explore the psychological outcomes of social presence on user experience. Taylor (2011) points out that the stronger this presence is, the greater users feel emotionally connected to the conversation agent. This results in positive opinion towards the chatbot and the organization it represents (Araujo, 2018), and favourable intentions and user loyalty (Go & Sundar, 2019).

In order for the chatbot to sound as humane and natural as possible the user face must not be restricted to a fixed set of commands and phases (Mctear, 2017). It should imply flexibility by being able to express a message in a variety of ways, meaning that chatbots should be able to use syntaxes variation, abbreviations and in some cases, slang (Pilato, Augello & Gaglio, 2011). In the case of medical chatbots, it is important that they are able to deliver complex information regarding diagnosis and treatments in a way that is understandable for the patients (Gennaro, Krumhuber & Lucas, 2020). Johnson, Patron and Lane (2007) state that conversations between users and chatbots become less reliable when the structure of the language feels unfamiliar. Additionally, the nature of the interaction should be mixed-initiative, both users and chatbots are responsible for contributing to the quality of the dialogue (Hill, Ford & Farreras, 2015).

In order to enhance the social presence and the ‘feeling of another’ during the chatbots-user interaction, the designers can focus on the frequency of the responses. According to Sundar, Go, Kim and Zhang (2015), higher messaging response and interactivity is correlated to the perceived humanness of the chatbot. High frequency interaction is expected to be associated with contingency, which is a typical characteristic of the interpersonal communication. Moreover, Liews, Tan and Ismail (2017) established a connection between social presence of a chatbot and the media richness theory. This means that in order to convey social presence a chatbot should demonstrate vividness and interactivity through their conversational style.

Language cues and visual cues are closely related to one another, in order for a chatbot to achieve positive results there should be cohesion between the two. Empirical research in the field of e-commerce has shown that the most optimal chatbot design combines characteristics that convey high social presence. A study by Keyzer, Dens and Plsmacker (2017) supports this idea, finding interaction effect between voice of an assistant and its appearance. And that the different combination would lead to either higher or lower UX. Despite the lack of scientific research on the topic, there is a possibility that the language used

by a chatbot may have an interaction effect with the social presence of the visual cues.

These findings explain the importance and effect of language design and will be further investigated in the context of medical chatbots. Chatbot language that integrates social cues, and is perceived to be more-human like is expected to result in more positive user experience. Similarly, the combination of anthropomorphic visual and language cues is expected to result in better UX.

2.3 Trust

Trust is an essential condition that needs to be met in order to have successful interpersonal, or online interaction. It is built upon communication and cooperation, and has been recognized by psychology and communication theories as a key attribute in the process of governing transactions (Arrow, 1974). More specifically, it influences the decision-making processes in the human-to-human, or human-machine interactions. The conceptualization of trust varies and depends on the field of research. According to Barber (1983; as cited in Madhavan & Wiegmann, 2007), it is the confidence in another, which is based upon the odds that they will behave favourably and cooperate. It implies that both of the involved parties are willing to be vulnerable, experience betrayal and extend goodwill (Friedman, Khan & Howe, 2000). In the context of the online environment however, trust does not emerge from physical, or emotional interaction, but rather relies on the UX (Oliveira, Alinho, Rita & Dhillon, 2017).

There are three factors that influence the information acceptance: predictability, dependency and faith in the information source (Rempel, Holmes & Zanna, 1985). The consistency and stability of the individuals' actions throughout a certain period of time influences the predictability. The amount of confidence one has in the information carrier is related to the dependency. And finally faith, which is based on the belief we have in regards to future actions of the information source. If all three are met it is expected that users would have full trust in this information source.

Trust plays a central role in the process of technological diffusion and adoption. Luhman (1979; as cited in Elofson, 2001) argues that trust usually begins where knowledge ends, and it serves a way to bridge the knowledge gaps people have. AI technologies have gone beyond what one could comprehend without technical sufficient knowledge, or expertise. Andras, Esterle, Guckert, Han, Lewis, Milanovic, and Wells (2018) conclude that as machines are getting smarter and more independent, people tend to delegate more of their daily responsibilities to them. Moreover, people have the disposition to apply the norms typical for the human-to-human interaction in their contact with intelligent machines (Madhavan & Wiegmann, 2007). This is why anthropomorphic design is essential when it comes to users' trust.

Research shows that most people have the tendency to over interpret their relationship with technology by adding moral or social depth, even if this action is not justified. Andras, et al. (2018) highlight that these users might even overestimate the level of humanness a technology possesses. Assuming that a machine may have its own mind, and will act in a way that furthers its own achievements. In the context of chatbots, trust is closely related to the

willingness of users to provide access to their personal information, accept and follow the personalized recommendations (Nordheim, Følstad & Bjørkli, 2019). This requires the system to have certain design characteristics that would communicate transparency and credibility, using both verbal and textual communication. Moreover, there should be cohesion and consistency between the selected design elements.

The successful implementation of a chatbot is determined by the users' perception and experience with it. Studies show that the users' response towards chatbot recommendation falls into two extremes. Either they do not trust them, or they expect the answers they receive from the machine to be better than those of specialists (Mugria, et al., 2016). Koh and Sunda (2010) found out that consumers have the tendency to assign full responsibility for the message's credibility to the messenger sources they trust. Chatbots have limited interaction capacity, this is why designers use visual and language cues to convey trust. As already mentioned, the presence of different social cues is associated with trustworthiness of the conversational system (Mcknight, Carter, Thatcher & Clay, 2011). Chatbot that integrates human-like language is more likely to elicit trust, in comparison to one that sounds robotic and generic (Zhou, Mark, Li, & Yang, 2019). According to Bartneck, Kulić, Croft and Zoghbi (2008), users tend to mistrust conversational agents that demonstrate patterns in their behaviour. Furthermore, Nordheim, Følstad and Bjørkli (2018) state that visual cues that convey social presence are implicitly pursuing the users' willingness to trust the chatbot.

2.4 Perceived intelligence

There is little empirical evidence investigating how design of chatbots relates to the perceived intelligence of the agent. According to Bartneck, et al. (2008) for many years virtual agents have struggled with imitating human behaviour, thus they held not value to the average user. However, most chatbots now are equipped with AI and NLP, which somehow turned the tables around. It is important to take into account that there is a thin line between being perceived as human, and going down the uncanny valley, by becoming too human and even smarter than them (Ciechanowski, et al., 2019). The language used by the chatbot plays an important role when forming the impression of intelligence and competence. For example, healthcare diagnosis and consultations include a lot of complex terminology and extensive explanations. This is why it is important that a chatbot is able to provide clear, non-ambiguous information to the patient. Thanks to NLP many modern chatbots are completely autonomous, and they have the ability to exhibit intelligence through their conversation style. They are able to translate this complex information to a more understandable, everyday language. Depending on their role some are able to not only appear knowledgeable but also emotionally intelligent to handle complex situations (such is the case of chatbots dealing with mental health).

2.5 User satisfaction

User satisfaction is the key for long-term success of any organization or product. It is a signal for the successful technological adoption, and it highly depends on the customers' perception of the product or service (Hult, Morgeson & Morgan, 2017; Mahmood, Burn, Gemoets & Jacquez, 2000). Similarly to the variable trust, there are many definitions found in the literature that explain what satisfaction is. Rust and Oliver (2000) define satisfaction as the customers' beliefs that the experience from using the product can create positive feelings and associations. Similarly, Wixom and Todd (2005) define it as a positive feeling, derived from the fulfilment of one's wishes, needs and assumptions that emerge from the interaction process. If the interaction process leads to positive feelings and associations it is expected to grow into customer loyalty (Anderson & Sullivan, 1993; Lee & Choi, 2017). In the case of online services, consumers' satisfaction can be indicated by their willingness to continue exploring the functions of the interface.

Chatbots have the immense ability to improve customer service for both the users and the provider. Toma (2010), highlights the importance of creating chatbots with coherent design characteristics. Elements, such as visuals and conversational cues, affect the perception of satisfaction, users' emotions and cognitive evaluation (Handro, 2018). On one hand, the conversational characteristics determine the quality and value of the interaction. McTear, Callejas and Griol (2016) outline three characteristics, which predict communication efficiency – speed, dialogue and smoothness. Similarly, Morrissey and Kikrakowski (2013) had found out that cues, such as friendly greeting and personality are perceived favourably by users. Therefore anthropomorphic linguistic design where the chatbot displays personality and moods is expected to increase user satisfaction. On the other hand, there is the visual design, and its relation to satisfaction. Most of the research on this topic focuses on chatbots in e-commerce and banking. In an empirical research Holzwarth, Janiszewski and Neumann (2006) found out that users were in general more satisfied when they were presented with a human looking avatar, than by just viewing textual information.

2.6 Willingness to use

Based on the social presence theory anthropomorphic cues are also associated with consumer brand, or product loyalty. Chatbots that use human images and appropriate language are associated with higher usability and hold greater value for the average consumers (Liew, Tan & Ismail, 2017). Moreover, the social presence of chatbots is associated with its capacity to transmit the feeling of 'warmth' and 'personality' (Hassanein & Head, 2007). Thus, enabling users to build emotional connection with the disembodied agent (Go & Sundar, 2019). Araujo (2018) links these social cues conveyed throughout the interaction process, with users' appreciation and willingness to build meaningful relationships and dependence, thus leading to consumer loyalty.

In the field of e-commerce research, chatbots that evoke the feeling of 'personality' are

assumed to also have cognition, and competency (Qian, et al., 2018; Wang & Siau, 2018). Therefore their responses are perceived to be more intelligent, satisfying and trustworthy (Cyr, Head, & Ivanov, 2009; Liew, et al. 2017). De Visser, Monfort, McKendrick, Smith, McKnight, Krueger and Parasuraman (2016) conclude that chatbots that are able to expand the limits of robotic conversational capacity, and create the impression of human-to-human dialogue result in higher positive response rate and longer engagement from the users. If users are able to treat the technology as a real person their trust and willingness to use it will be prolonged (Reeves & Nass, 1996).

2.7 Hypotheses

Following the literature analysis the following hypothesis are established:

H1: Users' perception of (a) trust, (b) perceived intelligence, (c) user satisfaction and (d) willingness to use is higher when interacting with e-Health chatbot that uses human avatar, compared to the one that uses logo.

H2: Users' perception of (a) trust, (b) perceived intelligence, (c) user satisfaction and (d) willingness to use is higher when interacting with e-Health chatbot that uses human language, compared to the one that uses robot language.

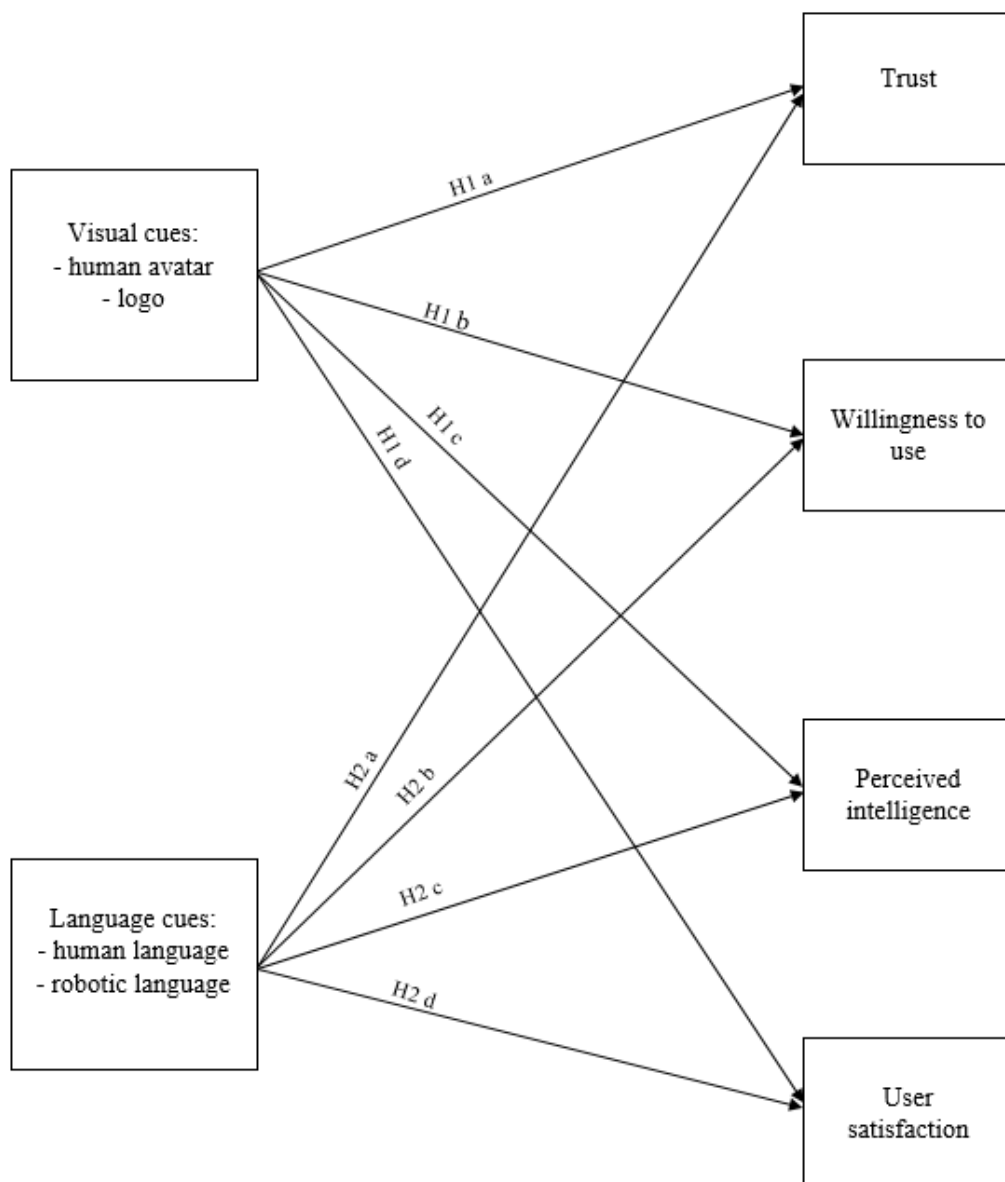
H3: Users' perception of (a) trust, (b) perceived intelligence, (c) user satisfaction and (d) willingness to use is higher when interacting with e-Health chatbot that uses human avatar and human language, compared to the one that uses logo and robot language.

2.8 Research model

To visualize the relationship between the variables the following model was created. It includes the independent variables (visual cues and language cues) and the dependent variables (trust, perceived intelligence, user satisfaction and willingness to use). The arrows represent the direction of the interactions that will be observed later on.

Figure 1.

2x2 Research model



3. Methods

3.1 Design:

This quantitative research consists of a 2x2 chatbot design study, which investigates the effect of language and visual design on users' perception of a medical chatbot. The four chatbot conditions vary in terms of either their language (human-like and robotic-like), or visual cues (human avatar and logo). The four conditions are presented in Table 1 (refer to Appendix 2 for more detailed representation).

Table 1.
2x2 chatbot conditions

		Visual cues:	
		Human language	Robot language
Language cues:	Human avatar	Video 1	Video 2
	Logo	Video 3	Video 4

3.2 Participants:

In total 130 participants completed the survey, out of which 120 were used for the data analysis. 10 respondents were removed from the data set due to a number of issues. Firstly, two of these participants did not complete the whole survey, leaving the last two blocks with little to no answers. Another five responses were removed after seeing that they completed the whole study for less than two minutes. Given the fact that each video lasted between 1.20 minutes and 1.38, it can be assumed that these participants either did not view the video materials in detail, or did not pay attention to what the survey questions. And lastly, three responses were removed to even out the number in all categories.

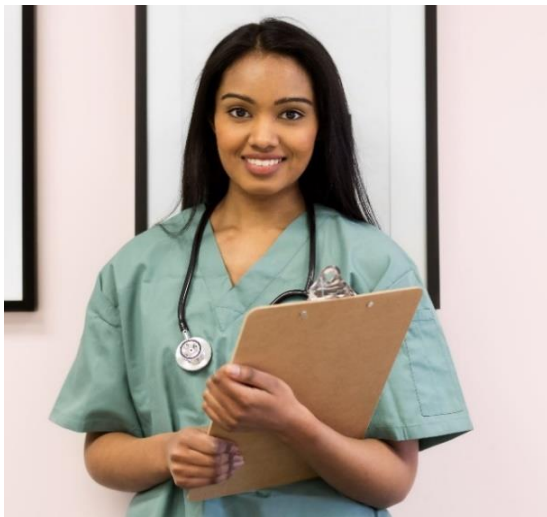
The participants' nationalities were mostly from Germany (15), the Netherlands (17) and Bulgaria (30). The age group of this study ranged between 18 and 52 ($M = 25$, $SD = 6.08$). Descriptive statistics also shows that 5 of the participants did not enter their age, thus are reported as missing value. Further, predominant part of the participants were female (81), compared to the male (36), and 3 participants were reported as missing values. Lastly, most of the respondents hold either bachelor degree (44), or high school diploma (39), followed by masters' degree (27), associate degree (11), Doctorate (2), and lastly below high school (1). Each video condition was viewed by an even number of participants (30 participants per video condition). Looking at the age mean per group there are no visible mean age differences between the groups, however due to the missing values only the group that viewed condition

4 had all 30 participants filling in the age. The groups that viewed video 1 and 2 have two missing values each, and for video 3 only one was recorded as missing.

3.3 Materials:

Below are the two visual designs used for this study can be seen. Picture 1 shows the human avatar design that used a picture of a female nurse. And Picture 2 depicts the minimalistic design that only consists of the hospitals' logo and name. Adam and Galinsky (2012) discuss the effect of enlothed cognition, and how clothes implicitly affect the credibility of the information shared by those who are dressed a certain way. Moreover, a female avatar was chosen based on a study from McDonnell and Baxter (2019). This study supports the assumption that users have a favourable attitude towards female avatars.

Picture 1. *Human avatar*

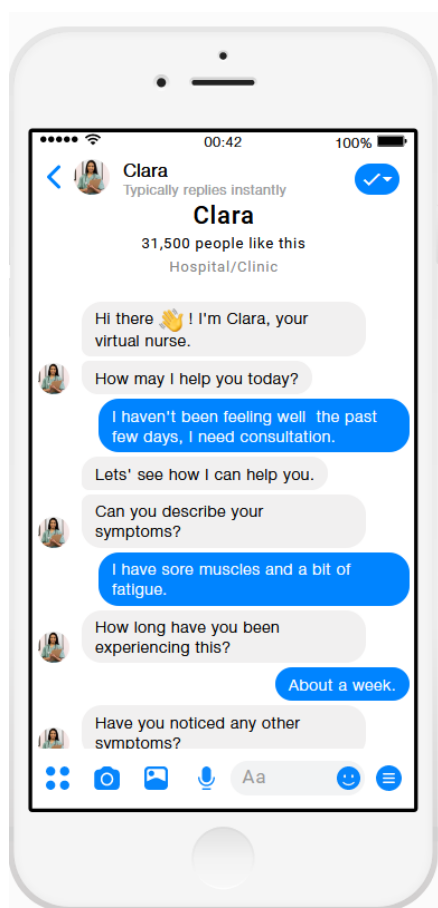


Picture 2: *Logo*

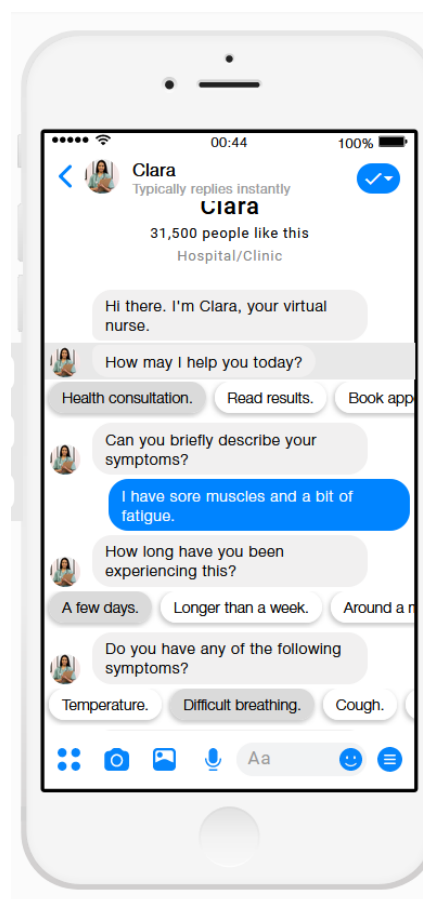


The second dependent variable is called Language cues. As mentioned, participants see only one of two video conditions. The first one shows the chatbot giving ‘human language’ responses to answer the questions (Picture 3). In this condition there were no limits in terms of what kind of input the patient could provide to the chatbot, and the chatbots’ responses do not limit the input that the users can provide. According to (Sundar, et al., 2015) this makes the interaction feel more human-like and vivid. The second language condition showed robotic language (Picture 4). This interaction limits the users by allowing them to only give their input by select button options (Kvale, Sell, Hodnebrog, & Følstad, 2020).

Picture 3. *Human language*



Picture 4. *Robot language*



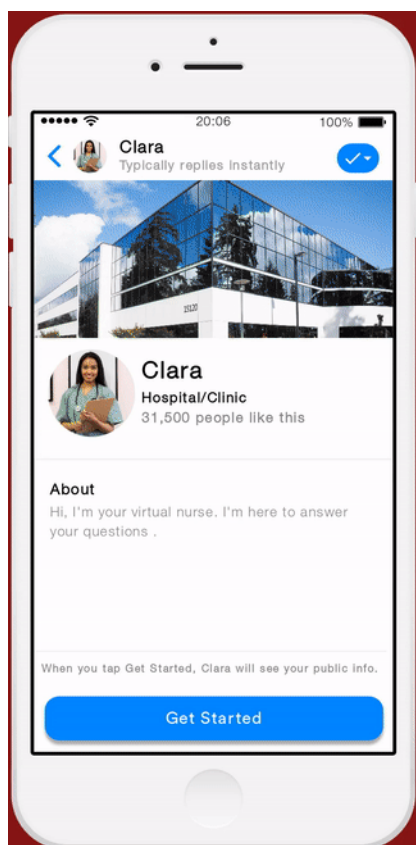
3.4 Pre-test:

Before starting with the main data collection a pre-test was performed. The purpose of this part of the research was to eliminate any possible errors, uncertainties and misunderstanding in the formulation of the survey. Moreover the pre-test was used as an opportunity to choose between two interface designs, this was an opportunity to find which one was more believable for the participants. (Pictures 5 and 6 show the two interface designs that were used for the pre-test).

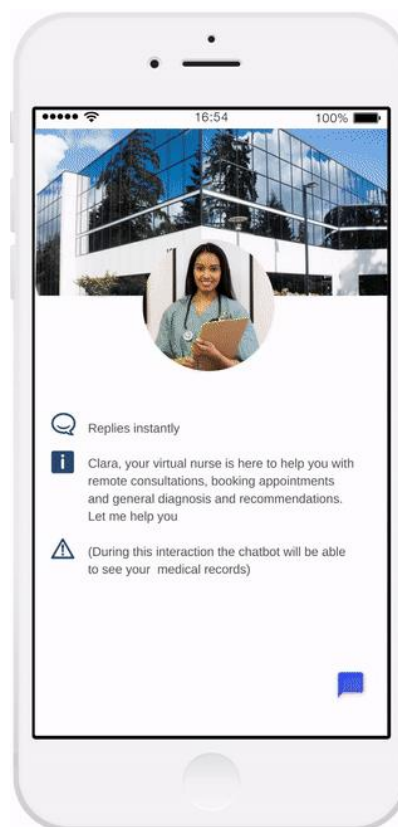
For this part of the research 12 participants were recruited, each of which was exposed to one of four conditions from the 2x2 designs. The aim was to see if they would be able to recognize and react differently to the stimuli that were prepared. For this procedure the participants were asked to pay close attention to not only to videos, but also the items in each construct, as well as the scale that was used. Finally, the first 8 participants were approached for a quick feedback session, where they discussed the different stages of the pre-test research. Following the discussion, they were presented with an alternative interface design (picture 6) and they were asked to express their preference.

As a result of this pre-test the first interface design (Picture 5) was used as final stimuli, the reason for this choice was higher readability of the text and better layout of the home screen. Thus the second option was eliminated and was not used for testing. Furthermore, the speed of the video was adjusted to be slower on the opening page, which according to the participants increased the readability. The final result of the pre-testing required the rephrasing of a question, to make them more explanatory and less ambiguous.

Picture 5. Survey design



Picture 6. Feedback design



3.5 Procedure:

Before starting with the data collection procedure a permission was requested from the University of Twente Ethical commission. After obtaining permission the pre-test and main study were conducted. Both were designed only in English therefore the participants that were recruited were of various nationalities. The four chatbot designs and videos were created using an online tool for chatbot mock-ups called BotSociety. After this they were exported to Qualtrics, which was the tool used to create both the pre-test and main research.

The participants were recruited through various social media platforms: Instagram, announcements, personal approach through WhatsApp and Messenger and Facebook groups created to help students find participants. Besides social media the SONA test subject system of the University of Twente was used to promote the study.

Before the start of the survey participants had to read a short introductory briefing. This is a standard procedure to ensure they have given their consent to the general terms and conditions of the study, that are in line with the data protection regulations outlined by the GDPR. In this section they were also introduced to the purpose of the study, and were provided with the contact e-mail of the researcher in case they had further questions, or wanted their data to be removed.

Following this online consent form participants were asked questions regarding their demographics: gender, age, nationality and education. Following this, they were asked to read carefully through a fictional scenario that would give background information regarding the next sanction. After reading the scenario, the participants had to view one of four short videos depicting ‘them’ interacting with the e-Health chatbot (Table 1 summarizes the four conditions). And finally they had to answer to fill in a short survey measuring the trust, perceived intelligence, user satisfaction and willingness to use.

3.6 Measurements

The survey was constructed using the program Qualtrics. A 7-point Likert scale was used to measure each independent variable and their effect on the dependents. The possible answers varied from ‘Strongly disagree’, which was coded as 1 (the lowest value), to ‘Strongly agree’, coded as 7 (the highest value). To construct this survey a few already existing measurement instruments were combined and modified. In total 28 items divided into 6 blocks of questions were used to measure the constructs. The choice of scales is justified as it follows:

Visual design:

The first block of items measured the level of anthropomorphism of the chatbot appearance. This block was constructed in order to see whether or not there was a significant difference between the human avatar and the logo design. To construct this block a pre-existing scale measuring the humans’ perception of the chatbot was modified and used. This scale is also known as the Godspeed questionnaire constructed by Bartneck, et al. (2008). The items in this

block induced the following definitions: ‘real person’, ‘life-like’, and ‘alive’. An example question from this block is: ‘The chatbots' picture made the interaction feel life-like.’

Language design:

The same scale from Bartneck, et al. (2009) was used to create the construct that measures the anthropomorphism of the language design. More specifically, to see if users can distinguish between human language and robot language condition. The items included wording, such as: ‘sensible’, ‘natural’, ‘human-like’ and ‘alive’. An example question of this block of items is: ‘The chatbots' responses felt human-like.’

Trust:

Anthropomorphism is closely related to the level of trust one has in a chatbot system. The human-likeness of the avatar and the more frequent and natural the language used by the bot, the more positively it is expected to be perceived (Go & Sundar, 2019). To see whether this is applicable to the case of medical chatbots items for this construct were developed using measurement instruments by Mcknight, et al. (2011) and Charalambous et al. (2015). The items taken from these studies were again modified to fit the flow of the survey. Four items were used to measure this construct, an example of which is: ‘The information received from the chatbot was credible:’

Perceived intelligence:

To test the second dependent variable a three question construct was formed using the Godspeed questionnaire from Bartneck, et al. (2008). This construct measures whether or not users see the chatbot as intelligent, and if there is a significant effect of the independent variables. The items included wording, such as: ‘intelligent’, ‘qualified’, ‘competent’. An example item of this block is: ‘This chatbot is qualified to provide medical assistance.’ (Kozak, et al., 2006).

User Satisfaction:

For the third dependent variable a general questionnaire related to human interface design was used. The original questionnaires were from Chun, Ko, Young and Kim (2018), and Cronin, Brady and Hult (2000). They were developed to measure users’ perception of e-commerce chatbots’ interaction and their intention to purchase. For this study they were modified to measure the user satisfaction following the medical assistance. An example statement from this block is: ‘I am satisfied with the assistance received from the chatbot.’

Willingness to use:

The last construct of this survey focuses on participants’ willingness to continue using e-Health chatbot after getting to know the system. This part consisted of four items that were previously tested in the research of Lee, Lee and Sah (2019) and Xu, Zhang and Li. (2011). An example statement from this block is: ‘I would recommend this chatbot to a friend’

3.6.1 *Validity*

An important part of the research process is measuring the validity of all constructs. In other words to check if all items measure the variables they were intended to measure in the first place.. To investigate further the construct validity of the manipulation a Factor Analysis was conducted using the program SPSS.

In total the 23 items were expected to separate between the 6 constructs that are measured. Table 2 shows the results of the Factor analysis. More specifically a Rotated Component Matrix was used to visualize the distribution of each item to the factor it measures. Based on this output a number of adjustments were made. The first factor showed that item 1 doesn't measure the construct "Visual cues", SPSS under the construct "Language cues". This item cannot be merged with the construct, therefore this item was not used for further analysis. Furthermore, one of the items originally measuring 'Trust' loaded into the construct 'Language cues', therefore this item was moved to the construct 'Language cues'. Lastly the two constructs 'Perceived intelligence' and 'User satisfaction' loaded into one factor. For further analysis these two factors were merged into a new construct 'Satisfaction with robot'. The only construct that was not adjusted after the Factor Analysis is 'Willingness to use'.

Moving further, the total explained variance of the model is 76.89% which scores rather high. This indicates that the model has statistically significant explanatory power. Furthermore, for each factor the eigenvalue was set to over and above 1. In theory every item with eigenvalue that is higher than 1 is perceived valid.

Table 2. *Validity - Factor Analysis*

	Component				
	1	2	3	4	5
Factor 1: Visual cues					
The chatbots' picture made the interaction feel as if it was with a real person:		.628			
The chatbots' picture made the interaction feel life-like:				.893	
The chatbots' picture made the interaction feel human-like:				.901	
The chatbots' picture made the interaction feel natural:				.852	
Factor 2: Language cues					
The impression of the chatbots' language felt alive.		.761			
The chatbots' responses felt human-like		.795			
The language that the chatbot used felt natural:		.737			
The chatbots' responses felt sensible:		.539			
Factor 3: Trust					
The information received from the chatbot was credible:					.800
The information I received from the chatbot is trustworthy;					.723
The chatbots' responses were reliable:					.746
The chatbot interaction felt believable:		.568			

Factor 4: Perceived intelligence

This chatbot is intelligent: .635

This chatbot is qualified to
provide medical assistance: .751

This chatbot is competent
enough to provide this
information: .740

Factor 5: User satisfaction

I am satisfied with the
responses of this chatbot: .785

I am satisfied with the way
the chatbot helped me: .751

I believe the chatbot provided
good responses: .690

I am satisfied with the
assistance received from the
chatbot: .594

Factor 6: Willingness to use

I would download this
application: .837

If I have symptoms I would
turn to this this chatbot: .832

I would recommend this
chatbot to a friend: .693

I would use this chatbot in the
future: .845

3.6.2 Reliability

To check whether or not all six variables are reliable a Cronbach's Alpha is calculated using SPSS. The aim is to check how closely related the items are in each construct. In order for a construct to be reliable the results of these tests must be at least $\alpha = .70$. First the two independent variables were tested both adjusted, the result showed that for *Visual cues* $\alpha = .91$, and *Language cues* $\alpha = .84$. For the dependent variables the results were *Trust* $\alpha = .89$. The constructs for *Perceived Intelligence* and *User Satisfaction* were merged into new dependent variable named *Satisfaction with the robot*, with $\alpha = .93$. And finally, *Willingness to use* $\alpha = .91$. All constructs showed values over and above the minimum $\alpha = .70$. Thus it can be confirmed that they are reliable and no further adjustments need to be made to the data.

4. Results:

4.1 Manipulation check

The first step of the result analysis is to do a manipulation check. This consists of performing two t-tests to check the two independent variables Visual cues and Linguistic cues, which had each two conditions. The goal is to see if there is a significant difference between the four different design conditions of the eHealth chatbot. This is a necessary step to ensure the internal validity of the 2x2 research design, and to further confirm if the participants are able to recognize the different design elements.

4.1.1 *Manipulation for Visual cues*

Each participant had to evaluate one of four stimuli that varied in terms of either Visual or Language cues. After viewing the assigned stimuli participants' had to evaluate four statements regarding their perception of the visual characteristics of the chatbot. First, the four items describe the visual characteristics into one construct called 'Visual cues' using SPSS. Following this an Independent Sample t-test was performed.

The results from the t-test showed substantial evidence that there is a significant difference between the two groups' human avatar and logo. Looking at the results for the two groups there is notable evidence that there are significant differences between human avatar ($M = 5.23$, $SD = 1.22$), and logo ($M = 3.47$, $SD = 1.51$) with $t(118) = 4.99$, $p < 0.001$. Given this, it can be concluded that the Human avatar scores higher in anthropomorphism compared to the Logo confirming the initial assumption.

4.1.2 *Manipulation for Visual cues*

The second independent variable that was examined was Linguistic cues. The participants viewed one of two conversational conditions (human language or robot language). There were 4 items measuring the construct of this independent variable, which were combined into one construct called 'Language cues. Following a t-test was performed to see if there are significant differences between the means of the two groups. The results show that there was a significant difference between the two groups human language ($M = 5.15$, $SD = 1.19$) and robot language ($M = 4.65$, $SD = .73$), with $t(118) = 9.99$, $p < 0.006$. Although the difference is not as vast as the other independent variable, it can still be concluded that the participants were able to distinguish between the two conditions. Therefore this assumption is confirmed.

4.2 Multivariate Analysis of Variance:

For this study the participants were presented with one of four conditions, which they had to view. Following this they answered 28 items measuring trust, perceived intelligence and user satisfaction (merged into one variable called robot satisfaction) and willingness to use. MANOVA analysis was conducted. This multivariate analysis of variance allows the analysis main and interaction effects of chatbot appearance and language on trust, satisfaction with robot and willingness to use are elaborated. The results are presented separately for each dependent variable. At the end of the section an overview of all hypotheses given.

4.2.1 Manipulation test for the independent variables

The first step of the MANOVA analysis is to observe the overall effect between the two independent variables (visual and linguistic cues). Table 3 shows the descriptive statistics for the Wilks' Lambda results. The first row shows the result for the main effect of chatbot language, there is a substantial evidence that there is an effect of chatbot language, with $\Lambda = .93$, $F(3, 114) = 2.88$, $p < .04$. Followed by the main effect of chatbot visual cues that is shown to be significant, with $\Lambda = .902$, $F(3, 114) = 4.13$, $p < .008$. Finally the interaction effect between the two independent variables called 'Visual cues * Language cues' came out to be non-significant with $\Lambda = .96$, $F(3, 114) = 1.56$, $p = .203$. Following the insignificance of the interaction effect H3 and H4 have to be rejected due to lack of proof.

Following the results of the Wilks' Lambda only the relation of the dependent variables Visual cues and Language cues on trust, the new variable satisfaction with the robots and willingness to use will be investigated further.

Table 3.
Wilks Lambda results

Effect		Value	F	Hypothesis		
				df	Error df	Sig.
Visual cues	Wilks' Lambda	.930	2.878	3.000	114.000	.039
Language cues	Wilks' Lambda	.902	4.127	3.000	114.000	.008
Visual cues * Language cues	Wilks' Lambda	.961	1.561	3.000	114.000	.203

4.2.2 Trust

In this section the main effect of visual and language cues on trust will be examined. Further, the interaction effect between the independent variables on trust will also be discussed. The SPSS analysis shows that there is a significant effect between visual cues and trust, with $p = .007$ ($F = 7.55$), and between language and trust with $p = .002$ ($F = 10.11$). However, as expected following the Wilks Lambda, there is no significant effect between the interaction of visual cues*language cues on trust ($p = .208$, $F = 1.602$).

Looking at table 4, it depicts the differences between the two groups: Human avatar and Logo. The variable 'Trust' was measured on a 7-point Likert scale from '1-Strongly disagree' to '7-Strongly agree'. The results of the descriptive show that the respondent presented with Human avatar ($M = 4.94$, $SD = .95$) are slightly more willing to trust the chatbot, compared to the ones presented with just the hospital Logo ($M = 4.44$, $SD = 1.09$). Hence, the assumption for H1 (a) is confirmed.

Similarly for the main effect of language cues on trust. The descriptive for the two groups for language cues show that participants presented with the Human language condition ($M = 4.97$, $SD = 1.09$) are slightly more willing to trust the conversational agent, compared to those that viewed the chatbot with the Robot language ($M = 4.40$, $SD = 0.92$). Meaning that H2 (a) can also be confirmed.

It was hypothesized that the combination of human language and human avatar will lead to higher level of trust. However, the results from the Multivariate analysis that there is no interaction effect of Visual cues * Language cues on trust. Due to this insignificance, H3 (a) cannot be confirmed, thus it is rejected.

Table 4.

Descriptive statistics Trust

Dependent variable	Independent variables	Condition	Std.		
			Mean	Deviation	N
Trust	Visual cues	Human avatar	4.94	.95	60
		Logo	4.44	1.09	60
	Language cues	Human language	4.97	1.09	60
		Robot language	4.40	.92	60

4.2.3 Satisfaction with the robot:

Further, the MANOVA analysis showed significant main effect of language on the new variable satisfaction with the robot, with $p = .04$ ($F = 4.29$). Table 5 depicts the two language groups that were designed and tested. The human language condition induced a higher likelihood of satisfaction ($M = 4.87$, $SD = 1.32$) than the robot language ($M = 4.44$, $SD = .95$). Meaning that anthropomorphic design of chatbot language has a significant impact on user satisfaction. This confirms hypothesis H2 (b, c).

It was hypothesized that visual cues would also have a main effect on satisfaction. However, the analysis showed this effect to be insignificant, with $p = .262$ ($F = 1.27$). It was expected that participants who were exposed to the human avatar condition would express higher satisfaction with the chatbot, compared to those who were exposed to the logo condition. Therefore, hypothesis H1 (b, c) has to be rejected due to the insignificance of the results.

Likewise, there was no interaction effect of visual cues*language cues on satisfaction, with $p = .857$ ($F = .033$). It was expected that the interaction between human avatar and human language would lead to higher satisfaction. However, since the interaction effect is insignificant H3 (b, c) is rejected.

Table 5.

Descriptive statistics Satisfaction with the robot

Dependent variable	Independent variables	Condition	Std.		
			Mean	Deviation	N
Satisfaction with robot	Language cues	Human language	4.87	1.32	60
		Robot language	4.44	.95	60

4.2.4 Willingness to use

In regards to the last dependent variable, the MANOVA analysis yields a significant main effect of language on willingness to use, with $p = 0.005$ ($F = 8.18$). Table 6 depicts the results of the main effect for the two groups. The survey respondents showed higher willingness to use human language chatbot ($M = 4.83$, $SD = 1.47$), compared to the robot language condition ($M = 4.14$, $SD = 1.13$). These results are in line with H1 (d), although the difference between the two groups is not large, human language is associated with higher willingness to use the chatbot.

It was hypothesized that there will be a main effect of visual cues on willingness to use. More specifically that human avatar condition will result in higher willingness to use, compared to the logo condition. However, the main effect between the variables is insignificant, with $p = .262$ ($F = 1.27$). Hence, due to the insignificance of the results H1 (d) is rejected.

Finally, it was hypothesized that there will be an interaction effect between visual cues*language cues on willingness to use. However there was no significant interaction ($p = .325$, $F = .977$). Thus, their effect cannot be examined further and H3 (d) is rejected.

Table 6.

Descriptive statistics Willingness to use

Dependent variable	Independent variables	Condition	Std.		
			Mean	Deviation	N
	Language cues	Human language	4.83	1.47	60
		Robot language	4.14	1.13	60

5. Hypothesis overview:

Table 7 depicts a visual summary of the results, and clarifies whether or not all hypotheses of this study are supported or rejected.

Table 7. *Hypotheses*

Hypothesis			Supported
H1	a	Users' perception of trust is higher when interacting with e-Health chatbot that uses human avatar, compared to the one that uses logo.	Yes
	b+c	Users' perception of satisfaction with the robot is higher when interacting with e-Health chatbot that uses human avatar, compared to the one that uses logo.	No
	d	Users' perception of willingness to use is higher when interacting with e-Health chatbot that uses human avatar, compared to the one that uses logo.	No
H2	a	Users' perception of trust is higher when interacting with e-Health chatbot that uses human language, compared to the one that uses robot language.	Yes
	b+c	Users' perception of satisfaction with the robot is higher when interacting with e-Health chatbot that uses human language, compared to the one that uses robot language.	Yes
	d	Users' perception of willingness to use is higher when interacting with e-Health chatbot that uses human language, compared to the one that uses robot language.	Yes
H3	a	Users' perception of trust is higher when interacting with e-Health chatbot that uses human avatar and human language, compared to the one that uses logo and robot language.	No
	b+c	Users' perception of satisfaction with the robot is higher when interacting with e-Health chatbot that uses human avatar and human language, compared to the one that uses logo and robot language.	No
	d	Users' perception of willingness to use is higher when interacting with e-Health chatbot that uses human avatar and human language, compared to the one that uses logo and robot language.	No

6. Discussion:

Whilst researchers have spent over two decades investigating the effect of chatbot design on users' perceptions, there is little empirical research regarding e-Health chatbots specifically. This paper seeks to answer the question: *'To what extent do the visual and conversational style of the virtual assistant affect the users' satisfaction, willingness to use and trust?'*. To do so, a 2x2 design study was conducted. This tested how visual and language characteristics affect users' trust, perceived intelligence, satisfaction and willingness to use an e-Health chatbot.

6.1 Discussion of results:

6.1.1 Main effect of Visual cues: (exploring H1)

The first effect that was tested was between visual cues and the dependent variables. According to the reviewed literature it was assumed that the participants will react more favourably towards the medical chatbot that was designed with anthropomorphic characteristics. To evoke a feeling of social presence the chatbot was given a human face and a name (Kim & Sundar, 2012; Kim & Sundar, 2012). Furthermore, according to McDonnell and Baxter (2019) users react more openly when presented with a female avatar. Based on these findings the 'human avatar' chatbot condition was designed to embody a female named 'Clara'. It was assumed that participants presented with this condition will score high on trust, satisfaction with the robot and willingness to use. However, this hypothesis was only partially confirmed. The only effect that was significant and was between the variables visual cues and trust. This was in line with the literature findings arguing that social cues convey presence, which results in higher willingness to trust a chatbot. These findings further support the idea that research used to evaluate e-commerce chatbots could also be applicable for the medical sector.

Literature finding also made connections between visual cues, and the variables satisfaction and willingness to use. It was assumed that users who viewed the human avatar, instead of just textual information would assume that the chatbot has personality. Hence, it was expected that these participants would score higher on satisfaction (Holzwarth, Janiszewski & Neumann, 2006). Moreover, high satisfaction is often associated with strong user loyalty and interest to use (Araujo, 2018). However, the finding showed that there are no main effects between the independent and dependent variables. Meaning, that this part of the hypothesis has to be rejected.

It is important that these findings are interpreted with caution. There are many factors that could have influenced the outcome of these results, and it is possible that further research may show different outcomes for this hypothesis. Another explanation that can be taken into account, is that the setting and the chatbot interaction process were highly unrealistic. The participants had to only imagine they were interacting with the robot. This may have caused frustrations and lack of attention when viewing the videos. Which in turn might have influenced their answers regarding satisfaction and future willingness to use. Moreover, there is the possibility that participants did not identify with the scenario that was displayed and this might explain the insignificant results of this interaction.

6.1.2 *Main effect of language cues (exploring H2):*

Similarly, the effect of language cues on trust, satisfaction with the robot and willingness to use was measured using the second hypothesis. As chatbots have limited capacity for interaction, they mainly rely on conversations to deliver the feeling of humanness (Go & Sundar, 2019). Some chatbots are able to appear as if they truly have their own ‘personality’ through their interactivity (Mctear, 2017; Sundar, et al., 2015), and familiarity of the language they use (Johnson, Patron & Lane, 2007). This ability is highly important when it comes to creating a medical chatbot that is expected to deal with complex and sensitive information. Moreover, the medical chatbot must be able to convey empathy and connect with users, in order to establish an emotional connection (Taylor, 2011). To make this chatbot sound more human like certain elements were added, such as emoticons and longer sentences when interacting with the user.

As expected, the participants that were exposed to the human language condition reported the higher result in all constructs. This further supports the importance of implementing social cues in the design of conversational agents as help establishing trust with patients. And also contributes to higher user satisfaction and willingness to use, which are theoretically related to user loyalty. However, it is again important to mention that there are a number of limitations that might have influenced the outcomes of this study. It is highly possible that the results may differ in a more realistic setting, or if the data was collected from a bigger population. These limitations will be discussed in detail below.

6.1.3 *Interaction effect (exploring H3):*

It was expected that there would be result differences between the groups that combined different design elements. This assumption came from a study done by Keyzer, Dens and Plsmacker (2017) that investigated the relationship between tone of voice and language used by a voice driven assistant, and its appearance. This inquiry was expected to confirm that the most beneficial combination of design characteristics for a medical chatbots would be human language and human avatar. Therefore it cannot be concluded that visual cues, by all means, have to be completely coherent with the language cues.

However, due to the insignificance of the interaction effect conclusion cannot be made, and the hypothesis was rejected due to lack of proof. It is highly possible that this was a result of the lack of personal interaction users had with the bot. They were not able to personally explore the design of the bot and overlooked some of the design characteristics. What is more, very little of the visual design was integrated in the video materials participants viewed.

6.2 Limitations:

Although this research provides new insights and contributes to the field of chatbot design, more specifically visual and linguistic design of medical chatbots. It is important to take into account all possible limitations and points for improvement. This would contribute to the planning, and preparation process for future research and findings in the field of e-Health chatbots.

First, it should be acknowledged that the participants were not able to personally interact with the medical chatbot. Instead, they had to imagine a fictional situation and view a video and think as if they were the one that were interacting with the medical chatbot. This limitation most likely had a great impact on the study outcomes, as the participants were not able to get to know the system properly. Moreover, the ‘user’ symptoms or explanations may have been another important point, where the participants were not able to identify with their role.

Second, due to the pandemic, the participants were approached mainly online, meaning that the research setting was not realistic. E-Health chatbots are meant to help people that might be in need of primary medical assistance. However, this study was conducted with participants that may have not been in the need of one. Meaning, the users were probably not in a situation where they could completely identify with the given scenario.

Thirdly, the survey intended to measure the concept of perceived intelligence and user satisfaction as two separate variables. However, due to the results of the factor analysis they had to be merged into one that only measured satisfaction. Thus, the effects between visual and linguistic design and perceived intelligence of a medical chatbot could not be measured.

Fourth, the construct of trust only measured the general trust in the medical chatbot. In the theoretical framework the concept of information privacy was discussed briefly, but no items were included to measure this. Future research on the topic may consider investigating further the context of privacy, in relation to trust in medical chatbots.

Fifth, the visual design of the chatbots was only limited to the avatar of the chatbot. Future research could focus on more elements of the interface designs, such as colors and positioning of different elements on the screen (buttons, etc.).

Finally, the e-Health chatbot was introduced as a service offered by a hospital. However, the interface of the chatbot was not designed from scratch. Instead, a Facebook Messenger template was used to design the four chatbot conditions. This may have left an impression on those that are familiar with the Messenger application, and thus affected their answers when asked about trust, satisfaction and willingness to use the chatbot in the future.

6.3 Implications

6.3.1 Theoretical implications:

The implementation of chatbots as part of the e-Health systems is still at its infant stages. This explains the lack of scientific research and information exploring the effects of design on patients. This paper expands the knowledge regarding the design implication of

anthropomorphic visual and language cues. It supports the idea that using anthropomorphic language designs results in better trust, satisfaction and willingness to use. Moreover, anthropomorphic visual design leads to higher user trust in the chatbot.

Social presence theory had a central role in justifying the choice of anthropomorphic design. Social cues conveyed through the chatbot language leads to higher trust, satisfaction and willingness to use the medical chatbot in the future (Go & Sundar, 2019; Lee & Choi, 2017). Taking this a step further, according to Liews, Tan and Ismail (2017) social presence is related to the media richness theory. The human language design of the chatbot contributes to the vividness of the interaction, and is perceived as more human-like. In this research, anthropomorphism was seen as an essential part for projecting social presence, and giving the chatbot a ‘personality’ in the online environment. Hence, the positive impact of the social presence theory is also applicable to the case of medical chatbots.

Next to this, this paper concludes that human-like features of medical chatbots do not create unsettling feelings among users. This goes against the uncanny valley theory, which suggests that anthropomorphic characteristics of technology and systems, such as chatbots may cause discomfort and upset users (Ciechanowski, et al., 2019). Thus, implementing human characteristics in the design process of medical chatbots would not result in negative UX. On the contrary, it rather creates an advantage.

6.3.2 Practical implications:

Despite the fact that some hypotheses were not supported by the results of this study, this paper still provides a number of practical implications. It serves as a helpful inside to future research in the field of developing medical chatbots. The outcomes can be of big influence to guide the work of interface designers, developers and researchers. The current design trend revolves around minimalistic design approach when creating interfaces. This study explores the practical implementation of anthropomorphic design of two of the main design characteristics of a chatbot – visual and language. The results of the study prove that social cues have a direct effect over UX. This sets the stage for further exploration of design elements and combinations.

The insides can be also used for direct practical implications in the field of medical development and research of e-Health automation. It be a useful inside, and inspiration for finding solutions, in the course of improving the quality of medical services and their impact on the quality of life. If designed and implemented correctly, chatbots are expected to be trusted enough by users to become the first point of contact with the hospital. Handling tasks such as result reading, checking medication and booking appointments, giving the patients an opportunity to receive medical assistance at all times. Moreover, this leaves room for more meaningful doctor-patient interaction.

6.4 Conclusion:

This research paper delved into the topic of anthropomorphic design of medical chatbots. The aim was to find out whether the assumptions designing more human-like conversational agents would lead to higher levels of trust, satisfaction and willingness to use among users. The findings of this study support the implementation of human language cues when designing the interface. What is more, the visual representation of a medical face contributes to the trust one has in the chatbot. However, there were no further results that could confirm that visual design would lead to better satisfaction and willingness to use. The health sector is one of the fields that is constantly going through technological developments, with the goal of improving the quality of life. Conclusively, this paper emphasizes on the importance of further investigation of the effect of design and its impact on users perceptions and experience with a system.

7. Acknowledgements:

Before wrapping up this study I would like to express special gratitude to the people who put as much effort and nerves into this product as I have. Firstly, to Dr. Joyce Karreman, who was my thesis supervisor the past few months. Despite the pandemic situation you have shown immense support and patience with my work process. Thank you for your positive and passionate supervision throughout the whole process. In addition, I would like to thank my family and friends for everything they have done to make the past three years at the University of Twente so memorable.

References:

- Adam, H., & Galinsky, A. D. (2012). Enclothed cognition. *Journal of Experimental Social Psychology*, 48(4), 918-925. doi:10.1016/j.jesp.2012.02.008
- Agrawala, M., Li, W., & Berthouzoz, F. (2011). Design principles for visual communication. *Communications of the ACM*, 54(4), 60-69. doi:10.1145/1924421.1924439
- Anderson, E. W., & Sullivan, M. W. (1993). The antecedents and consequences of customer satisfaction for firms. *Marketing science*, 12(2), 125-143.
- Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., & Wells, S. (2018). Trusting intelligent machines: deepening trust within socio-technical systems. *IEEE Technology and Society Magazine*, 37(4), 76-83. doi:10.1109/mts.2018.2876107
- Appel, J., Pütten, A. V., Krämer, N. C., & Gratch, J. (2012). Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Advances in Human-Computer Interaction*, 2012, 1-10. doi:10.1155/2012/324694
- Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85, 183-189. doi:10.1016/j.chb.2018.03.051
- Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2008). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71-81. doi:10.1007/s12369-008-0001-3
- Bates, M. (2019). Health care chatbots are here to help. *IEEE Pulse*, 10(3), 12-14. doi:10.1109/mpuls.2019.2911816
- Bibault, J., Chaix, B., Nectoux, P., Pienkowski, A., Guillemasé, A., & Brouard, B. (2019). Healthcare ex Machina: Are conversational agents ready for prime time in oncology? *Clinical and Translational Radiation Oncology*, 16, 55-59. doi:10.1016/j.ctro.2019.04.002
- Blascovich, J. (2001). Immersive virtual environment technology as a tool in psychological science. *PsycEXTRA Dataset*. doi:10.1037/e324232004-012
- Bouchagiar, G. (2018). Initial coin offering and cryptocurrencies: shifting trust away from human actors and toward a cryptographic system. *Journal of Financial Risk Management*, 07(04), 386-427. doi:10.4236/jfrm.2018.74022
- Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *Internet Science Lecture Notes in Computer Science*, 377-392. doi:10.1007/978-3-319-70284-1_30
- Cahn, J. (2017). *CHATBOT: Architecture, Design, & Development* (Master's thesis, University of Pennsylvania, 2017) (pp. 30-32). Pennsylvania: Department of Computer and Information Science.

- Chun, M., Ko, E., Joung, H., & Kim, S. J. (2018). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*. doi:10.1016/j.jbusres.2018.10.004
- 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. doi:10.1016/j.future.2018.01.055
- Cronin, J. J., Brady, M. K., & Hult, G. T. M. (2000). Assessing the effects of quality, value, and customer satisfaction on consumer behavioural intentions in service environments. *Journal of Retailing*, 76(2), 193-218
- Cyr, D., Head, M., & Ivanov, A. (2009). Perceived interactivity leading to e-loyalty: Development of a model for cognitive–affective user responses. *International Journal of Human-Computer Studies*, 67(10), 850-869. doi:10.1016/j.ijhcs.2009.07.004
- Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(5), 811-817. doi:10.1017/s1351324916000243
- De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331-349. doi:10.1037/xap0000092
- Denecke, K., Tschanz, M., Dorner, T. L., & May, R. (2019). Intelligent conversational agents in healthcare: hype or hope? *Healthcare of the Future*, 259(1), 91-134. doi:10.3233/978-1-61499-961-4-77
- Duran, N. D., Hall, C., McCarthy, P. M., & Mcnamara, D. S. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics*, 31(3), 439-462. doi:10.1017/s0142716410000068
- Elofson, G. (2001). Developing trust with intelligent agents: an exploratory study. *Trust and Deception in Virtual Societies*, 125-138. doi:10.1007/978-94-017-3614-5_6
- Etemad-Sajadi, R. (2016). The impact of online real-time interactivity on patronage intention: The use of avatars. *Computers in Human Behaviour*, 61, 227-232. doi:10.1016/j.chb.2016.03.045
- Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, 24(4), 38-42. doi:10.1145/3085558
- Friedman, B., Khan, P. H., & Howe, C. H. (2000). Trust online. *Communications of the ACM*, 43(12), 34-40. doi:10.1145/355112.355120
- Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-commerce and the importance of social presence: experiments in e-products and e-services. *Omega*, 32, 407-424
- Gennaro, M. D., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, 10. doi:10.3389/fpsyg.2019.03061

- Giard, F., & Guitton, M. J. (2010). Beauty or realism: The dimensions of skin from cognitive science to computer graphics. *Computers in Human Behavior*, 26(6), 1748-1752
- Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behaviour*, 97, 304-316. doi:10.1016/j.chb.2019.01.020
- Gustavsson, E. (2005). Virtual servants: Stereotyping female front-office employees on the internet. *gender, work and organization*, 12(5), 400-419. doi:10.1111/j.1468-0432.2005.00281.x.
- Handro, P. (2018). The role of customer experience in retail banking and the rise of fintechs. *'Constantin Brancusi' University of Targu - Jiu. Economy Series*, (1).
- Hassanein, K., & Head, M. (2007). Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping. *International Journal of Human-Computer Studies*, 65(8), 689-708. doi:10.1016/j.ijhcs.2006.11.018
- Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. *Computers in Human Behavior*, 49, 245-250. doi:10.1016/j.chb.2015.02.026
- Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266. doi:10.1126/science.aaa8685
- Holzwarth, M., Janiszewski, C. & Neumann, M.M.(2006).The influence of avatars on online consumer shopping behaviour. *Journal of Marketing*, 70(4), 9-36.doi: <https://doi.org/10.1509/jmkg.70.4.19>
- Hult, G.T.M., Morgeson, F.V., Morgan, N.A. (2017) Do managers know what their customers think and why?. *J. of the Acad. Mark. Sci.* 45, 37–54 (2017). doi:10.1007/s11747-016-0487-4
- Hussain, S., & Athula, G. (2018). Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. doi:10.1109/waina.2018.00170
- James Manyika (2018). AI, automation, and the future of work: Ten things to solve for. Retrieved April 22, 2020, from <https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for>
- Johnson, N., Patron, P., & Lane, D. (2007, June). The importance of trust between operator and auv: Crossing the human/computer language barrier. In *OCEANS 2007-Europe*(pp. 1-6). IEEE.
- Keyzer, F. D., Dens, N., & Pelsmacker, P. D. (2017). Don't be so emotional! How tone of voice and service type affects the relationship between message valence and consumer responses to WOM in social media. *Online Information Review*, 41(7), 905-920. doi:10.1108/oir-08-2016-0219

- Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1), 241-250. doi:10.1016/j.chb.2011.09.006
- Korthagen, F. A. (1993). Two modes of reflection. *Teaching and Teacher Education*, 9(3), 317-326. doi:10.1016/0742-051x(93)90046-j
- Kvale, K., Sell, O. A., Hodnebrog, S., & Følstad, A. (2020). Improving conversations: lessons learnt from manual analysis of chatbot dialogues. *Chatbot Research and Design Lecture Notes in Computer Science*, 187-200. doi:10.1007/978-3-030-39540-7_13
- Lee, S. Y., & Choi, J. H. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, 103, 95–105. doi: 10.1016/j.ijhcs.2017.02.00
- Lee, S., Lee, N., & Sah, Y. J. (2019). Perceiving a mind in a chatbot: Effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of Human-Computer Interaction*, 36(10), 930-940. doi:10.1080/10447318.2019.1699748
- Liew, T. W., Tan, S., & Ismail, H. (2017). Exploring the effects of a non-interactive talking avatar on social presence, credibility, trust, and patronage intention in an e-commerce website. *Human-centric Computing and Information Sciences*, 7(1). doi:10.1186/s13673-017-0123-4
- Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: the impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*. doi:10.1287/mksc.2019.1192
- Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), 277-301. doi:10.1080/14639220500337708
- Mahmood, M. A., Burn, J. M., Gemoets, L. A., & Jacquez, C. (2000). Variables affecting information technology end-user satisfaction: a meta-analysis of the empirical literature. *International Journal of Human-Computer Studies*, 52(4), 751-771
- Marcondes, F. S., Almeida, J. J., & Novais, P. (2018). Chatbot theory. *Intelligent Data Engineering and Automated Learning – IDEAL 2018 Lecture Notes in Computer Science*, 374–384. doi:10.1007/978-3-030-03493-1_40
- Marino, M. C. (2014). The racial formation of chatbots. *CLCWeb: Comparative Literature and Culture*, 16(5). doi:10.7771/1481-4374.2560
- Mcdonnell, M., & Baxter, D. (2019). Chatbots and gender stereotyping. *Interacting with Computers*, 31(2), 116-121. doi:10.1093/iwc/iwz007
- Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2(2). doi:10.1145/1985347.1985353

- McTear, M. F. (2017). The rise of the conversational interface: A new kid on the block? *Lecture Notes in Computer Science Future and Emerging Trends in Language Technology. Machine Learning and Big Data*, 38-49. doi:10.1007/978-3-319-69365-1_3
- McTear, M., Callejas, Z., & Griol, F. G. (2016). The conversational interface -talking to smart devices. *Springer international publishing (1st ed.)* doi: 10.1007/978-3-319-32967-3
- Morrissey, K., & Kirakowski, J. (2013). 'Realness' in chatbots: establishing quantifiable criteria. *Lecture notes in computer science: vol. 8007. International Conference on Human-Computer Interaction* (pp. 87-96). Berlin, Germany: Springer. doi: 10.1007/978-3-642-39330-3_10
- Müller, L., Mattke, J., Maier, C., & Weitzel, T. (2020). Conversational agents in healthcare: using QCA to explain patients' resistance to chatbots for medication. *Chatbot Research and Design Lecture Notes in Computer Science*, 3-18. doi:10.1007/978-3-030-39540-7_1
- Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An initial model of trust in chatbots for customer service—findings from a questionnaire study. *Interacting with Computers*, 31(3), 317-335. doi:10.1093/iwc/iwz022
- Oliveira, T., Alinho, M., Rita, P., & Dhillon, G. (2017). Modelling and testing consumer trust dimensions in ecommerce. *Computers in Human Behavior*, 71, 153-164. doi:10.1016/j.chb.2017.01.050
- Palanica, A., Flaschner, P., Thommandram, A., Li, M., & Fossat, Y. (2019). Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *Journal of Medical Internet Research*, 21(4). doi:10.2196/12887
- Pereira, J., & Díaz, Ó. (2019). Using health chatbots for behavior change: a mapping study. *Journal of Medical Systems*, 43(5). doi:10.1007/s10916-019-1237-1
- Pilato, G., Augello, A., & Gaglio, S. (2011). A modular architecture for adaptive chatbots. *IEEE Fifth International Conference on Semantic Computing*, Palo Alto, CA, 2011, pp. 177-180.
- Qian, Q., Huang, M., Zhao, H., Xu, J., & Zhu, X. (2018). Assigning personality/profile to a chatting machine for coherent conversation generation. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. doi:10.24963/ijcai.2018/595
- Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, 25(4), 145-182
- Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent Conversational agents. arXivpreprint arXiv:1704.04579
- Reeves, B., & Nass, C. (1996). The media equation: how people treat computers, television, and new media like real people and places. New York, USA: Cambridge University Press.
- Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust scale. *PsycTESTS Dataset*. doi:10.1037/t01317-000

- Reshmi, S., & Balakrishnan, K. (2018). Empowering chatbots with business intelligence by big data integration. *International Journal of Advanced Research in Computer Science*, 9(1). doi:10.26483/ijarcs.v9i1.5398
- Sansonnet, J.-P., Leray, D., & Martin, J.C. (2006). Architecture of a framework for generic assisting conversational agents. *Intelligent Virtual Agents Lecture Notes in Computer Science*, 145–156. doi:10.1007/11821830_12
- Shum, H., He, X., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1), 10-26. doi:10.1631/fitee.1700826
- Sundar, S. S., Go, E., Kim, H., & Zhang, B. (2015). Communicating art, virtually! Psychological effects of technological affordances in a virtual museum. *International Journal of Human-Computer Interaction*, 31(6), 385-401. doi:10.1080/10447318.2015.1033912
- Taylor, L. D. (2011). Avatars and emotional engagement in asynchronous online communication. *Cyberpsychology, Behavior, and Social Networking*, 14(4), 207-212. doi:10.1089/cyber.2010.0083
- Toader, D., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2019). The effect of social presence and chatbot errors on trust. *Sustainability*, 12(1), 256. doi:10.3390/su12010256
- Toma, C. L. (2010). Perceptions of trustworthiness online: the role of visual and textual information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 13-22). ACM
- Trivedi, J. (2019). Examining the customer experience of using banking chatbots and its impact on brand love: the moderating role of perceived risk. *Journal of Internet Commerce*, 18(1), 91-111. doi:10.1080/15332861.2019.1567188
- Wixom, B. & Todd, P. (2005). A theoretical integration of user satisfaction and technology acceptance. *Information Systems Research*, 16(1), 85-102. doi:10.1287/isre.1050.004
- Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3025453.3025496
- Xu, B., Zhang, W., & Li, D. (2011). Co-presence and its social-relational antecedents in computer-mediated communication: The case of instant messenger. *International Journal of Human-Computer Interaction*, 27(10), 991-1009. doi:10.1080/10447318.2011.555316
- Zhou, L., Gao, J., Li, D., & Shum, H. (2020). The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1), 53-93. doi:10.1162/coli_a_00368
- Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting virtual agents. *ACM Transactions on Interactive Intelligent Systems*, 9(2-3), 1-36. doi:10.1145/3232077

Appendix 1: Survey

Thank you for participating in this online study about e-Health chatbot design.

In order to answer the survey questions you will first read a fictional scenario and watch a short video interaction with the chatbot. Completing the whole study will take about 12 minutes.

Your participation in this study is voluntary, therefore you have the right to withdraw at any point during the process. Any data you provide will be handled anonymously and will not be used outside the purpose of this study.

In case of any ambiguities, or other questions, feel free to contact the researcher by email: v.petrova@student.utwente.nl

In compliance with the EU General Data Protection Regulations (GDPR) your consent is required.

I consent that I am:

- I am 18 or older;
- My participation is voluntary;
- I understand that I can withdraw from this study at any given time;
- I have been informed about my right to request access, clarification, and erasure of my data.

Block 1: Demographics:

What is your age:

What is your gender: Female/Male/Other/ Prefer not to say

Where do you come from: Netherlands/ Germany/Other

What is the highest degree you have completed: Bellow high school/ High school/ Associate/ Bachelor/ Master/ Doctorate/ Other

Block 2: Read carefully before proceeding:

Imagine that the hospital in your district has recently implemented a chatbot that is able to answer primary consultations questions. This includes suggesting general diagnosis, delivering and clarifying test results and booking appointments with specialists, based on your medical records and input.

You have been feeling unwell for a while but right now are too busy to go for a check so you decide to give this chatbot a try.

(Now you will play a video interaction with this chatbot, following this you will be presented with a short survey where you will rate your experience.)

Block 3: Assigned video condition:

Please watch the following video carefully. (All four conditions can be found in Appendix 2.)

Block 4: Chatbot manipulation:

The chatbot avatar showed a human picture:

Yes/No

The chatbot interaction was open (the answers were open and not per-defined by the chatbot):

Yes/No

Block 5: Visual cues:

The chatbots' picture made the interaction feel as if it was with a real person:

The chatbots' picture made the interaction feel life-like:

The chatbots' picture made the interaction feel human-like:

The chatbots' picture made the interaction feel natural:

Strongly disagree – Strongly agree

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Block 6: Linguistic cues:

The impression of the chatbots' language felt alive:

The chatbots' responses felt human-like:

The language that the chatbot used felt natural:

The chatbots' responses felt sensible:

Strongly disagree – Strongly agree

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Block 7: Trust:

The information received from the chatbot was credible:

The information I received from the chatbot is trustworthy;

The chatbots' responses were reliable:

The chatbot interaction felt believable:

Strongly disagree – Strongly agree

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Block 8: Perceived intelligence:

This chatbot is intelligent:

This chatbot is qualified to provide medical assistance:

This chatbot is competent enough to provide this information:

Strongly disagree – Strongly agree

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Block 9: Satisfaction:

I am satisfied with the responses of this chatbot:

I am satisfied with the way the chatbot helped me:

I believe the chatbot provided good responses:

I am satisfied with the assistance received from the chatbot:

Strongly disagree – Strongly agree

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Block 10: Willingness to use:

I would download this application:

If I have symptoms I would turn to this this chatbot:

I would recommend this chatbot to a friend:

I would use this chatbot:

Strongly disagree – Strongly agree

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

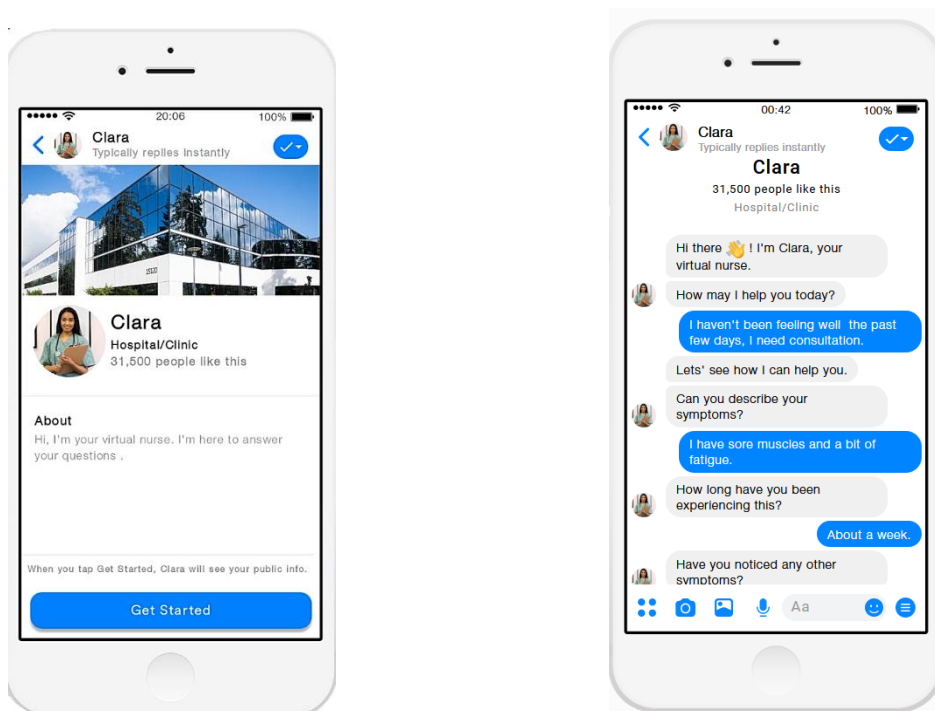
1 2 3 4 5 6 7

Appendix 2: Video conditions

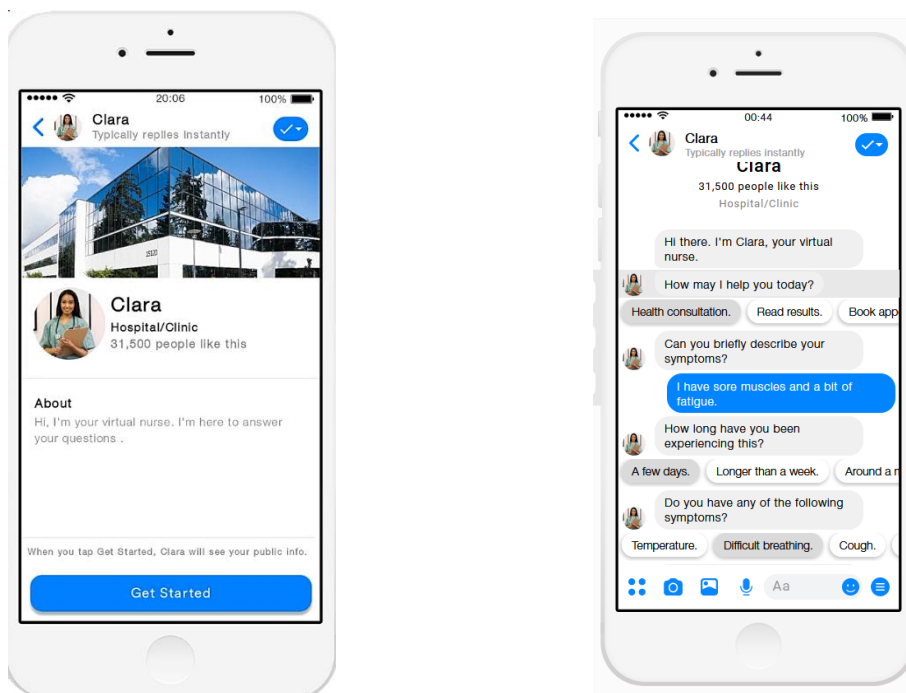
All videos are uploaded on the following link:

https://drive.google.com/drive/folders/1jNKzfuiUN_kD4OCA24_35wm-UIUMdA-f?usp=sharing

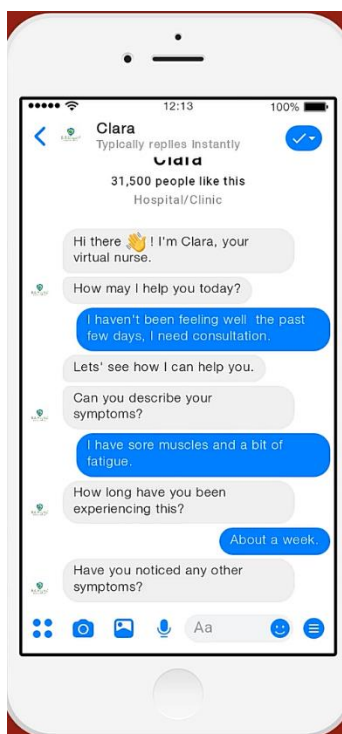
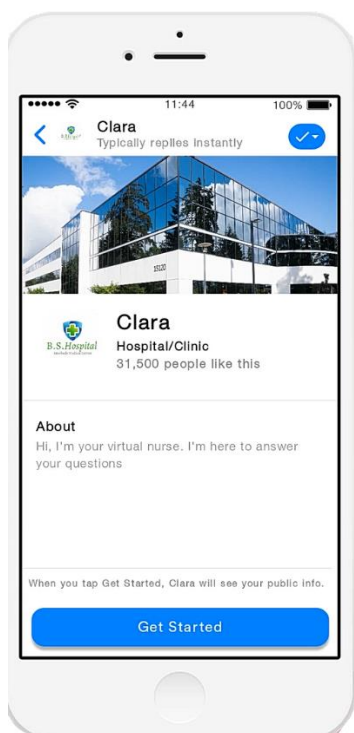
Video condition 1: human avatar x human language



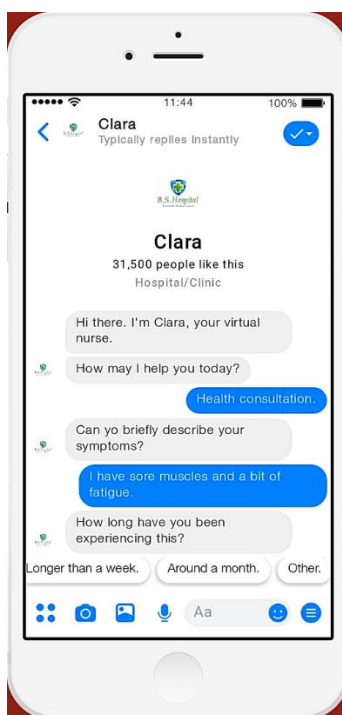
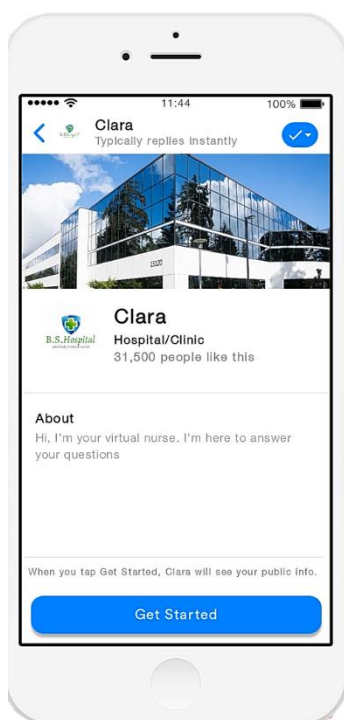
Video condition 2: human avatar x robot language



Video condition3: logo x human language



Video condition 3: logo x human language 4: logo x robot language



Appendix 3:

Source criteria:

Source type: scientific research articles

Date of publishing: 1850- 2020

Search log:

Constructs	Search terms and strategies	Broader terms	Narrower terms
Visual cues	Chatbot appearance Human design Anthropomorphic design Social cues e-Health visuals Human avatar Medical chatbot design	Interface design	Chatbot appearance
Language cues	Social presence Interactive design Human language Social cues Frequency of interaction Medical chatbot design Response design Natural language processing (NLP)	Human-computer interaction	Human language
Trust	Trust in chatbots Trust in human design Trust in e-Health Uncanny valley	Vulnerability	Predictability Dependency Faith
Perceived intelligence	Intelligent machines Credibility of medical information Uncanny valley Information complexity	Competence	
User satisfaction	Satisfaction in design Chatbots and UX Anthropomorphic design and satisfaction UX and design	Experience	Expectations Accomplishment
Willingness to use	Social presence theory Acceptance and adoption	Commitment	Loyalty Usability Value

Databases:

The main databases that were used for the literature research were Scopus and Google Scholar because they provide access to number of materials such as: thesis, scientific articles, and books and from various disciplines. Moreover, searching across all three platforms made resulted in large amount of scientific research on the concepts of the study, the articles were selected by means of specific term search and snowballing. The search was limited to documents published since 2016 so theta the findings are more relevant, however after using the snowball search method many articles preceding this period were included.

Search

Database	Search string	Total hits
Google scholar	'Chatbot' and 'design'	10,500
Google scholar	'Chatbots' and 'healthcare'	2,980
Google scholar	'Chatbot' and 'anthropomorphic design '	141
Scopus	(TITLE (chatbot) AND TITLE (design or anthropomorphism))	260
Google scholar	'Chatbot' and 'UX'	913
Google scholar	'Chatbots' and 'social presence'	428
Scopus	(TITLE (chatbot) AND TITLE (e-Health or medical or hospital))	107
Scopus	(TITLE (chatbots) and TITLE (UX))	14
Scopus	(TITLE (chatbots) and TITLE (trust))	30
Scopus	(TITLE (chatbots) and TITLE (adoption))	56
Google scholar	'AI' and 'perceived intelligence'	540
Google scholar	'chatbots' and 'UX '	851
Google scholar	'Chatbot' and 'willingness to use'	106

Found references:

- Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *Internet Science Lecture Notes in Computer Science*, 377-392. doi:10.1007/978-3-319-70284-1_30
- 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. doi:10.1016/j.future.2018.01.055
- Følstad, A., & Brandtzaeg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, 24(4), 38-42. doi:10.1145/3085558
- Handro, P. (2018). The role of customer experience in retail banking and the rise of finetechs. *'Constantin Brancusi' University of Targu - Jiu. Economy Series*, (1).
- McTear, M., Callejas, Z., & Griol, F. G. (2016). The conversational interface -talking to smart devices. *Springer international publishing (1st ed.)* doi: 10.1007/978-3-319-32967-3
- Toader, D., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2019). The effect of social presence and chatbot errors on trust. *Sustainability*, 12(1), 256. doi:10.3390/su12010256
- Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting virtual agents. *ACM Transactions on Interactive Intelligent Systems*, 9(2-3), 1-36. doi:10.1145/3232077