Universiteit Leiden
The Netherlands

Bachelor Informatica & Economie

Human Social Characteristics of Conversational Agents for Elderly Users: The case of Welzijn.AI

Bachelor Thesis

Casper van Wordragen
S2669994

Supervisor: **Prof. Dr. Marco Spruit**
Second reader: **T.D. Offerman MSc**

Abstract

With the demand for healthcare personnel increasing, the healthcare world is looking for options to decrease the necessity of people or support them with machines. One of the opportunities for the healthcare world is the use of e-health applications. One example is Welzijn.AI, an application that registers the physical and mental state of elderly patients by means of a questionnaire. This thesis explores two distinct parts. Firstly the current state of the application as perceived by elderly users was researched. Secondly, this thesis aimed to find what elderly deem the most important human social characteristics, in a conversational agent that administers a questionnaire in the application. The results in this exploratory research were obtained by conducting interviews with elderly people. This interview consisted of a general questionnaire, to elicit the current state of the application based on 6 variables, and a card sorting variation, to find which human social characteristics are most important in a conversational agent to elderly users and finally a segment where any further suggestions to improve the application were discussed. The questionnaire produced varying results, meaning that the interpretation of the current state of the application is not accurate, but what is clear is that the application needs to be improved to accommodate all elderly. The use of card sorting resulted in the discovery that there is a large variation in preferences by elderly. By using the trend for the ranking of the characteristics by the elderly, it was found that elderly find the expression or interpretation of emotion, paired with a distinct character, the most important characteristic in a conversational agent. The characteristic that was least important for a conversational agent was the addition of natural cues. Since most users have different preferences, and results were not unanimous, it was concluded that further research with a larger sample size is needed to confirm the current findings and improve upon the application.

Contents

1. Introduction

In 2018 experts from multiple health disciplines joined forces and framed a report concerning the improvement of patient care in the Netherlands and specifically elderly patient care (Rijksoverheid, 2018). One of the leading arguments in the report is that innovation will lead the Dutch medical care system to new heights, with the most important step in the innovation process being digitalization. The study states that if we continue down the path that we are headed without digitalization, one in four people in the Netherlands will end up working in health come 2040, adding to the already present shortage of people for other professions (Rijksoverheid, 2018).
In this modern world of technology however, many processes previously performed by humans can now be automated by machines or applications. Take for example the use of chatbots by web shops, reducing or removing the need for human customer support. Besides business, automation by application within the realms of education, finance or health is increasing (Dabney et al., 2013). This thesis will focus on the latter. To be more specific, automatization of standardized questionnaires performed on elderly people, registering their mental and physical state by means of a conversational agent.
Currently, many of these questionnaires are performed in person, by either a doctor in a hospital, nurse in a clinic or a general practitioner. This healthcare personnel might already have a high workload, which is proven correlate with a lower practice performance and could be harmful to patients and practitioners themselves (Van den Hombergh et al., 2009). It would impact patients and healthcare personnel positively if workload could be reduced, for example by taking a simple task like a questionnaire out of their hands. Studies have shown that doctors think the addition of e-health to the medical world can lead to an increase in efficiency (Mickan et al., 2013), meaning it is an interesting road to explore as a next step in the digitalization of healthcare. Problems with e-health applications however, are that just like with normal medication, people have to be adherent to them (Shruthi, 2016). And this is only the case if users perceive the system as useful and easy-to-use (Davis, 1989).
The goal of this exploratory study is the improvement of the application “Welzijn.AI”, largely focusing on what important characteristics are for a suggested conversational agent. Welzijn.AI is an e-health application being developed for doctors to track mental and physical health of elderly patients by means of a questionnaire. To achieve the goal of this study, this thesis tries to find what the current state of the application is and what implementations could improve the application. The research will consist of a questionnaire to test in what state the application is currently, according to 6 different variables, and a round of interviews to find out what the preferred characteristics of a conversational agent are. The envisioned goal of Welzijn.AI is to create an application to with high elderly adherence levels, for example by adding the suggested conversational agent adding to the fun factor of the application. In this case, the application is not yet complete, so adherence to the application cannot be measured exactly. However, adherence can be tested through a variable called ‘intent to use’, a study into user intent for information systems states that intent of use translates closely to actual use (Davis, 1989), which was later confirmed by another study (Heerink et al., 2008). This thesis aims is to answer the following research question: “With which characteristics and to what extent can a conversational agent be used to improve adherence in e-health applications such as Welzijn.AI for the elderly?” To answer this question, two separate sub-questions should be answered first. Namely:
Q1: “What is the perception of Welzijn.AI, an example of an e-health application for elderly users, and do they intend to use this system” and;
Q2: “What human social characteristics are most and least important in a conversational agent to elderly people?”.

In the following sections literature regarding e-health applications, elderly patient care and conversational agents will be discussed. Furthermore, the design of the study will be outlined. Then results from the performed study will be shown and discussed, outlining the limitations of the study

and answering the sub-questions. Finally a conclusion is formed.

2. Background

It is safe to say that e-health will be a big part of future medical developments, for treatment, diagnosis and after care (GSMA and McKinsey & Company, 2012). People like having quick access to information and sometimes do not like to go to the doctor, so for some, this is a development that can come sooner rather than later. Before arriving at the point where one application, program or machine can solve every medical mystery like in sci-fi movies, smaller steps have to be made in the automation direction first. One of these steps is the improvement of an application that this research is focused on, the 'Welzijn.AI' application. This is an e-health application that is focused on patient care, especially for the elderly generation. The goal of this application is to stay up-to-date with a patient's mental and partial physical state. Usually, a person's mental state is determined by talking to a doctor and filling in some questionnaires which give a score.
Take for example the EQ-5D questionnaire which is a standardized questionnaire that measures the wellbeing of a person on 5 levels, mobility, self-care, daily activities, pain/discomfort and fear/depression. This questionnaire produces a score, which the doctor can in turn use to describe a certain treatment plan. This is a process that has to be repeated biweekly to confirm diagnosis, track changes or foresee any future treatments.
The goal of many e-health applications, aside from helping patients, is improving efficiency and saving time (Elbert et al., 2014). For example, if the process that the 'Welzijn.AI" application is focused on takes 20 minutes for one patient each week and you multiply this by the number of patients a doctor has to keep track of. This results in the questionnaires not being filled in because there is no time available for doctors to facilitate this. Welzijn.AI seeks to automate this and provide the results of the questionnaires to the doctors. The task at hand now, is getting the elderly to use such an e-health application, for example by increasing the fun factor of the application with a conversational agent.

2.1 E-health applications

The amount of consumer-focused e-health applications that have entered the market during the past two covid dominated years has skyrocketed (RIVM, 2021). These applications give people the opportunity to be in charge of their own health. These applications help users keep track of their intake, get daily reminders to do healthy activities, or get information from symptom descriptions usingAIs (Noar et al., 2012), whilst not even having to go to a hospital or other medical facility. The opportunities that e-health offers give people direct access to medical knowledge, improvement in patient medication adherence by means of notifications (Kebapci et al., 2019), and more control over their own health (Ministerie van Algemene Zaken, 2016).

The ultimate goal of all combined e-health applications is to automate things. To make diagnosis more efficient, to free up doctors time and reduce error rates (Rowland et al., 2019). Seeing as elderly are the largest recipients of healthcare (World health organization, 2009), it is a logical next step to have elderly patients start using e-health.

2.2 Elderly patient care using e-health applications

An underrepresented group in the e-health sphere is the elderly generation. An elderly is defined as someone who is above the age of 65 (Orimo et al., 2006). Usually this is the audience with the most health problems, but with the least knowledge when it comes to mobile technology (Shahbazi et al., 2021). Elderly people usually don't like talking with machines, they like talking to real people and having meaningful conversations. A study by Shahbazi et al. (2021), concluded that the use of e-

health applications by elderly may present challenges as their understanding of technology is limited. But these challenges should not be the reason that elderly do not get access to better self-management and lessen need for family members to step in. Campen et al. (2011) states that due to the increase of mental health issues in elderly, the elderly are becoming more and more vulnerable. A suggestion to improve upon mental health issues is the use of e-health applications (Ebert et al., 2018). Having elderly patients use technology for any type of treatment however, can prove to be a difficult task, especially if they are not willing to learn. A transition from being treated face to face, to partially being helped with technology is not something that can be achieved easily. It is a big step which many elderly probably do not want to take, face-to-face care is preferred to face-to-machine care (Yuan et al., 2015). When, and if doctors choose to transition certain parts of patient care to an automated system, this has to be done incrementally and with support of staff that trains elderly. This step is part of change management and is essential when transferring systems. Elderly are not the quickest with technology and will need training in how to use these applications. Research conducted by Bagherian et al. (2021b) showed that there are technical, human and managerial challenges for getting elderly to use e-health applications. With the most important challenge being the human challenge, the study outlined that the users knowledge and acceptance of a system were the biggest challenges in this domain. Opportunities suggested by the study are a reduction of healthcare costs, no need for verbal visits and remote access to health information. Advantages of using e-health application are improvement in physical or mental e-health, with a disadvantage being a breach of privacy. Many of the challenges and disadvantages faced can be solved by education of elderly in the use of e-health systems (Bagherian et al., 2021b).

2.3 Conversational AI

To understand how a conversational AI might lead elderly users to use the Welzijn.AI application, one must first understand what a conversational AI actually is and how it functions.

The basic premise of a conversational AI or agent is a machine that can imitate a human-like conversation (Khatri et al., 2018). The 3 key principles that make a conversational AI work are:

- **Machine learning:** A simulated way for computers to learn the way that humans themselves learn as well. Every time it performs a certain action with new input data, it learns from that and improves upon itself (Rebala et al., 2019). Just like humans do.
- **Natural language understanding:** Comprehension or interpretation of what a person has said (Semaan, 2012).
- **Natural language generation:** Making sure an appropriate natural language response is formed (Reiter & Dale, 1997).

These 3 principles, when combined with the correct training data, will lead to a working conversational agent. Where machine learning will make the conversational AI be able to learn from previous conversations, by asking questions and defining intended perception of the input. Natural language understanding will help the AI to categorize the input and respond with correct information. And natural language generation will help the AI to respond in a human-like way.
The most common use cases for these types of agents are customer service services for online web shops, data collection and healthcare. This study focusses on the latter. A study implementing a conversational AI, studying the effects it has on mental wellbeing, produced promising results regarding its effectiveness in improving mental health (Inkster et al., 2018).

2.3.1 Rule based chatbots vs conversational AI

Besides conversational AI, a rule based chatbot is also a way to conduct a conversation. But this is a ‘dumb’ chatbot. Meaning it has a limited amount of responses it can give to inquiries done to the chatbot. It can only respond to questions or tasks it was programmed to respond to. So anything off-

topic cannot be responded to and will raise the error message that it did not understand the question. This is what the current structure of the Welzijn.AI resembles, a chatbot that can only take a questionnaire and then shuts down. As one can imagine an interactive conversation is preferred over a conversation that has no variation and just stops.

2.4 Anthropomorphism in conversational agents

Anthropomorphism is the act of adding human like features to a non-living entity (Guthrie, 1997). Usually with the intent to let the interaction feel as human as possible. Where a conversational agent can learn from previous conversation, just like a human would, a rule based chatbot would only be able to respond with a preprogrammed set of sentences. Adding anthropomorphism to a conversational agent makes a system feel more like a normal human interaction, making it more accessible to talk to.

2.4.1. Perception of a conversational agent

This study aims to find out what the key characteristics are that should be incorporated in a conversational agent if implanted in the Welzijn.AI application. A study that looked into "conventional" human-robot interaction (Fong et al., 2003), suggest that conversational robots should exhibit the following "human social" characteristics:

1. express and/or perceive emotions;
2. communicate with high-level dialogue;
3. learn/recognize models of other agents;
4. establish/maintain social relationships;
5. use natural cues (gaze, gestures, etc.)
6. exhibit distinctive personality and character;
7. may learn/develop social competencies

A study conducted by Heerink et al. (2008) explored human social characteristics in conversational agents and largely confirmed the above characteristics and explored them more in depth. Heerink et al. (2008) suggest that in addition to 1, 4, 5, 6, and 7 important social abilities are trust and responsibility. Characteristics 2 and 3 as defined by Fong et al. (2003) are very general, so these are replaced by the two extra suggested characteristics by Heerink et al. (2008). Below shows a comparative table of characteristics proposed by both studies, with the characteristics marked in green that were used for this study.

	Study	
	Fong et al. (2003)	**Heerink et al. (2008)**
Characteristics	Express and/or perceive emotions;	Express empathy
	Communicate with high-level dialogue	Cooperate
	Learn/recognize models of other agents	Show assertivity
	Establish/maintain social relationships	Exhibit self-control
	Use natural cues (gaze, gestures, etc.)	Show competence
	Exhibit distinctive personality and character	Show responsibility
	May learn/develop social competencies	Gain trust

s

These anthropomorphic characteristics are implemented in an example of a conversational agent, and tested to what extent they influence 6 distinct variables. These variables all have impact on whether a person will use an application and thus adherence. Accessibility is defined, because if an application is not accessible enough, there might be a barrier to use the application (Díaz-Bossini and Moreno, 2014). Usefulness and satisfaction of an application are defined because these have direct impact on the use of the application as well: use of a system is correlated to its perceived usefulness (Davis 1989) and there is a causal relationship between satisfaction and the use of an application (Atreja et al., 2005). Furthermore, intention to use a system leads to actual use of a system, as hypothesized by Davis and confirmed by Heerink et al. Finally trust in an application also directly impacts the adherence of a system Meng et al. (2022). Additional explanation of the variables that are being tested by questionnaire are described below.

Accessibility

The market for mobile applications is still growing exponentially, but this does not mean that the people that use these applications are evolving at the same rate. This is definitely not the case for elderly users. A systematic mapping of accessibility in applications for elderly people found that for many applications, accessibility is still an issue and many elderly are unable to use certain applications due to some kind accessibility barrier (Díaz-Bossini and Moreno, 2014). The top 3 accessibility barriers that were found most often are the use of a small font size, difficulty with touching the screen and difficulties with accessing menus.

Research into elderly accessibility for web design by Panayiotis, Z. et al. (2005) has resulted in 38 accessibility guidelines. Taking inspiration from these web-based guidelines, Díaz-Bossini and Moreno (2014) defined 6 main categories that are valuable in a mobile context. These categories are Target design, Use of Graphics, Browser Window Features, Content Layout Design, User Cognitive Design, Use of Color and Background. These beforementioned categories will be implemented into the "Welzijn.AI" example application and tested by means of a questionnaire, this will be expanded upon in the method section.

Usefulness

Adherence to a conversational agent, an e-health application, or any application for that matter is, among other things, largely correlated to its perceived usefulness. Davis (1989) suggests that there are two specific variables contributing to user acceptance of a system. These variables are the perceived usefulness and perceived ease of use. Ease of use is largely correlated to the accessibility of the application and its conversational agent. The Technology Acceptance Model as defined by Davis suggests that a person will only use an IT system if it will enhance job performance. Adjusting this metric to a mobile standard, this means that an elderly user will only willingly use the application if its perceived usefulness supersedes the usefulness of currently existing alternatives. Such as other applications or physical doctor appointments. This is an extremely important metric to test.

Satisfaction

The satisfaction of elderly people, or people in general, when using an e-health application is very important, since this has a causal relationship with the adherence of an application (Atreja et al., 2005). User satisfaction is regarded as one of the most important measures for the success of an information system. A common rationale is that more usage of an information system results in higher success of the system. This is only the case when the envisioned goal of the issuing party of the system aligns with the wishes of the end users (Gelderman, 1998). User satisfaction is defined as "extent to which users believe the information system available to them meets their information requirements" (Ives et al., 1983). As found by Gelderman (1998) the relation between user satisfaction and application performance was significant.

Intention to use

One of the important factors of the “Welzijn.AI” application is whether or not the elderly are intending to actually use the application. Previous research into the acceptance of a companion robot for elderly people by Heerink et. al (2008) has shown that higher enjoyment levels lead to a higher acceptance score of an application. As defined by the TAM model constructed by Davis (1986), the perceived ease of use and usefulness determine the behavioral intent to use a system. Heerink et. al (2008) confirmed an earlier hypothesized statement by Davis, that the intention to use a system leads to the actual use of a system.

Trust
Trust is an important factor in applications and systems in general. A study by Meng et al. (2022) shows that a higher trust in a system boosts the adherence to that said system. Meaning that if elderly users have higher trust in a system, they are more likely to use this system. Adding to this, Lee and Song (2013) have shown that trust directly correlates with the intention to use, which in turn has large correlation with the actual use.

3. Research design

3.1 Goal

This paper focusses on finding out if elderly patients wish to use an e-health application and specifically: which elements of a conversational AI are most important to elderly patients to make use of the ‘Welzijn.AI’. This study focusses on a group of elderly that are in a so called ‘huiskamergroep’, translated from Dutch this means ‘living room group’, and members of an elderly sporting club. These are elderly who seek social interaction with other elderly so they don’t get too lonely, seeing as loneliness is a large part in elderly people’s lives, this has a large effect on their mental state. The welzijn.AI application aims to map. All study participants are >67 years of age and will be surveyed first and then interviewed separately, so choices are not influenced by others. At the start of each interview, I will provide the scope of the project: that I am trying to figure out if they are more likely to use an application with a conversational agent and if that is the case, which defining features such a conversational agent should have. These results will be obtained using a card sorting method, which will be elaborated upon further in the card sorting interview section of this paper. The answer to the first question will give insight if patients actually want to use an e-health application. Once this is established, the next step is to figure out which defining features are most important to elderly participants for a conversation with a conversational agent.

3.2 Method

This study is divided into two parts. First some general questions in the form of a survey, then an interview guided by card sorting. An overview of the steps followed is listed below:

1. Elicit possible characteristics of a conversational agent, this is done in the previous chapter 2.4.1.
2. Start interviews:
    1. Show elderly participant a sample of a conversation in the Welzijn.ai application
    2. Conduct survey according to a sample conversation to find out state of application according to 6 different variables. These variables are mostly tested according to a 5-point Likert scale.
    3. Find out which characteristics are most important in a conversational agent for elderly participants. This is done using a card sorting method, elaborated further upon in 3.4.
3. Draw conclusion from gathered results.

The first step of the interview is showing elderly a sample conversation, the image below shows how this conversation looked. The full script of this conversation and the incorporated human social characteristics can be found in the appendix.

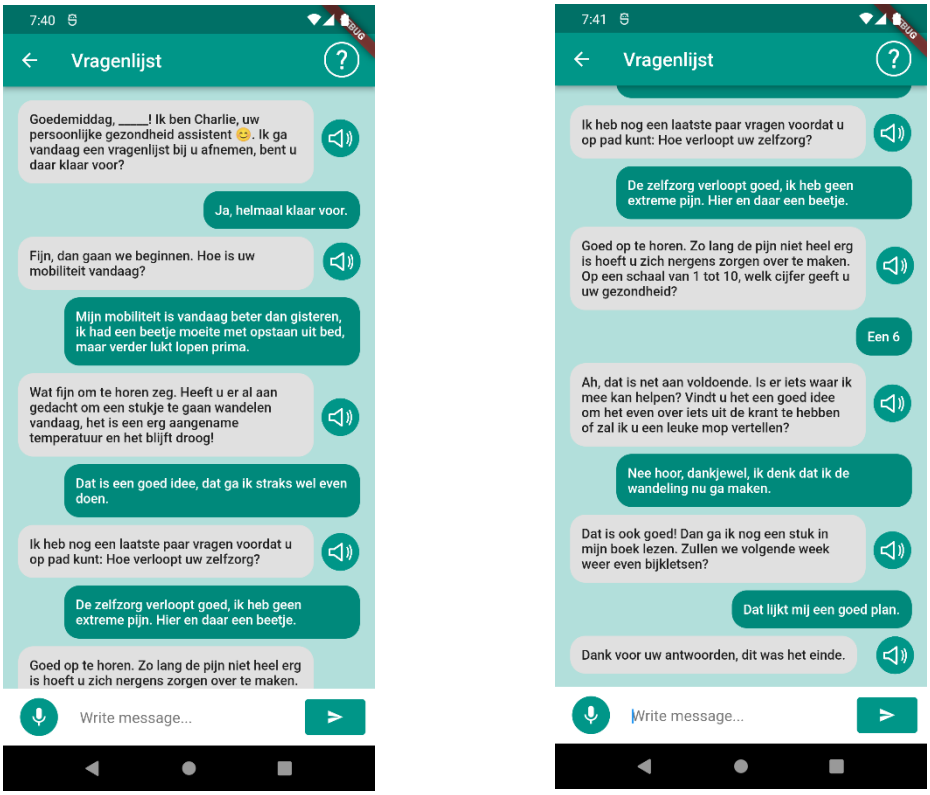

Figure 1: example of the presented application to elderly participants

Like mentioned above, a distinction is made between the variable testing part of the interview and the card sorting part. Both methods are described below.

3.3 Variable testing

As first part of the study a survey will be performed, based upon various questionnaires defined by other researchers. The previous chapter 2.3.1 suggests 6 different variables for the application to be tested on by the elderly participants. These variables are anthropomorphism, accessibility, usefulness, satisfaction, intent to use and trust. To test the current state of the app by using the 6 variables, a survey will be performed. This survey consists of 6 categories, one for each variable, and 31 questions in total. The survey can be found in the appendix. To ensure the correct demographic is being tested, as defined in chapter 2.2, introductory questions in regards to age and gender are asked. How each of the variables are tested is described below:

Anthropomorphism
Anthropomorphism as explained in the previous chapter 2.4, is the addition of human like features to something that is not alive. The point of the anthropomorphism part of the survey is to see if participants perceive a conversational agent as human-like or totally machine-like. The first part of the survey consists of 4 questions asking if the conversation in the application is perceived as: fake – real, machine-like – human-like, not conscious – conscious and artificial – natural. This questionnaire is called the Godspeed I: Anthropomorphism questionnaire, defined by Bartneck et al. (2008), and

tests human robot interaction. This questionnaire is a semantic differential scale, which measures people's perception of a certain concept (Summers, 1970).
Normally this questionnaire consists of 5 questions, but since the last question was geared towards the movement of a physical or on-screen robot, the choice was made to remove this question as there are no moving items in the application.

Accessibility
The study conducted by Oliveira et al. (2018b) found a list of accessibility issues in applications affecting elderly users. Adding to that, Díaz-Bossini and Moreno (2014), defined a list of accessibility features which developers should take into account when creating an application for the elderly. The applicable features for the "Welzijn.AI" application were used as part of the questionnaire and evaluated on a 5-point Likert scale. As opposed to the semantic differential scale, the Likert scale tests to what extent participants agree with statements (Joshi et al., 2015). In this case ranging from strongly agree to strongly disagree. Statements about the following topics, as suggested by Oliveira et al., are incorporated into the questionnaire: font size, help options, symbol identification, contrast and complexity of language. This part of the questionnaire consists of 8 questions. The entire questionnaire can be found in the appendix.

Usefulness
The perceived usefulness is tested by the system usability scale (Brooke, 1996). This questionnaire can be adjusted to test usefulness for mobile applications. This scale is evaluated on the 5-point Likert scale. Since the application was not actually in use, some of the questions in the standard SUS scale could not be used. So a selection of 8 questions from the scale were used in the proposed questionnaire. The assumption is made that removing these questions adds to the accuracy of the scale and does not lessen the validity. Questions were asked about the complexity of the system, ease of use and perceived usefulness. All questions can be found in the appendix.

Satisfaction
To measure satisfaction of participants for the application the QUIS (Stanton et al.) is used. This stands for 'Questionnaire for User Interface Satisfaction' and measures the satisfaction of an interface. This part of the survey consists of 6 questions on a semantic differential scale, asking if the system is: terrible – wonderful, easy – hard, frustrating – satisfying, satisfactory – not satisfactory, boring – fun and applicable – not applicable.

Intention to use
Since intention to use is largely correlated to perceived usefulness (Atreja et al., 2005), part of the usefulness questionnaire can be used. The first question in the SUS scale is if a person would use this system frequently, so this question was reused in this questionnaire. Participants were also asked if they would only use the system per doctor's recommendations, to see if there is a distinction between voluntary usage and prescribed usage. This part of the questionnaire consisted of 4 questions and can be found in the appendix.

Trust
To measure trust in the application a questionnaire was created based on the findings of McKnight et al. (2011), where labels are suggested for testing the trust of people in certain technologies. Questions were formed for the labels: trust in a specific technology, trust in technology in general and trust in a systems functionality, helpfulness and reliability. In total 8 questions were asked, these questions can be found in the appendix.

3.4 Card sorting

The questionnaire serves the purpose to find out how the current application is perceived and if elderly are intending on using such an application. The next step is carrying out an interview that is led by card sorting. Card sorting is a technique where different participants have to order certain content (Olmsted-Hawala, 2006) that is presented to them. In this case cards with all characteristics are presented to the participants, and they are in turn asked to order these. The cards will contain all 7 defined characteristics by Fong et al. (2003) and Heerink et al. (2008) as described in chapter 2.4.1. with an example of the characteristic in use. This will give the participants a clearer view of how a characteristic should be perceived.
Heerink et al. (2008) suggests combining certain social abilities so these can be tested. For example being attentive and polite (using abilities 1 and 5), or by using emoticons or a moving character on screen (using abilities 1, 5, 6). But this does not give a clear enough image of which characteristic is more important, only with a really large sample size and many examples of characteristic combinations. So all 7 characteristics will be tested separately.

To rank the social characteristics 7 different cards are created to display the characteristics. Elderly participants are asked to order these cards on perceived importance for a conversation. The general content of the cards contain the characteristics and an example of the characteristic, the following characteristics are tested:

1. Express and/or perceive emotions: the conversational agent will respond in an empathetic manner when presented with negative emotions and happy when presented with positive emotions.
2. Establish/maintain social relationships: an example here is that the conversational agent introduces itself and greets the patient like they already know each other.
3. Use natural cues (gaze, gestures, etc.): natural cues are difficult to show. Usually this is done by means of animation, but this is not an implemented feature. In this case the natural cues of a conversational agent are shown by means of emoticon.
4. Exhibit distinctive personality and character: the conversational agent introduces itself and tells the users something about itself.
5. May learn/develop social competencies: it shows that the conversational agent has developed social competencies, for example by talking about the weather.
6. Trust: Build a relation of trust with the user
7. Responsibility: admit mistakes and ensure integrity to the users. For example by being transparent and telling users how data is processed and stored.

The cards with the corresponding explanations can be found in the appendix.

4. Results

Results from this study consist of a general survey to find out what the current state of the application is, by means of variable testing and a ranking of a conversational agents' social characteristics. The intention of the survey was for elderly to read an example conversation and fill in a survey afterwards. However, this proved to be easier said than done. After conducting one test interview, it became clear that questions asked about information technology without further explanation are too difficult for elderly to understand. It was found that elderly understand the basic premise of the application, but cannot wrap their head around a "talking robot". Many terms used in the standard questionnaires for usefulness or satisfaction were too difficult to understand, reason for this is the often lower education levels for elderly people. This is also supported by the research of Joung et al. (2000), who predicted that only 18% of elderly aged 80-84 in the Netherlands in 2020 had received higher education. Many of the interviewees told us that they think their lower

education levels has impact on how well they can understand these 'new' applications. The majority of interviewed participants told us they only attended lower education or went to a school of trade.

This meant that all interviews, instead of only interviewing the elderly to rank characteristics, also consisted of walking through the questionnaire and explaining each question. The added benefit of walking through the entire questionnaire is that you not only get a datapoint, but you also hear the reasoning behind their choices. So besides just being able to say something about the perception of certain variables, suggestions as to improve upon these variables are gathered as well. Besides the survey terminology being too difficult, the characteristics cards were too difficult to understand as well, therefore these were adjusted accordingly. The characteristics were translated into simpler terms so they could be understood better. For example, the characteristic 'developing social competencies' became 'being able to act on social needs'. All characteristics and their corresponding ranking cards be found in the appendix. In total about 50 elderly were approached to partake in the study, more than half declined as they were too busy for something that they are not interested in. They were largely not interested in the study because they did not understand the application or never intend to ever use one.

This study was conducted on a total sample size of $N = 20$ elderly participants. Consisting of 12 female participants and 8 male participants. Ages ranged from a minimum age of 67 years old to a maximum age of 94 years old. The average age of all participants is $83{,}2$ ($\sigma = 8{,}1$) years old with a median age of 85.

4.1 Survey

The core functionality of this application was tested with a questionnaire divided into six parts: anthropomorphism, accessibility, usefulness, satisfaction, intention to use and trust. These variables are tested on 5-point Likert scales. All parts of the questionnaire are listed below. Here the Cronbach's alpha (α) of each partial questionnaire is given. This alpha tells us the internal consistency of the proposed questions (Sijtsma, 2009), thus telling us if the asked questions can be used to form a conclusion. In most situations an α of at least 0.7 is considered an acceptable internal consistency for a questionnaire. An alpha $0.6 < \alpha < 0.7$ is questionable and an $\alpha < 0.6$ means that there is poor internal consistency (Gliem & Gliem, 2003).
Besides the Cronbach's alpha, the score, standard deviation and mode are discussed as well. For each of the different variables suggestions that came directly from elderly participants, if applicable, will be outlined as well. From the 20 filled in questionnaires, all were used. A questionnaire would not be used when it was not filled in properly or the participant was not aged above 65.

Anthropomorphism
Internal consistency for this part of the questionnaire was calculated to be $\alpha = 0{,}92$. Meaning that the questionnaire asked closely related questions. This results in the scale being reliable. As for the results: 4 questions were asked in this part of the questionnaire, asking if the conversation seen is human- or machinelike. Rated on a scale from 1 to 5, meaning that if a 5 is scored the conversation is perceived as being human and if a 1 is scored the conversation is perceived as machinelike. From all completed questionnaires 53,75% scored the conversation more human like than machine like. The rest was either neutral or thought that the conversation was machine-like. The standard deviation for the answered questions was fairly high at $\sigma = 1{,}4$. This means that the data is not at all centered around the mean, and has a lot of variation. This was also picked up in multiple interviews, some elderly thought the conversation was very human like, but others thought it was way too uptight and did not resemble a human at all.

Accessibility

The accessibility part of the questionnaire consisted of 8 questions and resulted in a Cronbach's alpha of $\alpha = 0{,}78$. In this case the internal consistency is good. From the 8 questions 62,50% agreed or strongly agreed with accessibility statements. The standard deviation of 1.5 was very high. So just like before, on average the accessibility of the application is up to standard, but results vary a lot. Seeing as accessibility is one of the most important aspects of an elderly application (Díaz-Bossini and Moreno, 2014), it would be preferred to have more consistent results. Especially notable was that most questions regarding the icons displayed on screen were scored very poorly, with the mode score from these questions being a 1. If the icons are improved, the accessibility of the application would increase. All other accessibility questions regarding font size, contrast or complexity of language had a mode score of 4 or 5.
Another part of the Welzijn.AI application is the talking interface, which allows users to use speech-to-text capabilities to send a message. The accessibility of this function was only tested in the form of a question whether or not elderly were able to locate the microphone button. Other tests regarding the usage of the talking interface could not be conducted because it was not a live demonstration of the application.

Satisfaction
The third part of the questionnaire consisted of 5 questions regarding the perceived satisfaction of the application. With $\alpha = 0.89$, this questionnaire is internally consistent. 42,00% of participants were satisfied to some extent with the application, scoring 4 and 5. With the standard deviation being $\sigma = 1.3$. Same as before, the overall average score of satisfaction is alright but with a standard deviation this high that does not say a lot. The mode for all questions overall was a 3, with an exception for the second question where it was 5. This question asked if the application seemed easy- or hard to use, so elderly people think that the application seems easy to use. According to the other questions regarding satisfaction, elderly think that the application is average.

Usefulness
The 8 questions regarding the usefulness of the application had an internal consistency of $\alpha = 0.63$, meaning that the consistency is questionable. But seeing as the questionnaire is based on the system usefulness scale, the decision was made to still use the results of the questionnaire. The distribution derived from the 5-point scale showed that 32,50% agreed or strongly agreed with the usefulness statements. A standard deviation of $\sigma = 1.3$ was found. With the high standard deviation a definitive conclusion about the usefulness of the application cannot be drawn. Interesting to see however was that the mode of the score for questions 3, 5, and 8 were all 4. Meaning that elderly strongly agree with the position that they need help from a technical person to understand the application, that there is added benefit to having an e-health application for administering questionnaires and using the system frequently. Elderly do think that the application is complex and will need time to get used to and understand all the features, according to the mode of the scores given to the other questions being either a 1 or a 2.

Intention to use
The intention to use part of the questionnaire scores an internal consistency of $\alpha = 0.62$, which is debatable. 23,75% of the scores were a 4 or 5, meaning that the majority of the scores were 1, 2 or 3. A standard deviation of $\sigma = 1.2$ was found. This score is the lowest so far, and with a standard deviation this high again. Nothing regarding the intention to use can be said with certainty. Something that did become clear was that elderly gravitate to only wanting to use the application when the doctor has prescribed its use, the mode for the scores for this question was 4. The mode of the question if elderly would use the system in general was also a 4, so elderly do think they would use the application.

Trust

As last part of the questionnaire trust was measured. An internal consistency of $\alpha = 0.70$ was measured for this questionnaire, meaning it is on the edge of acceptable and questionable. But since this questionnaire is based on research by McKnight et al. using some of their example questions with minor changes, the decision to use the results and not revise the questions was made. 54,29% of the answers on the 5-point Likert scale were a 4 or 5, meaning that more than half of the answers were positive. A standard deviation of $\sigma = 1.1$ was scored. Modes of all the scores for the questions were 3 or 4 for all questions, so trust was scored high.

The answered questionnaires give insight into the application to an extent. Seeing as the average standard deviation for the entire questionnaire is $\sigma = 1.1$ on a scale of 5, this means that the data is very distributed. The different parts of the questionnaires did give insight which aspects definitely need improving and which aspects are already fairly well implemented.

4.2 Card sorting

The second part of the research focused on sorting 7 defining characteristics for a conversational agent. The initial idea was to define an order for these characteristics in each of 3 mental states: a positive mental state, a neutral mental state and a negative mental state. Whilst conducting the research however, it became clear that elderly participants are not good at imagining themselves to be in a different mental state than the state that they are currently in. In every interview the question was asked "If you imagine yourself to be in a different mental state, let's say a negative state, would you change the order of the characteristics?". Only 2 out of the 20 participants changed the order of the importance of characteristics. Due to the small sample size, the decision was made to only focus on the importance of characteristics in general, not in 3 different mental states. Since 2 entries are not enough datapoints for a conclusion as to the importance of characteristics in different mental states. The table below shows the aggregated picks for each characteristic at each rank.

Characteristics	**Rank (Combined ages)**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1. Express /perceive emotions	7	3	2	2	4	2	0
2. Maintaining social relationships	2	2	3	4	4	4	0
3. Natural signals	0	1	6	1	2	4	6
4. Distinctive character	5	2	3	4	1	2	3
5. Social competencies	1	6	2	3	3	2	4
6. Trust	4	4	2	2	4	1	3
7. Responsibility	1	2	2	4	2	5	4

Table 1: Amount of times that a certain characteristic is preferred at a certain rank

From the first 8 columns of this table it does not necessarily become clear which characteristic is most important, ordering the characteristics by amount of times picked per rank is not the most representative way to rank the characteristics. Assuming maximum amount picked as the ranking criterion the order for characteristics would be as follows: 1 – 5 – 2 – 4 – 6 – 7 - 3.
But this is not the most balanced way to conclude the order of characteristics. Take for example the *natural signals* characteristic, based on the maximum amount picked, one would say that this characteristic would be ranked at number 3. But looking at the rest of the data, it becomes clear that this representation is skewed. From all participants, half ranked the *natural signals* characteristic in the top 5 ranks, while the other half ranked them in the last two ranks, 6 and 7. Meaning that the pick-density is the same for the two increments.

A better way to define an order for the characteristics is looking at the aggregate score of the characteristics, this is what the last column in the table shows. This is calculating by taking the number of picks for a rank and multiplying this by a weight, in this case weight 7 for rank 1 and weight 1 for rank 7. Plotted in a graph, all the data still looks very messy, as can be seen in graph 1 on the next page. But by adding a scoring function the data becomes better interpretable in table format, these scores can be found in table 2.

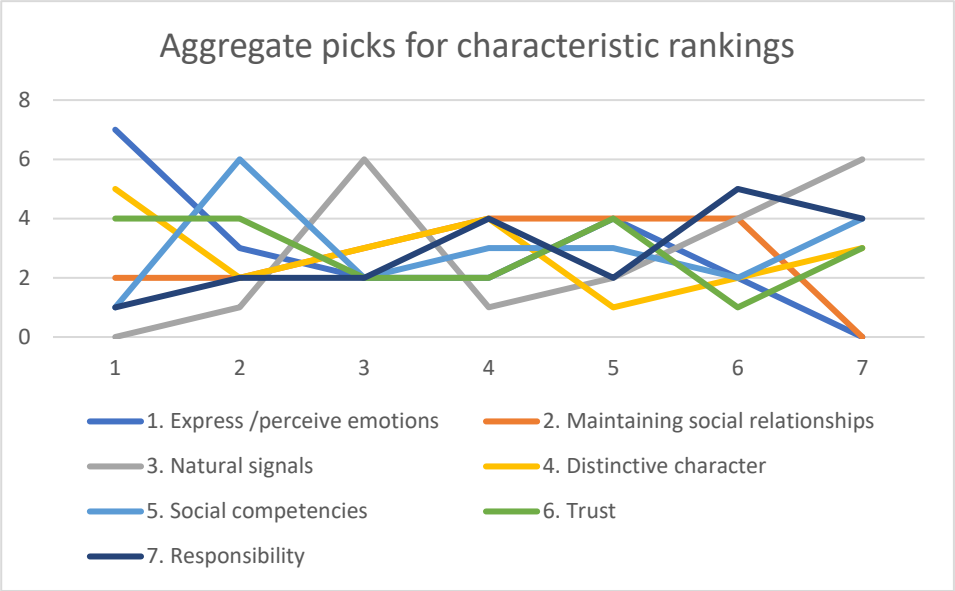

Figure 2: Aggregate picks for each characteristic at each rank

Characteristic	Score
1	101
2	77
3	60
4	88
5	82
6	87
7	65

Table 2: Aggregated scores of characteristics rankings

Converting the data of all participants into scores accounting for the weight of each rank results in a better representation of the importance of the characteristics. When using the aggregate score of the rank for each characteristic, the following order of importance can be read: 1 – 4 – 6 – 5 – 2 – 7 – 3. The highest score is attributed to *express/perceive* emotions, meaning this is – on average – the most important characteristic for the elderly participating in this study. The lowest score is that of the characteristic *natural signals*, meaning this is the least important characteristic. Results from ranking based on the aggregate score a contrary to the maximum number of picks made for a different result, but one that is more representative for all participants.
Some scores were not separated by large margins, meaning that it can not with certainty be said that this is the best order for the characteristics. This would require more extensive research.

4.2.1 Two age groups

When conducting the interviews, it became clear that every participant had different needs, while in an ideal world one could see a clear distribution of the characteristics over the ranks. But this was not the case, so an aggregate score had to be used. But seeing as the age increment is quite large, ranging from 67 to 94, it might be that a clearer distribution will become visible if this increment were smaller. The difference between some of the scores of the characteristic are quite small, meaning that the preference for these characteristics are similar. A distinction can be made but is not clear cut.
So in addition to the order of characteristics for all interviewed ages, two separate increments were created with about equal participants to see if difference in age had effect on the order of the characteristics. Just like for all ages, the aggregate score for the characteristic rankings had to be calculated as order preference for participants varied. The table below shows the aggregate scores in relation to all age increments and its corresponding characteristics.

Characteristic	Scores per increment		
	Age 67 - 81	**Age 82 - 94**	**Combined ages**
1. Express /perceive emotions	38	70	101
2. Maintaining social relationships	29	54	77
3. Natural signals	21	41	60
4. Distinctive character	43	50	88
5. Social competencies	29	56	82
6. Trust	29	59	87
7. Responsibility	35	34	65

Table 3: Scores for all characteristics at different age increments

From these scores the order for the characteristics can be determined for the separate age increments. The table below shows an overview of the age increments alongside the order for this increment and the amount of participants.

Age increment	Order	N
67 - 81 (μ = 75,9 (σ = 4,9))	4 - 1 - 7 - 2 / 5 / 6 - 3	8
82 - 94 (μ = 89,3 (σ = 3,0))	1 - 6 - 5 - 2 - 4 - 3 - 7	12
67 - 94 (μ = 84,0 (σ = 7,6))	1 - 4 - 6 - 5 - 2 - 7 - 3	20

Table 4: Derived order for the characteristics at different age increments

Interesting to see is that there is not a clearer distribution of the order of the characteristics. The differences between the scores are still minimal, same as for the combined ages increment. If anything, it is less obvious what is preferred, since 3 characteristics have the same score in the first age increment. The findings of this current experiment result in the order 4 - 1 - 7 - 2 / 5 / 6 - 3 for the first age increment, and the order 1 - 6 - 5 - 2 - 4 - 3 - 7 for the second age increment. Most notable is that 1, 2 and 3 have about equal positions, with 1 being ranked first at 1st or 2nd , 2 being ranked 4th and 3 being ranked last at either 6th or 7th. 7 is ranked 3rd in the first segment but ranked 7th in the second segment, meaning that elderly between 67 and 81 prefer a responsible conversational agent more than elderly in the 82 – 94 segment.

From the results above the conclusion can be drawn that characteristic 1 is most important, characteristic 3 least important and characteristic 2 has middle importance. The remaining characteristics differ in order for the two separate age increments, so it cannot be said with certainty if one characteristic is more important than the other. Combining the two age increments leads to the most complete representation of the characteristic ranking at this point in time, but has to be confirmed by performing an extra experiment with a larger sample size.

4.3 Suggestions for the application

Following the interviews containing the survey and card sorting, I inquired with the elderly participants if they had suggestions for the improvement of the application. A few suggestions were named:

- **Use better defined icons:** as was concluded by the questionnaire, the icons and what they meant were not clear enough. These icons really need to stand out and have more defining features as to what they mean. Just explaining them in the introduction of an application is not enough, since elderly have the tendency to forget things (Imhof et al., 2006). To battle

this forgetfulness, the application should reduce its complexity, meaning that the icons should be intuitive or labeled.

- **If natural cues are used, explain what they mean:** for many people the use of natural cues by information technology, such as animations, leads to a decline in accessibility (Díaz-Bossini and Moreno, 2014). So if a natural cues are used in this application in the form of emoticons, there should at least be a screen that shows what they mean. As the question was asked multiple times what the meaning of the natural cue (emoji in the text) was. Seeing that natural cues ranked last in the study, it is recommended that natural cues are not used.
- **Interactive tutorial to display the use of the application:** seeing as elderly are more prone to forgetfulness (Imhof et al., 2006) and have a harder time using modern technology than younger generations (Eurostat, 2017). This means that they might not be able to use an application, to its fullest capabilities that is. A big help here would be an interactive tutorial to explain what every function of the application does. This tutorial needs to be accessible from every screen and easy to spot.
- **Add contrast options:** many elderly are visually impaired. During this research this did not pose a problem. But the application should have an option to change to different contrast settings that can accommodate most visual problems that do not limit reading entirely.
- **Make sure the "Ask for help" button stands out:** as is concluded by the survey, the icons are not clear in their function and can sometimes be hard to find. Especially the question if elderly can find the 'ask for help' button had a low score. Since this is the button that is probably going to be pressed a lot, it needs to stand out and not disappear in the design of the application. When the button was pointed out, multiple elderly said the button did not stand out enough. So in the next version of the application, the button needs a bold color or better positioned.
- **Explain to users what can be sent to the application:** this ties in somewhat to the suggestion of adding an interactive tutorial. Elderly said that they do not know what they can and cannot say to the application, because they are not sitting of a real person. The goal of the application needs to be stated clearly and in simple terms, in other words it needs to be accessible.

5. Discussion

In this exploratory research the goal was to find if elderly people would use the e-health application Welzijn.AI and what the defining human social characteristics of the conversational agent in the application should be. During the course of this research certain limitations were discovered. What these limitations are and to what extent these impacted the research will be discussed in chapter 5.1. After discussing the limitations of the study, chapter 5.2 will focus on the found results and answers the two proposed sub-questions.

5.1 Limitations

This study encountered various limitations, which are best described when they are divided into two categories: participant selection and data.

Participant selection and interaction proved to be a difficult task for this research. Defining the requirements for the participants was easy enough, they only had to check one box, being an elderly citizen. But actually getting elderly to participate and finding a large enough group was more difficult. Participants were sourced from my own social network, leading to two groups of elderly, but finding other groups willing to participate led to no results. Even in the two groups who agreed to participate in the study, many elderly opted out from the study. Saying they do not have the attention span, time or interest to participate. The elderly that did partake varied a lot in their understanding of

technology in general. Some of the interviewed elderly did not know what an application, emoji or icon was. It meant that to get reliable results from the questionnaire, every question had to be explained in its intention and meaning. Leading to a 30 minute interview per participant to explain the questionnaire, help filling it in and sorting the characteristics of a conversational agent. Doing so the research was more time consuming than first anticipated. Having more resources to source elderly and conduct interviews could lead to more extensive results.

Data was also a limitation is this study. There was only a sample size of 20 elderly participating in the study, meaning that the results cannot be averaged out to resemble the elderly demographic as a whole. For a follow up study a larger sample size should be inquired. In addition to the small sample size, the results acquired from the elderly varied a lot. Meaning that all elderly had a different perception of the application and all had different preferences when it comes to the importance of characteristics in a conversational agent. Unfortunately there is not a conclusive answer as to the state of the application currently and the importance of characteristics in a conversational agent.

Another partial limitation of the study was the question list itself. All scales used were based upon already existing questionnaires or resembled them closely. But for some questionnaires certain questions, that did not correlate with the research, had to be removed. Removing these questions made it that the total score of a questionnaire part was more reliable, seeing as aspects that did not influence the research were not taken into account.

As to attest to the representativeness of the study participants: there are no exact numbers, but many interviewees expressed that they only had one mood and that was a bad one, caused by pain or inability to do something (walk, read or stand up for example). This refers back to the first intention of the research, wanting to study preference in 3 different moods. Going from the conversations with the participants, it seemed that many elderly were lonely and some slightly depressed. But this was never scientifically established, this was only my interpretation.

5.2 Sub-questions

The first proposed sub-question is “What is the perception of the Welzijn.AI application for elderly users, and do they intend to use this system?”. To answer this question a questionnaire was conducted based on 6 defined variables: anthropomorphism, accessibility, usefulness, satisfaction, intent to use and trust. Normally these variables say a lot as to the perception of an application, but in this case not so much. Due to the small sample size and varying answers to the questionnaire no definitive conclusion can be formed to speak to the current state of the application. On the one hand elderly thought the application was accessible, had a clear use and were satisfied with the proposed interaction with the application. But on the other hand, some elderly thought the exact opposite. While one cannot with certainty say what the exact state of the application is, one thing that can be said with certainty is that the application has to be improved. Since this is an e-health application it needs to be easily used by everyone, and that is clearly not how it currently is. This is especially important since this application is geared towards elderly users. This generation already has a harder time using technology when compared to a younger demographic, so if the system is also lacking refinement or functionality in certain categories, this will probably lead them to be non-adherent to the application.

The second proposed sub-question is: “What human social characteristics are most and least important in a conversational agent to elderly people?”. A card sorting method was used to try and answer this question. Since the rankings provided by the elderly were far from unanimous, the method to select the characteristic importance on count was dismissed. To order the characteristics, a trend in the order of the characteristics picked by the elderly were used. Using this trend it was

found that for the age segment 67 – 94 the order of 1 - 4 - 6 - 2 - 5 - 7 - 3 was preferred, where these numbers reflect the following characteristics: 1. Expression or perception of emotions, 2. Maintaining social relationships, 3. Using natural signals, 4. Displaying distinctive character, 5. Develop social competencies, 6. Trust and 7. Responsibility. Some of the trends were very similar however, so it cannot be said with certainty that this is the preferred order, this can only be tested with a larger sample size. To attain more distinct trends in the rankings of the characteristics, the age increment was split in two separate segments, 67 – 81 and 82 – 94. Here more distinct trends were found in the order of ranking the characteristics, but the results are still not unanimous. The respective orders are 4 - 1 - 7 - 2 - 6 - 5 - 3 for the first segment and 1 - 4 - 6 - 2 - 5 - 7 - 3 for the second segment. Further testing is required to confirm this order. What did become clear was that elderly find the expression or perception of emotions, together with a distinctive character, very important. The elderly found the use of natural signals the least important.

6. Conclusion

This research aimed to map the perception of the Welzijn.AI application and suggest improvements in the form of a human social characteristic importance in conversational agents. The research question of this paper was “With which characteristics and to what extent can a conversational agent be used to improve adherence in e-health applications such as Welzijn.AI for the elderly?”. The answer to the research question is that following this research it was concluded that for the age group 67 – 94 the most important characteristics were the expression or perception of emotions ranked first, displaying a distinctive character second, trust third, maintaining social relationships fourth, the development of social competencies fifth, responsibility sixth and the use of natural cues last. Adherence was tested to an extent but could not with certainty be determined, since the perception of the application was different to many elderly. To more accurately test this and confirm the findings for the characteristic ranking, additional research with a larger sample size should be conducted.

This question was answered by conducting interviews with elderly patients during which a questionnaire was completed and human social characteristics of a conversational agent were ranked on importance. Additionally, improvement suggestions from the elderly were outlined as well.

The questionnaire that was presented in an interview with elderly participants delivered inconclusive results. So it cannot be said with certainty what the current state of the application using the 6 variables anthropomorphism, accessibility, usefulness, satisfaction, intent to use and trust. The perception of the application varied a lot, this can be seen as the average standard deviation of the questionnaire was $\sigma = 1.3$ on a scale of 5. Seeing as the application is for elderly users, these numbers are not acceptable. Elderly already have a hard time using technology, and if the application they are using does not work as expected they will probably end up being non-adherent to it.

Ranking the characteristics for the implementation of a conversational agent was done with aggregated scores but was not unanimous, even for two different age groups. The eventual suggested importance of characteristics following the gathered results is: the expression or perception of emotions ranked first, displaying a distinctive character second, trust third, maintaining social relationships fourth, the development of social competencies fifth, responsibility sixth and the use of natural cues last.

It became clear from the research that every elderly has different needs and preferences, the goal is for the application to provide everything necessary to satisfy these needs and fulfill said preferences.

This can only be done by talking to the actual elderly users and gathering data as to what they think should be improved or added. During the conducted interviews, after the questionnaire and card sorting, the elderly outlined several suggestions that should be implemented in future versions of the application. These suggestions were: using better defined icons, the addition of an interactive tutorial that can be accessed from any page, improve upon contrast options and making the application accessible by explaining what can be said to it.

Further research should focus on the actual implementation of a conversational agent and the improvement suggestions that have been made. The same questions asked in the questionnaire should be posed again, but this time to a larger sample size, to find if the application has indeed improved in its perception. After this, the actual adherence to the application can be tested. Testing of adherence should be done with two applications, one with a conversational agent and one without. Results in adherence should be measured over increments of time instead of the variable *intent to use*. Another interesting study would be to make the *anthropomorphism* variable exactly measurable instead of just testing the perception of it. For example by conducting a Turing test, to test the actual perceived *anthropomorphism* in the system. To improve upon this application further, the suggested implementations should be incorporated. After which usability, accessibility and satisfiability need to be studied again to re-evaluate and cater the application to an elderly person's needs.

7. References

Anderson, R. E., & Srinivasan, S. S. (2003). E-satisfaction and e-loyalty: A contingency framework. *Psychology and Marketing*, *20*(2), 123–138. https://doi.org/10.1002/mar.10063

Bagherian, H., Shahbazi, M., Sattari, M., & Saghaeiannejad-Isfahani, S. (2021). The opportunities and challenges of using mobile health in elderly self-care. *Journal of Education and Health Promotion*, *10*(1), 80. https://doi.org/10.4103/jehp.jehp_871_20

Bagherian, H., Shahbazi, M., Sattari, M., & Saghaeiannejad-Isfahani, S. (2021b). The opportunities and challenges of using mobile health in elderly self-care. *Journal of Education and Health Promotion*, *10*(1), 80. https://doi.org/10.4103/jehp.jehp_871_20

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. https://doi.org/10.1007/s12369-008-0001-3

Brooke, J. (1996). Sus-a Quick and Dirty Usability Scale. Usability Eval. industry 189 (194), 4–7

Campen, C., Ross, J. A., & Van Campen, C. (2011). Kwetsbare ouderen. Sociaal en Cultureel Planbureau. http://docplayer.nl/2498379-Kwetsbare-ouderen-cretien-van-campen-red.html

Dabney, M. H., Dean, B. C., & Rogers, T. (2013). No sensor left behind. *Proceeding of the 44th ACM technical symposium on Computer science education - SIGCSE '13*. https://doi.org/10.1145/2445196.2445378

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

Díaz-Bossini, J. M., & Moreno, L. (2014). Accessibility to Mobile Interfaces for Older People. *Procedia Computer Science*, *27*, 57–66. https://doi.org/10.1016/j.procs.2014.02.008

Ebert, D. D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., Brugnera, A., Øverland, S., Trebbi, G., Jensen, K. L., Kaehlke, F., & Baumeister, H. (2018). Internet- and Mobile-Based Psychological Interventions: Applications, Efficacy, and Potential for Improving Mental Health. *European Psychologist*, *23*(2), 167–187. https://doi.org/10.1027/1016-9040/a000318

Edwards, N. (2002). Unhappy doctors: what are the causes and what can be done? *BMJ*, *324*(7341), 835–838. https://doi.org/10.1136/bmj.324.7341.835

Elbert, N. J., Van Os-Medendorp, H., Van Renselaar, W., Ekeland, A. G., Hakkaart-van Roijen, L., Raat, H., Nijsten, T. E., & Pasmans, S. G. (2014). Effectiveness and Cost-Effectiveness of eHealth Interventions in Somatic Diseases: A Systematic Review of Systematic Reviews and Meta-Analyses. *Journal of Medical Internet Research*, *16*(4), e110. https://doi.org/10.2196/jmir.2790

Eurostat. (2017). *Eurostat: Digital economy and society*. https://ec.europa.eu/eurostat/web/digital-economy-and-society/data/database

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3–4), 143–166. https://doi.org/10.1016/s0921-8890(02)00372-x

Gelderman, M. (1998). The relation between user satisfaction, usage of information systems and performance. *Information & Management*, *34*(1), 11–18. https://doi.org/10.1016/s0378-7206(98)00044-5

Gliem, J. A., & Gliem, R. R. (2003). Calculating, Interpreting, And Reporting Cronbach's Alpha Reliability Coefficient For Likert-Type Scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*. https://scholarworks.iupui.edu/handle/1805/344

GSMA and McKinsey & Company. (2012, march). *mHealth: A new vision for healthcare*. https://www.gsma.com/iot/wp-content/uploads/2012/03/gsmamckinseymhealthreport.pdf

Guthrie, S. E. (1997). Anthropomorphism: A definition and a theory. In R. W. Mitchell, N. S. Thompson, & H. L. Miles (Eds.), *Anthropomorphism, anecdotes, and animals* (pp. 50–58). State University of New York Press.

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2008). The influence of social presence on acceptance of a companion robot by older people. *Journal of Physical Agents (JoPha)*, *2*(2), 33–40. https://doi.org/10.14198/jopha.2008.2.2.05

Hussain, S., Ameri Sianaki, O., Ababneh, N. (2019). A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In: Barolli, L., Takizawa, M., Xhafa, F., Enokido, T. (eds) Web, Artificial Intelligence and Network Applications. WAINA 2019. Advances in Intelligent Systems and Computing, vol 927. Springer, Cham. https://doi.org/10.1007/978-3-030-15035-8_93

Imhof, L., Wallhagen, M. I., Mahrer-Imhof, R., & Monsch, A. U. (2006). Becoming Forgetful: How Elderly People Deal With Forgetfulness in Everyday Life. *American Journal of Alzheimer's Disease & Other Dementiasr*, *21*(5), 347–353. https://doi.org/10.1177/1533317506292499

Inkster, B., Sarda, S., & Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, *6*(11), e12106. https://doi.org/10.2196/12106

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975

Joung, I & Kunst, Anton & Imhoff, E & Mackenbach, Johan. (2000). Education, aging, and health: to what extent can the rise in educational level relieve the future health (care) burden associated with population aging in the Netherlands?. Journal of clinical epidemiology. 53. 955-63.

https://www.researchgate.net/publication/12319994_Education_aging_and_health_to_what_extent_can_the_rise_in_educational_level_relieve_the_future_health_care_burden_associated_with_population_aging_in_the_Netherlands

Kebapci, A., Ozkaynak, M., & Lareau, S. C. (2019). Effects of eHealth-Based Interventions on Adherence to Components of Cardiac Rehabilitation. *Journal of Cardiovascular Nursing*, *35*(1), 74–85. https://doi.org/10.1097/jcn.0000000000000619

Khatri, C., Venkatesh, A., Hedayatnia, B., Gabriel, R., Ram, A., & Prasad, R. (2018). Alexa Prize — State of the Art in Conversational AI. *AI Magazine*, *39*(3), 40–55. https://doi.org/10.1609/aimag.v39i3.2810

Lee, J. H., & Song, C. H. (2013). Effects of trust and perceived risk on user acceptance of a new technology service. *Social Behavior and Personality: an international journal*, *41*(4), 587–597. https://doi.org/10.2224/sbp.2013.41.4.587

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology. *ACM Transactions on Management Information Systems*, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

Meng, F., Guo, X., Peng, Z., Ye, Q. and Lai, K.-H. (2022), "Trust and elderly users' continuance intention regarding mobile health services: the contingent role of health and technology anxieties", *Information Technology & People*, Vol. 35 No. 1, pp. 259-280. https://doi.org/10.1108/ITP-11-2019-0602

Mickan, S., Tilson, J. K., Atherton, H., Roberts, N. W., & Heneghan, C. (2013). Evidence of Effectiveness of Health Care Professionals Using Handheld Computers: A Scoping Review of Systematic Reviews. *Journal of Medical Internet Research*, *15*(10), e212. https://doi.org/10.2196/jmir.2530

Ministerie van Algemene Zaken. (2016, 26 mei). *Benefits of eHealth*. eHealth (Telehealth) | Government.Nl. Geraadpleegd op 13 juli 2022, van https://www.government.nl/topics/ehealth/benefits-of-ehealth

Noar, S. M., & Harrington, N. G. (2012). eHealth applications. *EHEALTH APPLICATIONS*, 1.

Oliveira, C. D. D., Fioravanti, M. L., Fortes, R. P. D. M., & Barbosa, E. F. (2018b). Accessibility in mobile applications for elderly users: a systematic mapping. *2018 IEEE Frontiers in Education Conference (FIE)*. https://doi.org/10.1109/fie.2018.8659017

Olmsted-Hawala, E. (2006). Card Sorting, Information Architecture And Usability: Adding in Our Users' Perspective to Re-Design the Census Bureau Web Site. *2006 IEEE International Professional Communication Conference*. https://doi.org/10.1109/ipcc.2006.320360

Orimo, H., Ito, H., Suzuki, T., Araki, A., Hosoi, T., & Sawabe, M. (2006). Reviewing the definition of "elderly". *Geriatrics and Gerontology International*, *6*(3), 149–158. https://doi.org/10.1111/j.1447-0594.2006.00341.x

Rebala, G., Ravi, A., Churiwala, S. (2019). Machine Learning Definition and Basics. In: An Introduction to Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-15729-6_1

REITER, E., & DALE, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87. https://doi.org/10.1017/s1351324997001502

Rijksoverheid. (2018, april). *Rapport De juiste zorg op de juiste plek.* https://www.rijksoverheid.nl/documenten/rapporten/2018/04/06/rapport-de-juiste-zorg-op-de-juiste-plek

Rowland, S.P., Fitzgerald, J.E., Holme, T. *et al.* What is the clinical value of mHealth for patients?. *npj Digit. Med.* **3,** 4 (2020). https://doi.org/10.1038/s41746-019-0206-x

Semaan, P. (2012). Natural language generation: an overview. *J Comput Sci Res*, *1*(3), 50-57

Shruthi, R. (2016). A Study of Medication Compliance in Geriatric Patients with Chronic Illnesses at a Tertiary Care Hospital. *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH*. https://doi.org/10.7860/jcdr/2016/21908.9088

Sijtsma, K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika* 74**,** 107 (2009). https://doi.org/10.1007/s11336-008-9101-0

Stanton, N. A., Salmon, P., Walker, G. H., Baber, C., & Jenkins, D. P. (2005). *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate Publishing Ltd.

Summers, G. F. (1970). Chapter 14: The Semantic Differential and Attitude Research. In *Attitude measurement* (2de editie, pp. 235–253). Rand McNally.

Use of e-health increases but not always effective | RIVM. (2021). RIVM. Geraadpleegd op 12 juli 2022, van https://www.rivm.nl/en/news/use-of-e-health-increases-but-not-always-effective#:%7E:text=The%20coronavirus%20pandemic%20has%20boosted,they%20were%20before%20the%20pandemic

World Health Organization. Regional Office for Europe, Health Evidence Network, European Observatory on Health Systems and Policies, Rechel, Bernd, Doyle, Yvonne. et al. (2009). How can health systems respond to population ageing?. World Health Organization. Regional Office for Europe. https://apps.who.int/iris/handle/10665/107941

Yuan, S., Hussain, S. A., Hales, K. D., & Cotten, S. R. (2015). What do they like? Communication preferences and patterns of older adults in the United States: The role of technology. *Educational Gerontology*, *42*(3), 163–174. https://doi.org/10.1080/03601277.2015.1083392

8. Appendix

Questionnaire:

Below shows the questions asked in the questionnaire, sorted by category.

- **Anthropomorphism (Semantic differential scale)**:
  1. fake – real
  2. machinelike – humanlike
  3. not conscious – conscious
  4. artificial – natural

- **Accessibility (5-point Likert scale)**:
  1. De tekst en iconen zijn groot genoeg
  2. De "vraag hulp knop" makkelijk te zien
  3. Het is duidelijk wat de functie is van de iconen op het scherm
  4. Het taalgebruik is makkelijk te begrijpen

5. Het contrast tussen de achtergrond en de tekstvelden zijn duidelijk
6. Ik snap hoe ik een bericht terug moet sturen na de introductie van het computerprogramma
7. Ik snap waar ik moet drukken als ik een bericht in wil spreken
8. Ik snap waar ik moet drukken als ik een bericht wil beluisteren

- **Satisfaction (Semantic differential scale):**
  1. terrible – wonderful
  2. easy – hard
  3. frustrating – satisfying
  4. satisfactory – not satisfactory
  5. boring – fun

- **Usefulness (5-point Likert scale):**
  1. Het computerprogramma ziet er onnodig complex uit (*)
  2. Het computerprogramma ziet er uit als of deze makkelijk te gebruiken is
  3. Ik heb hulp nodig van een technisch onderbouwd persoon als ik dit computerprogramma wil gebruiken (*)
  4. Ik denk dat ik dit computerprogramma snel onder de knie kan krijgen
  5. Ik kan het nut van een computerprogramma als deze goed inzien
  6. Ik denk dat het computerprogramma moeilijk in gebruik is (*)
  7. Ik denk dat ik dit computerprogramma met vertrouwen kan gebruiken
  8. Ik denk dat ik dit computerprogramma frequent zal gebruiken

- **Intention to use (5-point Likert scale):**
  1. Ik ben van plan dit systeem te gebruiken
  2. Ik ben niet van plan dit systeem te gebruiken (*)
  3. Ik zou dit systeem alleen op aanraden van de dokter gebruiken
  4. Ik zou dit systeem aanraden aan een ander

- **Trust (5-point Likert scale):**
  1. Ik denk dat dit systeem betrouwbaar is
  2. Ik denk dat dat technologie over het algemeen betrouwbaar is
  3. Als dit systeem betrouwbaar is, dan levert deze goede resultaten
  4. Ik denk dat dit systeem goed kan doen wat het zegt dat het gaat doen
  5. Ik heb geen vertrouwen in het systeem (*)
  6. Ik denk dat dit systeem mijn behoeftes kan vervullen op het gebied van vragenlijsten invullen
  7. Ik denk dat het systeem uit zichzelf, zonder hulp van personen, kan werken
  8. Ik heb over het algemeen vertrouwen in het systeem

*Reversed items on the Likert scale.

Card sorting:

Below shows the example of one of the cards used for the card sorting, the other cards look the same but with different descriptions. The contents of these cards are listed below the image.

Figure 4: Example of card from card sorting method

Other used cards were as follows:

- **Het onthouden van vorige gesprekken**: Een voorbeeld hier van zou zijn dat de conversational AI zich in het begin aan u voorstelt en als deze al meerdere malen contact met u heeft gehad, uw naam gebruikt en eventuele hobbies met u bespreekt.
- **Het gebruik van gezichtsuitdrukkingen:** Een voorbeeld hier van zou zijn dat de conversational AI naast het sturen van alleen tekst ook een bijbehorende gezichtsuitdrukkingen laat zien aan de hand van emoticons of een animatie op het scherm.
- **Het introduceren van zichzelf en zijn of haar functie benoemen:** Een voorbeeld hier van zou zijn dat de conversational AI tijdens de start van de conversatie een introductie over zichzelf doet, dus hoe hij/zij heet en wat zijn/haar doel is.
- **In kunnen spelen op sociale behoeftes**: Een voorbeeld hier van zou zijn dat de conversational AI over alledaagse dingen praat met u. Dus dat deze bij mooi weer bijvoorbeeld voorstelt dat u een stukje gaat wandelen of het met u heeft over dingen die in de krant hebben gestaan.
- **Het ontwikkelen van vertrouwen:** Een voorbeeld hier van zou zijn dat de conversational AI een vertrouwensband met u creëert, door te vertellen dat deze aan uw kant staat en graag met u wil samenwerken.
- **Het nemen van verantwoordelijkheid:** Een voorbeeld hier van zou zijn dat de conversational AI uitlegt waar hij/zij toe in staat is en vertelt wat er met uw gegevens gaat gebeuren.

Example script:

Below shows the script used as preliminary for filling in the questionnaires, with markings where certain characteristics are used.

1. Express /perceive emotions
2. Maintaining social relationships
3. Natural signals

4. Distinctive character
5. Social competencies
6. Trust
7. Responsibility

AI:

Goedemiddag, naam! Ik ben Charlie, uw persoonlijke gezondheid assistent 😊. Ik ga vandaag een vragenlijst bij u afnemen, bent u daar klaar voor?

Patiënt:

Ja

AI:

Fijn, dan gaan we beginnen. Hoe is uw mobiliteit vandaag?

Patiënt:

Mijn mobiliteit is vandaag beter dan gisteren, ik had een beetje moeite met opstaan uit bed, maar verder lukt lopen prima.

AI:

Wat fijn om te horen zeg. Heeft u er al aan gedacht om een stukje te gaan wandelen vandaag, het is een erg aangename temperatuur en het blijft droog!

Patiënt:

Dat is een goed idee, dat ga ik straks wel even doen.

AI:

Ik heb nog een laatste paar vragen voordat u op pad kunt: Hoe verloopt uw zelfzorg?

Patiënt:

De zelfzorg verloopt goed, ik heb geen extreme pijn. Hier en daar een beetje.

AI:

Goed op te horen. Zo lang de pijn niet heel erg is hoeft u zich nergens zorgen over te maken. Op een schaal van 1 tot 10, welk cijfer geeft u uw gezondheid?

Patient:

Een 6

AI:

Ah, dat is net aan voldoende. Is er iets waar ik mee kan helpen? Vindt u het een goed idee om het even over iets uit de krant te hebben of zal ik u een leuke mop vertellen?

Patient:

Nee hoor, dankjewel, ik denk dat ik de wandeling nu ga maken.

AI:

Dat is ook goed! Dan ga ik nog een stuk in mijn boek lezen. Zullen we volgende week weer even bijkletsen?