Bachelor thesis

Artificial Intelligence

Radboud University

The influence of personalized interaction on the user experience of educational chatbots

Author:
Annemiek van der Leest
s1021076

First supervisor:
M.A. De Sá Siqueira
PhD student - Behavioural Science Institute
marianna.desasiqueira@ru.nl

Second supervisor:
dr. ir. M.H. Vastenburg
Behavioural Science Institute
martijn.vastenburg@ru.nl

January 24, 2022

Abstract

Chatbots have become increasingly popular in recent years. Chatbots are one of the main focuses to improve e-learning processes. However, a reason for dissatisfaction towards chatbots is that they may be too generic. They often follow a one-size-fits-all approach. In this study, an educational chatbot was developed that teaches the user about cybersecurity. Two versions of the chatbot were created, one with personalization aspects and one without. A between-subjects design was used to measure the user experience of both chatbots. The user experience was measured by the User Experience Questionnaire (UEQ) together with some additional questions about Social Presence and User Satisfaction. The results show that, in the setting of this research, personalization does not have a statistically significant effect on the user experience of chatbots. The findings also show that Social Presence and Attractiveness appear to be good determinants of User Satisfaction. Besides, research about the UEQ is supported by our findings. Since this study is quite limited, however, it can be considered as a first encounter for investigating personalization. Future research needs to be done to better understand the effects of content personalization in the usage of chatbots.

Contents

Chapter 1

Introduction

Chatbots have become a popular means of communication in recent years. However, they have not always been so popular. The main reason why the use of chatbots has increased is that the communication style has changed [1]. Through the development of the internet and social network sites, people now interact through brief, typed messages.

Chatbots can be used in a variety of fields, for instance, in customer service, the financial industry [2], and the educational field [3]. As chatbots are increasingly being used for communication, it is interesting to see what aspects are relevant to the user. An important factor for measuring the relevance of the aspects is user experience. The user experience can be used to measure the quality of technology, and thus of chatbots.

This research focuses on how the user experience of chatbots could be improved. A possible reason for dissatisfaction towards chatbots is the fact that they may be too generic [4]. Chatbots often follow a one-size-fits-all approach [5]. According to Fan & Poole [6], personalization has an influence on how users experience systems. It can bridge the gap between the user and the computer. Personalization refers to the change of a system in order to increase its personal relevance to an individual or a group of individuals [6]. There can be a change in content, interface, functionality, and information access, or a combination of these. It is, however, still not clear whether users prefer a personalized chatbot or not. The research of Duijst [7] suggests that there is no significant effect on the user experience of a financial chatbot with respect to personalization. However, chatbots are used in other fields as well. Therefore, it is interesting to see the effects of personalization on the user experience in other domains. In this research, chatbots in the educational field are considered. Therefore, the aim of this research is to answer the following research question:

What is the effect of personalized interaction on the user experience of educational chatbots?

To give an appropriate answer to this research question, two versions of the educational chatbot were created. There is one general chatbot that is the same for everyone (non-personalized) and one chatbot that adapts to the user (personalized). For the personalized chatbot, the content of the dialogue was adapted to the user to make it more personal. The chatbot is used for educational purposes, since using chatbots in education may improve e-learning processes [8]. The chatbot teaches the user about cybersecurity. The use of computers and networks has increased enormously and therefore also the risks of possible attacks [9] [10]. So, it would be good to make people more aware of how to protect themselves against these risks.

In order to compare the two versions of the chatbot, a between-subjects design was used. The experiment was analyzed by the User Experience Questionnaire (UEQ) [11] together with some additional questions about Social Presence [12] and User Satisfaction [7].

To answer the research question, firstly some background information is provided. Afterwards, the research model and hypotheses are defined. Next, the methods are explained. Then, the analysis is done and the results are shown. Lastly, the results are discussed.

Chapter 2

Background

2.1 Chatbots

Chatbots refer to any software program with which humans can have a conversation using natural language [1]. This can be done through typed messages, but also through speech. Chatbots are used as a communication tool for users of online stores. They are also used in, for instance, customer service, marketing, and the entertainment sector [13]. Additionally, they can be used in the educational domain [3]. As chatbots are used in a variety of fields, they are developed for many different purposes. They can serve as an entertainment tool or they can be developed to provide certain information [14].

Chatbots can introduce benefits for both the companies that implemented the chatbot and the users [15]. People use chatbots for a variety of reasons, of which the most common one is productivity [16]. Users can quickly and efficiently access support or information when using chatbots. Other motivations for using chatbots are entertainment, social factors, and curiosity. In companies, chatbots are implemented because they lower costs and increase the number of clients they can help at the same time [17].

2.1.1 History of chatbots

Chatbots have a long history. In 1950, Turing proposed the question "Can machines think?" [18]. He described the problem in terms of an "imitation game", in which an interrogator should ask questions to distinguish between a machine and a human. The Turing test is passed, when the interrogator is not able to make this distinction. This test is regarded as a source of inspiration for chatbots.

ELIZA was one of the first programs that passed the Turing test [19]. It was developed in 1966 by Joseph Weizenbaum [20]. ELIZA served as a psychotherapist that asks users to express their feelings. Although ELIZA was far from perfect, it served as an inspiration for the creation of later chatbots [17].

Parry is another well-known chatbot that passed the Turing test [19]. It was introduced by Kenneth Mark Colby in 1972. Parry has the opposite role of ELIZA, it behaved as a schizophrenic patient [13]. Parry is seen as more sophisticated compared to ELIZA because it has a personality [17].

Experiments with ELIZA and Parry showed that interactions between humans and computers were possible and therefore should be investigated further [19]. Although these first chatbots served more as an entertainment tool, being aware of the history of chatbots can help to comprehend them in the present [13]. Even the most recent chatbots are still based on the pattern matching technique used in ELIZA [21].

2.1.2 Chatbots in the Educational Field

In recent years, the learning experience in the educational field has improved through the usage of e-learning methods [8]. Chatbots are one of the main aspects for improving e-learning processes. Actually, using chatbots is the most innovative approach to close the gap between technology and education [22].

Bob Heller et al. developed the Freudbot [23]. The Freudbot is a chatbot that talks about Freudian notions and theories. The chatbot talks in the first person to mimic Sigmund Freud. The researchers wanted to explore the experience of the students interacting with the Freudbot. The results suggest that the usage of well-known persons in chatbots is a promising e-learning method.

Another study assessed 47 chatbots in the educational domain to see how well they supported learning [24]. They suggest that educational chatbots are still in the early phases of development. The findings show that the responses of most chatbots were inadequate, they could not even answer simple questions like "How are you?". The findings also show that Quizzes and Question & Answer modules in educational chatbots may improve the learning processes. The International Financial Reporting Standards (IFRS) Rookies Messenger bot was one of the chatbots that ended in the top ten, based on the evaluations of the quality of supporting learning. This bot teaches about International Financial Reporting Standards [25]. Many of the participants were satisfied with the chatbot and many of them agreed that their engagement with the chatbot was higher than it would be in a traditional classroom.

A study about chatbots for corporate security training suggests that people view security more positively when they are trained through the use of chatbots rather than standard e-learning technologies [26]. In this research, an educational chatbot that teaches the user about cybersecurity is implemented. Cybersecurity is defined as: "the protection against malware and hacker attacks" [27]. In recent years, the use of computers and networks has increased enormously. They are being used in almost every aspect of our daily lives [9]. Because of this, the risk of possible attacks has increased as well [10]. Therefore, it would be good to make people more aware of how to protect themselves against malware and attacks.

More specifically, students are increasingly using the internet. This is due to the rise of e-learning. A study has shown that most students are unaware of the relevance of cybersecurity [10]. Thus, there is a need to make students more aware of it, which is why they are the target audience of this research.

2.2 Personalization

The definition of personalization differs per research field. Research suggests that most definitions consist of the following: a goal of personalization, what is personalized, and the target of personalization [6]. This has resulted in the following definition:

> "a process that changes the functionality, interface, information access and content, or distinctiveness of a system to increase its personal relevance to an individual or a category of individuals." [6]

Personalization can be done in various ways. Fan & Poole [6] provide a personalization scheme, which consists of three aspects:

- What is personalized: Four elements can be personalized, i.e. content, user interface, functionality, and channel/information access.
- To whom to personalize: Personalization can be done to a particular individual or to a group of individuals (e.g. women).
- Who does the personalization: The personalization can be implicitly or explicitly applied. Implicit personalization refers to personalization that is automated by the system. This personalization is based on information retrieved from the behavior of the users and the context of the messages. Explicit personalization refers to personalization that is based on choices made by the user or information given by the user.

Research found that personalization enhances the relationship, collaboration, and engagement with agents [28]. Besides, personalization improves user satisfaction [29] and trust in agents [30].

2.2.1 Personalization in Chatbots

One of the reasons users could be unhappy with chatbots is the fact that they can be too generic [4]. The interaction is rarely personalized and chatbots often follow a one-size-fits-all approach [5]. However, personalization has an influence on the user experience of systems [6]. A personalized interaction in chatbots can improve its social intelligence [4]. This is because the chatbot can become more aware of the context. Additionally, the chatbot can tweak its responses and features for every single person.

As every user is different, a social chatbot should be able to personalize the messages for individual users [31]. The messages should be empathetic, supportive, and tailored to the user's interests. The chatbot should lead the dialogue and it should make the user feel understood.

Chaves & Gerose [4] propose three advantages of personalization in chatbots. Firstly, it enriches the relationship between the user and the chatbot. People want the chatbot to respond differently to various users. Secondly, personalized interactions offer unique experiences. It leads to an increase in the value of the information that is given by the chatbot. Thirdly, personalization reduces the number of interactions that are broken down. Through personalization, the interface can be adapted to the user. This can be used in a way that encourages users to continue talking.

In the educational field, personalization in chatbots is seen as an important aspect [8]. It is one of the most relevant aspects with respect to learning since it can enhance the learning process. Therefore, the chatbot should have the ability to identify the behavior and needs of the user.

In a study about a chatbot that serves as a study buddy for students, it is suggested that a personalized interaction might be the best way to enhance the study habits of students [32]. The personalization was achieved by providing (more supportive) tips appropriate for a specific individual or by e.g. sending more reminders to students who have troubles with sticking to schedules.

Another study investigated the extra value of personalization for the user experience of chatbots in the financial field [7]. Two chatbots were developed (non-personalized vs. personalized) on which users could perform a simple task or a complex task. The study showed that personalization has no significant effect on the user experience for both tasks. It would be interesting to see if this is also the case in the educational field.

Research suggests that the personality of a chatbot is an appropriate aspect for personalization [33]. Due to a lack of time, however, personality will not be taken into account in this study. The concepts of personality and personalization are often intertwined. Personality can be defined as: "a complex organization of mental and biological systems that uniquely characterize an individual's behavior, temperament and emotion attributes" [34]. In the context of chatbots, personality is described as the characteristics that shape the communication style and the character of the agent [4]. Personalization is concerned with adapting to users' preferences and needs, whereas personality focuses on personality characteristics [33].

Aside from the personalization scheme provided in the section above, the chatbot could learn from and about the user in order to personalize the interaction in chatbots [4]. It could learn from the information provided in the interaction, for instance, information about the culture or behavior of the user. Besides, if there are previous conversations the chatbot should remember the user's preferences. However, users should be able to choose if the chatbot remembers these preferences. Another option for personalizing chatbots is to let users customize the chatbot by, for instance, changing its appearance and personality [4]. Furthermore, visual elements could be added to make the interaction more personalized. Visual elements shape the conversation, for instance, quick replies.

2.2.2 Personalization and Privacy

A challenge when adding personalization is the protection of privacy [35]. In order to personalize the interaction, the chatbot needs to remember the user's preferences and personal information. Collecting this information introduces concerns about privacy.

Solutions to this problem may include the ability to indicate to the chatbot that something is private or the ability to delete the history of the chatbot [35]. In this research, the privacy problem is partially addressed in various ways. Firstly, it is not required to provide your name and/or age. Besides, it is not possible to link the name to the other activities the participant performed in the experiment, i.e. the questionnaire. Also, participants have to sign an informed consent form. Lastly, IBM Watson Assistant is used to develop the chatbot. IBM Watson also provides some security. It protects the data by using a three-layer model which shares the information in a bottom-up manner.

2.3 User Experience

User experience is a relevant concept for evaluating conversational technologies [36]. However, there is no unified definition of user experience [37]. According to the International Organization for Standardization (ISO), the definition of user experience is: "a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" [38]. Although it is difficult to obtain a shared definition, a study suggests that the majority of researchers agree that user experience is seen as dynamic, subjective, and context-dependent [39]. According to Peras, user experience has four determinants: usability, performance, affect, and satisfaction [40]. In this research, the focus is on satisfaction. Multiple questionnaires are used to measure user satisfaction. Compared to a single questionnaire, the usage of multiple questionnaires may lead to a better evaluation of user satisfaction [36].

2.3.1 User Satisfaction

User Satisfaction refers to "user's pleasure arising from the comparison of their expectations and chatbot performance" [40]. It is a common metric for evaluating chatbots [41]. User Satisfaction is a subjective evaluation method. It is based on the user's expectations and needs [42]. As mentioned before, User Satisfaction is a determinant of the user experience.

Various questionnaires can be used to measure User Satisfaction. Duijst used a combination of four questions from various questionnaires to assess the User Satisfaction of chatbots [7]. One question comes from the Chatbot Evaluation Questionnaire [43] and the other three questions come from the happiness metric in the HEART (Happiness, Engagement, Adoption, Retention, Task success) framework [44]. The questions that are being asked are about if the user is unsatisfied with the chatbot, if functionalities are missing, if the chatbot is fun to use, and if the user would recommend it. Half of the questions are positively asked and half of them are negatively asked. This prevents response bias [45].

To measure User Satisfaction in this study, the questions of Duijst are used. A seven-point Likert scale is set for answering the questions since this measures the user's true attitude more precisely than a five-point Likert scale [46].

2.3.2 User Experience Questionnaire

The User Experience Questionnaire (UEQ) can be used to measure the overall impression of interactive products and systems. It enables a quick and instant evaluation [47]. Feelings about the product can be expressed immediately. The UEQ focuses on the overall effect the product has on the user and not just on one single aspect (e.g. usability). It takes into account both the Pragmatic and the Hedonic Quality. Pragmatic Quality relates to task-oriented quality features, like efficiency and learnability. Hedonic Quality relates to non-task-oriented quality features, like stimulation and aesthetic impression. [48]

The UEQ can be used to compare two products [11]. This is useful in this study as two versions of the chatbot (personalized vs. non-personalized) are tested and evaluated. Comparing the two versions can be done by a statistical comparison. Besides, the UEQ can be used to evaluate the User Satisfaction [49]. The UEQ has already been used to evaluate chatbots in previous studies [50] [51].

The User Experience Questionnaire contains six scales:

- **Attractiveness**: the overall impression towards the product.
- **Perspicuity**: the ease of use of the product.
- **Efficiency**: the fastness and efficiency of the product.
- **Dependability**: the user's level of control over the product and the safety of the interaction.
- **Stimulation**: how interesting, exciting, and motivating the product is.
- **Novelty**: how creative, innovative, and attention-getting the product is.

The relationship between these scales is shown in Figure 2.1 [11]. The Attractiveness scale consists of both the Pragmatic Quality and Hedonic Quality. Perspicuity, Efficiency, and Dependability are related to Pragmatic Quality. Stimulation and Novelty are related to Hedonic Quality. The overall impression is captured by the Attractiveness scale, which is thus affected by the other five scales [52]. Attractiveness is one of the dimensions to measure User Satisfaction [49] [53]. This may indicate a positive correlation between them.

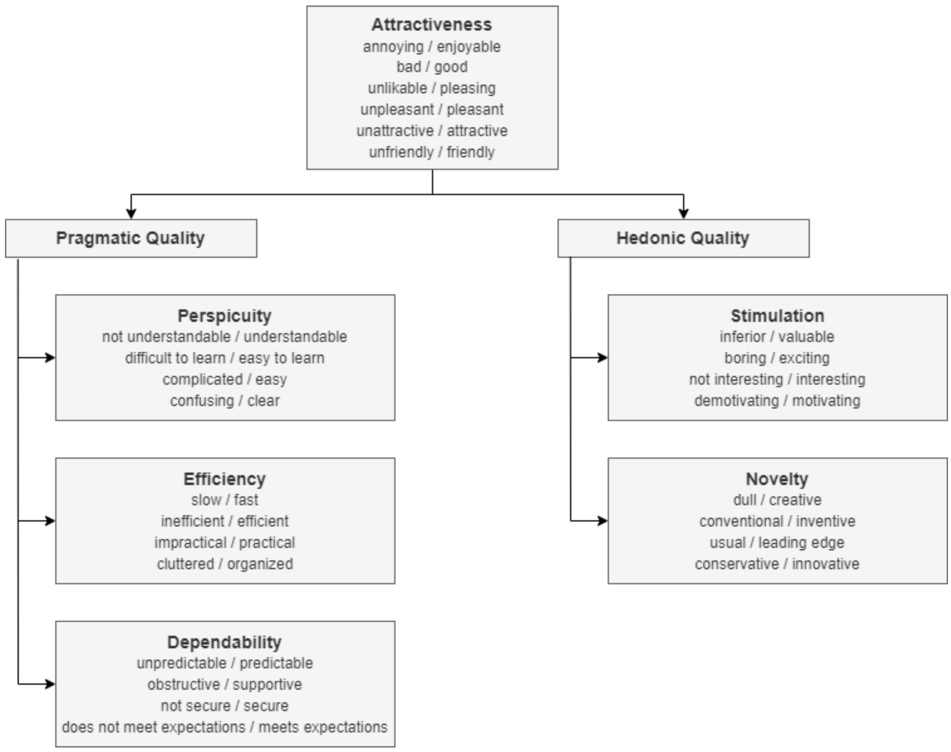


Figure 2.1: Dependency of the UEQ scale.

The six scales contain 26 items in total that are rated on a seven-point Likert scale. Each item consists of a pair of opposing attributes. Half of the items start with the positively worded attribute and the other half start with the negatively worded attribute. This is done in a randomized order [52]. The scores of the items are converted to a scale from -3 to +3 [48]. The -3 denotes the most negative response and +3 denotes the most positive response. It should be taken into account that most people avoid giving extreme responses like -3 and +3. Thus, most responses will be between -2 and +2.

2.3.3 Social Presence

Social Presence refers to "the feeling that the conversational partner is living in the same world, and is also capable of reacting to human queries" [12]. It is a relevant concept in the world of human-computer interactions. Through Social Presence, a relationship can be established between the agent and the user [54]. Social Presence depends on the user since it is based on how the messages are being interpreted [55]. According to research, both Social Presence and personalization are influenced by friendliness and expertise in online services. Besides, personalized greetings are said to improve the sense of Social Presence [56].

Research shows that Social Presence enhances the consumer's trust, enjoyment, and perceived usefulness [12] [57]. Another study suggests that Social Presence improves the experience in e-learning environments [55].

In the online service, Social Presence is found to have a positive effect on consumer satisfaction [56]. In addition, the findings of another study suggest that Social Presence and User Satisfaction are positively correlated in instant messaging [58].

For the measurement of Social Presence in chatbots, Toader et al. used five items [12]: human contact, human warmth, sociability, source of comfort, and sense of support when in need. These items are also used to measure Social Presence in this research. A seven-point Likert scale is set for those items.

2.4 Research Model and Hypotheses

For analyzing this experiment, the User Experience Questionnaire was used together with some questions about User Satisfaction and Social Presence. Two versions of the chatbot were tested with a difference in how personalized the interaction between the chatbot and the user is. The personalized chatbot adapts the content of the dialogue to the user. The non-personalized chatbot is the same for everyone and does not adjust the content to the user. Users tested one of the two versions and filled out a questionnaire afterwards. The aim is to find out if personalization improves the user experience of chatbots. The corresponding research model can be seen in Figure 2.2.

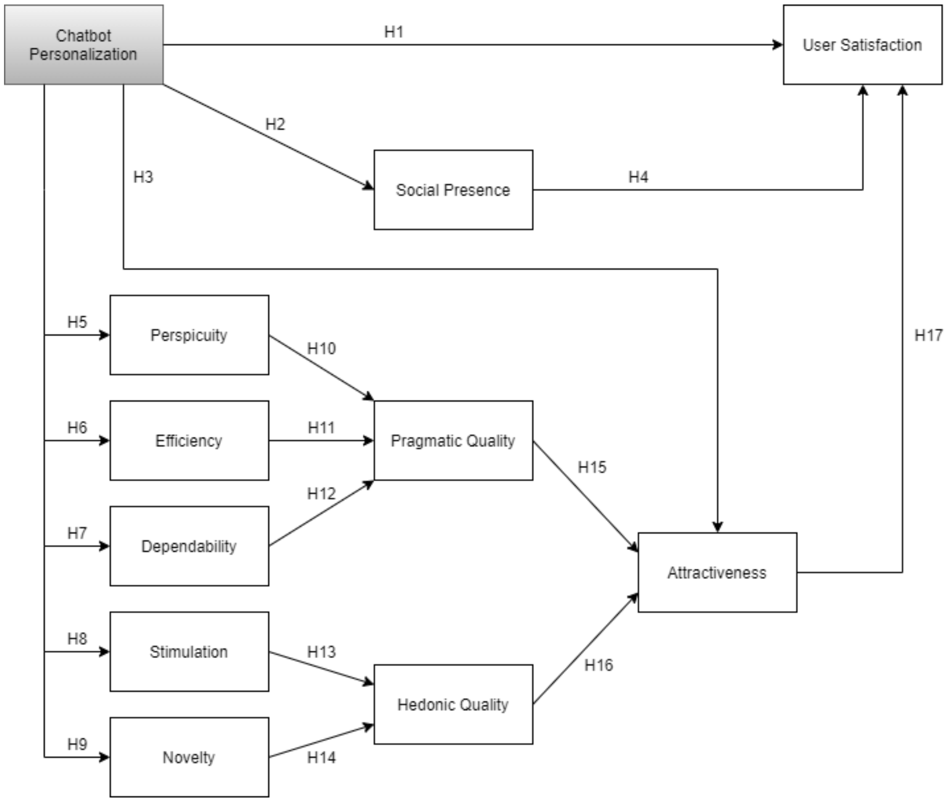


Figure 2.2: Research Model.

The following hypotheses are tested:

H1: A personalized chatbot has a higher User Satisfaction than a non-personalized chatbot.
H2: A personalized chatbot has a higher Social Presence than a non-personalized chatbot.
H3: A personalized chatbot has a higher Attractiveness than a non-personalized chatbot.
H4: The higher the Social Presence, the higher the User Satisfaction.
H5: A personalized chatbot has a higher Perspicuity than a non-personalized chatbot.
H6: A personalized chatbot has a higher Efficiency than a non-personalized chatbot.
H7: A personalized chatbot has a higher Dependability than a non-personalized chatbot.

H8: A personalized chatbot has a higher Stimulation than a non-personalized chatbot.
H9: A personalized chatbot has a higher Novelty than a non-personalized chatbot.
H10: The higher the Perspicuity, the higher the Pragmatic Quality.
H11: The higher the Efficiency, the higher the Pragmatic Quality.
H12: The higher the Dependability, the higher the Pragmatic Quality.
H13: The higher the Stimulation, the higher the Hedonic Quality.
H14: The higher the Novelty, the higher the Hedonic Quality.
H15: The higher the Pragmatic Quality, the higher the Attractiveness.
H16: The higher the Hedonic Quality, the higher the Attractiveness.
H17: The higher the Attractiveness, the higher the User Satisfaction.

Chapter 3

Method

An educational chatbot is developed to test the influence of personalized interaction on the user experience. The chatbot conducts a conversation about cybersecurity. Two versions of the chatbot were created: a personalized chatbot and a non-personalized chatbot. A between-subject design was conducted to compare the two versions. The participants were asked to have a conversation with one of the two chatbots and fill out a questionnaire afterwards. The sections that follow provide more details about this.

3.1 Materials

3.1.1 IBM Watson Assistant

The chatbot was developed by using IBM Watson Assistant, which is available in the IBM Cloud. It is an intelligent virtual agent that can be implemented into any device, application, or channel. It uses Artificial Intelligence to understand what users are saying. Watson Assistant is chosen because it provides a simple user interface that is easy to use. Besides, it offers a convenient chat user interface for websites. Watson Assistant is well-developed and has a good background in natural language understanding (NLU).

The chatbot was developed together with Serah Sommers, Sanne Janssen, Spence van Asperdt, and Andy Huang. Together we thought about the general topic and the sub-topics the chatbot would cover. After this, we worked together on the chatbot, where everyone focused on different features. Everyone built on this general chatbot by individually developing the experimental condition. In this research, the experiment condition refers to the personalized chatbot, while the control condition refers to the non-personalized chatbot.

3.1.2 Dialogue Design

The chatbot has to be able to have a conversation with the user about cybersecurity. As the conversation should not take too long, four topics about cybersecurity are selected: passwords, phishing, public Wi-Fi, and (illegal) downloading. These topics are relevant for students, the population of interest in this study. Due to the rise of the internet and e-learning methods, students often have to deal with these kinds of topics. For instance, phishing is the most commonly used attack and research shows that phishing mainly occurs because of the unawareness about it [10]. Thus, it would be good to make students more aware of phishing.

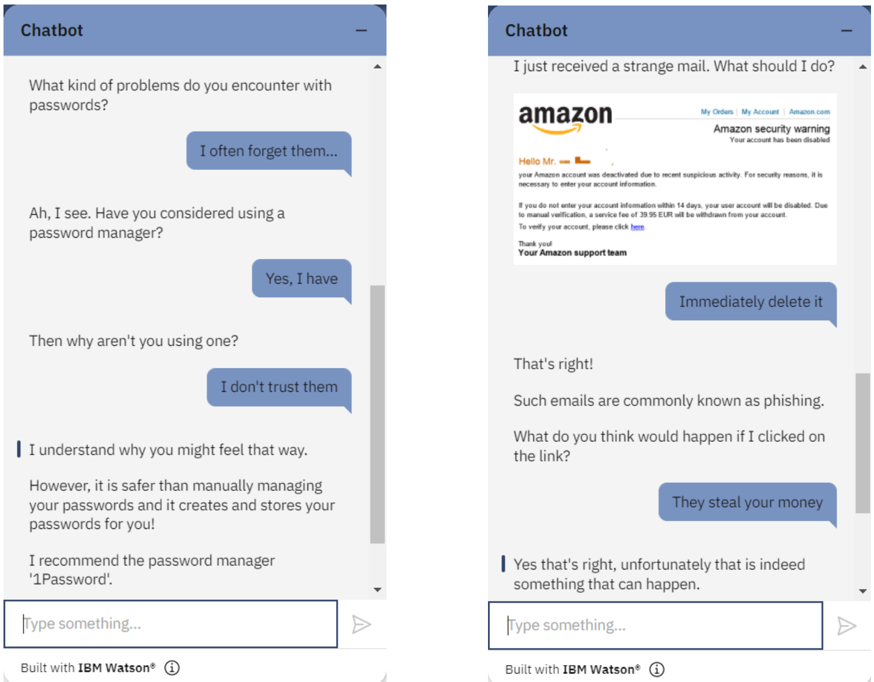
Aside from providing information on the various topics, the chatbot involves the user in the conversation. This is done by asking the user questions about their problems, knowledge, and behavior. To give proper answers to the user, the chatbot has to recognize different types of responses. These different types are created by 'intents'. Intents are the different intentions that the user might express. For example, an intent can be created for 'yes', which contains many examples of how the word 'yes' could be phrased ('yes', 'yeah', 'absolutely', etc.). The dialogue flow can be modified depending on the intent being recognized. An intent can be used in the condition of a dialogue node. Depending on if the condition is met, the node is processed or not.

Covering all possible intents is a challenge and due to time limits, we focused on the most important intents. This can result in situations where the chatbot does not recognize what the user is saying. In situations like that, the chatbot asks the users to rephrase their response. This may frustrate the users.

Instead of recognizing an intent, a condition can also be set to true. If this is the case, the target node is evaluated instantly regardless of what the user says. Several conditions are set to true to prevent the user from rephrasing their answers too many times. When conditions are set to true, each user gets the same response from the chatbot regardless of their input. This is another factor that may frustrate users.

The order of the topics is fixed, meaning that the dialogue flow is mostly the same for every participant, with the exception of some side branches due to recognizing different intents. Thus, most questions are always asked in the dialogue and each topic is discussed. After the welcome message, the chatbot immediately starts with the password topic. The chatbot asks if the user ever has any problems with passwords and if yes, it asks for the kind of problems. This way the chatbot involves the user in the conversation. Then, the chatbot asks if the user has ever considered a password manager. If users did consider it, they can give a reason why they are not using it right now. Then, the chatbot provides information about password managers in general. An example dialogue can be seen in Figure 3.1a.

After the password topic is discussed, it will move on to the next topic: phishing. The chatbot shows a (phishing) e-mail it just 'received' and asks the user what it should do and what could happen if the user clicks on the link. After the user answered these questions, the chatbot explains a bit more about phishing. The user also gets the option to learn more about how to recognize phishing or what should be done after clicking a phishing link. An example dialogue about the phishing topic is shown in Figure 3.1b.



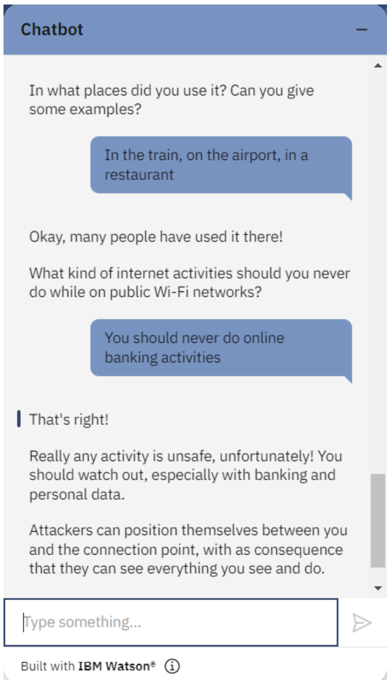
(a) Password topic

(b) Phishing topic

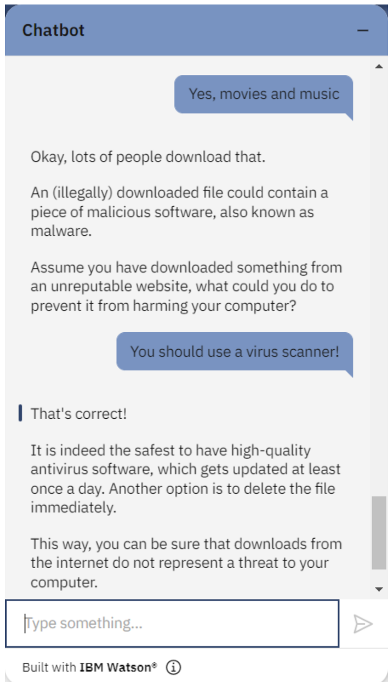
Figure 3.1: Example dialogues about the password and phishing topics.

The next topic that is discussed is public Wi-Fi. The chatbot asks if the user has ever used a public Wi-Fi network and in what places it is generally used. Then, it asks what kind of internet activities never should be done while on public Wi-Fi networks. After the user answers, the chatbot explains more about the dangers of public Wi-Fi networks. An example dialogue is shown in Figure 3.2a.

The last topic that is discussed is (illegal) downloading, of which an example dialogue can be seen in Figure 3.2b. The chatbot asks if the user has ever (illegally) downloaded something on the internet and if so, what it was. The chatbot makes clear that it could be harmful to your computer and asks the user what could be done to prevent a downloaded file from harming the computer. Afterwards, the chatbot explains a bit about anti-virus software.



(a) Public Wi-Fi topic



(b) Downloading topic

Figure 3.2: Example dialogues about the public Wi-Fi and (illegal) downloading topics.

Sometimes, the chatbot gives the option to get more information about a topic by providing links. This is done because the chatbot must not give too much information. Providing too much information makes the interaction less conversational. By showing links, however, users have the ability to learn more if they wish.

At the end of the discussion about a topic, the chatbot asks if everything is clear about the topic. This is done to make the dialogue even more conversational. If something is not clear, the chatbot also gives a link to a website on which the user can read more information about that topic.

The link to the questionnaire is provided by the chatbot at the end of the conversation. This is done to ensure that users go through the entire dialogue flow before filling out the questionnaire, without stopping earlier or not having the conversation at all.

3.1.3 Personalized Condition

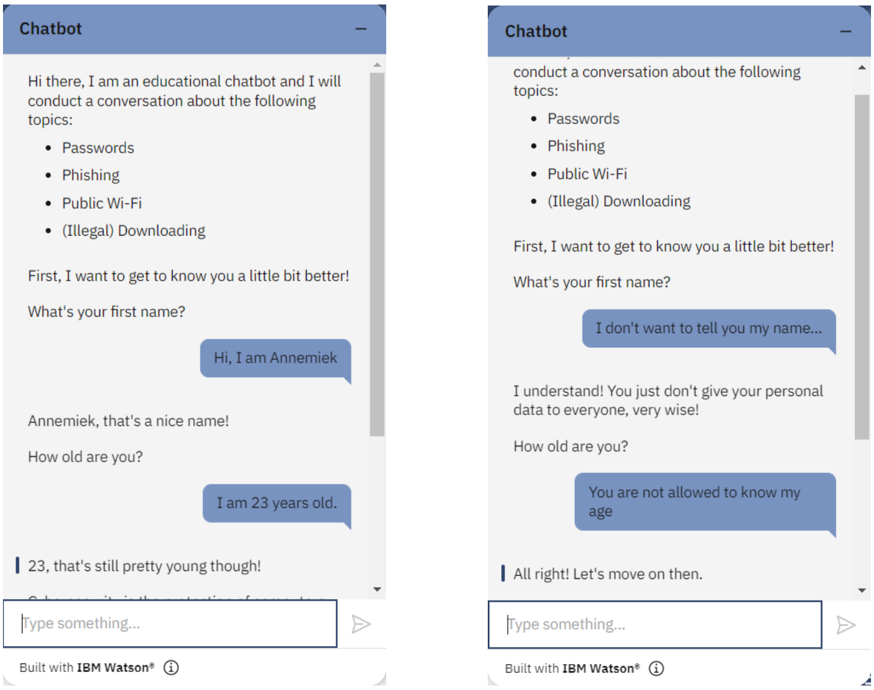
A comparison is made between the non-personalized chatbot and the personalized chatbot. The way the interaction is personalized is based on the research discussed in the background section. Research proposes three aspects of personalization that should be considered: what is personalized, to whom to personalize, and who does the personalization [6]. In this research, the focus is on personalizing the content of the messages rather than the interface, channel/information access, or functionality. The content is relatively easy to change with IBM Watson Assistant, while the other three aspects are harder to change. Also, when multiple aspects are chosen instead, it is difficult to say which of these aspects would cause a significant difference. With respect to the second aspect, personalization is done to a specific individual, rather than a group of individuals. Furthermore, explicit personalization is implemented, instead of implicit personalization. Explicit personalization is personalization that is tailored to the choices made by users or information given by users. Although research suggests that users experience implicit personalization as more human-like than explicit personalization [6], explicit personalization is more convenient to implement in Watson Assistant.

The content is adapted to a particular individual through explicit personalization in a few ways. The name and age of the user are asked, as well as their level of knowledge about cybersecurity. The messages include the name and are tailored to this level of knowledge. Besides, the user can choose the order in which the topics are discussed. More details about this will be given below.

Another study suggests that the chatbot should learn from and about the user and it should remember the user's preferences from previous conversations [4]. However, our participants only have one conversation with the chatbot. Therefore, it is chosen to learn from and about the user during the interaction.

Name and Age

At the beginning of the conversation, the chatbot is personalized in a way that it gets to know the user. The user's name is prompted to create the feeling that this dialogue is aimed at this individual user. The name is used several times in the dialogue to strengthen this feeling. For the same reason, the age of the user is asked. The age is repeated once, to show the user that the chatbot not only asks for the age but also remembers it. By mentioning the name and age, the content is adapted based on the information given by a particular individual (explicit personalization). An example dialogue can be seen in Figure 3.3a. It is not required for the user to tell their name or age. This is shown in Figure 3.3b. It is personal information, thus it should not be mandatory to provide this information.



(a) Tell name and age (b) Do not tell name and age

Figure 3.3: Example dialogues with the personalized chatbot at the start of the conversation.

In order to recognize the name and age in the input, intents and entities are both used. Entities identify information in the user input that is appropriate for the user's purpose. Intents can be seen as the action a user wants to pursue (verbs) and entities as the context in which that action takes place (nouns). Entities can help to respond in a more targeted way. Watson Assistant contains built-in system entities, e.g. 'sys-number'. Sys-number can be used to extract numbers (digits or written out as numbers) out of the user input, which makes it appropriate for recognizing the age. For example, the number 23 is extracted from "I am twenty-three years old".

There used to be an entity 'sys-person'. However, it was removed from Watson Assistant in 2020. Therefore, a 'name' entity had to be created. A list of all the names occurring in the United States was imported into this entity. This CSV file with all the names was made available in Dropbox [59]. In addition, Dutch names were added manually, since most participants were Dutch. As there is the possibility that users make a whole sentence ("My name is Annemiek"), instead of only telling their name, an intent had to be created as well. This intent contains many examples of how users might mention their name in a sentence. To teach Watson Assistant the entity (name) that should be extracted from the sentence, annotations could be used. By annotating the name in an example, Watson Assistant can identify names it has not seen before, but that appear in similar sentences. For example, in the sentence "My name is Eliza", the word 'Eliza' is annotated.

The name and age that the chatbot recognizes are assigned to context variables. Context variables can pass information through the dialogue. These context variables can be used as conditions for nodes. For example, if the name is known, the chatbot can provide a sentence including this name. This can be done at any time in the dialogue, as the value of a certain variable is saved.

Knowledge Cybersecurity

To learn even more about the user, the chatbot asks how much the user knows about cybersecurity. The chatbot gives a definition of cybersecurity and the user can choose between two options: having only a little knowledge or having quite some knowledge. An example dialogue can be seen in Figure 3.4.

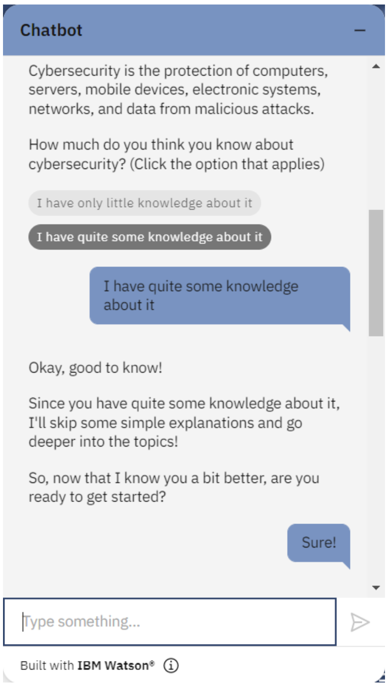
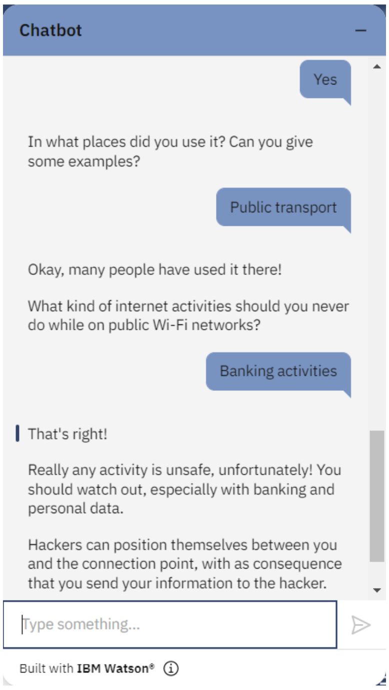
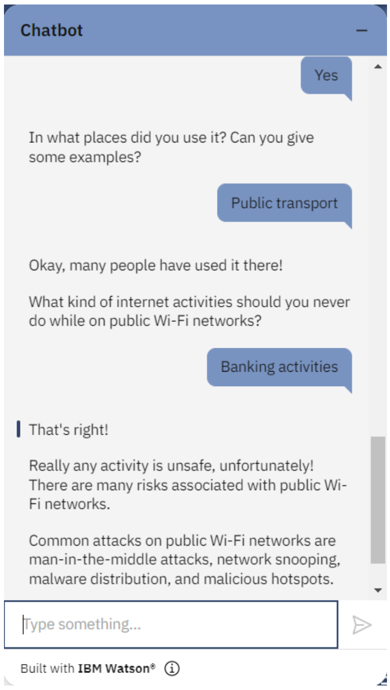


Figure 3.4: An example dialogue with the personalized chatbot for finding out the level of knowledge.

Depending on which option they pick, the messages are tailored to the knowledge level. When they only know a little bit about cybersecurity, the chatbot will provide understandable messages. If the user has more knowledge about cybersecurity, the chatbot assumes that the user already has some knowledge about the basic cybersecurity topics and will skip explanations about this. The messages will provide more details and more complex terms. For instance, it talks about ‘encryption’, ‘man-in-the-middle attacks’ and ‘spear phishing’. In the Figure below, the difference in messages between the two levels of knowledge can be seen. The structure of the dialogue and the questions the chatbot asks remain the same in both cases. The difference is in the explanations given by the chatbot.



(a) Only little knowledge



(b) Quite some knowledge

Figure 3.5: Example dialogues with the personalized chatbot for different levels of knowledge.

By tailoring the messages to this knowledge level, the content is adapted based on choices made by that specific user (explicit personalization). The chatbot takes into account the user’s preferences and the content has been changed in a way that increases the relevance of the topics to each individual. If the content would not have been adapted to the level of knowledge, the conversation may feel less relevant as the chatbot is saying a lot of things the user may already know or does not understand.

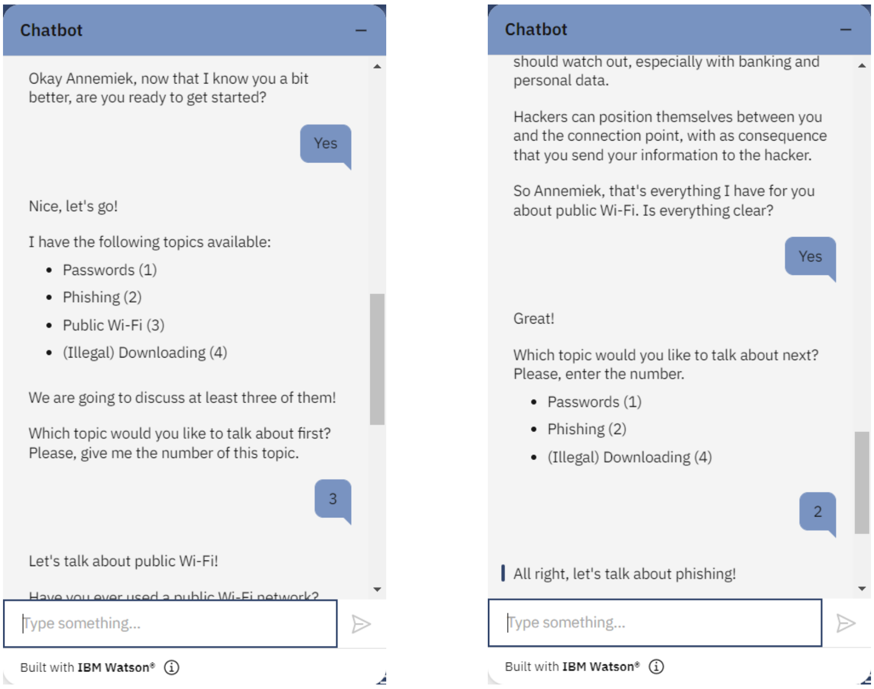
The chatbot could also have asked for the user’s knowledge about technology instead of cybersecurity. However, the term ‘technology’ is very broad. Students grew up in a world full of technology, so they might think they know a lot about it even if they do not really know. Besides, it is hard to relate the level of knowledge about technology to cybersecurity topics. Thus, the term cybersecurity is more appropriate given the topic of the chatbot.

To implement the messages based on the knowledge level, the sys-number entity and context variables are used. The value 1 is assigned to the option “I have only a little knowledge about it” and the value 0 is assigned to “I have quite some knowledge about it”. Thus, when one of the two options is chosen by the user, the value will become 1 or 0. This can be recognized by using the sys-number entity and this value is then assigned to a context variable ($number). Watson Assistant has the ability to enable multiple conditioned responses. This allows the chatbot to give different messages to the same input, depending on the different conditions. This way, at certain places in the dialogue, the chatbot provides the information either for people who only have little knowledge ($number == 1) or for people who have quite some knowledge ($number == 0).

Order of Topics

Lastly, the chatbot is personalized in a way that the user can control the order of the topics the chatbot provides. The chatbot talks about four different topics: passwords, phishing, public Wi-Fi, and (illegal) downloading. The user can choose the number of the topic it would like to talk about first. This is shown in Figure 3.6a. After the first topic is discussed, the user can indicate which of the remaining topics he/she would like to talk about next. An example dialogue of this can be seen in Figure 3.6b. After three topics, the user has the option to continue with the last topic or to end the conversation.

By controlling the order of the topics, the user can indicate which topics he/she would like to know more about. Besides, if they already know a lot about one topic, they can skip this. This way, each conversation is different and adapted to the order that the user prefers. The order of the content is adapted based on the choices made by that individual (explicit personalization).



(a) All topics available (b) Remaining topics available

Figure 3.6: Example dialogues with the personalized chatbot for selecting the topics.

To enable this feature, the entity sys-number and many context variables are used. The number that the user enters, which is recognized through the sys-number entity, becomes the value of the context variable. At the beginning of every topic, the assistant checks the conditions. If the context variable is set to the number of a topic, the assistant will go into that specific topic. After the topic is discussed, that context variable is set to null such that the same topic is not discussed again. A new context variable is created for the number of the next topic the user wants to talk about. By managing the context variables this way, the chatbot does not repeat topics it already discussed and users cannot choose these topics anymore.

3.1.4 Non-personalized Condition

The non-personalized chatbot does not contain all the features mentioned above. It does not ask for the name, age, or level of knowledge of the user. Besides, the user cannot determine the order in which the topics are discussed. The chatbot immediately starts with the first topic and discusses the topics in a fixed order. An example dialogue is shown in 3.7.

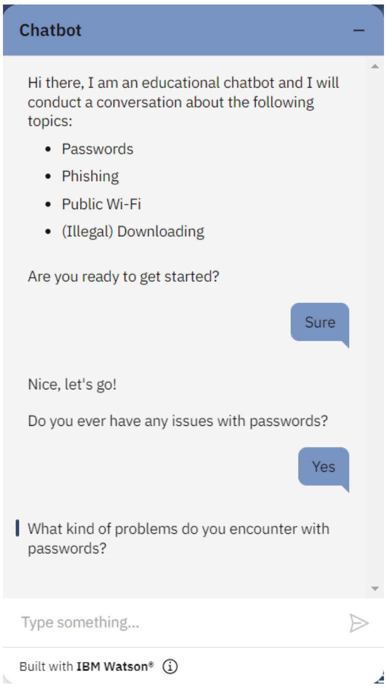


Figure 3.7: An example dialogue with the non-personalized chatbot.

3.1.5 Website Design

For the development of the website on which the chatbot is implemented, Blogger was used. This system is specially made for blogs that are hosted by Google through the subdomain blogspot.com. By adapting the layout, most blog functions could be turned off. This resulted in a simple, one-page website. A print screen of the website can be seen in Figure 3.8

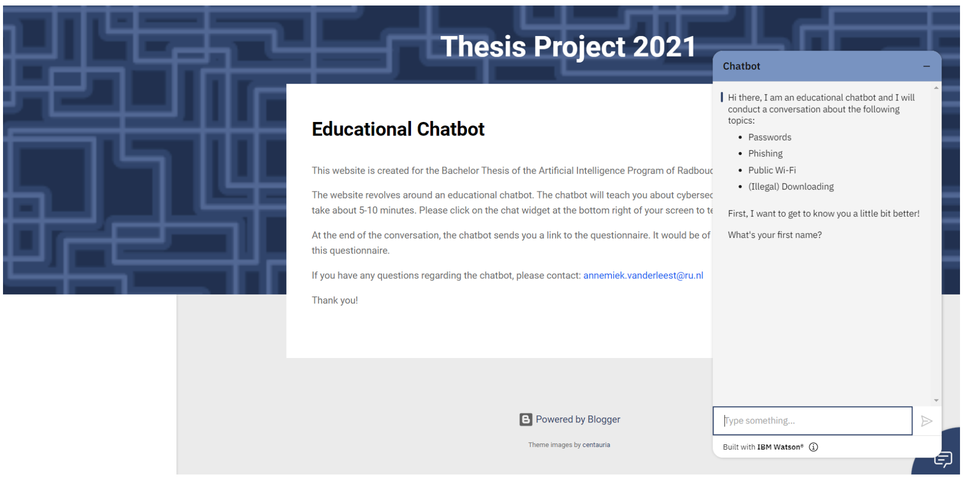


Figure 3.8: A print screen of the website.

Two websites were created: one in which the personalized chatbot is implemented and one in which the non-personalized chatbot is implemented. The website contains some basic information, which can be found in Appendix A.6. The chatbot becomes available by clicking on the chat widget at the bottom right of the screen. The chatbot is embedded in the website by HTML code made available by IBM.

3.1.6 Instructions and Questionnaire

Informed Consent and Task Instructions

The informed consent form, non-personalized chatbot task instructions, and personalized chatbot task instructions can be found in Appendix A.1, A.2, and A.3 respectively. Participants only get access to the task instructions when they consent. The task instructions provide some instructions about how to use the chatbot and include the link to the chatbot. Two nearly identical forms were created. The only difference is that the link leads to the personalized chatbot in one form and to the non-personalized chatbot in the other form. The informed consent and task instructions are implemented through Qualtrics, which is a survey tool available for free to students of Radboud University.

Questionnaire

The questionnaire was also developed using Qualtrics. It can be found in Appendix A.5. The questionnaire is provided at the end of the conversation with the chatbot. Two identical questionnaires were created, one for the personalized chatbot and one for the non-personalized chatbot. This way, the control and experiment conditions were separated from each other.

The questionnaire starts with some general questions about the user's experience with chatbots and their knowledge about technology and cybersecurity. Experience with a system has a positive effect on the ease of use [60], which refers to Perspicuity in the UEQ [61]. Besides, research suggests that people who use a system more often seem to have a higher user satisfaction [62].

The general questions are followed by the User Experience Questionnaire. First, an explanation of this questionnaire is provided. The questionnaire contains 26 items which are rated on a seven-point Likert scale (1...7). Each item consists of a pair of contrasting attributes. The Attractiveness scale contains six items, whereas the other scales all contain four items. The other scales are Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. After the UEQ, the questions for measuring User Satisfaction and Social Presence are provided. User Satisfaction contains four items and Social Presence contains five items. The items of both concepts are measured by a seven-point Likert scale as well (from 1="Strongly disagree" to 7="Strongly agree").

Then, three open questions are asked for qualitative research. Questions are asked about what they like or did not like about the chatbot. Besides, a question is asked about if the user had the feeling that the chatbot adapted its content to him/her individually. This is done to see if the participants experienced a difference in personalization between the two chatbots. The open questions are followed by some demographic questions (age, gender, nationality, education level) to get an idea of the different characteristics of the participants. A short overview of the questions in the questionnaire can be found in Appendix A.4.

Debriefing

The questionnaire ends with a debriefing text, which can be found in Appendix A.5.7. This text contains an explanation of the aim of this research. It explains which conditions were tested and gives a short definition of personalization. Apologies are made if the user still experienced some mistakes. Additionally, the users are informed about the confidentiality and release date.

3.2 Procedure

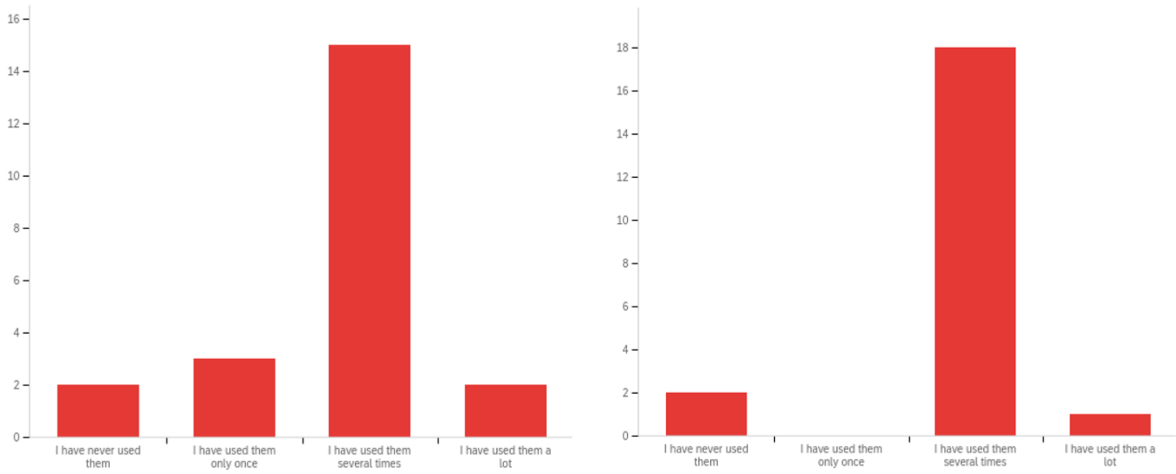
For this research, a between-subject design was conducted. The participants were all acquaintances. They were randomly split into two groups. One group tested the control condition and the other group tested the experiment condition. The control condition refers to the non-personalized chatbot and the experiment condition to the personalized chatbot. After developing the chatbots, the websites, the questionnaires, the informed consent form, and task instructions, a message was sent to the participants with a link to the informed consent form. When consent was given, the task instructions were provided. The participants were redirected to the website containing the chatbot. The participants were instructed to have a conversation of about 5-10 minutes with the chatbot until the chatbot provided a link to the questionnaire. This link was given at the end of the conversation. After the conversation, the participants had to fill in the questionnaire. The User Experience Questionnaire, questions about User Satisfaction, questions about Social Presence, open questions, general questions, and demographic questions were included in the questionnaire. Filling in the questionnaire took about 4-10 minutes. After all the data had been collected, the data analysis was performed.

3.3 Participants

In total, there were 42 people who participated in this research. All of them were acquaintances. In each condition, there were 21 participants who tested the chatbot and filled out the questionnaire. On average, the participants were 22 years old. In the control condition, the youngest participant was 19 years old and the oldest participant was 26 years old. The average age of the participants in this condition was 22 years. In the experiment condition, the average age was also 22 years, with the youngest participant being 20 years old and the oldest participant being 25 years old. Regarding gender, the majority of the participants were female. 61.90% of the participants in the control condition were female and 38.10% were male. In the experiment condition, 71.43% of the participants were female and 28.57% were male. As for nationality, all of the participants were Dutch.

The majority of the participants indicated secondary school as the highest level of education they have completed, namely 47.62% in the control condition and 42.86% in the experiment condition. Furthermore, in the control condition, the distribution was as follows: 14.29% higher education bachelor (HBO), 28.57% university (WO) bachelor, and 9.52% university (WO) master. The experiment condition was distributed as follows: 4.76% higher education bachelor (HBO), 42.68% university (WO) bachelor, and 9.52% university (WO) master.

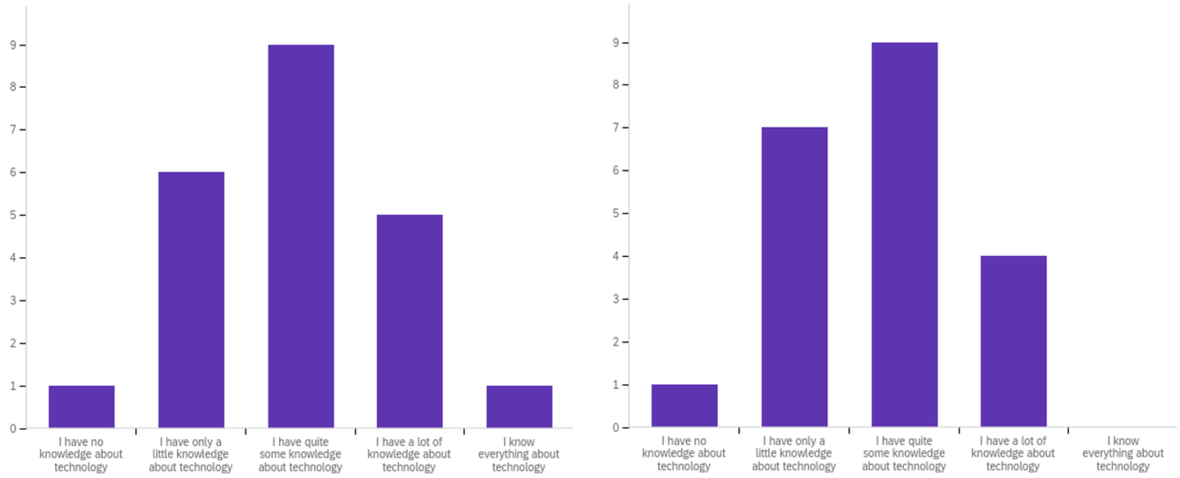
Some questions were asked about the user's experience with chatbots and their knowledge about technology and cybersecurity. The statistics in Figure 3.9 show that most participants have used chatbots several times.



(a) Control condition. (b) Experiment condition.

Figure 3.9: Experience Statistics.

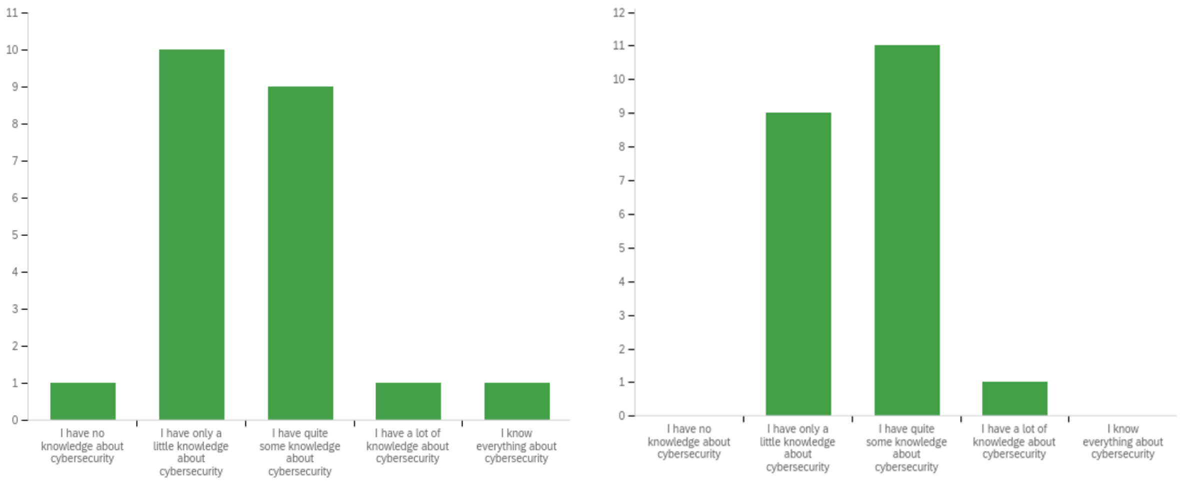
The answers for the user's knowledge about technology are quite divided. The statistics of this question are shown in Figure 3.10. In both conditions, most participants have quite some knowledge about technology. However, there is also a considerable portion that has only a little knowledge about it or a lot of knowledge about it.



(a) Control condition. (b) Experiment condition.

Figure 3.10: Technology Knowledge Statistics.

The statistics of the question about the user's knowledge of cybersecurity can be seen in Figure 3.11. The majority of the participants have either only a little knowledge or quite some knowledge about cybersecurity.



(a) Control condition. (b) Experiment condition.

Figure 3.11: Cybersecurity Knowledge Statistics.

Chapter 4

Results

In this section, the results of this research are presented. The data from the questionnaire is used for the analysis. First, the quantitative results will be presented and afterwards the qualitative results.

4.1 Quantitative Analysis

The User Experience Questionnaire, questions about User Satisfaction, and questions about Social Presence were analyzed for the quantitative results. All the items in these questionnaires were evaluated by means of the Likert scale. For each concept included in the questionnaires, a composite score was computed. This was done by taking the average of the scores of all items belonging to that concept. The UEQ consists of six concepts: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. Aside from these concepts, the concepts of User Satisfaction and Social Presence were measured as well, using their corresponding questionnaires. In addition, composite scores were computed for Pragmatic Quality and Hedonic Quality. Pragmatic Quality depends on Perspicuity, Efficiency, and Dependability. Hedonic Quality is based on Stimulation and Novelty. One-way ANOVA and Pearson's Correlation coefficient were performed on the scores for each concept per participant. One-way ANOVA was used to compare the personalized and non-personalized chatbots. Pearson's Correlation Coefficient was used for analyzing the correlation between two concepts. There has been some discussion about the usage of ANOVA and Pearson's correlation coefficient in combination with the Likert scale. However, according to [63] and [64], both parametric and non-parametric tests can be used for analyzing Likert scale data. Thus, ANOVA and Pearson's correlation coefficient (both parametric tests) are suitable.

ANOVA can be used to compare the means of two (or more) groups. One-way ANOVA is ANOVA where only one independent variable is considered [65]. Therefore, one-way ANOVA is suitable for this research. Three assumptions should be met when conducting ANOVA: the data is normally distributed, the distributions have equal variances, and the data of the two conditions are sampled independently [66]. ANOVA is robust against violations of these assumptions [67] [68]. The significance level was set to 0.05, a typically used value in many fields [69].

Pearson's correlation is a measure for the linear correlation between two variables [70]. It has values between -1 and +1, where a minus sign means a negative correlation and a plus sign means a positive correlation. A value of zero means that the two variables are uncorrelated. Pearson's correlation is extremely robust against assumption violations [68] [71].

The analysis was done by using Excel, Python, and Jupyter Notebook. The modules NumPy, Pandas, Scipy, Seaborn, and Matplotlib were used. Before analyzing the results, the scale was converted and possible outliers were removed.

4.1.1 Converting Likert scale

The User Experience Questionnaire is rated on a seven-point Likert scale from 1 to 7. To prepare the data for analysis, the scores of the items are converted to a scale from -3 to +3 [48]. The -3 refers to the most negative response (e.g. boring) and +3 refers to the most positive response (e.g. exciting).

In order to compare this with the concepts of User Satisfaction and Social Presence, the scores of these Likert-scale items should be converted to a scale from -3 to +3 as well. This can be done as the response scale is a linear scale [72] [73]. The scores of the negatively stated items are reversed (-2 will become 2), such that the attitude is interpreted in the right way [45] [74]. An example can be seen in Table 4.1. When someone rates the statement "The chatbot is fun to use" with "Agree", the corresponding score is 6. This number will be converted to the number 2. The statement "I miss functionalities in this chatbot" is negatively stated, meaning that a score of 7 will be converted to a value of -3 to show the real attitude. This is done for all the items of User Satisfaction and Social Presence.

Item	**Rating**	**Score**	**Converted score**
This chatbot is fun to use	"Agree"	6	2
I miss functionalities in this chatbot	"Strongly Agree"	7	-3

Table 4.1: Examples of converting the scale.

4.1.2 Outlier Removal

A common outlier removal technique is the Interquartile Range (IQR) method. It is a well-known and easy-to-use method. This method can be applied even when the data does not follow a normal distribution [75], as it makes use of the median. This makes it an appropriate method for this research since Likert scale data does not (necessarily) result in a normal distribution.

The interquartile range is the difference between the third quartile (Q3) and the first quartile (Q1), where the third quartile is the 75th percentile of the data and the first quartile is the 25th percentile of the data. The interquartile range is multiplied by a constant, which is usually 1.5 [76]. Outliers are detected when they are outside the range of Q1 - 1.5*IQR to Q3 + 1.5*IQR.

After detecting outliers, there are different methods to handle the outliers [77]. The outliers could be kept and treated in the same way as any other data point. The outliers could be removed by dropping all data from that participant. Lastly, the outliers could be replaced by values that are closer to the other data points (e.g. mean or median). The choice is made to remove the outliers when only a few outliers are detected (three or less). If there are more outliers, the outliers are kept and treated as normal data points. This way, not too much data will be removed, as the data size is already small.

With this technique, outliers were detected for five participants. Four of these participants were in the control condition. Outliers were found for the concepts Perspicuity, Dependability, and two times for Stimulation. The other participant was in the experiment condition. For this participant, outliers were detected in the concepts of Attractiveness, Social Presence, and User Satisfaction. As more than three outliers were found, the outliers were kept and treated as normal data points. Thus, the analysis is done with all data.

4.1.3 Analysis

Through the conversion of the Likert scale, the data is now scaled from -3 to $+3$. The descriptive statistics of each concept are shown in Table 4.2. It can be seen that for Novelty the mean in the experiment condition is considerably higher than in the control condition. For User Satisfaction and Hedonic Quality, this is also the case, but to a lesser extent. The other concepts do not show a noticeable difference.

	Control	Experiment
User Satisfaction	0.583 ± 1.210	0.940 ± 0.993
Social Presence	-0.314 ± 1.379	-0.038 ± 1.264
Attractiveness	1.174 ± 0.676	1.278 ± 0.870
Perspicuity	1.917 ± 0.704	1.952 ± 0.522
Efficiency	1.452 ± 0.744	1.489 ± 0.777
Dependability	1.202 ± 0.551	1.155 ± 0.490
Stimulation	0.536 ± 0.906	0.560 ± 0.925
Novelty	-0.071 ± 1.102	0.619 ± 0.947
Pragmatic Quality	1.524 ± 0.516	1.532 ± 0.449
Hedonic Quality	0.232 ± 0.843	0.589 ± 0.872

Table 4.2: Means and Standard Deviations of each concept (Mean ± Std).

For the hypothesis H1 until H3 and H5 until H9, one-way ANOVA is used to analyze the results. ANOVA is conducted to see if a personalized interaction has a significant effect on each concept in these hypotheses. The concepts are User Satisfaction, Social Presence, Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. First, the three assumptions were checked before conducting a one-way ANOVA.

The first assumption is that the data is normally distributed. This can be checked by using the Shapiro-Wilk test. For the concepts of Perspicuity in the experiment condition and Efficiency in the control condition, the p-values are below 0.05. This means that for these concepts the null hypothesis of the Shapiro-Wilk test is rejected and therefore these concepts are not normally distributed. However, assumptions may be violated when using ANOVA. For all other concepts, the p-values are above 0.05, meaning that these concepts are normally distributed.

The second assumption is that the distributions have equal variances. To see if this is the case, the Levene test can be performed. The Levene test is robust against non-normality [78]. The p-values for all concepts are higher than 0.05, meaning that for all concepts the conditions have equal variances.

Independence is assumed because the two conditions were randomly allocated to the users and a between-subject design was conducted.

One-way ANOVA

After checking the assumptions, one-way ANOVA was conducted to compare the effect of personalization on the concepts. In Table 4.3, the computed degrees of freedom, F statistics, and corresponding p-values of the test are shown.

	$df_{between}$, df_{within}	F statistic	p-value
User Satisfaction	1,40	1.093	0.302
Social Presence	1,40	0.458	0.503
Attractiveness	1,40	0.184	0.670
Perspicuity	1,40	0.035	0.853
Efficiency	1,40	0.023	0.880
Dependability	1,40	0.087	0.769
Stimulation	1,40	0.007	0.933
Novelty	1,40	4.743	0.035

Table 4.3: One-way ANOVA statistics

Each concept has a p-value higher than 0.05, except for Novelty ($F(1,40) = [4.743], p = 0.035$). This means that for Novelty, the null hypothesis of ANOVA is rejected. For the other concepts, the null hypothesis of ANOVA is accepted. The null hypothesis of ANOVA states that two conditions have the same mean. Therefore, the hypotheses H1, H2, H3, H5, H6, H7, H8 have to be rejected and hypothesis H9 is accepted.

Correlation

The hypotheses H4 and H10 until H17 are analyzed by Pearson's correlation coefficient. Pragmatic Quality is based on the concepts of Perspicuity, Efficiency, and Dependability. The correlation between each of these concepts and the Pragmatic Quality is measured. The coefficients are shown in Table 4.4. A positive value indicates a positive correlation. This means, for example, the higher the Perspicuity, the higher the Pragmatic Quality. The higher the coefficient, the stronger the correlation. As can be seen, the concepts of Perspicuity and Pragmatic Quality are highly positively correlated. The same holds for Efficiency and Pragmatic Quality. Dependability is only moderately positively correlated with Pragmatic Quality.

	Pragmatic Quality	
	Correlation Coefficient	P-value
Perspicuity	0.806	1.158e-10
Efficiency	0.821	2.671e-11
Dependability	0.626	9.117e-6

Table 4.4: Pearson Correlation Coefficients and the corresponding p-values between Perspicuity, Efficiency, Dependability and Pragmatic Quality.

The concepts of Stimulation and Novelty determine the Hedonic Quality. Thus, the correlation between each of these concepts and the Hedonic Quality is measured. The coefficients are presented in Table 4.5. Both concepts are highly positively correlated with Hedonic Quality.

	Hedonic Quality	
	Correlation Coefficient	P-value
Stimulation	0.852	8.666e-13
Novelty	0.897	8.740e-16

Table 4.5: Pearson Correlation Coefficients and the corresponding p-values between Stimulation, Novelty and Hedonic Quality.

Attractiveness is influenced by both the Pragmatic and the Hedonic Quality. The correlation between those concepts is shown in Table 4.6. As can be seen, both concepts are moderately positively correlated with Attractiveness.

	Attractiveness	
	Correlation Coefficient	P-value
Pragmatic Quality	0.674	1.016e-6
Hedonic Quality	0.674	9.807e-7

Table 4.6: Pearson Correlation Coefficients and the corresponding p-values between Pragmatic Quality, Hedonic Quality and Attractiveness.

Besides, the correlation between Attractiveness and User Satisfaction is computed, as well as the correlation between Social Presence and User Satisfaction. The correlations between those concepts are shown in Table 4.7. Attractiveness is highly positively correlated with User Satisfaction, whereas Social Presence is only moderately positively correlated with User Satisfaction.

	User Satisfaction	
	Correlation Coefficient	P-value
Attractiveness	0.736	2.728e-8
Social Presence	0.527	0.334e-3

Table 4.7: Pearson Correlation Coefficients and the corresponding p-values between Social Presence, Attractiveness and User Satisfaction.

The information in the tables above is supported by the correlation plots shown in Figure 4.1. The plots show that all variables have a positive correlation. The slopes give information about the strength of the correlation. The plots show that the correlation between Social Presence and User Satisfaction is the lowest. The correlation between Novelty and Hedonic Quality is the highest.

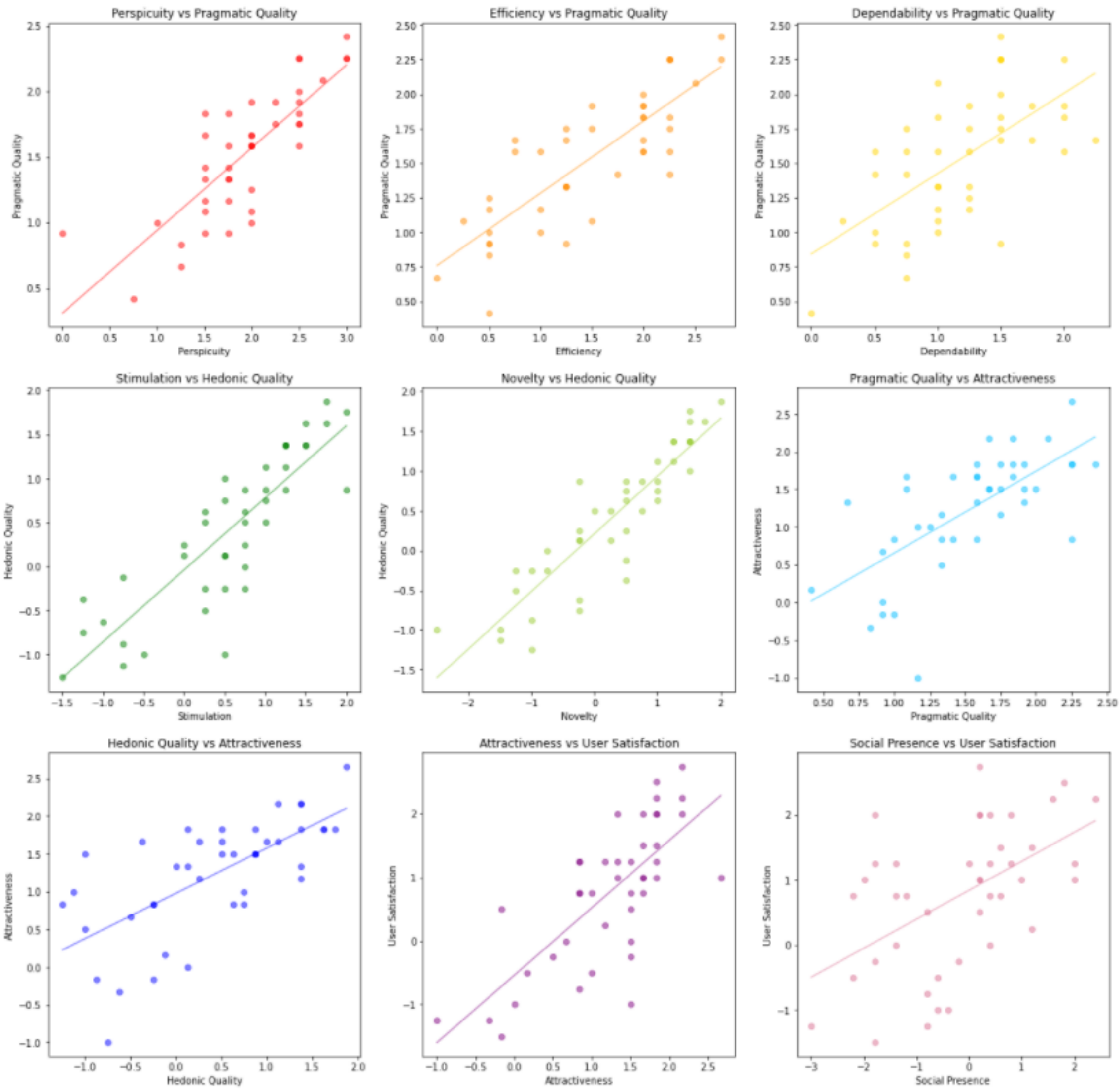


Figure 4.1: Correlation Plots.

The hypotheses are accepted or rejected based on the analysis above. As can be seen in the tables above, the p-values for the relevant hypotheses are all below 0.05. This means that the hypotheses H4, H10, H11, H12, H13, H14, H15, H16, and H17 are all accepted.

Figure 4.2 shows which hypotheses in the research model are accepted and rejected. The hypotheses in green are accepted and the hypotheses in red are rejected.

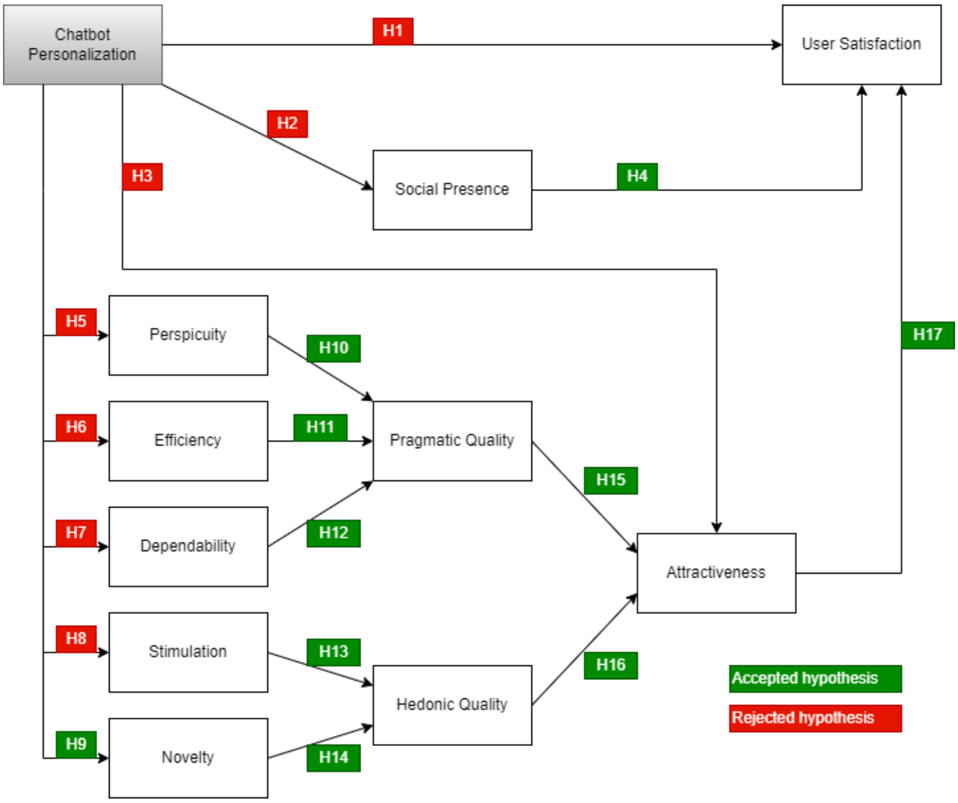


Figure 4.2: Research Model with accepted and rejected hypotheses.

4.2 Qualitative Analysis

The qualitative data was collected through the open-ended questions in the questionnaire. Some notable observations can be drawn from the responses to these questions:

- In both the control and experiment conditions, participants found the chatbot fast, friendly, supportive, and motivating. For instance, some participants in the control condition mentioned the following: "It was supportive and cheerful, responded quickly ..." and "It was really friendly and motivating". In the experiment condition, someone noted that she really liked the fast reply and the cheeriness: "The fast reply, the cheeriness (a lot of 'great!', 'indeed!') ...". Another participant mentioned: "It was clear, fast and friendly".

- Both conditions noted that the information was valuable and clearly explained. In the control condition, a participant noted: "It gave clear and elaborate answers about the subjects.". A participant in the experiment condition wrote the following: "It responded quickly and gave valuable and useful information that was easy to implement."

- According to some participants in both conditions, the right amount of information was given: "The chatbot was able to give the right amount of information.". Others mentioned that the chatbot provided too much information at once. The possibility to get more information by clicking on links was liked by some participants: "There was nothing I didn't like, yet it gave a lot of information after answering one question. Maybe linking to another page could change that.". However, there are also some participants who wanted the chatbot to explain this extra information: "It referred me to external websites. I wanted the chatbot to have the information".

- In the experiment condition, participants mentioned that they liked it that the chatbot was using their name. They got the feeling that the chatbot spoke to them specifically: "I liked how it spoke to me specifically (using my name)." and "It felt more personal because it repeated my name.".

- Quite some participants in the experiment condition also liked that they could pick the topics that interested them the most: "It was nice getting some knowledge, and choosing what I wanted to know.". However, a few others mentioned that they wanted the opportunity to come back to previous questions and that picking themes by numbers felt less human-like. Someone in the control condition suggested the following: "Maybe shuffling the topic content a bit would make the conversation more dynamic (now it was more like: topic 1, done, topic 2, done, etc)".

- Some participants in both conditions noted that the chatbot parsed most of the responses correctly: "It parsed most of what I said correctly and the sentences did show some emotion, as far as that is possible.". Others mentioned that they did not like the fact that the chatbot could only respond well to straightforward answers: "He did not respond very well to answers such as 'not sure'. Only very straightforward answered he did respond to well".

- Participants in both conditions mentioned that they did not like the general, standardized response. They did not get the feeling that the chatbot was reacting to their messages and they got the idea that every user followed the same path. In the control condition, there were twice as many participants who did not like this compared to in the experiment condition. In the control some participants mentioned: “It could not move from the path it chose.” and “That it answers in a very non-specific way”.

- In the control condition, a lot of participants (10 out of 21 participants) did not have the feeling that the content was adapted to them individually. They believed that everyone received the same messages in a pre-determined way: “No, because the way the bot displayed the content was quite general and felt more like a basic decision tree, where every decision lead to the same answer”.

  There were, however, six participants in the control condition that had the feeling that the content was adapted to them individually. They noted that the chatbot responded well to their answers: “Yes, because I gave answers that weren’t only yes or no and still the chatbot gave answers to my questions perfectly.”.

  Also, five participants noted that they only slightly felt that the content was adapted to them individually: “it did react to what I said back but because it were easy answers I feel like the answers from the chatbot could be said to multiple answers”.

- A part of the participants (9 out of 21) in the experiment condition noted that they had the feeling that the content was adapted to them individually. Many of them mentioned that this was because the chatbot was using their name, asked them for their level of knowledge about cybersecurity, and asked for the order of the topics: “Yes, it used my name and made it feel like a conversation. Also, giving me the choices of difficulty and order makes it more personal.” and “yes, because you had to make the choices yourself and it adapted to your wishes.”. A few mentioned the same reason as in the control condition, i.e. that the chatbot provided the right messages to their answers.

  Five of the participants in the experiment condition did not have the feeling that the content was adapted. They had the feeling that the messages were general and standardized: “No, the messages seemed very much standard. My name was integrated, sure, but all messages seemed like a simple if-then situation”.

  Besides, seven participants only slightly felt that the content was adapted to them individually: “A bit, since the answers were adapted with synonyms to my answer, so it understood well what I was saying. However I don’t know if the content changed much because of my answers”.

Chapter 5

Discussion

The participants in this study interacted with either the personalized or the non-personalized chatbot and this gave us interesting insights. The results show that there is no statistically significant difference in user experience between the two chatbots.

For the concepts of Perspicuity and Efficiency, no statistically significant difference was found between the control and experiment conditions. This means that hypotheses H5 and H6 are rejected. In the open questions, many participants mentioned that the chatbot was pretty fast and clear. However, this was found in both conditions, meaning that this was not caused by the personalized interaction.

The results also did not show a statistically significant difference between the personalized and the non-personalized chatbot for Dependability and Stimulation. Thus, hypotheses H7 and H8 are rejected as well. The open questions supported this because participants in both conditions found the chatbot supportive and motivating. In addition, they found the information given by the chatbot valuable.

In contrast, a statistically significant difference between the control and experiment conditions was found for Novelty (in favor of the experiment condition), meaning that H9 is accepted. This indicates that participants found the personalized chatbot significantly more creative, innovative, and attention-getting than the non-personalized chatbot. The open questions show that participants in the experiment condition liked it that the chatbot used their name and that they could pick the topics that interested them the most. These factors could be related to Novelty, as they can make the conversation more creative, innovative, and attention-getting. The fact that messages were tailored to the user's level of knowledge could also be related to Novelty. Participants barely mentioned this in the open-ended questions, but this may be because participants could not specifically observe that there were some differences in messages.

For both conditions, Social Presence was slightly below zero, meaning a neutral answer. The results show that personalization did not have a statistically significant effect on Social Presence. Therefore, hypothesis H2 is rejected. Many participants mentioned that in both conditions the messages of the chatbot were general and standardized. The chatbot could only respond to straightforward answers. Perhaps, using less general messages in the experiment condition could result in a higher Social Presence. It could even result in a significant difference between both conditions. Less general messages could be given when more different intents are used and fewer conditions are set to true.

As expected, hypotheses H10 until H16 are all accepted. A positive correlation was found for these concepts. These hypotheses are based on the User Experience Questionnaire [11], which is a validated model that has been well explored. The hypotheses H4 and H17 are also accepted. These hypotheses were not based on the UEQ, but on other research [53] [56] [58]. This research and the research about the UEQ are supported by our results.

Attractiveness was measured directly by six Likert-scale items. The results show that no statistically significant difference was found for Attractiveness between the control and experiment conditions. Therefore, H3 is rejected. This is partially supported by the qualitative research, as it shows that for both conditions the chatbot was considered friendly. Besides, Attractiveness is influenced by the concepts of Perspicuity, Efficiency, Dependability, Stimulation, and Novelty, of which most did not show a significant difference. As mentioned before, hypotheses H15 and H16 are accepted. This means that there was a positive correlation between Pragmatic Quality and Attractiveness and between Hedonic Quality and Attractiveness. Furthermore, a positive correlation between Attractiveness and User Satisfaction was found. The higher the Attractiveness, the higher the User Satisfaction. Therefore, hypothesis H17 is accepted as well.

The User Satisfaction was measured directly and indirectly. It was indirectly measured by the terms Social Presence and Attractiveness. As mentioned above, hypotheses H4 and H17 are accepted. Besides, User Satisfaction was directly measured through four Likert-scale items. The results show that there was no statistically significant difference between the control and experiment conditions for User Satisfaction, meaning that hypothesis H1 is rejected.

Overall, the findings in this study suggest that no statistically significant difference was found between the personalized chatbot and the non-personalized chatbot. The concepts in the research model are positively correlated and only a significant difference between the control and experiment conditions was found for Novelty. As there was no significant difference caused by personalization for the majority of the concepts, there was no significant difference caused for User Satisfaction. Thus, it can be concluded that, in the setting of this research, a personalized interaction does not improve the user experience of a chatbot. This finding is supported by the research of Duijst [7], which suggests that there is no significant effect on the user experience of a financial chatbot with respect to personalization. Furthermore, the findings show that Social Presence and Attractiveness appear to be good determinants of User Satisfaction. In addition, research about the UEQ is supported by our findings. However, this study also has some limitations, which are mentioned in the next section.

5.1 Limitations

Several limitations emerged over the course of this research. Firstly, only 21 participants for each condition conducted the experiment, which could have an impact on the findings. The focus of this study was on quantitative research based on the research model. Through this quantitative research, an overall impression of the user satisfaction of the chatbots could be obtained. The qualitative research supports the findings of the quantitative research. As user satisfaction is subjective, it would be better to have a larger sample size for quantitative research. This way, a better overall impression of the chatbots could be obtained. A good sample size would be 60 participants for each condition [7].

Because of the small sample size, the outliers were not removed. They were treated as normal data points. This may also have an impact on the results. Furthermore, the participants may not be representative of the student population, as all the participants were Dutch. Thus, it can be hard to generalize the findings.

Another limitation is that, by no means, all participants in the experiment condition had the idea that the content was adapted to the user individually. This may indicate that some participants did not experience the interaction with the chatbot as personalized. It is not necessary to perceive the personalization, since there still is a difference between the two chatbots (although the users do not know this). However, it could indicate that the personalization should be better implemented. This is hard, as people consider various things as personal. In addition, some participants in the control condition did have the feeling that the content was adapted to them individually. Thus, it could also be the case that the interaction in the control condition contained too many personalization features. The difference between the two conditions might not have been big enough, which may affect the results. This is something that should be investigated further.

As mentioned before, it is difficult to cover all possible intentions that users might express and due to time limits, the focus of this research was only on the most important intents. It was often noted by participants in both conditions that the chatbot responded quite general and standardized, which they did not really like. This general response was often the reason why participants did not think the chatbot adapted its responses to the user individually. This limited interaction could influence the findings of this research.

This study is also limited in a way that the participants only interacted with the chatbot for about 5-10 minutes. Research suggests that when adding personalization, the chatbot should learn from and about the user and that it should remember the user's preferences from previous conversations [4]. In this study, there was only one interaction, thus the chatbot could not remember the preferences. In addition, ten minutes may be fairly short to really get to know the user. It may be that the dialogues were not rich and extensive enough to properly investigate the results of personalization.

In literature, there is no unified definition of user experience. In this research, the focus was on User Satisfaction, which is one of the determinants of user experience. However, it is disputable if User Satisfaction is sufficient to measure the overall user experience.

To get a more complete impression of the User Experience, the interactions of the users with the chatbot could be analyzed as well. Due to time limits, this was not done in this research. By doing this, however, it would be possible to see if there are differences in participants' behavior between the two chatbots.

5.2 Future work

Chatbots in the educational field are becoming popular for improving learning processes. However, educational chatbots in combination with personalization is still a fairly unexplored area. Due to time limits, not all aspects of educational chatbots in combination with personalization could be investigated in this study. In this section, some aspects for future research are proposed.

First of all, not all participants in the experiment condition had the feeling that the content was adapted to them individually. So, they may feel that the interaction was not personalized that much. As mentioned in the limitations, it is not necessary to perceive this personalization. However, it would be interesting to investigate if perceiving the personalization would improve the user experience of chatbots. In case the chatbot was just not personalized enough, more personalization features could be added. Adding a larger extent of personalization can be done by covering more of the possible intentions that users might express and setting fewer conditions to true. For example, when the chatbot asks for examples of public Wi-Fi places, no matter what the user answers, the chatbot responds with "Okay, many people have used it there.". When more intents are created (e.g. one with examples of public transport), more different conditions could be set and thus the chatbot could react more specifically. This way, the messages are more specific to what the user is saying. This may result in users experiencing the interaction as more personalized. Exploring more possible intentions can be achieved by extending the testing phase.

Besides, everyone considers different aspects as personalized. Therefore, it may help to explore what aspects people think fall under personalization. This information can be gathered by means of a survey before developing the personalized chatbot.

Another way to add more personalization is by taking into account more of the aspects proposed in research. Not only the content can be personalized, but also the user interface, channel/information access, or functionality. It would be interesting to look at these aspects as well. Additionally, implicit personalization could be done instead of explicit personalization. It would also be interesting to see how this affects the experience of users, since implicit personalization is experienced as more human-like than explicit personalization.

In addition, research proposes that the chatbot's personality may influence personalization. Adding a personality to the chatbot is something that can be further explored in the field of personalization. It may also be interesting to ask some questions about the user's personality (by the chatbot or in a survey beforehand) and adapt the interaction based on this personality.

During the interaction, the users could choose which topic they want to know more about by choosing the corresponding number. However, it may feel more human-like if they can enter the topic in words instead of entering the number. This is something that could make the interaction more personalized as well.

To properly investigate the results of personalization, a richer and more extensive dialogue could be developed. This could, for instance, be done by having a longer conversation or by having more than one interaction with the chatbot, such that the chatbot could remember the user's preferences.

Another suggestion for future research is to measure more aspects aside from User Satisfaction. For instance, Usefulness, Perceived Enjoyment, or Anthropomorphism could be measured. Besides, other research models could be used to measure the user experience, for instance, the Technology Acceptance Model or the Unified Theory of Acceptance and Use of Technology model.

Besides, further research is necessary to generalize the findings for all kinds of chatbots. In this research, the focus was on the educational field. However, personalization should be investigated in other fields to generalize the findings, e.g., in healthcare or e-commerce. In the educational field, it would be interesting to investigate the effects of making the chatbot more goal-oriented. The chatbot in this research is quite general about cybersecurity. It could be made more specific for the cybersecurity aspects covered within universities (or within companies).

In this study, the focus was on students. However, it would also be interesting to focus on another target audience, like working people. Finally, a study with more participants could be conducted, possibly with a more diverse population (e.g. not only Dutch students).

Chapter 6

Conclusions

In this study, the influence of a personalized interaction on the user experience of educational chatbots is researched. The findings suggest that no statistically significant difference is found between the two chatbots. This means that, in the setting of this research, a personalized interaction does not improve the user experience of a chatbot. The findings also show that Social Presence and Attractiveness appear to be good determinants of User Satisfaction. Furthermore, research about the UEQ is supported by our findings. As this study is quite limited, however, it can be seen as a first encounter for investigating personalization. Future research needs to be done to better understand the impacts of content personalization in the use of chatbots. To further investigate the effect of personalization on the user experience of chatbots, a study with more participants could be conducted. This could be done in either the same domain or in a different domain. Besides, a richer dialogue with less general and standardized messages could be applied, which could be achieved by extending the test phase. Additionally, the length of the conversation could be extended or the amount of conversations could be increased. Finally, additional personalization features could be added to the personalized chatbot.

Chapter 7

Acknowledgements

First of all, I would like to thank Marianna De Sá Siqueira for her guidance and assistance as supervisor of this research. Besides, I would like to thank all the participants who took part in this experiment. I also want to thank Serah Sommers, Sanne Janssen, Spence van Asperdt, and Andy Huang for developing the chatbot together. Lastly, I would like to thank Martijn Vastenburg for being the second supervisor.

Bibliography

[1] Robert Dale. The return of the chatbots. *Natural Language Engineering*, 22(5):811–817, 2016.

[2] Takuma Okuda and Sanae Shoda. Ai-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal*, 54(2):4–8, 2018.

[3] Sharob Sinha, Shyanka Basak, Yajushi Dey, and Anupam Mondal. An educational chatbot for answering queries. In *Emerging Technology in Modelling and Graphics*, pages 55–60. Springer, 2020.

[4] Ana Paula Chaves and Marco Aurelio Gerosa. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758, 2021.

[5] Asbjørn Følstad and Petter Bae Brandtzæg. Chatbots and the new world of hci. *interactions*, 24(4):38–42, 2017.

[6] Haiyan Fan and Marshall Scott Poole. What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202, 2006.

[7] Daniëlle Duijst. Can we improve the user experience of chatbots with personalisation. *Master's thesis. University of Amsterdam*, 2017.

[8] Mohammed Bahja, Rawad Hammad, and Mohammed Hassouna. Talk2learn: a framework for chatbot learning. In *European Conference on Technology Enhanced Learning*, pages 582–586. Springer, 2019.

[9] Richard A Kemmerer. Cybersecurity. In *25th International Conference on Software Engineering, 2003. Proceedings.*, pages 705–715. IEEE, 2003.

[10] Sreenivas Sremath Tirumala, Abdolhossein Sarrafzadeh, and Paul Pang. A survey on internet usage and cybersecurity awareness in students. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, pages 223–228. IEEE, 2016.

[11] Martin Schrepp. User experience questionnaire handbook. *All you need to know to apply the UEQ successfully in your project*, 2015.

[12] Diana-Cezara Toader, Grațiela Boca, Rita Toader, Mara Măcelaru, Cezar Toader, Diana Ighian, and Adrian T Rădulescu. The effect of social presence and chatbot errors on trust. *Sustainability*, 12(1):256, 2020.

[13] Mgr Tomáš ZEMČÍK. A brief history of chatbots. *DEStech Transactions on Computer Science and Engineering*, (aicae), 2019.

[14] Yi Fei Wang and Stephen Petrina. Using learning analytics to understand the design of an intelligent language tutor–chatbot lucy. *Editorial Preface*, 4(11):124–131, 2013.

[15] Darius Zumstein and Sophie Hundertmark. Chatbots–an interactive technology for personalized communication, transactions and services. *IADIS International Journal on WWW/Internet*, 15(1), 2017.

[16] Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In *International conference on internet science*, pages 377–392. Springer, 2017.

[17] Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.

[18] Alan Mathison Turing. Mind. *Mind*, 59(236):433–460, 1950.

[19] Jeffrey Epstein and WD Klinkenberg. From eliza to internet: A brief history of computerized assessment. *Computers in human behavior*, 17(3):295–314, 2001.

[20] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[21] Luka Bradeško and Dunja Mladenić. A survey of chatbot systems through a loebner prize competition. In *Proceedings of Slovenian language technologies society eighth conference of language technologies*, pages 34–37. Institut Jožef Stefan Ljubljana, Slovenia, 2012.

[22] Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale, Antonio Pietrosanto, and Saverio Lemma. Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research*, 7(5):528–533, 2018.

[23] Bob Heller, Mike Proctor, Dean Mah, Lisa Jewell, and Bill Cheung. Freudbot: An investigation of chatbot technology in distance education. In *EdMedia+ Innovate Learning*, pages 3913–3918. Association for the Advancement of Computing in Education (AACE), 2005.

[24] Pavel Smutny and Petra Schreiberova. Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, 151:103862, 2020.

[25] Astrid Schmulian and Stephen A Coetzee. The development of messenger bots for teaching and learning and accounting students' experience of the use thereof. *British Journal of Educational Technology*, 50(5):2751–2777, 2019.

[26] Stewart Kowalski, Katarina Pavlovska, and Mikael Goldstein. Two case studies in using chatbots for security training. In *Information Assurance and Security Education and Training*, pages 265–272. Springer, 2013.

[27] Morten Bay. What is cybersecurity? *French Journal for Media Research*, 6, 2016.

[28] Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. What can you do? studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 acm conference on designing interactive systems*, pages 264–275, 2016.

[29] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research*, 21(11):e15360, 2019.

[30] Sherrie YX Komiak and Izak Benbasat. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly*, pages 941–960, 2006.

[31] Heung-Yeung Shum, Xiaodong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *arXiv preprint arXiv:1801.01957*, 2018.

[32] Xiaoyi Tian, Zak Risha, Ishrat Ahmed, Arun Balajiee Lekshmi Narayanan, and Jacob Biehl. Let's talk it out: A chatbot for effective study habit behavioral change. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–32, 2021.

[33] Sofia Rönnberg. Persuasive chatbot conversations: Towards a personalized user experience, 2020.

[34] Zoraida Callejas, Ramón López-Cózar, Nieves Ábalos, and David Griol. Affective conversational agents: the role of personality and emotion in spoken interactions. In *Conversational agents and natural language interaction: Techniques and effective practices*, pages 203–222. IGI Global, 2011.

[35] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'neill. How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. In *IFIP Conference on Human-Computer Interaction*, pages 441–459. Springer, 2017.

[36] A Baki Kocabalil, Liliana Laranjo, and Enrico Coiera. Measuring user experience in conversational interfaces: a comparison of six questionnaires. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*, pages 1–12, 2018.

[37] Effie Law, Virpi Roto, Arnold POS Vermeeren, Joke Kort, and Marc Hassenzahl. Towards a shared definition of user experience. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 2395–2398. 2008.

[38] Alexander G Mirnig, Alexander Meschtscherjakov, Daniela Wurhofer, Thomas Meneweger, and Manfred Tscheligi. A formal analysis of the iso 9241-210 definition of user experience. In *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*, pages 437–450, 2015.

[39] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold POS Vermeeren, and Joke Kort. Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 719–728, 2009.

[40] Dijana Peras. Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97, 2018.

[41] Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, pages 111–119, 2019.

[42] Zaihasriah Zahidi, Yan Peng Lim, and Peter Charles Woods. Understanding the user experience (ux) factors that influence user satisfaction in digital culture heritage online collections for non-expert users. In *2014 Science and Information Conference*, pages 57–63. IEEE, 2014.

[43] Silvia Quarteroni and Suresh Manandhar. A chatbot-based interactive question answering system. *Decalog 2007*, 83, 2007.

[44] Kerry Rodden, Hilary Hutchinson, and Xin Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2395–2398, 2010.

[45] James T Croasmun and Lee Ostrom. Using likert-type scales in the social sciences. *Journal of Adult Education*, 40(1):19–22, 2011.

[46] Kraig Finstad. Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of usability studies*, 5(3):104–110, 2010.

[47] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Applying the user experience questionnaire (ueq) in different evaluation scenarios. In *International Conference of Design, User Experience, and Usability*, pages 383–392. Springer, 2014.

[48] Maria Rauschenberger, Martin Schrepp, Manuel Pérez Cota, Siegfried Olschner, and Jörg Thomaschewski. Efficient measurement of the user experience of interactive products. how to use the user experience questionnaire (ueq). example: Spanish language version. 2013.

[49] Stefan Werrlich, Austino Daniel, Alexandra Ginger, Phuc-Anh Nguyen, and Gunther Notni. Comparing hmd-based and paper-based training. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 134–142. IEEE, 2018.

[50] Serena Di Gaetano and Pietro Diliberto. Chatbots and conversational interfaces: Three domains of use. In *Fifth International Workshop on Cultures of Participation in the Digital Age, Castiglione della Pescaia, Italy*, volume 2101, pages 62–70, 2018.

[51] Dario Fiore, Matthias Baldauf, and Christian Thiel. " forgot your password again?" acceptance and user experience of a chatbot for in-company it support. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, pages 1–11, 2019.

[52] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.*, 2017.

[53] Sung H Han, Kwang J Kim, Myung H Yun, Sang W Hong, and Jongseo Kim. Identifying mobile phone design features critical to user satisfaction. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 14(1):15–29, 2004.

[54] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 633–644, 2019.

[55] Benjamin Kehrwald. Understanding social presence in text-based online learning environments. *Distance Education*, 29(1):89–106, 2008.

[56] Tibert Verhagen, Jaap Van Nes, Frans Feldberg, and Willemijn Van Dolen. Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication*, 19(3):529–545, 2014.

[57] Khaled Hassanein and Milena Head. 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, 2007.

[58] Wei Wang, JJ Po-An Hsieh, and Baoxiang Song. Understanding user satisfaction with instant messaging: An empirical survey study. *International Journal of Human-Computer Interaction*, 28(3):153–162, 2012.

[59] FetchaSquad LLC. allUSNames.csv — dropbox.com. `https://www.dropbox.com/s/6eac0mrw3aj17ip/allUSNames.csv?dl=0`, 2020. [Accessed 15-Nov-2021].

[60] Timothy J Willis. *An evaluation of the technology acceptance model as a means of understanding online social networking behavior*. University of South Florida, 2008.

[61] Kevin Andika Lukita, Maulahikmah Galinium, and James Purnama. User experience analysis of an e-commerce website using user experience questionnaire (ueq) framework. In *Prosiding Seminar Nasional Pakar*, pages 347–355, 2018.

[62] Simone Borsci, Stefano Federici, Silvia Bacci, Michela Gnaldi, and Francesco Bartolucci. Assessing user satisfaction in the era of user experience: comparison of the sus, umux, and umux-lite as a function of product experience. *International journal of human-computer interaction*, 31(8):484–495, 2015.

[63] Gail M Sullivan and Anthony R Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013.

[64] Jacqueline Murray. Likert data: what to use, parametric or non-parametric? *International Journal of Business and Social Science*, 4(11), 2013.

[65] Prabhaker Mishra, Uttam Singh, Chandra M Pandey, Priyadarshni Mishra, and Gaurav Pandey. Application of student's t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia*, 22(4):407, 2019.

[66] Rudolf N Cardinal and Michael RF Aitken. *ANOVA for the behavioral sciences researcher*. Psychology Press, 2013.

[67] M José Blanca Mena, Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré, and Rebecca Bendayan. Non-normal data: Is anova still a valid option? *Psicothema, 2017, vol. 29, num. 4, p. 552-557*, 2017.

[68] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

[69] Michael Lewis-Beck, Alan E Bryman, and Tim Futing Liao. *The Sage encyclopedia of social science research methods*. Sage Publications, 2003.

[70] Philip Sedgwick. Pearson's correlation coefficient. *Bmj*, 345, 2012.

[71] Larry L Havlicek and Nancy L Peterson. Robustness of the pearson correlation against violations of assumptions. *Perceptual and Motor Skills*, 43(3_suppl):1319–1334, 1976.

[72] Satyendra Nath Chakrabartty. Scoring and analysis of likert scale: Few approaches. *Journal of Knowledge Management & Information Technology*, 1(2), 2014.

[73] J Robert Warmbrod. Reporting and interpreting scores derived from likert-type scales. *Journal of Agricultural Education*, 55(5):30–47, 2014.

[74] Katherine A Batterton and Kimberly N Hale. The likert scale what it is and how to use it. *Phalanx*, 50(2):32–39, 2017.

[75] Steven Walfish. A review of statistical outlier methods. *Pharmaceutical technology*, 30(11):82, 2006.

[76] Jiawei Yang, Susanto Rahardja, and Pasi Fränti. Outlier detection: how to threshold outlier scores? In *Proceedings of the international conference on artificial intelligence, information processing and cloud computing*, pages 1–6, 2019.

[77] Dhiren Ghosh and Andrew Vogt. Outliers: An evaluation of methodologies. In *Joint statistical meetings*, volume 2012, 2012.

[78] David W Nordstokke, Bruno D Zumbo, Sharon L Cairns, and Donald H Saklofske. The operating characteristics of the nonparametric levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research, and Evaluation*, 16(1):5, 2011.

Appendix A

Appendix

A.1 Informed Consent Form

Information

Purpose of the study
This study is part of a Bachelor Thesis of the Artificial Intelligence Program of Radboud University. The study involves an educational chatbot that will teach the user (you) about cybersecurity. You will converse with the chatbot and afterwards you will fill out a questionnaire.

Voluntary Participation
Your participation should be voluntary. During the study, you can indicate at any time that you want to quit participating. You do not have to give a reason.

Anonymity
All collected information and data is processed anonymously: the results cannot be traced back to you later on. This also means that we cannot inform you about your personal results when the study is completed. However, if you are interested, we can inform you about the results of the complete study. If you wish to be informed about these results, please let us know.

Data Usage
The (anonymous) data that is collected about you will be used as part of data sets, articles, and presentations. It will be accessible to other researchers for a period of at least 10 years. However, as this is anonymous data, it cannot be traced back to you. The data is collected from the chat history and the questionnaire forms.

Contact
If you have any further questions about this research or if you are interested in the results of the study, you can contact us at the following email address: annemiek.vanderleest@ru.nl

Consent
By giving consent as a participant of this study, you agree with the following statements:

- I have read the provided information about this research study and understand it.
- The aim of this research study has been made clear to me.
- I have been given the opportunity to ask questions regarding the research study.
- I participate in this research study on a voluntary basis.
- I understand that I am allowed to stop at any time during this research study, without having to give a reason.
- I understand that the data gathered from this research study is anonymous.
- I understand how the data gathered from this research study will be stored and used.
- I consent to participating in this research study.

By clicking on 'I consent' below, you consent to participating in this study.

○ I consent
○ I do NOT consent

A.2 Task Instructions Non-personalized Chatbot

In this study, you are going to have a short conversation for about 5-10 minutes with a chatbot. The chatbot is an educational chatbot that aims to teach you about cybersecurity. Your task is to conversate with the chatbot and fill out a questionnaire about the chatbot. Thus, the whole experiment will take about 15 minutes. The link to the chatbot is given after the task instructions. Please read the following instruction rules very carefully:

- Both mobile phones and computers/laptops support the participation of this research, but we advise you to use your computer/laptop.

- Please make sure you are in an environment that is distraction-free. Close every unnecessary program on your computer, interact with no other person, and put your mobile phone on silent mode.
- Keep this instruction page open at all times, so you can re-read the instructions if things are unclear to you.
- Do NOT close the chatbot window during the conversation session.
- Interact with the chatbot until you receive a link to the questionnaire. This link will be sent by the chatbot at the end of the conversation.
- If you click on any link given during the conversation it opens in a new tab.
- As this is an English chatbot, it expects English responses from the user. Make sure you ONLY answer in English. You are allowed to quickly lookup translations.
- Make sure not to respond to the chatbot before it is done typing (indicated by the '. . . ').
- Only respond with ONE message at a time.
- Do not wait too long (more than a minute) to respond to the chatbot, as the dialog will otherwise shut down.
- If you do not receive the link to the questionnaire within 15 minutes of testing the chatbot, please contact the researchers.

Now that you have read the instruction rules, you can start the actual research. The link to the chatbot is `https://chatbot-thesis-2021.blogspot.com/`. It will open in a new tab.

If you still have any questions regarding this study, please contact the researcher at: annemiek.vanderleest@ru.nl

A.3 Task Instructions Personalized Chatbot

In this study, you are going to have a short conversation for about 5-10 minutes with a chatbot. The chatbot is an educational chatbot that aims to teach you about cybersecurity. Your task is to conversate with the chatbot and fill out a questionnaire about the chatbot. Thus, the whole experiment will take about 15 minutes. The link to the chatbot is given after the task instructions. Please read the following instruction rules very carefully:

- Both mobile phones and computers/laptops support the participation of this research, but we advise you to use your computer/laptop.

- Please make sure you are in an environment that is distraction-free. Close every unnecessary program on your computer, interact with no other person, and put your mobile phone on silent mode.
- Keep this instruction page open at all times, so you can re-read the instructions if things are unclear to you.
- Do NOT close the chatbot window during the conversation session.
- Interact with the chatbot until you receive a link to the questionnaire. This link will be sent by the chatbot at the end of the conversation.
- If you click on any link given during the conversation it opens in a new tab.
- As this is an English chatbot, it expects English responses from the user. Make sure you ONLY answer in English. You are allowed to quickly lookup translations.
- Make sure not to respond to the chatbot before it is done typing (indicated by the '. . . ').
- Only respond with ONE message at a time.
- Do not wait too long (more than a minute) to respond to the chatbot, as the dialog will otherwise shut down.
- If you do not receive the link to the questionnaire within 15 minutes of testing the chatbot, please contact the researchers.

Now that you have read the instruction rules, you can start the actual research. The link to the chatbot is `https://chatbot-thesis.blogspot.com/`. It will open in a new tab.

If you still have any questions regarding this study, please contact the researcher at: annemiek.vanderleest@ru.nl

A.4 Short Overview Questionnaire

Measurement	Number of Questions	Questions
General Questions	3	How much experience do you have with chatbots? How much knowledge do you have about technology? How much knowledge do you have about cybersecurity?
Attractiveness	6	Annoying/Enjoyable Good/Bad Unlikable/Pleasing Unpleasant/Pleasant Attractive/Unattractive Friendly/Unfriendly
Perspicuity	4	Not understandable/Understandable Easy to learn/Difficult to learn Complicated/Easy Clear/Confusing
Efficiency	4	Fast/Slow Inefficient/Efficient Impractical/Practical Organized/Cluttered
Dependability	4	Unpredictable/Predictable Obstructive/Supportive Secure/Not secure Meets expectations/Does not meet expectations
Stimulation	4	Valuable/Inferior Boring/Exciting Not interesting/Interesting Motivating/Demotivating
Novelty	4	Creative/Dull Inventive/Conventional Usual/Leading edge Conservative/Innovative
User Satisfaction	4	The chatbot is fun to use I miss functionalities in this chatbot I would recommend this chatbot to a friend I am unsatisfied with this chatbot
Social Presence	5	I felt a sense of human contact when interacting with the chatbot Even though I could not see the chatbot in real life, there was a sense of human warmth When interacting with the chatbot, there was a sense of sociability I felt there was a person who was a real source of comfort to me I felt there was a person who is around when I am in need
Open Questions	3	What did you like about the chatbot? What did you not like about the chatbot? Did you have the feeling that the chatbot adapted its content to you individually? Why (not)?
Demographic Questions	4	What is your age? What is your gender? What is your nationality? What is the highest level of education you have completed?

Table A.1: Short overview of the questionnaire.

A.5 Questionnaire

The questionnaire can be found on Qualtrics. The link of this can be provided when requested.

A.5.1 General questions

How much experience do you have with chatbots?
◯ I have never used them
◯ I have used them only once
◯ I have used them several times
◯ I have used them a lot

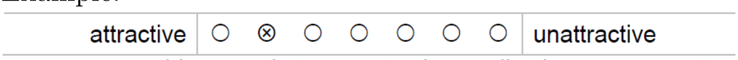
How much knowledge do you have about technology?
◯ I have no knowledge about technology
◯ I have only a little knowledge about technology
◯ I have quite some knowledge about technology
◯ I have a lot of knowledge about technology
◯ I know everything about technology

How much knowledge do you have about cybersecurity?
◯ I have no knowledge about cybersecurity
◯ I have only a little knowledge about cybersecurity
◯ I have quite some knowledge about cybersecurity
◯ I have a lot of knowledge about cybersecurity
◯ I know everything about cybersecurity

A.5.2 User Experience Questionnaire (UEQ)

This part of the questionnaire consists of pairs of contrasting attributes that may apply to the chatbot. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

Example:



This response would mean that you rate the application as more attractive than unattractive.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular product. Nevertheless, please tick a circle in every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Please assess the chatbot now by ticking one circle per line.

	1	2	3	4	5	6	7	
annoying	○	○	○	○	○	○	○	enjoyable
not understandable	○	○	○	○	○	○	○	understandable
creative	○	○	○	○	○	○	○	dull
easy to learn	○	○	○	○	○	○	○	difficult to learn
valuable	○	○	○	○	○	○	○	inferior
boring	○	○	○	○	○	○	○	exciting
not interesting	○	○	○	○	○	○	○	interesting
unpredictable	○	○	○	○	○	○	○	predictable
fast	○	○	○	○	○	○	○	slow
inventive	○	○	○	○	○	○	○	conventional
obstructive	○	○	○	○	○	○	○	supportive
good	○	○	○	○	○	○	○	bad
complicated	○	○	○	○	○	○	○	easy
unlikable	○	○	○	○	○	○	○	pleasing
usual	○	○	○	○	○	○	○	leading edge
unpleasant	○	○	○	○	○	○	○	pleasant
secure	○	○	○	○	○	○	○	not secure

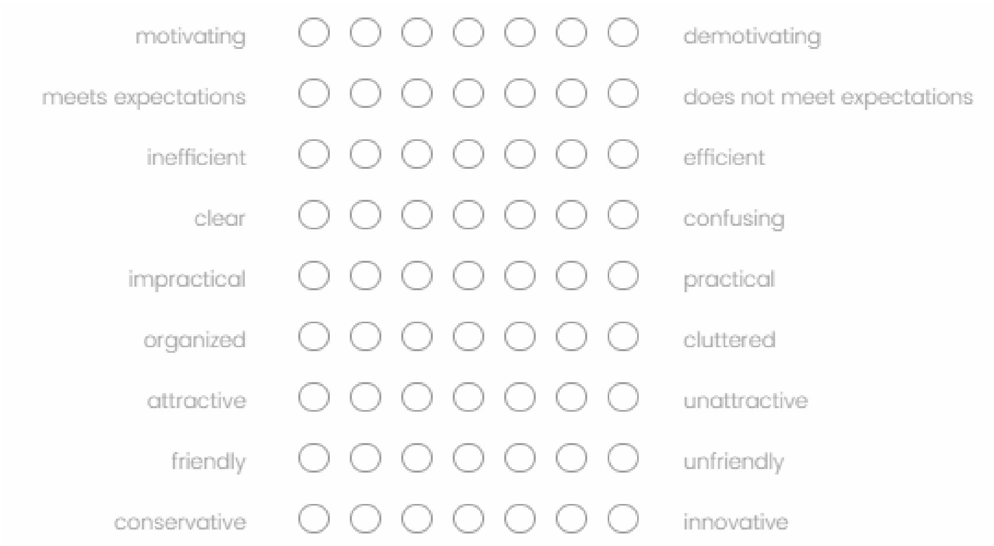


Figure A.1: User Experience Questionnaire

A.5.3 User Satisfaction Questionnaire

	Strongly disagree	Disagree	More or less disagree	Neither agree nor disagree	More or less agree	Agree	Strongly agree
This chatbot is fun to use	○	○	○	○	○	○	○
I miss functionalities in this chatbot	○	○	○	○	○	○	○
I would recommend this chatbot to a friend	○	○	○	○	○	○	○
I am unsatisfied with this chatbot	○	○	○	○	○	○	○

Table A.2: User Satisfaction

A.5.4 Social Presence Questionnaire

	Strongly disagree	Disagree	More or less disagree	Neither agree nor disagree	More or less agree	Agree	Strongly agree
I felt a sense of human contact when interacting with the chatbot	○	○	○	○	○	○	○
Even though I could not see the chatbot in real life, there was a sense of human warmth	○	○	○	○	○	○	○
When interacting with the chatbot, there was a sense of sociability	○	○	○	○	○	○	○
I felt there was a person who was a real source of comfort to me	○	○	○	○	○	○	○
I felt there was a person who is around when I am in need	○	○	○	○	○	○	○

Table A.3: Social Presence

A.5.5 Open Questions

What did you like about the chatbot? ______________________
What did you not like about the chatbot? ______________________
Did you have the feeling that the chatbot adapted its content to you individually? Why (not)? ______________________

A.5.6 Demographic Questions

What is your age? ________________

What is your gender?
○ Female
○ Male
○ Prefer not to answer
○ Other: ________________

What is your nationality? ________________

What is the highest level of education you have completed?
○ None
○ Primary school
○ Secondary school
○ Higher Education (HBO) Bachelor
○ Higher Education (HBO) Master
○ University (WO) Bachelor
○ University (WO) Master
○ Other: ________________

A.5.7 Debriefing text

Your response has been recorded.

Thank you for participating in this research. The aim of this research is to find out what aspects affect the user experience of chatbots. One condition was tested against a neutral condition: personalized chatbot vs. general chatbot. A personalized chatbot is a chatbot that adapts to the user. In this research, the focus is on adapting the content of the messages. The general chatbot gives the same messages to everyone. Apologies if you still perceived some mistakes during the conversation with the chatbot.

Your participation will remain confidential. Individual results will not be available, as all results will be grouped together. The final results of this research will be available at the end of January 2022. If you would like to read about the research or if you have remaining questions, please contact: annemiek.vanderleest@ru.nl

A.6 Website text

Educational Chatbot

This website is created for the Bachelor Thesis of the Artificial Intelligence Program of Radboud University.

The website revolves around an educational chatbot. The chatbot will teach you about cybersecurity. The conversation will take about 5-10 minutes. Please click on the chat widget at the bottom right of your screen to test the chatbot.

At the end of the conversation, the chatbot sends you a link to the questionnaire. It would be of great help if you fill out this questionnaire.

If you have any questions regarding the chatbot, please contact: annemiek.vanderleest@ru.nl

Thank you!