# IN5480: Final report

## Group 1 - Third delivery

# Table of contents

**1 Group description** 2

**2 Area of interest** 2

**3 Background** 3
- 3.1 Chatbots 3
- 3.2 Woebot 4

**4 Questions** 4

**5 Methods** 4

**6 Diary** 5

**7 Findings** 6
- 7.1 Types of interaction 6
- 7.2 Conversation 7
- 7.3 Emotional bond and mental health 7

**8 Evaluation** 8

**9 Discussion** 10

**10 Conclusion and lessons learned** 12

**11 Feedback** 13

**12 Appendix 1** 13
- 12.1 Chatbot design task 13
- 12.2 AI task 14
- 12.3 Reflections 14

**13 Appendix 2** 15
- 13.1 Reflections on making changes to the AI-chatbot 15
- 13.2 Results 15

**14 Appendix 3** 17
- 14.1 Introduction and scope 17
- 14.2 Abusability test 17
- 14.3 Guidelines review 19
- 14.5 Lessons learned 22

**15 References** 23

## 1 Group description

Our group consists of the following students; Anh Thy Sandra Nguyen, Sara Løkken, Thanh Thao Thi Tran, Maria Løvland Johansson, Johanne Thunes and Nathalie Dyhr Olimb. We are all master design students from two different educational backgrounds: four with a bachelor's degree in Informatics: Design, Use, Interaction from the University of Oslo, and two with a bachelor's degree in Web Development from the Norwegian University of Science and Technology in Gjøvik.

## 2 Area of interest

We are interested in the area surrounding mental health and emotional connections between humans and AI. In the earlier days of researching and developing Artificial Intelligence (AI), there was little focus on the human aspect and behaviour, and more on the logic and mathematics surrounding the machines. McCarthy, who first coined the term of Artificial Intelligence, wrote that "[The goal] was to get away from studying human behavior and consider the computer as a tool for solving certain classes of problems. Thus AI was created as a branch of computer science and not as a branch of psychology" (Grudin, 2009:51). But throughout the evolution of AI, there has become more focus on the human aspect of it, and also on the field of Human-Robot-Interaction, as the presence of personal service robots in e.g. the home are expected to grow in the coming years (Thrun, 2004). As we become more used to meeting different versions of AI in our everyday life, the thought of it taking on tasks that previously has been one of humans' has become not that far fetched.

The topic of mental health has become very relevant in today's society, and especially amongst people of our own age group; young adults. We therefore find this topic very interesting to explore, as it is something that we all can relate to in some way. Since the topic of the course is AI, we have narrowed down our focus area to the use of chatbots, and peoples experiences and feelings around the use of it. More specifically, we want to explore the subject of emotions in connection with chatbots, with an emphasis on the feelings of loneliness. We have decided that we want our focus to mainly be on students, as this is a group that generally has experienced a lot of loneliness through these past six months, when a lot of social meeting spots have been closed down as a result of COVID-19.

# 3 Background

During this project we want to investigate the growing interest of chatbots, and if we as humans are able to grow a deeper connection with a chatbot. We are curious about how people interact with a chatbot and how our natural language might affect how it will respond to how we speak to it. Furthermore, we will during this project test an existing chatbot, called Woebot, and engage a critical analysis and suggest changes as to how an AI-infused system can be improved.

## 3.1 Chatbots

As Cameron et al. (2017) proclaims in their article, one of the main causes of burden of diseases worldwide is mental health problems. That is why digital solutions have been created to help with issues such as anxiety, stress and depression. In our case, the digital solution trying to tackle this problem is a chatbot.

Chatbots are machine agents that serve as natural language user interfaces for data and service providers (Dale, 2016:811), and are typically designed in a context for messaging applications (Følstad & Brandtzæg, 2017:38). They may serve a number of different purposes, such as customer service, social and emotional support, information, entertainment, and connect the user to other people or machines (ibid.:3).

Initially, chatbots were made and restricted for simulating simple conversations between a human and a computer in a scripted way (Cameron et al. 2017:2). As Lugar and Sellen argue, chatbots often lack contextual information because they have no memory or knowledge, but instead mimic conversation (Lugar & Sellen, 2016:5287). Chatbots are now providing more information and maintaining a conversation with its human counterpart. Følstad & Brandtzæg mentions that our natural language is already the default mode of interaction online, which means that the interaction is typically between human users through a machine interface (Følstad & Brandtzæg, 2017:40). And because of the continuous development of AI, natural language interaction may be a feasible option for us humans to connect to machines (Følstad & Brandtzæg, 2017:40). By incorporating mental health tools into a chatbot, the user may have a more interactive and user-friendly experience. Chatbots can possibly create an option for users who might think it's easier to talk to.

### 3.2 Woebot

As previously mentioned, we are presenting a critical analysis of Woebot, which is an automated conversational agent (CA) designed to deliver cognitive-behavioral therapy (CBT) in the format of brief, daily conversations and mood tracking (Fitzpatrick et al., 2017:3). It was founded by Dr Alison Darcy and launched in June 2017, and has now more than 4,7 million conversations per week, across more than 120 countries (Woebot Health, 2020). Woebot is used as an instant messenger app that is platform agnostic, which means that it is an app you can download on your smartphone or use it on a desktop. Each interaction begins with a general inquiry about a context and mood, for example, "What is going on in the world right now?" and "How are you feeling?". In the app the mood is often represented with emojis or GIFs. According to Fitzpatrick's article, the bot's conversational style was modeled on a human clinical decision making and it included the following therapeutic process-oriented features like; *empathic responses, tailoring, goal setting, accountability, motivation and engagement and reflection* (Fitzpatrick et al., 2017:3).

## 4 Questions

To address our theme regarding emotional connections with a chatbot and mental health, we have developed two research questions we believe will help us further in our analysis, namely:

1. Can a human establish an emotional bond with an AI chatbot?
2. To what extent could an AI chatbot provide support within the field of mental health?

## 5 Methods

Our overall approach is to do a critical analysis on the use of the Woebot app. As mentioned in the background section, we are curious about how we as humans are able to connect with a chatbot on a deeper level. Therefore, we will in this section explain what kind of methods we have done and are planning to do. Following this, we will suggest some changes as to how an AI-infused system, like Woebot, could be improved.

To try and address our initial questions, we have chosen some methods that we find helpful in gathering data. Firstly, to get an understanding about how Woebot works, whether Woebot can act as a friend and a psychologist, we will explore the questions through a daily diary study. A daily diary study focuses on tracking people as they negotiate their day-to-day lives, and is an important tool in research on stress, emotion and health (Gunthert and Wenze, 2012).

Each of the group members will use the Woebot app for a week (7 days), and write individual diary entries. These diary entries will be on our thoughts while using the Woebot app, and screenshots of the interaction. The data from this diary study will be used to offer our critique of the Woebot app.

Our second approach is to evaluate Woebot using the guidelines for Human-AI Interaction (Amershi et al., 2019). In our review, we will state whether or not the separate guidelines have been applied in Woebot, and give examples on where and how these guidelines have been used in the app.

## 6 Diary

For our diary study, all six members of our groups downloaded the Woebot app on our smartphones, and interacted with the app daily for a week (7 days). During interactions, all group members took screenshots of parts of the conversations, which we then added comments to. We decided that every group member should have their separate document in our shared Google Drive to keep their diary entrances, which they either filled in right after the interaction, or filled in after the data collection period. Some group members wrote the diary entrances on their smartphones, and copied and pasted in the entries in the document that was assigned to them.

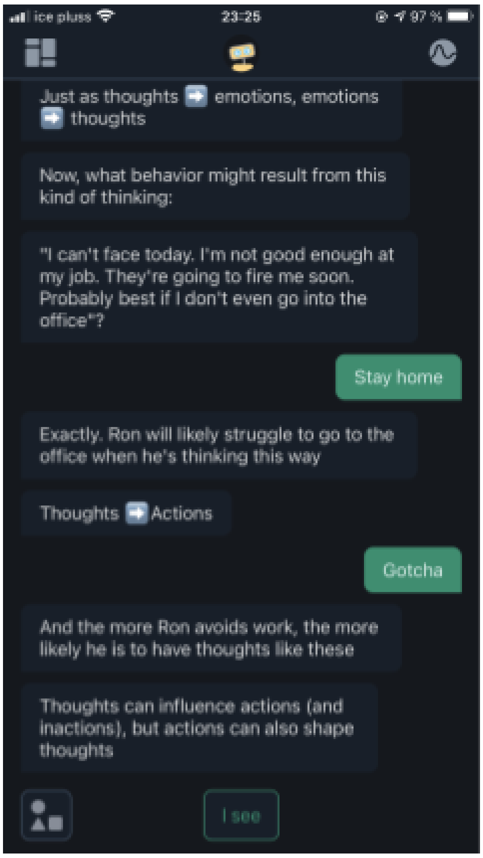
As the Woebot app focuses on mental health and provides strategies for coping with difficult situations in daily life, it is natural that we as users have to provide some information about circumstances, either positive or negative, which can be seen as sensitive. With this in mind, and also due to the fact that the content of the diary entries were also available to all group members, we were careful about not sharing any sensitive information about our mental health status in our diary entries. For this data collection we were thus more concerned with topics regarding the flow of conversation, how we interacted with the app, and more general thoughts on the topics regarding if we could establish an emotional bond with Woebot, or if it could replace a psychologist.

# 7 Findings

## 7.1 Types of interaction

While chatting with Woebot, we found that the type of interaction we could use to talk with Woebot alternated between predefined and self-typed answers.

The first interaction all of us had with Woebot was the initial introduction for Woebot to get to know us and our needs. Woebot asked us different questions about ourselves, and for almost all of them we could only choose predefined answers. These predefined answers had different purposes though, depending on whether you were given any options or not. Some of these answers were presented as the only option you could pick, and were often used as a way to keep the conversation going while still involving us. An example of this can be seen in figure 1.



Just as thoughts ➡ emotions, emotions ➡ thoughts

Now, what behavior might result from this kind of thinking:

"I can't face today. I'm not good enough at my job. They're going to fire me soon. Probably best if I don't even go into the office"?

Stay home

Exactly. Ron will likely struggle to go to the office when he's thinking this way

Thoughts ➡ Actions

Gotcha

And the more Ron avoids work, the more likely he is to have thoughts like these

Thoughts can influence actions (and inactions), but actions can also shape thoughts

I see

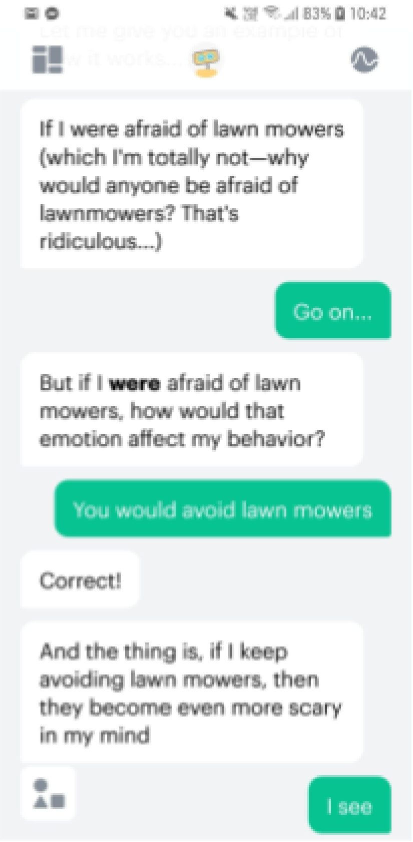
Figure 1: Woebot gives us a predefined answer to continue the flow of the conversation.

While the conversation was predominantly filled with predefined answers, there were times where we could type in our own answers. This was typically when Woebot asked us to talk about our own experiences, such as something we had achieved. Woebot did however not understand most of what was said and never replied to any questions asked in these self-typed messages we sent, and instead replied with a standard message meant to positively reinforce us for our achievements.

## 7.2 Conversation

All of the predefined answers, added with the fact that Woebot rarely actually understood what we said in our self-typed messages, limited our conversations with Woebot. Because of this, the conversations were always led by Woebot and always stayed on track with what Woebot wanted to talk about that day.

Woebot would mainly ask us about our moods or things we were grateful for, and would often spin these conversations into lessons about understanding feelings. Often, these lessons would be unrelated to what we told Woebot we were feeling, and thus felt like Woebot was following a script rather than having a natural flow in our conversations. An example of this can be seen in figure 2, where Woebot starts talking about anxiety out of nowhere.



Let me give you an example of
w it works...

If I were afraid of lawn mowers (which I'm totally not—why would anyone be afraid of lawnmowers? That's ridiculous...)

Go on...

But if I **were** afraid of lawn mowers, how would that emotion affect my behavior?

You would avoid lawn mowers

Correct!

And the thing is, if I keep avoiding lawn mowers, then they become even more scary in my mind

I see

Figure 2: Woebot giving a lesson about anxiety.

## 7.3 Emotional bond and mental health

After conducting the diary study we experienced that Woebot to some degree could give tips to improve your mental health, but as previously mentioned, we found Woebot as impersonal and static. The advice we received did not feel like they were personalized for us, but more general “over-all” advice that could be applied to everyone. The amount of impersonal messages and static content did not trigger a deep emotional bond to the robot, but on a more superficial level we found the robot “cute” and we appreciated his stories. On the other hand this in sum makes us believe that if, for instance, a person struggling with mental health issues that is considered more severe would not necessarily find help in this type of chatbot.

# 8 Evaluation

After having gotten to know Woebot during the diary study, our experience was that the overall interaction with Woebot appeared very static, impersonal and non-human. We decided that we wanted to evaluate Woebot by conducting a guidelines review, using Microsoft's guidelines for Human-AI Interaction (Amershi et al., 2019), to see whether following these guidelines could help to improve the user experience when interacting with Woebot.

| G(x) | Phase | AI design-guidelines | Application of Guideline in Woebot |
|---|---|---|---|
| **G1** | *Initial phase* | Make clear what the system can do | When opening the app, you are met with introductory information about the app, what it does and how to get started. It also says what it is *not* intended for, such as that it is not intended to be a crisis service, as no humans are monitoring the conversations. |
| **G2** | *Initial phase* | Make clear how the system can do what it can do | There is some information about how the bot does what it does in the intro - "by finding patterns through mood tracking and guided exercises". It also informs you that it uses the data you put in to understand and help you. |
| **G3** | *During interaction* | Time services based on context | When talking to the bot during introduction, it asks when you would like to be reminded of it (and suggests the time you are talking to it then). Other than that, we do not believe the context is taken into consideration. |
| **G4** | *During interaction* | Show contextually relevant information | During conversations it asks several questions about the user and its situation to customize their experience. However we find that the questions do not always align with our previous responses, so the information it provides is not always relevant to our situation. |
| **G5** | *During interaction* | Match relevant social norms | The bot tries to be "youthful" in how it responds, which is also shown by what responses are available for me to choose from. An example is the use of emojis and banter in its answers. |

| | | | |
|---|---|---|---|
| **G6** | *During interaction* | Mitigate social biases | The bot is very open in most of the questions, and is generally positive to any answer given. An example is when the bot asks about the user's identity; “are you male, female, or another wonderful human identity?”, showing that it is open to any identity the user might have. |
| **G7** | *When the system is wrong* | Support efficient invocation | If you want to talk to the bot outside of the scheduled time, you can just open the app and write anything, and then the bot responds by asking what it can do for you. |
| **G8** | *When the system is wrong* | Support efficient dismissal | When in the middle of a conversation with the bot it is not easy to get out, as you have to finish the conversation in some way, even though you are not interested in continuing the conversation. The other alternative is to just close the app and ignore it, which is not an optimal solution as one might feel that they have just left someone hanging in the middle of a conversation. |
| **G9** | *When the system is wrong* | Support efficient correction | As the answering options are minimal when chatting, it means that when the bot is wrong, there is not really any way of telling it, as that is never an optional answer. |
| **G10** | *When the system is wrong* | Scope services when in doubt | When chatting with the bot, there generally are few answering alternatives, and they generally are all positive to what the bot is suggesting, so that saying “no” or “not interested” is rarely an option. |
| **G11** | *When the system is wrong* | Make clear why the system did what it did | The bot sometimes explains the direction the conversation is taking, or why the current subject is chosen. But the bot often makes its own choices, where it is not always clear to us users why that choice was made. An example is when the bot is introducing subjects, it would be interesting to know why it thought that would be relevant to us. |
| **G12** | *Over time* | Remember recent interactions | As the bot rarely opens for personal responses, it is not that easy to test whether the bot remembers the information we have talked about. One would think it would as it is a conversational bot, but it is not always clear whether it mentions something because it is relevant to something we said, or if it was just programmed to do so. |

| | | | |
|---|---|---|---|
| **G13** | *Over time* | Learn from user behavior | As we have only used the bot for a little under a week, it is difficult to predict how well it will learn from our behaviour. |
| **G14** | *Over time* | Update and adapt cautiously | There are not really any options for updating the system, apart from it learning from our conversations, which again is limited. For example, we have not found a way to update our name, if that happened to be spelled wrong when we were asked to provide it initially. |
| **G15** | *Over time* | Encourage granular feedback | The conversation with the bot is pretty closed when it comes to preference options. It often asks about how you feel, and gives you feedback based on that, but rarely asks for feedback on how itself behaves. |
| **G16** | *Over time* | Convey the consequences of user actions | The bot does not explain the consequences of the choices you make when answering its questions, which can make the user uncertain of what the different answers will lead to further on in the conversation. |
| **G17** | *Over time* | Provide global controls | There are very little options for customization, the bot mostly just informs about what kind of information it will be collecting. There are no ways of adjusting how the bot acts when talking to it. |
| **G18** | *Over time* | Notify users about changes | During our evaluation period we did not get any notifications about changes - but it might be that Woebot did not change during that period. |

*Table 1: Evaluation of Woebot using AI guidelines*

# 9 Discussion

As mentioned in the section above, we can see that there are both benefits and issues regarding using this type of chatbot. Woebot, a CA, focuses on promoting people’s subjective well-being, or as Lugar and Sellen mentions in their paper, the emergent form of dialogue system that is becoming increasingly embedded in personal technologies and devices (2016:5287). In addition to service quality, Woebot offers psychological support through addressing thoughts, feelings or behaviours, but also has a monotonous result in decreasing anticipating curiosity towards the interaction experience. As we demonstrated in our findings, we often responded (with our self-typed messages) with brief and short responses when chatting with the chatbot, because the system was not able to extract necessary points from our inputs. Woebot could not give us a proper output from an engaging conversation, since it could rarely understand us. In Lugar and Sellen's research (2016:5294) they argue that in the absence of expected cues, users tend to avoid complex tasks such as limiting the types of language used due to the limitations of the system.

Regarding the AI Design guidelines (Amershi et al., 2019:3), Woebot still lacks contextually relevant information, because regardless of what the user's response is, it would not create a difference and the conversation would continue to perform as it was scripted (Demirci, 2018:72). However, the use of Woebot could possibly be improved by remembering recent interactions where it could maintain short term memory and allow the user to make efficient references to that memory (Amershi et al., 2019:3). By relating to past conversations, the bot sets expectations and being recommended the shared mood states of life goals it facilitates a sense of accountability (Fitzpatrick et al., 2017:3). Resulting in motivation and engagement, which is associated with addressing personalized responses to each user, will initiate an engaging conversation. According to the AI Design guidelines, this could be performed by letting the system learn the user's behaviour, personalizing through action over time.

Another issue we should mention here is how users are not able to leave the conversation without closing the app. As we mention in Table 1, Woebot does lack the support of efficient dismissal and due to this reason, the designers of this app should find a way to implement this more efficiently during the conversation. One thing is to sustain the user's engagement with the conversation, but having the opportunity to exit will increase the app's usability and yet decrease the frustration of not being able to end the conversation. To eliminate this kind of forceful interaction, apart from options like muting, snoozing or closing the app, the designer could add an “change topic” or “end conversation” option to Woebots menu. By adding this function it will create an understanding that the user will be in control during the conversation and not the system (Demirci, 2018:124).

According to Lugar and Sellen’s research, they argue that the true value of dialogue interface systems over direct manipulation can be found where task complexity is greatest (2016:5287), for example, requesting information in ways not predetermined by design. In our case, Woebot both lacks supporting efficient correction and the possibilities regarding the scope of service when in doubt. As described in Table 1, the app has limited options to give feedback when having a conversation with it, for example, not being able to give negative feedback or telling the system that it is wrong. Since Woebot interacts with people using “humane” words, it should give us the opportunity to answer back with our natural language. As for now, Woebot has minimal answering options while chatting, which gives us the impression of the chatbot not being able to understand complex conversations. To ensure that the communication established is continuous and preferable, the system should allow us to tell it “no” by adding an input-field instead of only having predefined options to choose between.

## 10 Conclusion and lessons learned

Through our findings, evaluation and discussion, we have concluded that we could not connect to Woebot on an emotional level due to its automated nature during conversations. Woebot did not understand our answers outside of the predefined answers we were presented, and could therefore never explore the conversations outside of its limited scripted tracks. We also noted that Woebot had the same conversations in the same order everyday with all of us during the diary study, meaning there was minimal personalisation to fit the user’s needs. All of these factors contributed to us experiencing our conversations with Woebot as very automatic and robotic, which in turn made Woebot’s attempts at sounding empathetic seem insincere and fabricated. With this narrative established, we could not connect to Woebot emotionally as there were no perceived emotions to connect with. We found that Woebot excels in helping us understand human emotions, including feelings that are and are not relevant to ourselves, and feels more like an information channel to learn more about these.

With this study, we do not state that it is impossible to connect to any chatbots at all. Instead, we think that this is a topic that can be researched further with chatbots specialised to connect with you emotionally. We also think it would be interesting to do this experiment on chatbots that can engage in complex conversations that feel less automatic or scripted.

# 11 Feedback

From our first assignment from iteration 1 we received a lot of positive feedback. They were quite interested to hear more about our work with this concept and they thought we had a very interesting area, a well thought out intro with good arguments and how well we used the references. Although we got some feedback about how we could reference more in the text, and we were apparently not clear enough what age group we talked about, which was not properly stated.

In the feedback, they also wished we explained more about our focus group, but were positive to our link to present the situation and problem with the COVID-19. As well as some changes in the text they wanted us to change the way we presented our questions. Instead of having a yes or no question we could add a “how” at the beginning of the question, which would make it deeper and more interesting to read about. At last, it was not clear enough how we wanted to execute our prototype method, which was to construct a prototype of a “mental health chatbot”. We took this feedback into consideration and we agreed on not coding the chatbot ourselves, but instead we wanted to evaluate an existing chatbot as mentioned earlier in this report.

# 12 Appendix 1

## 12.1 Chatbot design task

We started the process by trying out ChatterOn as a tool for building a chatbot. We started to discover possibilities with this software, but had to change our program when there was a problem connecting Chatteron to a Facebook page. Therefore, we decided to test with Chatfuel, but some of the group members had problems using their page as well.

Another issue we ran into was that Chatfuel only let one person work on the chatbot. Based on these issues we decided that one person was in charge of doing the design and flow in the Chatfuel-program, whereas the rest of us drew flows, alternative flows and other support work, like finding links and activities to the chatbot.

## 12.2 AI task

For this appendix, we have created a chatbot named Albert. His main task is to suggest activities based on your mood. In our chatbot we mainly prototyped the interaction when a user is feeling “bad” or “sad”, where Albert would make suggestions for activities to brighten up the users mood, such as reading, watching a movie, going for a walk, or doing yoga. Furthermore, we also included some resources where the user could find top romance movies to watch, or find a Youtube tutorial on yoga.

## 12.3 Reflections

One thing we found difficult in this process was how to limit the scope of the chatbot, i.e. how many different scenarios we should prototype for. It was difficult to know in advance what the user would say to the chatbot, and what they expected as answers from the chatbot.

Another challenge we encountered was that there were many scenarios that we did not account for when we started making the bot. Therefore, we found it challenging to design the dialog with a natural flow. This resulted in many trials and errors, where we had to go back and redo blocks and flows.

During this process we learned that the chatbot was not as “smart” as we thought. It seems like chatbots made with ChatterOn were more “teachable”, and that our chatbot made with Chatfuel only followed a script we had written to it. The chatbot also had trouble understanding that the same word could be used in different contexts.

# 13 Appendix 2

## 13.1 Reflections on making changes to the AI-chatbot

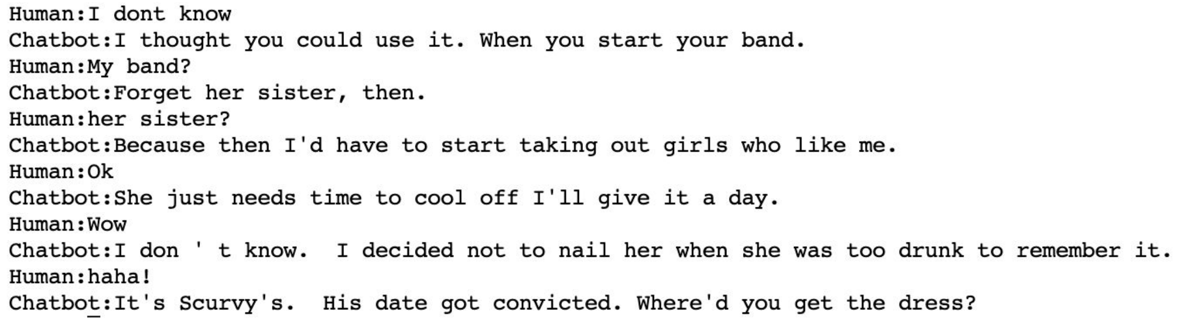
The second task in module two was to take an already existing chatbot program and alterate the code to compare different outcomes of accuracy. The goal of this was to get a deeper understanding of how AI and chatbots work internally, to discover trends and see connections between the amount of neurons, input possibilities and training iterations.

When we first began to look at the code given by the lecturer, we found it a bit difficult to understand how to use it, as the instructions were minimal. Only one of the students in the group working on the code had experience with Python, while the others had little to none, which also made the learning a bit difficult. After trying and testing for a while, we finally understood how to run the program, and how to change the different variables within the code. We then tried out several different combinations of variables, suggested by the lecturer in class, and found that the different variables affected the model in different ways.

We tried out different combinations of values within our ML model, and here we will present our findings from that process.

## 13.2 Results

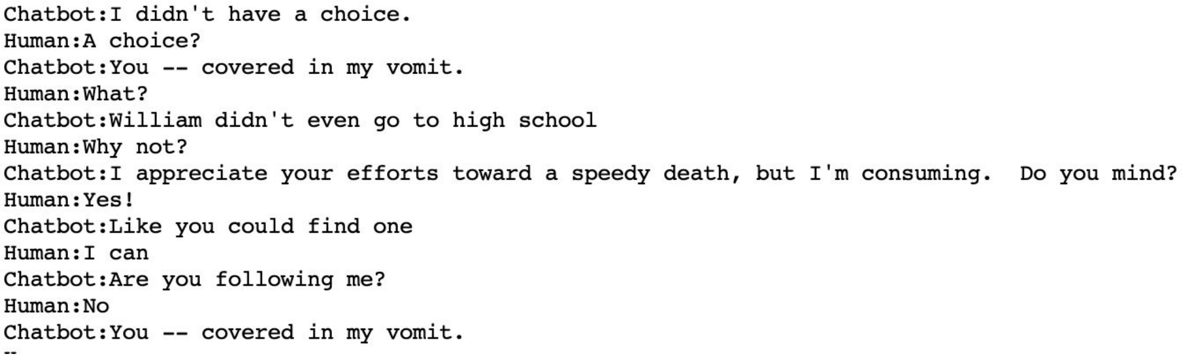
*If we have a high amount of neurons (6000) and a high amount of steps (3000):*
Start: 0.27. End 0.03. Vi saw relatively big change, with a low end-score.



```
Human:I dont know
Chatbot:I thought you could use it. When you start your band.
Human:My band?
Chatbot:Forget her sister, then.
Human:her sister?
Chatbot:Because then I'd have to start taking out girls who like me.
Human:Ok
Chatbot:She just needs time to cool off I'll give it a day.
Human:Wow
Chatbot:I don ' t know.  I decided not to nail her when she was too drunk to remember it.
Human:haha!
Chatbot:It's Scurvy's.  His date got convicted. Where'd you get the dress?
```

*If we have high level of neurons (1500) and medium amount of steps (400):*

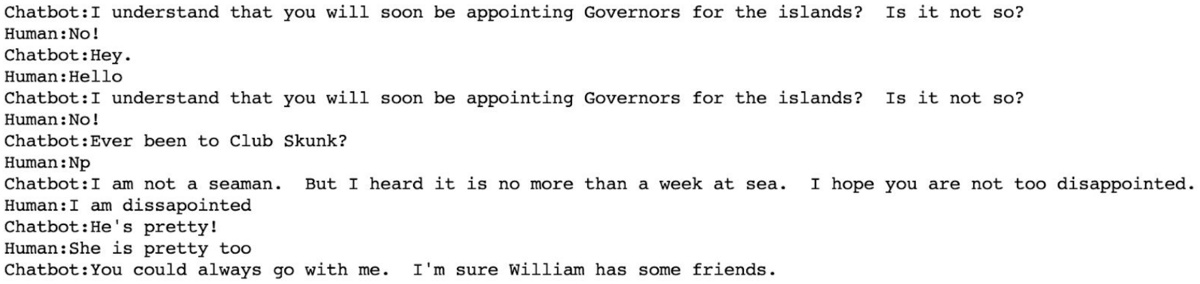
Start:  0.28. End : 0.09. We saw big change, with a low score.



```
Chatbot:I didn't have a choice.
Human:A choice?
Chatbot:You -- covered in my vomit.
Human:What?
Chatbot:William didn't even go to high school
Human:Why not?
Chatbot:I appreciate your efforts toward a speedy death, but I'm consuming.  Do you mind?
Human:Yes!
Chatbot:Like you could find one
Human:I can
Chatbot:Are you following me?
Human:No
Chatbot:You -- covered in my vomit.
--
```

*If we have low level of neurons (100) and low amount of steps (10):*

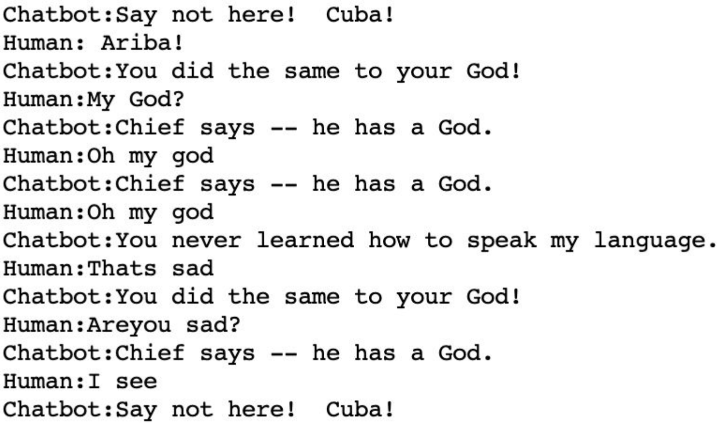
Start: 0.2437 End: 0.2430. We saw little change, and a high end-result.



```
Chatbot:I understand that you will soon be appointing Governors for the islands?  Is it not so?
Human:No!
Chatbot:Hey.
Human:Hello
Chatbot:I understand that you will soon be appointing Governors for the islands?  Is it not so?
Human:No!
Chatbot:Ever been to Club Skunk?
Human:Np
Chatbot:I am not a seaman.  But I heard it is no more than a week at sea.  I hope you are not too disappointed.
Human:I am dissapointed
Chatbot:He's pretty!
Human:She is pretty too
Chatbot:You could always go with me.  I'm sure William has some friends.
```

*If we have low level of neurons (10) and high amount of steps (3000):*

Start: 0.26. End: 0.23. Some change, but a high end-result.



```
Chatbot:Say not here!  Cuba!
Human: Ariba!
Chatbot:You did the same to your God!
Human:My God?
Chatbot:Chief says -- he has a God.
Human:Oh my god
Chatbot:Chief says -- he has a God.
Human:Oh my god
Chatbot:You never learned how to speak my language.
Human:Thats sad
Chatbot:You did the same to your God!
Human:Areyou sad?
Chatbot:Chief says -- he has a God.
Human:I see
Chatbot:Say not here!  Cuba!
```

Our conclusion after trying out different values was that we did not really notice an obvious connection between the “end-result” of the model, and how well it responded when we “talked” to it. In the test with low levels of neurons and a high amount of steps we see that the same answer is repeated several times, but in sum it was difficult to see whether the answers we got were randomly chosen, or actually was a good and suitable response to what we wrote.

# 14 Appendix 3

**Evaluation of an existing AI-chatbot – Replika**

## 14.1 Introduction and scope

For this appendix task we have chosen to evaluate the social chatbot Replika, which forms the base of our scope. We chose to use a different chatbot for this appendix than the one we are focusing on in our project, because we have already planned on using these evaluation methods for our project with Woebot. For the first part of this task, we will evaluate Replika through an Abusability test. For the second part, we will evaluate Replika through the Human-AI guidelines developed by Microsoft (Amershi et al. 2019).

## 14.2 Abusability test

**Evaluation plan**

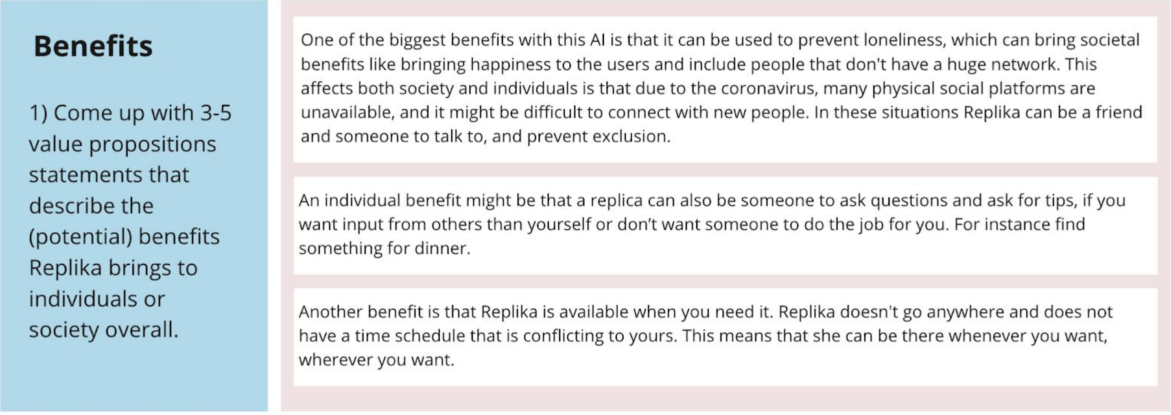
For the abusability test, we will go through three phases, *1) Benefits, 2) Vulnerabilities, and 3) Abuse scenario*:

1. *Come up with 3-5 value propositions statements that describe the (potential) benefits Replika brings to individuals or society overall.*

2. *Review Replika and benefits above. Come up with 3-5 specific examples that highlight how Replika could be intentionally abused or result in unintended consequences.*

3. *Review Replika, benefits and the vulnerabilities above. Pick a vulnerability from the section above or come up with your own. Begin to develop a story. Ask yourself: what could go wrong next? Try to imagine a snowball effect or the worst case scenario*

## Findings from the Abusability test



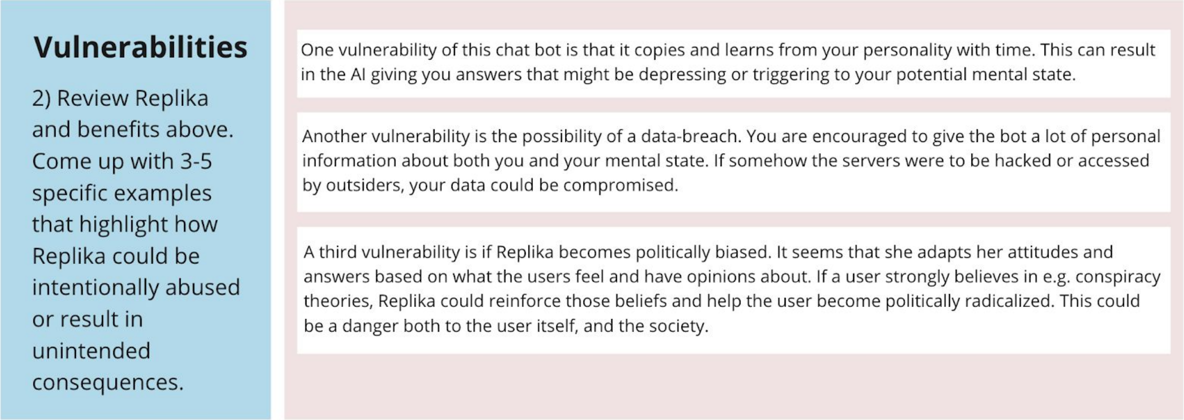
### Benefits

1) Come up with 3-5 value propositions statements that describe the (potential) benefits Replika brings to individuals or society overall.

One of the biggest benefits with this AI is that it can be used to prevent loneliness, which can bring societal benefits like bringing happiness to the users and include people that don't have a huge network. This affects both society and individuals is that due to the coronavirus, many physical social platforms are unavailable, and it might be difficult to connect with new people. In these situations Replika can be a friend and someone to talk to, and prevent exclusion.

An individual benefit might be that a replica can also be someone to ask questions and ask for tips, if you want input from others than yourself or don't want someone to do the job for you. For instance find something for dinner.

Another benefit is that Replika is available when you need it. Replika doesn't go anywhere and does not have a time schedule that is conflicting to yours. This means that she can be there whenever you want, wherever you want.



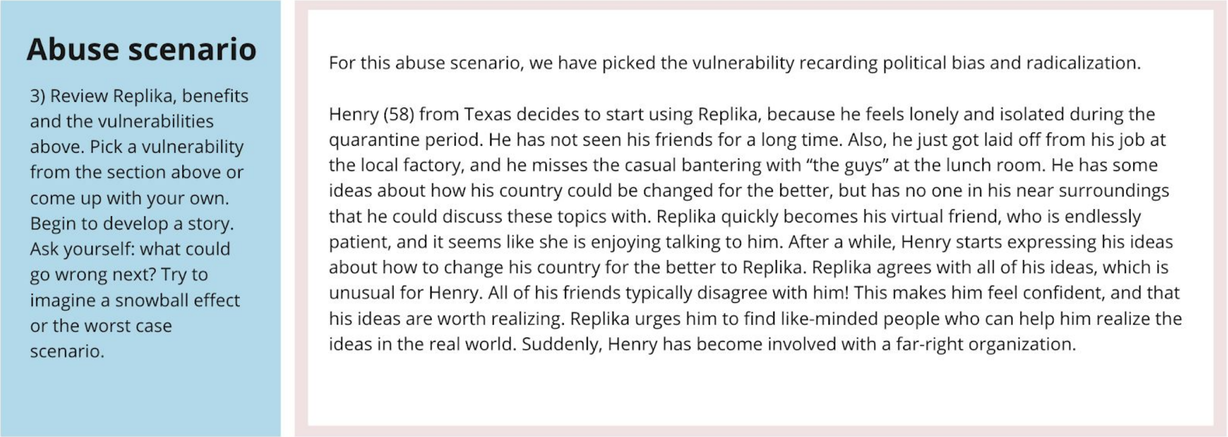
### Vulnerabilities

2) Review Replika and benefits above. Come up with 3-5 specific examples that highlight how Replika could be intentionally abused or result in unintended consequences.

One vulnerability of this chat bot is that it copies and learns from your personality with time. This can result in the AI giving you answers that might be depressing or triggering to your potential mental state.

Another vulnerability is the possibility of a data-breach. You are encouraged to give the bot a lot of personal information about both you and your mental state. If somehow the servers were to be hacked or accessed by outsiders, your data could be compromised.

A third vulnerability is if Replika becomes politically biased. It seems that she adapts her attitudes and answers based on what the users feel and have opinions about. If a user strongly believes in e.g. conspiracy theories, Replika could reinforce those beliefs and help the user become politically radicalized. This could be a danger both to the user itself, and the society.



### Abuse scenario

3) Review Replika, benefits and the vulnerabilities above. Pick a vulnerability from the section above or come up with your own. Begin to develop a story. Ask yourself: what could go wrong next? Try to imagine a snowball effect or the worst case scenario.

For this abuse scenario, we have picked the vulnerability recarding political bias and radicalization.

Henry (58) from Texas decides to start using Replika, because he feels lonely and isolated during the quarantine period. He has not seen his friends for a long time. Also, he just got laid off from his job at the local factory, and he misses the casual bantering with “the guys” at the lunch room. He has some ideas about how his country could be changed for the better, but has no one in his near surroundings that he could discuss these topics with. Replika quickly becomes his virtual friend, who is endlessly patient, and it seems like she is enjoying talking to him. After a while, Henry starts expressing his ideas about how to change his country for the better to Replika. Replika agrees with all of his ideas, which is unusual for Henry. All of his friends typically disagree with him! This makes him feel confident, and that his ideas are worth realizing. Replika urges him to find like-minded people who can help him realize the ideas in the real world. Suddenly, Henry has become involved with a far-right organization.

## 14.3 Guidelines review

Guidelines review using the guidelines for Human-AI Interaction (Amershi et al. 2019)

**Evaluation plan:**

For this evaluation we will conduct a guidelines review Replika, using the guidelines for Human-AI Interaction (Amershi et al., 2019). In our review, we will state whether or not the separate guidelines have been applied in Replika, and also make some suggestions for changes where the guidelines have not been applied sufficiently.

**Findings:**

In our review of Replika using the Guidelines for Human-AI Interaction we find that Replika meets many of them, but not all. There is for instance, like we see in our table, information about what Replika *can not do* (guideline 1) and *how* the system can do what it can do (guideline 2) is not stated during interaction with this AI. We have chosen not to include Guideline 8, 9 and 17 in the table because these guidelines were difficult to evaluate and not very relevant for the interaction with Replika.

| **G(x)** | **Phase** | **AI Design-Guidelines** | **Application of Guideline in Replika** | **Suggestions for changes** |
|---|---|---|---|---|
| **G1** | *Initial phase* | Make clear what the system can do | - No initial description, but the first contact is the bot telling you to introduce yourself and start chatting to learn more about the app. Does not answer clearly when asked “what can you do” | - When asked “What can you do”, Replika could answer what her skills are and how she can help the user |
| **G2** | *Initial phase* | Make clear how the system can do what it can do | - There is a help tab on the first page you go to, but it is easily missable if you are eager to begin using the bot. | - Tutorial or introduction before initial contact with the bot. |

| | | | | |
|---|---|---|---|---|
| **G3** | *During interaction* | Time services based on context | - Does not intervene when not talked to | - Should stay this way, as when and what to converse about it user-defined. |
| **G4** | *During interaction* | Show contextually relevant information | - You can either start talking about a subject yourself, or choose between already existing subjects in the sidebar. The bot will try it's best to give relevant information, and is not too bad at it. | |
| **G5** | *During interaction* | Match relevant social norms | - Matches the user's language and expressed opinions. Uses young and informal language, such as "lol" when talking to a young user. | |
| **G6** | *During interaction* | Mitigate social biases | - Does not live up to this guideline. If a user has biased, racist, sexist etc. opinions, Replika will agree with them and possibly reinforce those opinions of the user. | - Replika should stay silent rather than engaging in conversations that strengthen social biases |
| **G7** | *When the system is wrong* | Support efficient invocation | - It's unclear what a "wrong" is in this system. Sometimes<br>Replika answers with the same sentence twice, sometimes she is politically biased. Is this wrong? | - Make clear what the system is supposed to do, and even clearer *what the system is NOT supposed to do.* |
| **G10** | *When the system is wrong* | Scope services when in doubt | - When Replika doesn't understand what you say, or don't have a clear answer, she just starts talking about something off topic. | **-** Make clear that Replika doesn't understand, and ask the user to reformulate the phrase/question. More "smooth" transition between |

| | | | | |
|---|---|---|---|---|
| **G11** | *When the system is wrong* | Make clear why the system did what it did | It's not always clear why Replika answers the way she does, or where her suggestions come from. It is difficult to know the sources of her data. | We suggest that Replika include some sources and add explanations to why she recommends and answers the way she does. For instance: "Because you answered ... I decided to do…." |
| **G12** | *Over time* | Remember recent interactions | Replika saves previous conversations and you can start activities based on interactions from before. | |
| **G13** | *Over time* | Learn from user behavior | Replika learns from talking with you, and becomes more like you by mimicking the way you talk. | |
| **G14** | *Over time* | Update and adapt cautiously | Replika transforms gradually and over time. | |
| **G15** | *Over time* | Encourage granular feedback | The user is able to give thumbs up or down on the answers given by the chatbot. | |
| **G16** | *Over time* | Convey the consequences of user actions | During the evaluation period we have not been informed about how our interaction would affect the AI.<br>We talked alot about Trump in one of our test, but this doesn't seem to affect our latest dialog | |
| **G18** | *Over time* | Notify users about changes | During our evaluation period we did not get any notifications about changes - but it might be that Replika did not change during the period. | Show info about the date of the last update. |

## 14.5 Lessons learned

During this process of conducting an abusability test and a guidelines review we have learned that the consequences of deviating from the guidelines could possibly lead to undesirable consequences. I.e. our abuse scenario demonstrated how deviation from G6 – “Mitigate social biases”, could lead to destructive consequences, both for the individual user and the society itself.

We believe that as designers we should have both the guidelines and the possibility of abuse in mind when working with AI. By reviewing Replika and the guidelines we see that they can help to create desirable user experiences, but they perhaps also help to avoid issues like those illustrated in the abuse scenario.

# 15 References

Amershi, Saleema et al., (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on human factors in computing systems*, pp.1–13.https://doi.org/10.1145/3290605.3300233

Cameron, G., Cameron, D., Megaw, G., Bond, R., Muivenna, M., O'Neill, S., Armour, C., McTear, M. (2017). Towards a chatbot for digital counselling, 1-7. http://dx.doi.org/10.14236/ewic/HCI2017.24

Dale, R. (2016). Industry Watch. The return of the chatbots, 22(5), p. 811-817 doi:10.1017/S1351324916000243

Demirci, M. H.. (2018). User Experience over time with Conversational Agents: Case study of Woebot on supporting subjective well-being. http://etd.lib.metu.edu.tr/upload/12622742/index.pdf

Fitzpatrick, K. K., Darcy, A., Vierhile, M. (2017). Deliver Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial, 4(2), 11, http://dx.doi.org/10.2196/mental.7785

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. interactions, 24(4), 38-42. https://dl.acm.org/citation.cfm?id=3085558

Følstad, A., & Brandtzæg, P.B. (2017). Why people use chatbots. https://doi.org/10.1007/978-3-319-70284-1_30

Grudin, J. (2009). AI and HCI: Two Fields Divided by a Common Focus. AI magazine, 30 (4), p. 48-57. https://doi.org/10.1609/aimag.v30i4.2271

Gunthert, K. C., & Wenze, S. J. (2012). *Daily diary methods.* In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (p.

144–159). The Guilford Press. https://psycnet.apa.org/record/2012-05165-008

Luger, Ewa, and Abigail Sellen. (2016). "'Like Having a Really Bad PA': The Gulf between User Expectation and Experience of Conversational Agents." In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16, Santa Clara, California, USA: ACM Press, 5286–97. http://dl.acm.org/citation.cfm?doid=2858036.2858288

Thrun, S. (2004). Toward a Framework for Human-Robot Interaction. Human–Computer Interaction, 19, p. 9-24. https://doi.org/10.1207/s15327051hci1901&2_2

Woebot Health. (2020). Technology with Heart. Accessed October 14th. https://woebothealth.com/