Final report - third delivery

IN5480 - Fall 2020

Group 5:

Mariann Gundegjerde, Snorre Ødegård, Thea Aksdal Nordgulen,
Barbro Årnes, Claudia Sikora and Linda Østerberg

Table of content

1 About us

We are a group of 6 students:

Mariann Gundegjerde, margunde@uio.no
Snorre Ødegård, snorreod@uio.no
Thea Aksdal Nordgulen, theaaano@uio.no
Barbro Årnes, barbrora@uio.no
Claudia Sikora, claudisi@uio.no
Linda Østerberg, lindaeo@utio.no

2 Area of interest

Our area of interest has led us in many directions. We wanted to research what the attitude towards AI is among users and how this could change based on the language and behaviour of the AI. Do users have expectations for how an AI should act and what the limitations of these actions are. Are these expectations based on earlier experiences, things they have seen in the media or on TV. We were hoping to get a clearer view of the practice and knowledge among users, and the research done around this topic.

Furthermore, our area of interest is concerned with how an AI behaves, and whether this behavior can be modified to appear more human-like. What is a human-like behavior and how could this be translated to a chatbot or a physical robot? And again, what were users' experience and feelings toward this.

2.1 User group

Due to the chosen area of interest, we wanted to focus the scope on an appropriate user group. We discussed who might be relevant to the chosen area, and decided that young adults/ students would be a suitable user group as informants. Most young adults use chatting services on a daily basis and are familiar with how to communicate online with other humans. In addition to this, many young adults are familiar with the concept of chatbots, which might make it easier to collect valid data from the informants due to reduction of the hawthorne effect. Students are also quite an accessible group, as all the group members are students themselves.

3 Background

Since we live in a society where interacting with machines and AI is becoming part of our everyday life, the emerging field of AI and machine-behaviour has drawn an increased interest the past few years. This is a field not only including machine and data science, but also aspects of sociology, ethnography and phycology. Reflecting on questions such as, what trust do we put in AI, what do we expect and how does the interaction and use affect the society and the people as individuals? (Rahwan et al. 2019).

Research has been done on humans' emotional response when encountering non-human technologies. Shank et al. (2019) writes about how humans " emotionally process the gap between nonhuman technologies and having a mind, essentially feeling our way to machine minds". In other words the interaction between human and machine could be seen as equally as complex as the subjective feelings of the human that is interacting.

People's social responses to interacting with technology has also been of great interest to many. "The media equation" is a theory developed by the research of Clifford Nass and Byron Reeves. This theory claims that people respond socially to computers, the same way they would treat humans - the rules of human-human interaction apply to human-computer interaction (Reeves & Nass, 1996, p. 23). This also holds true for experienced computer users, such as IT-experts and the like - this is an innate reaction beyond our conscious control. Research has also shown that factors such as gender and ethnic stereotypes also apply to AI-systems (Nass & Moon, 2000, p. 81) respond to flattery and that the formulation of error messages affect how "friendly" they find the system to be.

This may be explained by evolutionary psychology (Nass & Gong, 2000, p. 38). To our brains, there is no differentiation between a robot and a live being. Humans have complex cognitive systems dedicated to understanding speech and other forms of incoming communication, which leads to several implications in the design of human-AI systems. For instance, when one fails to be understood, people tend to hyperarticulate their speech. The implication of this with human-AI interaction is that if the AI in question learns from human input, this may lead to the AI learning from speech that is not natural to the person normally. This behavior may likely happen in text-based communication as well, with users simplifying their choice of words and grammar during moments of frustration when a chatbot fails to understand the users' communication.

Personality factors are also mentioned - people generally prefer both people and systems with personalities similar to their own, and have greater levels of trust towards systems they believe are similar to themselves (Nass & Moon, 2000, p. 92). The solution may not be so straightforward as to simply letting users choose a "personality profile", since many people do not necessarily know themselves very well, and are often not aware of what their personality is like. It is important to note that social responses towards technology are more likely to happen the more human-like characteristics the technology possesses (Nass & Moon, 2000, p. 97). This can inform the design of chatbots, depending on what kinds of interactions one wants the user to have - a chatbot meant to assist in tasks and a conversational chatbot for home entertainment may benefit from using this principle in different manners.

"Presence" is a topic of interest to many researchers, referring to "a psychological state in which virtual (para-authentic or artificial) objects are experienced as actual objects in either sensory or nonsensory ways" (Lee, 2004, p. 37). According to Lee (2004), there are three different types of presence: physical, social, and self presence. A user experiences physical presence when said user does not recognize the artificial nature of a virtual object. Social presence is experienced when a user interacts with a virtual social agent and perceives them as real. Self presence occurs when a user interacts with a virtual representation of themselves within a virtual world. Presence is an important concept when developing chatbots, because the feelings of presence affects the psychological fidelity of a system (Sharples & Wilson, 2015, p. 208). Psychological fidelity may affect whether or not the chatbot is taken seriously, which in turn affects how the user behaves with the chatbot.

Developing an AI that responds to human nuances and manners really understanding the intent behind their words, are much more demanding and time consuming than developing an AI able to interact in a litterall and straightforward way. Communicating using this type of direct language would often be perceived as rude if the dialog were performed by two humans, while if it was performed with an AI it might be a question of effectiveness. As ISO 9241-11 states, usability concerns the "Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."

With this in mind we have tried to further explore users expectations and perception of AIs personality related to efficiency and the context of use. And, also tried to compare our findings to what other researchers have found during their work on these topics. Our goal was to see if there were connotations between our findings and the research we have presented above. This comparison and discussion will come later, in *section 7.*

4 Research questions

During our second iteration of the project we changed one of our research questions. The reason for this was based on our findings from the two interviews we have conducted. The previous question was "Is it possible to make an AI more human in the way it acts?". The participants expressed that they saw no need for making an AI-system act in a human-like way, and this made us change our question to "Do users expect/prefer AI to behave more like humans?".

"What are the expectations of how an AI should behave?"
With a rapidly increasing amount of different AI's made for different purposes the expectation for what a given system is able to do, varies depending on the users earlier interactions with similar systems. We want to examine what the expectations of users are and how interacting with AI-systems behaving differently than what they expect affects them.

"Do users expect/prefere AI to behave more like humans?"
We are wondering if users expect or prefer AI to behave like humans in the terms of creating an illusion of the AI having a personality and opinions. Does it for example make it more approachable for new users or is it distracting? Things we could test could be if the language of the AI was more human-like, for examples using dialects and giving the AI more of a unique personality. Here we would also have to examine what makes an AI seem more human and what differentiates for example a conversation with a human and a conversation with a chatbot.

5 Methods

Throughout our project we have gathered knowledge from existing research by conducting a literature review and gathered empirical data by having 2 interviews and 4 user tests. In the next sections we will go through how we conducted and planned our conducted methods.

5.1 Literature review

A literature review is a method where one can acquire the knowledge needed to research an area of interest. This was what we used our literature review to do. We started out with ideas, but wanted to familiarize ourselves within the topic. Mostly we wanted to research the area of expectations towards human-computer interaction. We looked up words and topics that we had discussed in our first brainstorming meeting and that we thought could yield interesting information for our research questions. Some examples of the search terms we used are AI-personality, human trust in computers, chatbot interaction, human-computer interaction and others connected to the same themes. We also got feedback after our first iteration of the project to see if there was any literature within the field of psychology, so we extended our knowledge during the second iteration to also include this topic.

Since there wasn't all that much time to analyse the information we found, we spent most of our time analysing the research we found compared to our own findings and how we saw that the information matched in some instances and were opposites in others. We found that there were many areas where our findings differed from the existing research, and this will be discussed later in the report. Our findings from this literature review is what accumulated into being our reports' background section.

5.2 Interview

Further we wanted to examine how users actually experience interacting with different kinds of AI. The best way to examine this is through interviews, and we ended up conducting 2 in depth interviews. Interviews are an effective way of gathering rich information about users, which may include information about unexpected topics (Lazar et al., 2017, p. 188). Our interview tried to find out what the users expectations were when using an AI, and why they had these expectations. We also had in mind that different users have different knowledge about AI and have different amounts of experience with using AI, so we made sure to customize the questions based on this. The interview guide we created and used can be found in *Appendix 4*, but we also go through the main topics of the interview in the next paragraph.

We started the interview with some basic questions about the user's knowledge surrounding AI, and asked them to describe what they knew, how they thought of the future of AI, and what their earlier experiences with AI was. Further we asked about their expectations of what an AI should do, how it should act and in general what they expect should happen when they use an AI. Lastly we tried asking about how one can make an AI act more human-like, and what they would describe as human-like behaviour. As our findings will describe, this last part made us change our focus a bit. Even though both participants found it interesting to talk about what makes an AI seem human-like, they both agreed that this was not something they expected nor wanted. Therefore, we ended up changing one of our research questions, to focus more on how a user would react if an AI acted differently than what was expected, and we used this change when planning our user test.

5.3 User test

To get some answers to our research questions, we conducted a comparative test. Based on insights from previous interviews, we wished to deepen the understanding by putting the AI we were to examine in a context of use. The purpose of the test was to explore users expectations and perception of AIs personality related to efficiency and the context of use. We wished to see how the language and level of professionality from the chat-bot affect the users perception of use and the helpfulness. This was moderated in a comparative form, conducted with the help of two different versions of a chatbot that provides the user with dinner inspiration. Because of the limited interaction available in the chatbots, the users were introduced to a chatbot through acting out a given scenario in a user test. After this interactive session they were asked questions about their experience and their perception of potential use. The semi-structured questionnaire leading the interview was a combination of open- and closed-ended questions with possibility for complementing. The questionnaire is found in *Appendix 5*.

5.4 Prototypes

We made two different chatbots in Dialogflow, where one acted “normal” and one acted “unexpected”. These were based on our findings from the interviews on what a user expects an AI, or in this case a chatbot, to act like. Both of them have a limited interaction with only one flow because of the amount of work with making two functioning chatbots. They both recommend the same dish to the user, but in a different way for an easier comparison between them. The normal chatbot acts in a professional way - how many expect a chatbot to act, while the unexpected chatbot is more casual and silly - kind of like a human. In the making of the unexpected chatbot we experienced that it was hard to design the personality. It was easy to exaggerate with jokes and hard to find the right balance so that we actually tested what we wanted to test. Screenshots of the prototypes can be found in *Appendix 6*.

5.5 Ethics

We have tried to keep a focus on the ethical ramifications of our work during the different stages of this group project. Making sure that we both collect, analyze and present information in a way we feel accurately convey our findings. But there is of course always room for improvement. Covid-19 has for example presented some practical problems for us concerning our interviews. This has led us to interview and user tests with our participants more remotely instead of face to face as we would usually do. This has both positive and negative implications as some subjects may be more comfortable doing it in this fashion, seen from both the perspective of a reduced risk of infection but also perhaps a reduced social pressure in answering the questions. But seen from the perspective of actually observing the participants in a way where we could pick up on physical cues, the remote interviews have probably had some negative effects. The remote interviews and user tests have also meant that we have not gotten signatures on our consent form as we have not actually met them physically, so we have instead had to rely on their verbal consent to participate in the beginning of the interviews and user test. We feel this together with our effort to anonymize our participants should be enough to say that we have tried to protect our participants. But we are of course aware that it would be better for both us and the participants to be able to have it in writing.

6 Findings

6.1 Findings from interviews

During our first round of interviews, we focused on getting knowledge about different types of AI users, and their experience with using AI. Our questions were open, and let the participant elaborate on their own experiences and thoughts around AI. We saw it as important to remember that not all users have the same knowledge as us, and it was therefore important not to lead the participants in directions we found interesting, but rather let them lead the conversation. We have conducted 2 interviews, and this is not enough to start drawing conclusions of peoples thoughts and expectations towards AI, but it has at least given us some interesting insights in some of the things we wanted to explore. A list of all findings can be found in *Appendix 7*.

The participants of the interviews have been what we call novice users, meaning people that may use AI from time to time but who do not currently use it in a professional setting or have any sort of education that would give them any special insights into AI´s. Examples of AI´s centered more towards novice users that the participants mentioned that they had used during the interviews are Siri, Google Home and a couple of different service chatbots.

It was clear that the participants had different expectations of different types of AI. One participant said that AI usually have little to no emotions and gives you the answer to what you're looking for and nothing more. While another participant said he had different expectations for an AI meant for solving work tasks and one meant for consumer use.

To further examine the participants' expectations of AIs we asked if they were afraid of getting replaced in their occupation in the near future, both of the participants answered that they did not think that AI would be able to replace them. Both stated that though AI can help effectivise a lot of their workload connected to things such as research and filling out paperwork, it would not be able to replace the human element needed in their occupation. In their opinion most AI-systems lack the skills to replace a human-to-human interaction, which is an important part of many occupations.

Perhaps the most interesting finding from our interviews was when we asked our participants how they would go about making an AI more human-like. To our surprise both of the participants answered they did not see a reason why there was a need for this. For example, why should a chatbot communicate with us in a more human way, when all they wanted was for the chatbots to be quick and effective, and do the task at hand. The participants stated that having to talk with a chatbot in the same way you would talk with a human would not be effective.

6.2 Findings from user tests

All of the participants expressed that they preferred the first chatbot when looking for dinner inspiration. They reasoned that by describing it as more professional and some said it was because it thereby was “easier” to conversate with. We also found that when comparing the two chatbots, addressing chatbot 2, the participants focused more on the personality of the bot, which some of them did not like, instead of its actual functionality. When they talked about chatbot 1 they talked more about the content in the sentences and came up with serious purposes for improvement for how it could be even more helpful. There were also indications that the users have different kinds of expectations for the more professional chatbot, they did not accept e.g. spelling mistakes that they did not notice or comment on in chatbot 2. Some expressed the importance of balance with the chatbot not being overly engaging, because this can be distracting and annoying as they are using the bot only to get dinner inspiration and nothing more. The context for use is thus very important for the personality of the chatbot. It is also important to hit the correct target group with the personality, and it can therefore be risky to not give the chatbot a general language.

7 Discussion

“The media equation” (Reeves & Nass, 1996) claim that people respond the same way to computers as they would to humans. Based on our findings this seems unlikely, since the participants from the interviews clearly stated that they mostly used AI to solve simple tasks and that they usually used a simple and straightforward language. To answer our own research question based on our findings, users don’t expect an AI to behave like a human, and it’s likely to say that humans would then also respond differently to AI than they would to humans. AI lacks the skill to replace human-to-human contact one participant said, which means one could not compare this to human-to-computer interaction. As we only talked to young novice users, with little experience using AI, this may not apply to the bigger picture of human-computer interaction. Though, it is interesting seeing how simple the expectations were for the participants.

Nass and Moon (2000) talk about people wanting systems to have personalities similar to their own, while our participants wanted the AI-system to be effective and straight to the point. One participant mentioned that when they used a service chatbot they always altered their language, instead of expecting the chatbot to accommodate their behaviour. As explained by evolutionary psychology (Nass & Gong, 2000), humans have the ability to adapt their way of interacting, and even though our brain may only register the interaction with a chatbot as a "normal interaction", the mind also adapts because of the already established expectations the user has.

A participant from the interviews described talking to an AI as "talking to a person in a different language than your mother tongue", . Also, the participants of the user test disliked the chatbot that we tried to create to mismatch expectations, yet they still had issues with the more "normal" chatbot. Which shows that there is a thin line between what users will accept from an AI. Even though they expect to interact "like they're talking to a person in a different language", the interaction can't be too complicated or distracting. It seems like the participants we have tested and talked to have most interest in an effective interaction with an AI, and their usual context of use is a setting where they have a request and want a quick answer, not a conversation.

Lastly, Lee (2004) describes the concept of presence, and how people perceive virtual objects. In our case it can be compared with how our participants experienced our chatbot prototypes. Chatbot 1 was created to act closely to a normal person, with a happy and helpful attitude. This was in a way an attempt to achieve social presence, and could be part of the reason why it was the chatbot the participants liked the best. Chatbot 2 was also created with a human-like manner, but it was not in the way the participants expected. It could be that this chatbot also achieved a sense of social presence, but at the same time it wasn't taken seriously by the participants and therefore it achieved a low level of psychological fidelity.

8 Conclusion

During our project we have read a lot of literature with discoveries we didn't find to match our own research. If we were to study this further, it would be important to conduct more research on our own to really see if our findings were accurate. It is also important to mention that we were all motivated and excited to study our research questions, but due to Covid-19, it was difficult to conduct data gatherings with users and we wished we could have presented more data to back up our findings. The topic of what users expect from AI will in our opinions need a lot more research and we can see that the general expectations will also change as AI becomes more commonly available to the average user. We didn't initially describe our user group as novice users, but due to our findings it became clear that they didn't have a lot of experience in using AI. If we were to research this further, it could also be useful to conduct our interview and user test on more experienced users, who most likely would have stronger expectations and maybe a bigger variation of expectations.

9 References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM.

Goodwin, Dr Morten. "*Interacting with Artificial Intelligence*" (2020). Centre for Artificial Intelligence Research. University of Oslo.

International Organization for Standardization. (1998). "*Ergonomics of human-system interaction. Part 11, Usability: Definitions and concepts*". ISO 9241-11. New York: American National Standards Institute.

Karahasanovic, A. Interacting with AI Module 3. (October 27, 2020). SINTEF. Sustainability & Design Lab UiO. URL: https://www.uio.no/studier/emner/matnat/ifi/IN5480/h20/Undervisningsmateriale/in5480-module-3---27-november-2020-amela-%281%29.pdf.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). "*Research methods in human-computer interaction*". Morgan Kaufmann.

Lee, K. M. (2004). "*Presence, explicated*". Communication Theory, 14, 27-50.

Nass, C., & Gong, L. (2000). "*Speech interfaces from an evolutionary perspective*". Communications of the ACM, 43(9), 36-43.

Nass, C., & Moon, Y. (2000). "*Machines and mindlessness: Social responses to computers*". Journal of social issues, 56(1), 81-103.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, JF., Breazeal, C., Crandall, J., Christakis, N., Couzin, I., Jackson, M., Jennings, N., Kamar, E., Kloumann, I., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D., Pentland, A., Roberts, M., Shariff, A., Tenenbaum, J., Wellman, M. (2019). "*Machine behaviour. Nature". * Vol.568 (7753), p.477-486

Reeves, B., & Nass, C. I. (1996). "*The media equation: How people treat computers, television, and new media like real people and places*". Cambridge university press.

Shank, B., Graves, C., Gott, A., Gamez, P., Rodriguez, S. (2019) "*Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence".* Computers in Human Behavior, Volume 98, Pages 256-266,

Sharples, S., & Wilson, J. R. (2015). "*Evaluation of human work*". CRC Press.

Appendix 1: Chatbot design task

An assignment during module two was to create a chatbot. It was a two weeks task and the group decided on creating a bot for dinner-inspiration. We ended up making a chatbot we called “Middagshjelp” (Dinner help).

The process

As mentioned, the purpose of the chatbot was to provide inspiration for choosing what to eat for dinner and the response from the chatbot was thought to be based on the user's personal preferences. After agreeing on the purpose of the chatbot the group performed a joint brainstorming concerning basic functionality necessary for a first prototype. The key tasks, like greeting the user, explaining functionality, asking for preferences and recommending a dish, were identified and some basic flows were sketched. At first all group members had a try at working with their own chatbot to get an understanding and some inspiration, for thereafter discussing further implementations to be made on the prototype chatbot. Amongst available tools for developing chatbots, different group members tested Chatteron and Dialogflow. The group decided on Dialogflow since we found it to have a more intuitive developing-UI and that it did not require connecting the bot to a Facebook page. The chatbot was tested and altered during development to improve user interaction. Some of the adjustments made were dividing tasks into separate steps and improving the language.

Reflection

We found the assignment interesting and ended up with a chatbot we were happy about. One frustrating part of the project was to cooperate on making the chatbot. Because of the current situation with the pandemic, many of the group members work from home and are used to having the ability to cooperate on tasks online. There were minimal options for cooperating on the chatbot-task, since everyone had to work on their chatbot individually. So we ended up letting just a few members of the group finish the chatbot, and then share the result with the rest of the group to give feedback. This worked out well in the end, but it would have been more instructive and educational if more of the group members could have worked more closely together to create the chatbot.

Lastly we reflected on the limitations we had when making our chatbot. Our finished product was much easier than what we first prototyped and discussed. The chatbot only had one path, and a very predetermined path. This of course is due to our abilities, and the abilities of Dialogflow. Although this was frustrating at times, it gave us the basic knowledge of what goes into developing a working chatbot.

Appendix 2: Machine learning task

In this appendix, we explored different capabilities of the python-script 'MovieChatbot.py' and the accompanying text-file 'movie_lines.txt'. The aim of this assignment was to customize our own model of the Chatbot-script, by appending or subtracting the number of neurons, as well as changing the input-text. We did this by editing the script in JupyterLab provided by Cair-hub. Throughout this assignment, we got to explore the different functionalities of machine learning by using the programming language Python and knowledge conveyed in lecture on the subject "Interacting with Artificial Intelligence" (Goodwin, 2020).

The process

We started the process by preparing the JupyterLab with copy and paste from the assignment description. The next step in this process was to understand how the script and the text file was connected and what kind of interaction we could expect when executing 'MovieChatbot.py'. By exploring the default version of the program, we got to understand what the different numbers presented to us in the terminal meant, which was the starting point of getting to know the algorithm. We noted the results of the first execution and discussed expectations of how changed in the script would affect the flow of the system.

The very first change we made was editing the amount of neurons in the function 1 (fc1) and function 2 (fc2). We decided that we would start by appending even more neurons to the functions, and discovered that the loss decreased from 0.23 to 0.20. In the lecture, we learned the loss should aim for being as close to zero as possible and accordingly we concluded that the decrease in neurons increased the loss of the algorithm. We increased the number of neurons by adjusting the algorithm multiple times, where the outcome confirmed our assumption. An increase in the number of neurons would make the algorithm gain more knowledge through each iteration.

An interesting aspect of the discovery of the increase in loss, was how adding more neurons when reaching a significant number (in our case 100 000) made a smaller impact on the learning of the algorithm. There was moderately variety in how the chatbot responded as well through the increase, but this observation might be biased by us not knowing what the algorithm actually does. We wanted to explore the concept of overfitting, but as the amount of neurons was increased we did not see the loss increasing again after decreasing. There is probably a way to provoke overfitting in this particular case, but we did not figure out how to make this happen. Maybe by adding more layers?

Further we tried to experiment by decreasing the amount of neurons. The findings of the process was how having quite few neurons would affect the learning curve of the algorithm by making it vastly shallow. The first loss number would be similar to the results of the algorithm handling many different neurons, but the number would not increase throughout the iterations of the system. By discussing this change and through knowledge about machine learning, we assume that this shallow learning curve is caused by fewer progrations and paths through the network of layers. Though the response of the chatbot was not perfect in the first place (with many neurons), the appearance of the chatbot was even more off and random now than ever.

Reflection

In this assignment, we addressed the challenge of trying to configure a system which deals with a machine learning algorithm. The main challenge of the exploration of the system was understanding what was really happening. We tried to understand how the algorithm was learning by revisiting lectures, watching videos on the subject and doing research online. We felt like we got to understand the concept of machine learning, but not how it was done in practice. This is a challenge mentioned in the curriculum, where a significant amount of research on this field is based on designers not really understanding the capabilities of artificial intelligence and machine learning. Thus we understood more by doing research, but it was challenging to understand exactly what was happening without the technical competence of machine learning.

Appendix 3: Evaluation

In this Appendix we have evaluated a chatbot named ‘Mats’ provided by the webpage and company ‘MatPrat’ where the theme is cooking - relating to our two prototypes. We have sketched an evaluation plan which will evaluate different aspects of the chatbot through an examination of whether or not the chatbot satisfies the 18 Guidelines presented in Amershi et al (2019). In the next section, we will take the chatbot through an abusability test (Karahasanovic, 2020, p.36) where we will identity benefits and vulnerabilities of the chatbot, as well as describing a potential abuse scenario which could occur in the use of ‘Mats - the chatbot of MatPrat’.

Findings

In the evaluation in satisfaction of the 18 guidelines, we went through the 18 guidelines step by step through the phases of ‘Initially’, ‘During interaction’, ‘When wrong’ and ‘Over time’ (Amershi, 2019, p.3). When you first enter the website which provides the gateway for chatting with Mats through Facebook’s Messenger, MatPrat provides a list of what the system is able /not able to perform. It also informs the user that Mats is not completely flawless yet, and describes what challenges this might promote. Thus, we conclude that the chatbot satisfies guideline 1 and guideline 2 in the Initially-phase (Amershi, 2019, p.3).

Mats also partly checks Guideline 4 in presenting different recipes and meal-related videos to the user as well as following social norms as in guideline 5. At the same time, Mats do not fulfill the potential of these guidelines perfectly, but in a sufficient way responding in a correct manner if the context is not quite complicated. The chatbot is quite polite and appears like it wants to help the user. Guideline 3 is also fulfilled in a way where Mats only replies when the user is sending a text to him. From our experience, Mats also ‘Mitigate social biases’ (Amershi, 2019, p.3) which is Guideline 6.

You can also trust Mats in providing support for efficient innovation, which relates to guideline 7 (Amershi, 2019, p.3). The chatbot is available at all times through the chatting service Messenger. In our experience, it is also easy to reject “wrong” suggestions done by Mats by ignoring them or asking for new suggestions. This relates to both guideline 8 and 9 (Amershi, 2019, p.3), which we think the chatbot handles relatively well. At the same time, Mats does not really provide any information why the system did what it did in providing answers to the user except telling the user that it is a popular suggestion, meaning there is a slightly insufficient approach to guideline 10.

In the phase ‘Over time’ we have uncertainties in how the algorithm of Mats really works. There are filters which you can apply to filter what content you wish for the chatbot to provide. There might be signs of Mats being able to remember interaction, but this might just be a coincidence since it does not happen frequently, which applies an uncertain and partly satisfaction of guideline 12 and guideline 13 (Amershi, 2019, p.3).

There is not a clear indication whether or not Mats is continually updating and adapting itself, so we conclude that guideline G14 is not completely implemented in this chatbot (Amershi, 2019, p.3). We also felt like we could not evaluate if Mats did “...convey the consequences of user actions” (Amershi, 2019, p.3), due to us not getting any feedback in case of adaption or updates, which is Guideline 16. During the use of Mats we did not get any notification about changes, related to guideline 18, but the webpage does not provide any information if the chatbot offers this service. Lastly, Mats do provide global controls as in Guideline 17, where it allows the user to put on filters for the interaction with ‘Mats - the chatbot of MatPrat’.

The benefits with Mats are that he can inspire and motivate users to make different and new meals, so that they can learn to make and also taste new dishes. Mats also helps with using up ingredients, something that can result in less food waste and helping the environment. Users get to save favorite recipes, making Mats better at coming up with personal suggestions to each individual user and they can thus save time when finding out what to cook. Users may therefore get better at cooking after using Mats for a while.

The vulnerabilities with Mats are that there is a lack of filters for for example vegetarians and muslims. This can make them feel undesired, that their preferences to food are not important enough or even that they are discriminated against. After using Mats for a while, the user can also get obsessed with eating a particular way, for example healthy or unhealthy, because Mats will learn what the user prefers over time and suggest more similar dishes. There are also privacy concerns that the chatbot is connected to Facebook: as a guest you agree that Facebook is collecting, using and sharing data from the interaction with Mats, and with logging in to Facebook MatPrat can see all the info that is made public on your Facebook profile.

Abuse scenario: Mats learns over time that Sophie likes to eat healthy and therefore recommends more and more of this to her. She starts to get obsessed with eating this way and Mats is excessively trusted by her being the one that chooses what she eats for every meal. Sophie doesn't want recommendations from anyone else anymore because she is now addicted to Mats, feeling a big pressure to only eat particular food and starts developing an eating disorder. One day Sophie experiences problems with the WiFi and her phone is nowhere to be found. She goes in panic mode because she is dependent on Mats, she doesn't want to improvise with the meals and she therefore refuses to eat. Sophie ends up fainting and going to the hospital, and has to get treatment for a long time to recover from the disorder.

Reflection

When we evaluated the chatbot with the AI design guidelines, we noticed that there is a lot to think about in relation to the design of a chatbot to include all guidelines. A lot of the potential issues with the interaction are “hidden” and only appear in some few use cases. The chatbot had more vulnerabilities than expected, maybe because these are not so clear at first glance. You had to talk with the chatbot for a while trying out different kinds of words, sentences and buttons, putting yourself in the perspectives of potential users and reading the terms closely. The abuse scenario shows that things can go surprisingly bad and is not something you really think about when using AI. We therefore thought this was very interesting and eye-opening.

Appendix 4: Interview guide

Semistrukturert intervju

Hva vi ønsker å finne ut av gjennom dette intervjuet:

- *What are the expectations of how an AI should behave?*
- *Is it possible to make an AI more human in the way it acts?*

Introduksjon:

- Hvis jeg sier AI/Artificial intelligence/Kunstig intelligens - hva tenker du på da?
    - Hvordan syns du AI fremstilles i media?
        - Føler du dette påvirker synet ditt på AI?
    - Noen tv-serier eller filmer du har sett der ai har vært i fokus?
        - Hvordan var AI i denne serien/filmen?
            - Er dette realistisk etter din mening?
        - Føler du dette påvirker synet ditt på AI?

- Hva er dine erfaringer med AI?
- Hvor ofte har du opplevd å bruke/interagere med en AI?

- Hva liker du å bruke AI til? Hvorfor?
    - Hva liker du ikke å bruke AI til? Hvorfor?

- Hva er det du har lyst til å jobbe/jobber med?
    - Tenker du at AI kan erstatte deg i denne stillingen i nær fremtid?
        - Hvis nei
            - Hvorfor ikke?
            - Er dette en av grunnene til at du valgte yrket?
        - Hvis ja
            - Hvordan ser du for deg at AI kan erstatte deg?
            - Hvorfor ser du for deg at dette er en mulig fremtid?

Forventninger til oppførsel:

- Har du noen spesielle forventninger til hvordan en AI skal "oppføre seg"?
    - Hvordan vil du beskrive annerledes/uforventet oppførsel fra en AI?
    - Hvordan tror du at du hadde reagert på en AI som oppførte seg utenfor det du forventet?

- For eksempel en chatbot:
    - Hvilke forventninger har du til en chatbots oppførsel i forhold til når du snakker med et menneske?

Hvordan gjøre AI mer menneskelig:

- Syns du det er vanskelig å vite/se forskjellen på et menneske og en AI?

  - Hva vil du si er de største forskjellene på det å interagere med et menneske og en AI?
- Hvordan ville du endret en AI for å gjøre den mer lik et menneske?
  - For eksempel en chatbot:
    - Hvordan ville du endret språket for at den skulle virke mer som om du snakket med en ekte person?

Avslutning:

- Takk for at du ville være med på intervjuet vårt.
  - Noen flere kommentarer eller tanker rundt AI du vil nevne?

Appendix 5: User test questionnaire + findings

Test nr: 1

Previous experience of chatbots: Ikke så mye, i banken og kundeservice hos telefonselskaper og ulike typer kundeservice

Alder, kjønn: Kvinne 31

Questionnaire:

What are your thoughts about the interaction with chatbot 1?	**Skrivefeil irriterende, virker litt dum, men fikk gjort jeg skulle**				
What are your thoughts about the interaction with chatbot 2?	**Litt morsom, også litt cringy,**				
Please describe your perception of the following. Where 1 is negative and 5 is positive.	**1**	**2**	**3**	**4**	**5**
General experience from interaction with bot 1			**X**		
General experience from interaction with bot 2			**X**		
Helpfulness of bot 1					**X**
Helpfulness of bot 2					**X**
Likeliness I would use bot 1	**X**				
Likeliness I would use bot 2	**X**				
Which chatbot do you prefer? **Why?** **Hva tenker du om en bot med personligt språk hvis det hadde passet deg bedre?**	Den første hvis den hadde vært litt smartere. Den andre var mest cringy og jeg følte den var useriøs. Snakket som en ung gutt så følte meg ikke helt i målgruppa. Føler ikke det er relevant hvis jeg vil ha en matoppskrift.				

Test nr: 2

Previous experience of chatbots: Kundeservice, liker mer å snakke i telefon

Alder, kjønn: kvinne 31

Questionnaire:

What are your thoughts about the interaction with chatbot 1?	Virker fint den, men kan jeg bare svare det du sa? Mye tekst og litt unødvendige spørsmål. Den trenger jo ikke å spørre hvis jeg vil ha spørsmål - det er jo derfor jeg bruker den. Burde spørre hvilke ingredienser jeg vil bruke ikke hva som er min favoritt. Ikke helt så intuitivt hva jeg skal svare				
What are your thoughts about the interaction with chatbot 2?	Jeg vet ikke. Føler jeg ble irritert!				
Please describe your perception of the following. Where 1 is negative and 5 is positive.	**1**	**2**	**3**	**4**	**5**
General experience from interaction with bot 1			x		
General experience from interaction with bot 2			x		
Helpfulness of bot 1				x	
Helpfulness of bot 2				x	
Likeliness I would use bot 1			x		
Likeliness I would use bot 2	x				
Which chatbot do you prefer? **Why?** **Further questions/suggestions?**	Den første. Den andre var mest tullete. Føler ikke den treffet meg helt.				

Test nr: 3

Previous experience of chatbots: testet en psykolog bot, kundeservice

Alder, kjønn: Mann 27

Questionnaire:

What are your thoughts about the interaction with chatbot 1?	It worked, maybe it could be shorter questions? Is there more examples than tikka? Can I ask for something else if i don't like that?				
What are your thoughts about the interaction with chatbot 2?	I liked the first one better i think, but this was a bit funny with the "favorite animal"				
Please describe your perception of the following. Where 1 is negative and 5 is positive.	**1**	**2**	**3**	**4**	**5**
General experience from interaction with bot 1				x	
General experience from interaction with bot 2			x		
Helpfulness of bot 1				x	
Helpfulness of bot 2			x		
Likeliness I would use bot 1		x			
Likeliness I would use bot 2			x		
Which chatbot do you prefer? **Why?** **Would it be more okay if it was funny if it was some other type of bot?**	Maybe nr 1, but did they do the same thing? Nr 1 had longer messages but it was more explaining. Nr 2 was a bit funny sometimes but i think it might be annoying if the conversation was longer. Maybe i just want a quick meal, that's why I ask a bot instead of using google by myself. Yes if it was a "friend-bot" or so, but then it had to be really funny and I guess thats hard to know.				

Test nr: 4

Previous experience of chatbots: Very little, service bots (flights, opening hours), try to avoid it because I like to talk over the phone instead

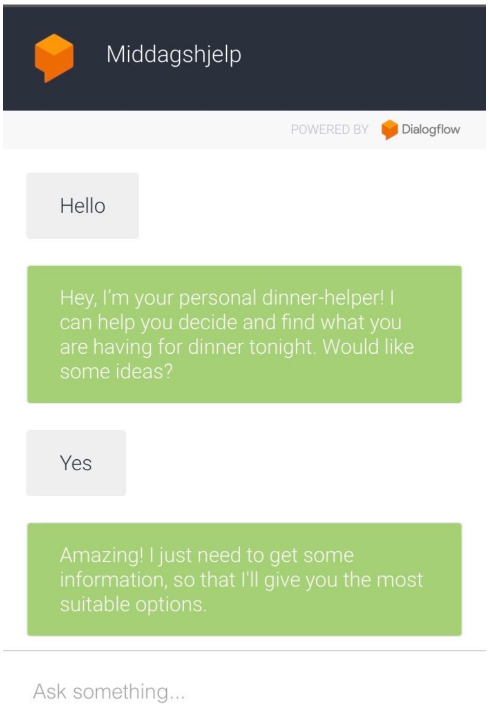
Alder, kjønn: Kvinne, 30

Questionnaire:

What are your thoughts about the interaction with chatbot 1?	Hyggelig og prøvde å være personlig engasjerende mest positiv, men ikke for mye jippi, støtte og overengasjement - balanse er viktig, ikke for mye tekst				
What are your thoughts about the interaction with chatbot 2?	Latterlig på en negativ måte, funker dårlig å være in the hood kompis over chat, mindre flyt, skeptisk, får man et seriøst svar når det er så mye tull?, distraherende				
Please describe your perception of the following. Where 1 is negative and 5 is positive.	**1**	**2**	**3**	**4**	**5**
General experience from interaction with bot 1				x	
General experience from interaction with bot 2		x			
Helpfulness of bot 1				x	
Helpfulness of bot 2			x		
Likeliness I would use bot 1				x	
Likeliness I would use bot 2	x				
Which chatbot do you prefer? **Why?** **Further questions/suggestions?**	1 Mer imøtekommende og positiv, bedre å opprettholde en dialog med, mer seriøs, virker som den ønsker å gi brukeren et bra resultat Kontrasten var litt vel stor, så tilbakemeldingene blir kanskje ikke så nyansert, for lett å like den ene og for lett å ikke like den andre				

Appendix 6: Prototypes

Screenshots from prototype 1:

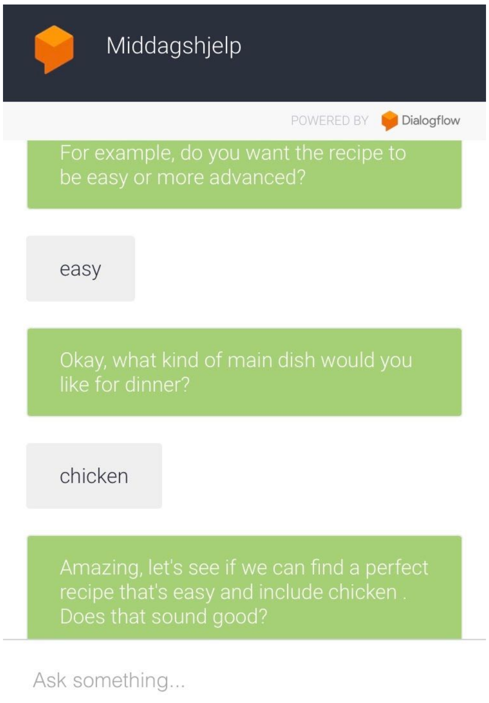


Hello

Hey, I'm your personal dinner-helper! I can help you decide and find what you are having for dinner tonight. Would like some ideas?

Yes

Amazing! I just need to get some information, so that I'll give you the most suitable options.



For example, do you want the recipe to be easy or more advanced?

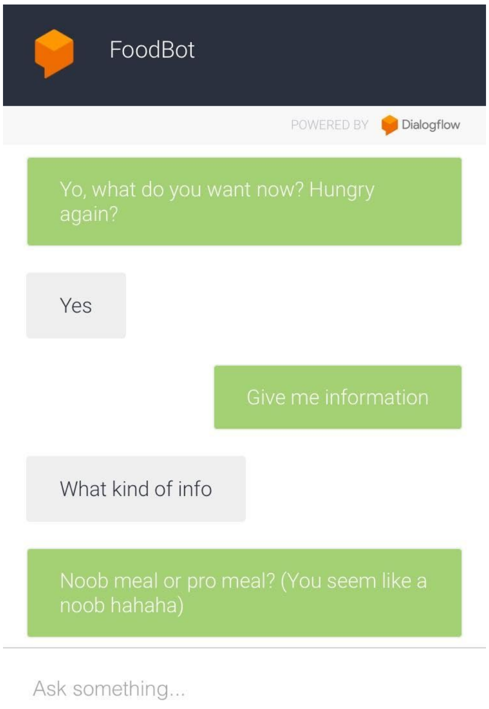
easy

Okay, what kind of main dish would you like for dinner?

chicken

Amazing, let's see if we can find a perfect recipe that's easy and include chicken . Does that sound good?

Screenshots from prototype 2:



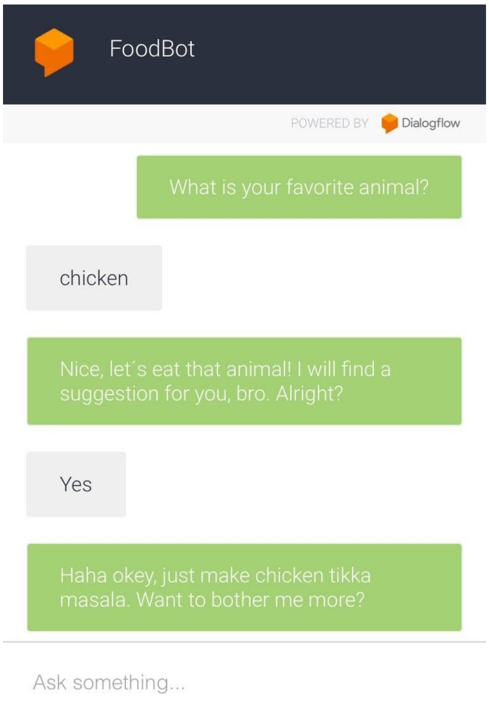
Yo, what do you want now? Hungry again?

Yes

Give me information

What kind of info

Noob meal or pro meal? (You seem like a noob hahaha)



What is your favorite animal?

chicken

Nice, let´s eat that animal! I will find a suggestion for you, bro. Alright?

Yes

Haha okey, just make chicken tikka masala. Want to bother me more?

Appendix 7: Detailed findings from interview

Spørsmål	**Intervju 1**	**Intervju 2**
Generelt erfaringer og tanker rundt AI	-Har brukt Siri, Google Home - Ser på disse som troverdige kilder -Beskriver dem som snakkende versjoner av Google -Tror man bruker AI mer i hverdagen enn man tenker over -Tror ikke roboter kan ta over verden, men det kan være med på å effektivisere arbeid -Ser på AI først og fremst som en hjelpefunksjon -Liker at service-chatboter kan henvise deg videre - føler sjeldent at de kan tilby den informasjonen som man faktisk ser etter	-Har brukt brukt enkelte chatboter. Ikke så fan av AIer som siri og google home. -Finnes 2 typer AI. Et smart program og en faktisk kunstig intelligens.
Fremstilling av AI i media og på film	-AI fremstilles alltid i filmer og serier som noe som skal overgå mennesker og gå imot sin skaper/leder.	-Synes medier skriver litt misvisende og sensasjonelt om det. -Har sett “Her” og “Ex Machina”
Kan AI erstatte deg på arbeidsplassen	-Ser ikke for seg å bli erstattet at AI i jobben som sykepleier. -AI kan ikke erstatte menneske-til-menneske kontakt. Og syns ikke AI har evnen til å håndtere etiske problemstillinger. -Mange problemstillinger i arbeidsdagen som kan være	-Tror ikke AI kan erstatte arbeide han gjør. Dette fordi AI mangler de menneskelige kvalitetene knyttet prioritering som trengs for yrket. Men at den kan gjøre jobben lettere ved å fullføre enkelte oppgaver mer effektivt enn et menneske hadde klart.

	vanskelig å programmere på forhånd.	
Forventninger til oppførsel	-Skal svare på det du lurer på	-Vil at AIer ment for arbeid skal være så effektive som

	-Nøytralt og enkelt språk, med lite engasjement - kan beskrives som monoton	mulig. Mens AIer ment for hjemmebruk som google home og sånt kan ha litt mer personlighet.
Hvordan gjøre AI mer menneskelig	-Hører at Siri er kunstig, er noe med måten setninger blir sagt på. Høres ut som et kunstig menneske -Når jeg snakker med AI bruker jeg kortere setninger og snakker tydeligere og roligere. -Blir litt som å snakke med en person på et annet språk. -For å gjøre AI mer menneskelig burde man endre stemmen så den ikke høres så robot-aktig ut. Burde være mer avslappet og mindre service-vibes. -Vet ikke hvordan en chatbot kan gjøres mer menneskelig - har allerede forventinger til hvordan man snakker med en chatbot og tilpasser meg ut i fra det. Skriver ikke lange kompliserte spørsmål. -Trenger den i det hele tatt å bli gjort mer menneskelig. Den skal bare gi informasjon, det trenger ikke å være en lang samtale.	-Synes ikke det er nødvendig. Fordi det ikke er en hensiktsmessig prioritering med tanke på hva AIs styrker faktisk er.