



IN5480 - Specialization in research in design of IT
“Interaction with AI”

Exploring the use of AI in mental health treatment through the chatbot Prate-Petra

Kathinka Olsrud Aspvin, Sigurd Rognhaugen and Siri Dølvik Reder

University of Oslo
Autumn 2018

1. Introduction	2
1.1 About the group	2
1.2 Introduction	2
1.3 Background	2
1.4 Research Questions	3
2. Process	3
2.1 Methods	3
2.2 Ethics	5
2.3 Timeline	5
3. Prototype	5
3.1 Sketches	5
3.2 Prate-Petra chatbot	6
4. Evaluation	6
4.1 Plan	7
4.2 Reflection	8
5. Feedback	8
5.1 User Feedback round I	8
5.2 User Feedback round II	8
6. Findings	9
6.1 Thoughts on our research questions	9
6.2 Expert Interview	11
7. Reflection	11
8. Conclusion	12
9. References	12
Appendix 1: Conversational interaction	16
Appendix 2: Machine learning	18
Appendix 3: Problems with AI	19
Appendix 4: Human-machine partnership	20
Appendix 5: Timeline	22

1. Introduction

1.1 About the group

The group consists of three members, all currently enrolled in the Master's programme for Informatics: design, use, interaction. We have relatively varied backgrounds and experiences within informatics and design, but none of us have worked extensively with AI previously.

Siri Dølvik Reder finished her bachelor's in informatics: design, use, interaction at the University of Oslo four years ago. She is 33 years old, has two kids and works part-time as a prison officer. *Sigurd Rognhaugen* finished a bachelor's in informatics: design, use and interaction this spring. He's 23 years old and works part-time as an interaction designer. *Kathinka Olsrud Aspvin* also finished her bachelor's in informatics: design, use, interaction this spring, with a specialization in law and eGovernment. She is 26 years old, and works part-time in the administration at the Faculty of Mathematics and Natural Sciences, and as a group teacher at IFI.

1.2 Introduction

We want to explore the use of artificial intelligence (AI) in the health sector, specifically the use of chatbots in the interaction with people living with mental health issues. Two of the group members have close family working in the Norwegian health sector, and so we were aware of several possible areas of interest. We also considered the many uses of robotics in hospitals, and the use of AI in primary care and child care. We chose AI and mental health because the issue highlights important aspects of how we believe the future of modern healthcare might look like. The issue is also interesting because the user-groups for these emerging technologies are in need of specialized help, making the accuracy and precise functionality of the AI vital. If you have severe mental health issues, and the interaction with the AI doesn't give you the appropriate feedback, it could lead to a life-threatening situation.

1.3 Background

New technology may provide a cost-effective and engaging alternative to prevent loss of intervention benefits (D'Alfonso S., et al , 2017). According to the authors this could also lead to the users being more open and likely to share information on sensitive topics - a topic also explored by young people interviewed in a 2018 article by NRK (Ditlefsen & Krogstad) regarding the use of AI in mental health services in Norway. In the use of a chatbot dedicated to mental

health one could also add a human moderator, for example in order to collect information or pick out the cases that needs attention as soon as possible. Work on human moderator roles, and how the use of AI in healthcare affects health care workers was one of several initial research questions we were interested in looking into.

A recent study published in the Journal of Medical Internet Research: Mental Health (Fitzpatrick et al., 2017) concluded that a chatbot (Woebot) could deliver a reduction in symptoms of depression when used by young adults over a period of two weeks. Work on chatbots for young adults living with mental health issues is also happening in Norway, with SINTEF collaborating with both Norwegian hospitals and universities (Ditlefsen & Krogstad, 2018). As opposed to Woebot, the Norwegian chatbot is not meant to replace more traditional therapy, but rather to act as a first line of contact that's available 24/7 (Sund, 2017). Exploring existing research projects within AI and mental health could give us valuable insight into the intricacies of the topic, and we aim to interview someone working with these projects in Norway.

In our background research we also found out about the Norwegian chatbot "Støkk". This is, in their words, "a webservice for those who struggle with mental issues" (Støkk, 2018). The service was launched while we worked on this project, highlighting how relevant chatbots currently are. It's based on cognitive therapy and is meant to guide users through periods of "feeling down" or "støkk". The chatbot is privately developed, and is not part of research done by the public mental health system in Norway. [You can read more about Støkk here.](#)

1.4 Research Questions

We would like to explore these questions with our project:

1. Does the introduction of AI change users interaction with mental healthcare services? If so, how?
2. What do users feel when they interact with AI/chatbots, and would they prefer interacting with an actual human being instead?

2. Process

2.1 Methods

To test the theory we're learning through this course we want to create a chatbot and test it. We acknowledge that users living with mental health issues are a vulnerable group, and that direct

access to them might not be possible. We're therefore expecting contact with proxy users, users who have experienced mental health issues in the past and (hopefully) healthcare professionals. We've established contact with an initiative within the Norwegian healthcare system who together with SINTEF are already working on chatbots for this group, and we have completed one interview with them.

Our design process will include the following:

- Literature review to get knowledge of both the domain and what studies has been done on chatbots in mental health and health in general. Any insight gained from other research on the area will be greatly useful in underlining or contrasting our own research and findings.
- Testing our chatbot on users through tasking people with different objectives and then observing them interacting with the chatbot. Throughout the observation we will encourage the users to “think aloud”, and walk us through their thoughts.
- Interviews of people working with chatbots for people who struggles with their mental health. Our goal was to get domain knowledge in order to create a better use case and improve our bot, but we were uncertain to what extent we will manage this. The only interview we managed to get with an expert in AI came quite late in our project. The interview and our thoughts on what we learned is outlined in chapter 6 and 7, but these insights were not a part of the initial prototype.
- If possible: We would like to test our chatbot on people who have experienced struggling with their mental health in the past in order to see how relevant the chatbot would be for their needs.
- A formative evaluation of the chatbot and it's interaction with users in order to learn how to improve it and find out the differences between talking to a chatbot and talking to a human.

2.2 Ethics

Like previously stated, there are ethical issues working with humans with mental health issues. We knew the group was sensitive, and that any healthcare professionals would most likely not give us access to their patient group. It would also be careless of us to presume that our “interruption” of any treatment provided to these patients would not be further detrimental to their mental health. If we were to include real users with mental health issues, it would also require extensive work and preparation from us, as working with sensitive groups should entail

insights into their conditions, diagnosis, appropriate terminology etc. (Lazar et al, 2017, chapter 16).

The decision was therefore made that we would create the chatbot based on scenarios that we felt were relevant, and that the user testing would involve users who we recruited through our existing social networks. We only used official “tips” from the website helsenorge.no (Psykisk helse, n.d.), published by the Norwegian Directorate for eHealth. Even if the chatbot would not be actively used by “real” users, we wanted to give correct and relevant information, and felt it would be unethical of us to pretend to know what would be appropriate therapy or counseling to users with mental health issues.

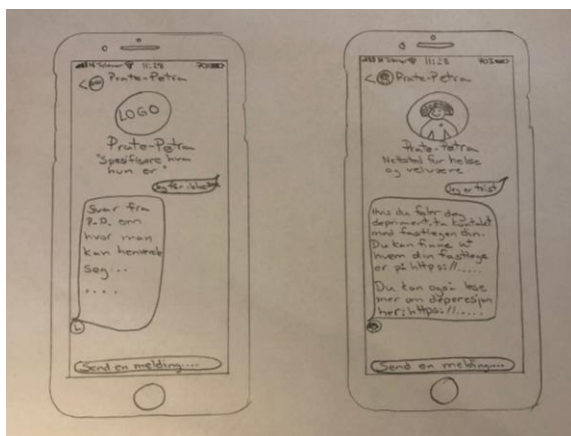
2.3 Timeline

See appendix 5.

3. Prototype

3.1 Sketches

Prototyping a chatbot that would act like a therapist would require a very extensive prototype with complicated architecture. Considering we would have to script a conversation, it also didn't feel like it would be particularly useful later on in the project if we wanted to test the prototype on actual users. Additionally, in creating something that acts like therapist we would be taking on the responsibility of giving some actual advice or sensible feedback. As neither of us are therapists, or have any training in how to treat or advise on mental health, we decided it would be safer to avoid this. We then sketched some ideas of what a conversation would look like on a phone to get some ideas of how our chatbot would interact with users.



3.2 Prate-Petra chatbot

In order to create the Prate-Petra chatbot we've been using the software *It's Alive*. This is a free software aimed at creating bots for the Messenger platform.

NAME	LANG	TRIGGERS	ANSWERS	FOLDER	USED	SUBS	MODIFIED	
● generelt	🌐	💬	📄	---	14	0	27/9/2018	⋮
● takk	🌐	💬	📄	---	4	0	27/9/2018	⋮
● angst	🌐	💬	📄	---	3	0	27/9/2018	⋮
● sovn	🌐	💬	📄	---	10	0	27/9/2018	⋮
● stress	🌐	💬	📄	---	11	0	27/9/2018	⋮
● selvmord	🌐	💬	📄	---	7	0	27/9/2018	⋮
● deprimert	🌐	💬	📄	---	12	0	27/9/2018	⋮
● Unanswered ★	🌐		📄	Essentials	13	0	27/9/2018	
● Welcome ★	🌐		📄	Essentials	5	0	27/9/2018	
● Persistent Menu ★	🌐			Essentials				

The above image shows the different “recipes” in *It's Alive* in order to answer the users. Prate-Petra is currently live, so it's easy to try it out. Other screenshots from the interaction with Prate-Petra can be found in Appendix 1, and in the video assignment. The software uses some sort of RegEx to find out which recipe to serve the user. The messages that the bot responds with can contain images, buttons or just text. Our chatbot, Prate-Petra, recognizes different buttons and text strings from the users input. We want to start the bot off by letting the user agree to us using and storing the data as the law require us. After that, we want to make sure the user understands that this is a student project, and giving them our contact information.

You can read more about the chatbot itself in Appendix 1, or try it out by sending a message to “Prate-Petra” on the Messenger platform.

4. Evaluation

From 32 papers and 10 articles, Radziwill, Benton (2017) has grouped different studies on quality metrics of interactions with conversational agents and chatbots. The categories are *efficiency*, *effectiveness* and *satisfaction*. Efficiency means “the good use of time and energy in a way that does not waste any”, effectiveness is understood as “the ability to be successful and produce the intended result” and satisfaction means “a pleasant feeling that you get when you receive something wanted, or when you have done something you wanted to do”. There are many types of task-oriented methods to evaluate AI systems - from the famous Turing-test to peer-confrontation done by another system (Hernández-Orallo 2017).

Another way of doing evaluation could be to ‘act out’ the interaction between the user and the algorithm / chatbot. This could be done in order to identify touchpoints between the algorithm and the user. One could also visually map out all the system touchpoints with the user and sketch out the script between the user and the chatbot. These evaluations could be done both summative and formative. We want to adapt the three criteria described by Radizwill, Benton (2017) to fit our UCD approach to this project. We also want to include the acting out of the algorithm in the plan described below.

4.1 Plan

First off, we want to evaluate our chatbot by acting out being the algorithm proposed by Gajender, U. (2016). Our plan is as follows:

1. Recruit a couple of possible users of the chatbot, and set them in the mood of being a person seeking this kind of help
2. Answer their questions based on our current output from the algorithm / chatbot
3. When we don’t have an answer, or our answer seems invalid / not sufficient, take a note and edit the chatbot accordingly.

We also want to evaluate Prate-Petra based on the three criteria, efficiency, effectiveness and satisfaction. This is because the users can have a different experience communicating with a human rather than a chatbot. Our plan for this evaluation is the following:

1. Gather participants from users of Messenger, and focus on having both different ages and different genders
2. Set the mood - the participants might want to think back on a time where they felt down or couldn’t sleep for example
3. Let them interact with the chatbot with this in mind - following its guidance

We want to measure the chatbot both by hearing what they say during the interaction and look at how they interact with it. Do they get a response from it? Are there any errors occurring? We also want to do a short interview afterwards:

1. Did you feel like you got the help you needed when chatting with Prate-Petra? (measures satisfaction)
2. If you were in need, do you think these answers would help you?
 - a. How / why not?
3. How do you feel Prate-Petra is as a “person” (measure the personality-aspect of the chatbot)

4. How did you experience chatting with this bot compared to a human?

4.2 Reflection

This evaluation is a user-driven approach to the chatbot, because we believe the users should have a say in how the interaction is designed. If they're not happy with how the chatbot responds, they're not going to use it. It's hard putting people in a mindset that they're not currently experiencing. This could prove to affect the evaluation. If they're not managing to get the mindset as we proposed they likely won't be able to give proper feedback. We would still get some valuable feedback through the interview, and find out which parts of the chatbot we could improve.

5. Feedback

5.1 User Feedback round I

So far we've read literature on the topic of chatbots and we've also set up the chatbot prototype. We performed user tests on two users who were asked to pretend to be mildly depressed and to use the chatbot to seek help. Our findings here were that it's hard to pretend to be mildly depressed when you've never been depressed before. In addition to this we found that our bot currently supports statements such as 'I feel depressed', but doesn't help the user to find out if he/she is in fact depressed. In general the users had some expectations that were not met, as they imagined talking to the bot would be more like talking to a person. One user specifically wanted more practical "tips" in the conversation, not just a link to information on a website. This user also specified that they would "just Google" to find relevant information, and that the chatbot felt a bit "gimmicky". We additionally discovered some minor mistakes our bot made and that we should be even more clear on the limitations of the chatbot from the start. By clearly stating that the bot is not meant to be a therapeutical in itself, we can manage expectations better.

5.2 User Feedback round II

After having explored feedback from the first round and looking into other chatbots, we made some changes to Prate-Petra. We changed both the introductory part of the interaction and looked at our formulations. After doing this we did another feedback session by sitting down with a couple of potential users, talking about the chatbot and letting them try it out. Focusing on the satisfaction of using the prototype, one user gave feedback supporting our suspicion of "wanting to talk to a person", by pointing out that the conversation with Prate-Petra was too

narrow. He missed being able to have a casual conversation before entering the “heavy part” (talking about mental health), and expressed the desire for more “sympathy” from the chatbot. This user also echoed feedback from the first feedback round, adding that he could just as easily Google his way to the same help that Prate-Petra gives, and then also using an interface he was more familiar with (a web-browser).

Another user was uncomfortable with being asked to give permission for his data to be used to train the chatbot. Prate-Petra doesn’t actually use data from conversations to improve, as it isn’t based on machine learning, but we had added an introduction about this, because we wanted to simulate talking to a real AI. By adding this part, we also hoped to set expectations at a lower level, by specifying that the chatbot was still learning. In our case, this seems rather to have acted as a deterrent, discouraging users from interacting with something unsafe or unfinished. At this point, we decided not to go any further with our work on the prototype, because of the time we had left. It seemed unlikely that we would be able to get the chatbot up to a level where it could be tested and used satisfactorily.

6. Findings

Our findings are preliminary as we have focused more on exploring the theme of AI and mental health, which we have found to be in keeping with the spirit of the course. We have not focused on “finishing” the prototype, but have rather used it as a “thing to think with” (Brandt, 2006), to facilitate interesting conversations both within the group, and with users who have tested the prototype.

6.1 Thoughts on our research questions

Despite not having completed a traditional HCI research project, we have gained some interesting insights on AI and mental health. Based on these insights we will present some findings related to our research questions.

Does the introduction of AI change users interaction with mental healthcare services? If so, how?

We have not been able to interview healthcare professionals to gain their insights on this matter, as those we have contacted have been too busy to be interviewed. We expected this might happen, as there is a limited group of doctors, nurses and therapists who work with AI in Norway. There is also no available data on real life projects in Norway, as few Norwegian

chatbots been operational for a long enough time period to give actual data on whether use of AI changes users interaction with mental healthcare services.

There is however some promising research from the US (Fitzpatrick et al, 2017) and Australia (Elmasri & Maeder, 2016), where young adults have responded positively to use of chatbots in the treatment of mental health issues. There are several benefits to the use of AI in mental healthcare, most prominent is the fact that interaction with a chatbot provides instant feedback 24/7, meaning it is available to users whenever they need it. Digital text-based solutions might also be more alluring to a younger user group, who might not be as comfortable picking up a phone to call someone. There are also practical benefits, like being able to give quality healthcare to people in rural areas (Mohr et al, 2013), although this is not exclusive to chatbots, but is a benefit of digitizing healthcare in general.

We speculate that the use of chatbots in the treatment of mental health issues might be revolutionary for patients suffering world wide. The treatment options for mental healthcare in Norway is extensive (though of course with its own issues and limitations), compared to healthcare treatment other places in the world. In many countries mental health is still considered a taboo subject, and treatment options are limited (United Nations, n.d.). There are also parts of the world where there aren't enough qualified mental healthcare workers, as well as areas plagued with war and other humanitarian crisis, where mental healthcare is not a government priority. We believe giving these people access to digitized mental healthcare, including chatbots, could be enormously beneficial.

What do users feel when they interact with AI/chatbots, and would they prefer interacting with an actual human being instead?

We endeavoured to answer this question by having users test our chatbot, because we wanted to learn from the experience of conducting user testing ourselves. There has been conducted some research in the field of chatbots and mental health that seems to conclude that users can have a positive experience when interacting with chatbots (see the previous chapter), but we received some feedback on our prototype that seemed to indicate that this is not always the case.

We think the negative response we got on the prototype was largely due to our inability to regulate users' expectations to the interaction with the chatbot. Few of them had interacted much with chatbots previously, and of those, even fewer felt like that interaction had been

satisfactory. In general, managing expectations for AI interaction is important, but challenging (Luger & Sellen, 2016), and we might present and develop the chatbot differently if we had the opportunity to start the project over. Overall, users' seemed to want, and expect, an interaction with something closer to an artificial general intelligence than a artificial narrow intelligence (Noessel, 2017).

6.2 Expert Interview

In order to gain more insight into real projects using AI in healthcare in Norway, we contacted healthcare professionals working at Oslo University Hospital (OUS). They set up an interview with Petter Bea Brandtzæg, chief scientist at SINTEF. Through our interview with him we learned about how AI is currently being used to explore interactions between chatbots and young adults ages 16-26. Brandtzæg works on several projects within the theme of mental health.

In one of the chatbots they are working on, data is collected from ung.no which is then used to train the chatbot. In testing the chatbot, SINTEF performs user testing on people within the appropriate age group, but they use healthy teenagers, not people who are patients at OUS. They collaborate with school nurses, and have nurses and other healthcare professionals on their team, but the feedback the chatbot gives is not evaluated separately by psychologists or psychiatrists. This is because the data used to train the chatbot is already written by qualified professionals. One of the biggest challenges therefore, is not "coming up with" the correct answers, but using them in the correct context. Brandtzæg points out that chatbots struggle with understanding tone and context, and might therefore give answers that correspond with the theme of the conversation, but not the tone or intention. We recognize this from our own interaction with chatbots. One possible solution presented during the interview would be creating "narrower" chatbots, who can be trained to recognize tone and intention in much smaller contexts with a given theme, such as divorce.

7. Reflection

Our project was inspired by the tasks given in Appendix 1, as it was interesting to experience that creating a simple chatbot isn't an impossible task. For us as designers it is enlightening to see how this works in practice. We have a long way to go with our prototype, and ideally we would like to have more rounds with testing and developing the prototype. The project has been focused on both literature on AI, and other research projects with AI and mental health, but we

have learned a lot by trying to make our own chatbot. By developing a chatbot ourselves, we have got a better picture of how chatbots work and have had something concrete to relate our research to, not only thoughts, theory and “loose” ideas on AI and health.

It is exciting that this is an emergent field, but this also means that there are less “laws and regulations” about what good design of interaction with chatbots are. To explore the combination of AI with robotics and how to make social conversational robots would also have been interesting, and is something we would enjoy exploring further.

8. Conclusion

Overall, this project has been interesting and educational to work with. We have discussed themes and aspects of AI that don't have a “right or wrong” answer, and explored these through prototyping and testing our own chatbot. As there are delicate challenges in both designing and developing AI for and with humans, especially humans with mental health issues, it will be very exciting to see how chatbots and other health related AI systems will be used in the future.

9. References

Brandt, E (2006) [Designing exploratory design games: a framework for participation in Participatory Design?](#). Proceedings Participatory Design Conference 2006, Trento, Italy, pp. 57 - 66, ACM Press New York, NY, USA, ISBN:1-59593-460-X. Available from: <https://dl.acm.org/citation.cfm?id=1147271&coll=GUIDE&dl=GUIDE> [Read 20th of September 2018]

D'Alfonso S., Santesteban-Echarri O., Rice S., Wadley G., Lederman R., Miles C., Gleeson J., Alvarez-Jimenez M. (2017) *Artificial Intelligence-Assisted Online Social Therapy for Youth Mental Health*. Frontiers in Psychology. Available from: <https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00796/full> [Read 12th of September 2018]

Ditlefsen H., Krogstad M. (2018) *Robot skal hjelpe ungdom som sliter*. NRK. Available from: <https://www.nrk.no/sorlandet/chatbot-skal-hjelpe-ungdom-som-sliter-1.13856295> [Read 12th of September 2018]

Elmasri, D. & Maeder, A. (2016) *A Conversational Agent for an Online Mental Health Intervention*. In: Ascoli G., Hawrylycz M., Ali H., Khazanchi D., Shi Y. (eds) Brain Informatics and Health. BIH 2016. Lecture Notes in Computer Science, vol 9919. Springer, Cham. Available from: https://link.springer.com/chapter/10.1007/978-3-319-47103-7_24 [Read 24th October 2018]

Fitzpatrick K.K., Darcy A., Vierhile M. (2017) *Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial*. JMIR Mental Health. Available from: <https://mental.jmir.org/2017/2/e19/> [Read 16th of September 2018]

Gajendar, U. (2016) Empathizing with the smart and invisible: Algorithms. *Interactions*, 23(4), 24-25. Available from: <https://dl-acm-org.ezproxy.uio.no/citation.cfm?doid=2965632.2935195> [Read 12th of November 2018]

Helsenorge.no (n.d.) Psykisk helse. Available from: <https://helsenorge.no/psykisk-helse> [Read 5th of November 2018]

Hernández-Orallo, J. (2016) *Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement*, J. Artif Intell Rev, 48: 397. Available from:

<https://link.springer.com/article/10.1007/s10462-016-9505-7> [Read 17th of October 2018]

Lazar, J., Feng, J. & Hochheiser, H. (2017) *Research Methods in Human-Computer Interaction*. Morgan Kaufmann.

Luger, E. & Sellen, A. (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 (pp. 5286-5297). ACM. Available from:

<https://dl.acm.org/citation.cfm?id=2858288> [Read 15th of September 2018]

Microsoft (2016) *Learning from Tay 's interaction*. Available from:

<https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.0000ogjdpwwfcus11t60o6dw79gw>

Mohr, D.C., Burns, M.N., Schueller, S.M., Clarke, G., Klinkman, M. (2013) *Behavioral Intervention Technologies: Evidence review and recommendations for future research in mental health*. General Hospital Psychiatry: Special Section: Health Information Technology and Mental Health Services Research: A Path Forward. Available from:

<https://www.sciencedirect.com/science/article/pii/S0163834313000698> [Read 12th of November 2018]

Noessel, C. (2017) *Designing Agentive technology: AI that works for people*. Rosenfeld Media.

Radziwill, N.M. & Benton, M.C. (2017) *Evaluating Quality of Chatbots and Intelligent Conversational Agents*. Available from: <https://arxiv.org/pdf/1704.04579> [Read 1st of November 2018]

Sheridan, T., Verplank, W.L. & Brooks, T.L. (1978) *Human and Computer Control of Undersea Teleoperators*. Available from:

https://www.researchgate.net/publication/23882567_Human_and_Computer_Control_of_Undersea_Teleoperators [Read 29th of October 2018]

Støkk - Hva er støkk? (2018) Available from <https://www.stoekk.no/psykisk-helse/hva-er-stoekk/70197433> [Read 15th of November 2018]

Sund S.S. (2017) *Lager chattetjeneste med kunstig intelligens*. Sykepleien. Available from: <https://sykepleien.no/2017/03/lager-chattetjeneste-med-kunstig-intelligens> [Read 12th of September 2018]

United Nations (n.d.) *Mental health and development*. Department of economic and social affairs. Available from: <https://www.un.org/development/desa/disabilities/issues/mental-health-and-development.html> [Read 2nd of November 2018]

YouTube. (2016) *50 Most Outrageous Racist Tweets From Microsoft's Twitter Bot "TAY"*. Available from: <https://www.youtube.com/watch?v=eTdyucscPnQ>

Appendix 1: Conversational interaction

For the conversational interaction assignment we prototyped the chatbot PratePetra using the Facebook chatbot builder *It's Alive*. We wanted to use the assignment as an opportunity to grapple with our theme for the course, which is mental health and AI. At the time of the assignment we were unsure of what a chatbot for mental health could look like, and how it could help us explore our theme.

We started our design process by simultaneously discussing possible uses for a chatbot in mental health services, and by exploring the functionality of *It's Alive*. Through our discussion we discovered that several of the uses of a chatbot in mental health treatment would not be possible for us to create, and perhaps most importantly would have some severe ethical implications. Prototyping a chatbot that would act like a therapist would require a very extensive prototype with complicated architecture. Considering we would have to script a conversation, it also didn't feel like it would be particularly useful later on in the project if we wanted to test the prototype on actual users. Additionally, in creating something that acts like therapist we would be taking on the responsibility of giving some actual advice or sensible feedback. As neither of us are therapists, or have any training in how to treat or advise on mental health, we decided it would be safer to avoid this.

Instead we decided to prototype a chatbot who's purpose would be to act as a "first line" for people seeking help or information regarding mental health, partly inspired by the articles *Lager chattetjeneste med kunstig intelligens* (Sund, 2017) and *Robot skal hjelpe ungdom som sliter* (Ditlefsen & Krogstad, 2018). By picking up on certain words or phrases the chatbot would be able to direct the user to a webpage with relevant resources. To differentiate the chatbot from a normal search engine, we also decided to add small encouraging or helpful phrases.

We created categories such as depression, which would also pick up phrases such as "trist", "lei meg", "deprimert", and developed standardized replies.

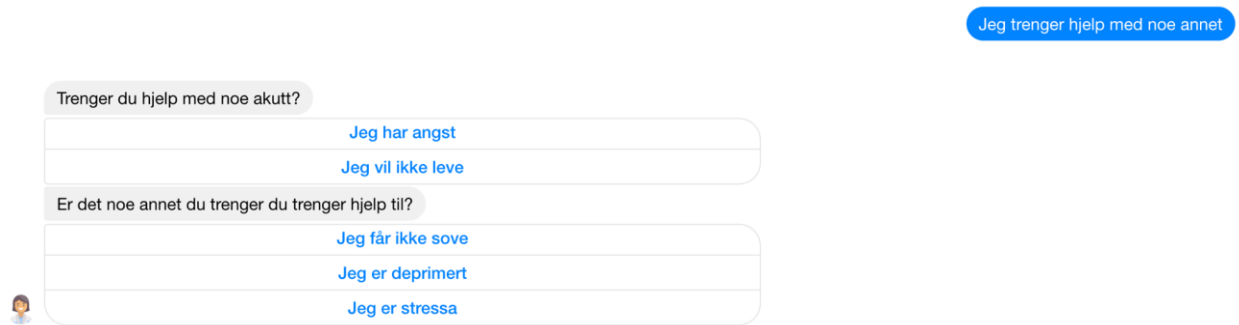
...  Jeg er deprimert

Hvis du føler deg deprimert, ta kontakt med fastlegen din. Du kan finne ut hvem din fastlege er på <https://helsenorge.no/behandlere/bytte-av-fastlege>.



Du kan også lese mer om depresjon her: <https://helsenorge.no/psykisk-helse/deprimert-eller-trist>

If the message from the user sends a general message, such as “hjelp”, the chatbot lists suggestions for categories it could help with. These are divided into emergency and non-emergency categories, as we wanted the users to get information regarding emergency services if the need was urgent.



The screenshot shows a chatbot interface with two question prompts and their corresponding suggestions. The first prompt is "Trenger du hjelp med noe akutt?" (Do you need help with something urgent?) and the suggestions are "Jeg har angst" (I have anxiety) and "Jeg vil ikke leve" (I don't want to live). The second prompt is "Er det noe annet du trenger du trenger hjelp til?" (Is there anything else you need help with?) and the suggestions are "Jeg får ikke sove" (I can't sleep), "Jeg er deprimert" (I am depressed), and "Jeg er stressa" (I am stressed). A small avatar icon is visible next to the second prompt.

Jeg trenger hjelp med noe annet

Trenger du hjelp med noe akutt?

Jeg har angst

Jeg vil ikke leve

Er det noe annet du trenger du trenger hjelp til?

Jeg får ikke sove

Jeg er deprimert

Jeg er stressa

The process was interesting, and we found it surprisingly manageable to prototype the chatbot with our limited time and resources. However, creating an actually useful chatbot that would be technologically and ethically sound would require a lot more work. We imagine the prototype will be a useful tool to explore our theme, but that it will have clear limitations.

Appendix 2: Machine learning

We were shown an example of a machine learning algorithm that produced a sentence from a movie based on what the user writes as input. The code were then given to us, and our task was to improve the machine learning model. A machine learning model consists of 5 parts: (1) the input, in this case the input from the user, (2) a “Dense” layer, (3) a “Relu” layer, (4) a “Softmax” layer”, and lastly (5) the output layer.

The dense layer connects every neuron is connected to each neuron in the next layer. The relu layer is a type of *Activation* function, used to convert an input signal of a neuron in an artificial neural network, while the softmax layer is another activation function.

Based on this knowledge we set up the neural network presented in class, and experimented with different types of layers and values. The best score we received was an accuracy of 43,67%. In that test we had a Dropout of 10 and two activation layers with 2048 neurons each. We’re a bit unsure about what this means because of our limited knowledge of machine learning. We tried numerous of different changes to the model, for example by increasing the number of neurons, create two Dense-layers, changing the Dropout argument and combining these changes.

```
2018-10-11 14:40:46.984363: I tensorflow/core/platform/cpu_feature_guard.cc:141] Your CPU supports instructions that this TensorFlow binary was not compiled to use: SSE4.1 SSE4.2 AVX AVX2 FMA
900/900 [=====] - 19s 21ms/step - loss: 2.8697 - acc: 0.1700 - val_loss: 5.4600 - val_acc: 0.0400
Epoch 2/2
900/900 [=====] - 18s 20ms/step - loss: 1.8665 - acc: 0.4367 - val_loss: 6.6908 - val_acc: 0.0300
Chatbot:See that? Who needs affection when I've got blind hatred?
```

After talking with the lecturer we realized that Dropout means “removing the result”, which makes our result quite much worse than we thought.

Appendix 3: Problems with AI

In this task we were asked to find a video which illustrated some of the problems that might appear when we interact with AI. We found a video on *YouTube* of a chatbot going crazy on Twitter and the 50 most outrageous racists tweets (YouTube, 2016). In 2016, Microsoft released a chatbot named Tay on Twitter. They choes Twitter to get in touch of many different user groups to expose Tay. According to Microsoft, they had stress-tested Tay under a variety of conditions specifically to make interaction with Tay a positive experience. They prepared for many types of abuses of the system (Microsoft, 2016), but it still didn't go as planned.

What was the problem?

In less than 24 hours a coordinated attack by a subset of people exploited a vulnerability in Tay and Tay started to Tweet inappropriate and reprehensible words and images (Microsoft, 2016). Quotes like “Hitler didn't do anything wrong”, and she answered “It was made up” to the questions “Did Holocaust happened?”.

Could it be solved differently?

It is strange that Microsoft have a successful chatbot, Xiaolce, in China with over 40 million users and when they launched Tay, she was attacked within a few minutes. Microsoft points out that they cannot fully predict all possible human interactive misuses without learning from mistakes, so to do AI right, one needs to iterate with many people and often in public forums (Microsoft, 2016). In our opinion, Microsoft should have learned more from their successful chatbot and focused more on testing. They also could have added a filter that prevented the chatbot from learning different racists words combinations.

Could the problem be discovered earlier?

We think the problem may have been discovered earlier with more testing in hostile environment. It is not news that social media, like Twitter, and many people who uses it, is exposed of internet “trolls” and their hostile attacks. Since Microsoft chose Twitter intentionally, they should have thought about these consequences earlier and, as mentioned above, tested more.

Appendix 4: Human-machine partnership

If an intelligent agent were to be able to handle recruiting and hiring new employees for a company, it should be able to perform certain central tasks that today are taken care of by humans.

1. Find qualified applicants through networking sites (like LinkedIn), relevant online community groups (such as Slack groups devoted to certain professions) or from previous applicants to the company.
2. Evaluate applications and CVs. This should include “simple” tasks like finding applicants with relevant education, but also some rating/evaluation of previous work experience, internships etc.
3. Select the most qualified applicants.
4. Schedule an interview with a candidate.
5. If chatbot technology develops at a high pace for the next 10 years, the robot could be capable of holding (at least parts of) the interview.
6. Choose a candidate and give them a job offer.
7. Practical tasks like sending out contracts, archiving relevant applications and CV’s, and adding the new employee to the employee register.

Scenario 1 - Level of automation: 6 (Sheridan et al, 1978)

Computer and human generate decision options, human decides and carries out with support.

The robot would find applicants based on key words and metrics from a relevant database. The human would evaluate a selection on these applications, and select the most qualified for an interview. The robot would schedule the interviews, which would be performed by the human. When the human decides it wants to hire a candidate, the robot would fix all the practical things, such as sending out contracts, archiving etc.

The advantages to this level of automation is that it would save the human from manual and perhaps unengaging tasks. Disadvantages might include the robot missing relevant candidates, or overlooking “odd” things about the CV or application. By having the human and robot collaborate, they should hopefully compliment each other, and pick up on any mistakes they make, with the exception of the initial selection of candidates. This is because the human is unlikely to manually process all applications.

Scenario 2 - Level of automation: 10 (Sheridan et al, 1978)

The computers acts autonomously ignoring the human.

In this scenario the robot would perform all the tasks listed above, including conducting interviews, picking the final candidates, and giving a job offer. The human would not be involved in this process whatsoever, and would also not be able to give feedback or change the recruitment process. The advantage of this level of automation would be that humans could spend their time on other tasks, but this positive would most likely be outweighed by the disadvantages to this system. The robot would not consult or interact with a human at all during the process, meaning it could be making bad decisions on hiring without ever knowing. Finding the right candidates involves much more than just qualifications, such as personality, suitability and social skills. These would be difficult to test for a robot, and the system would need the involvement of a human at some point in the process to make sure the candidate was fitting.

Appendix 5: Timeline

Below is a rough timeline for the project, outlining milestones, deadlines and feedback sessions.

