FINAL REPORT - INFORMATION BASED CHATBOT

In5480: Specialization in research in design of IT

Autumn 2018

Written by:

Vilde Mølmen Høst - *vildehos*

Marte Rimer - *martrim*

Anna Sofie Schei - *annassc*

Table of content

1 . Introduction

Our names are Marte Rimer, Anna Sofie Schei and Vilde Høst, we are all first-year master students on Design, use and interaction. We know each other from the interaction design bachelor here at 'Institute For Informatics' hereby referred to as IFI. We all think AI as a field is very interesting and are looking forward to having a lot of professional discussions about the topic through our project work.

1.2 Description

In our project we explore how a chatbot can give information to students about school-related information. In the first iteration of the project we created a chatbot for giving students information about where to get coffee etc. at IFI. One of our hypothesis was that information given by chatbots would be useful for new students at IFI, giving them information about things that we consider to be important when you're a first year students. In the second iteration we wanted to explore the use of chatbots through theory and used this in combination with testing to learn more about how a chatbot for this context should be. In the final iteration, iteration three, we improved and changed the chatbot based on the results from the last iteration and made a plan for evaluate the chatbot. The plan was then executed with five participants. In our conclusion we discuss the results from the evaluation in the light of our research question.

2. Questions: Using a chatbot in a school context

We wanted to investigate users' trust in an AI system such as a chatbot. We therefore designed a research questions we wanted to look further into.

"How will helpfulness affect trust in chatbot technology for students at IFi when it comes to school-related information?"

A chatbot needs a purpose, and if we consider that if this purpose is to be helpful, it also needs to gain trust from the users. There is no need to ask a chatbot for help if you don't trust the information it gives you. With this in mind we consider the first question to be a bit too ambiguous and large for us to investigate in this course. We have therefore used this question as a guideline for what we can actually manage to explore in this course and what we can find on the existing literature in this field. Trust is an important factor for reliance on and implementation of technology (Lee & See, 2004). In relationships trust means being reliable, having confidence in the other person both physically and emotionally (Lewicki & Bunker, 1995). So one can say that trust will also play a role in the interplay between human and machine. The problem with systems taking control is that it's often hard for people to rely upon it appropriately. Because people respond to technology socially, trust influences dependence in it. So trust will inevitably guide reliance when we are faced with complex and unanticipated situations. When we use systems to navigate and make decisions about our health, finances, relationships, and future — they must be trustworthy. In human-technology interaction trust is an example of the important influence of affect and emotions. Emotional feedback in technology is not only important for acceptance, but can also make a fundamental improvement regarding safety and performance (Lee & See, 2004).

To make the project more feasible we wanted to explore the following questions:

1. **How useful is information given by a chatbot compared to a human counsellor?**
2. **Does students find information given by a chatbot trustworthy?**

By exploring these questions we hoped to get indicators on how students experience interacting with a chatbot contra interacting with a human, and address if the students prefer one communication format over the other. This was done via selected methods in the design process, see chapter 4. Due to time constraints we later in the project had to focus our efforts more on the second question.

3. Background

Chatbots has emerged as a hot topic in the latest years, and it is used by numerous companies in various areas - help desk tools, automatic telephone answering systems, e-commerce and so on. Even though the technology has been around since the 60's (Atwell & Shawar, 2007). Why are we suddenly so interested in this technology now? This can likely be explained by the recent year's advancements in messaging applications and AI technology (Brandtzaeg & Følstad, 2017).

In the article *Chatbots: Are they really useful?* Atwell and Shawar provide real-life examples of different chatbots in different contexts. One of the examples is Sophia, a robot that was developed to assist in mathematics at Harvard by answering students questions. This turned out to be applicable in many other contexts. Living in Norway you have probably noticed “Kommune Kari”. A chatbot that many of the municipality have available on their web-pages. Kari is there to answer “easy” questions like “when will the garbage truck come?” and “where can I find available jobs?”. Kari's goal and the job is to provide information so that you as a user don't have to navigate the “massive information flow” (Schibevaag, 2017). This way of using a chatbot is a part of the Question Answering (QA) field which is a combination between AI and information retrieval (Molla & Vicedo, 2007). QA can be defined as:

“... the task whereby an automated machine (such as a computer) answers arbitrary questions formulated in natural language. QA systems are especially useful in situations in which a user needs to know a very specific piece of information and does not have the time—or just does not want—to read all the available documentation related to the search topic in order to solve the problem at hand”. (Molla & Vicedo, 2007).

Sophia and Kari are examples of chatbots that operate in "very specific" domains. This means that if you were to ask Kari about math and Sophia about when the garbage truck comes none of them would know the answer - because the question is outside of their domain. Chatbots have what is called a natural language user interface and therefore communicate with users via natural language — how a human would talk on a regular basis (Brandtzaeg & Følstad, 2017). Therefore they use what is called natural language processing (NLP) where the chatbot uses computational techniques to analyze text, where the goal is to produce a human-like answer based on a linguistic analysis (Hirschberg & Manning, 2015).

For a chatbot to be especially useful to a certain domain some criteria have to be met. Minock (2005) proposes the following criteria for a domain to be successful in answering domain-specific questions: a domain should be circumscribed, complex and practical. This is summarized in the table below.

Criteria	**Description**
Circumscribed	Clearly defined knowledge sources and comprehensive resources available (a database etc.)
Complex	If you could develop a simple FAQ then it would not be useful with a QA system. There has to be some level of complexity in the domain while still being able to meet the circumscribed criteria.
Practical	Should be of use to a large group of people in the domain and take into account: how the users will formulate questions, what is commonly asked and how detailed the answers should be.

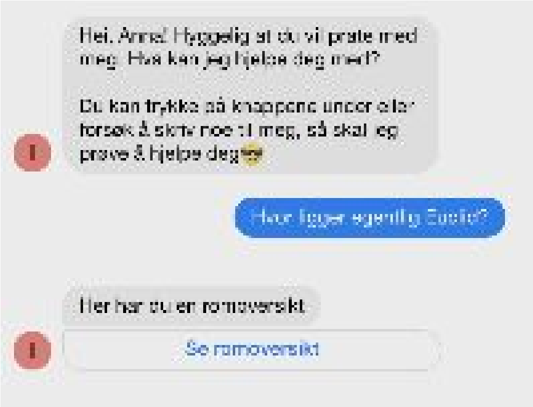
When designing an intelligent system that provides decision support one must consider the human as something outside the system, but also as an integrated system component that in the end, will ultimately determine the success or the failure of the system itself (Cumming, 2004).

4. Design process and methods

For the project, we wanted to have a simplified user-centred approach (hereby referred to as UCD). UCD is an iterative design process in which designers focus on the users and their needs in each phase of the design process (Interaction design foundation, unknown). UCD calls for involving users throughout the design process via a variety of research and design techniques so as to create highly usable and accessible products for them. The reason why we wanted to have a UCD design approach is to use the chatbot to explore how the users can, wish and needs to use the chatbot to achieve their goals.

Our goal was to facilitate user involvement through interviews and to learn about their context. The interviews was small where we tried to understand people's opinion about the subject. They were not only a conversation between the us and the participant but we also asked participants to execute some tasks interacting with a chatbot. Afterwards we asked them questions about the experience.

5. Prototype



Hei, Anna! Hyggelig at du vil prate med meg. Hva kan jeg hjelpe deg med?

Du kan trykke på knappene under eller forsøk å skriv noe til meg, så skal jeg prøve å hjelpe deg😎

Hvor ligger egentlig Euclid?

Her har du en romoversikt

Fig 1: first draft of our prototype

We made a chatbot that we used as a prototype to investigate the research questions. The chatbot was originally made for appendix 1. But we wanted to further use this in our project. During the design process we improved and tested the prototype. We tried to make it as helpful as we could manage within the time frames of the project by iterating multiple times.

5.1 How the chatbot meets Minock's three criteria:

Circumscribed - the information given to first year students are usually dispersed on differents sites and information channels. The information are usually given in a way where the students have to perform workarounds to retrieve the information. A lot of information is not written and usually learned and retrieved from other older students. This somewhat contradicts the goal of the system being fully circumscribed. Most of the information is found at the UiO webpage which we see as a "circumscribed source " but we also want to include the more verbal information.

Complex - the UiO webpage has many versions of FAQ´s but is in our experience sometimes to general. Because of the dispersed information and the different types of information a fully function chatbot in a school context should have, this could not be realised by a simple FAQ. Making a chatbot that is more advanced than a FAQ is not feasible in our project. But is rather a reason for using a chatbot in a school context, such as IFI.

Practical - Our chatbot is designed to meet the needs of a large group of students at IFI. We believe that it is practical in the sense that it detects short questions like: “I am hungry” and “Food” or “Where is Epsilon?” and “I can’t find my classroom”. Which in turn can reduce the time it takes for the students to locate this information. This can also be used as a way to gather data on the information that students are interested in.

5.2 Persona

In the making of the prototype we also formed a persona for the chatbot to make the chatbot consistent in its language. This worked as a guideline in the design of the chatbot and was very helpful since it gave us a common understanding of the chatbots characteristics. We focused on building the chatbot as an engaging partner with a “happy tone” and a sense of humor, including GIFs to make the experience more fun and intriguing.

6. Early testing and findings

In the beginning of our project we wanted to test the first version of our chatbot (from appendix 1) on first year students. This was late in the fall and most of the first year students were familiar with a lot of the answers our chatbot could provide. We therefore developed a scenario to help the participants imagine the context of use (see figure 2). We wanted to test this early version of the prototype to get input on what the chatbot could and could not answer in the future. After the test was completed we had a short interview with the participants. The main purpose for this test was to see how the participants interacted with the prototype and find out if a chatbot could be suitable to find the information they needed. Before the testing we also carried out a pilot test to find immediate flaws in the plan.

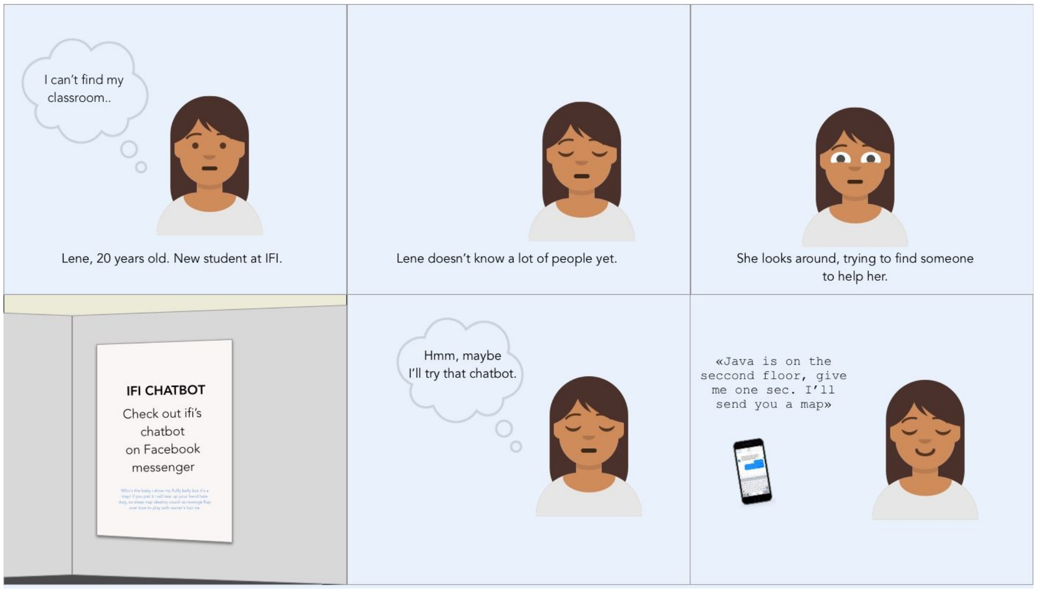
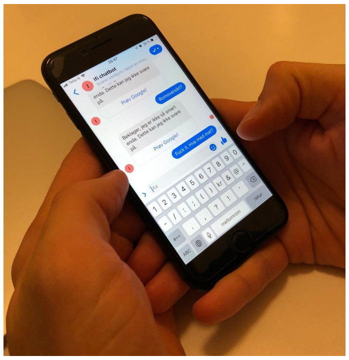

I can’t find my classroom..
Lene, 20 years old. New student at IFI.
Lene doesn’t know a lot of people yet.
She looks around, trying to find someone to help her.
IFI CHATBOT
Check out ifi’s chatbot on Facebook messenger
Hmm, maybe I’ll try that chatbot.
«Java is on the seccond floor, give me one sec. I’ll send you a map»

Fig 2: Scenario for use case

6.1 Results from the first testing



The first participant enjoyed talking to the bot, but stressed the fact that you had to “talk like “a dummy” for it to understand what you were asking. The participant pointed out that this really would have come in handy in his first weeks at the university, as he didn’t always know who to ask - especially if he was in a hurry. He pointed out that the prototype needs to get more features like tell you exam dates, or “ifi life-hacks, like get your coffee before all of the students have their break”.

The second participant was a bit frustrated that the chatbot wasn’t flexible enough (Fig.3). “I don’t like having to guess what questions to ask”. He would liked more instructions to know how to get more out of the chatbot.

The third participant had also problems with understanding what the chatbot could do. When given a hint for what the chatbot could do, the chatbot did not function properly. Here we tried to restart the system and then the chatbot displayed it´s welcome message— what it could do. Afterwards it was more clear what the participant could ask it, but the chatbot did not always give the response that the participant wanted.

6.3 Re-design of the prototype

This findings gave us a lot of insight in where the chatbot needed to be changed. E.g. adding a proper welcome message, defining the chatbots’ limitations and presenting this to the user. Luger & Sellen (2016) argues that it’s important to define goals and expectations so that your chatbot has a clear purpose. Knowing the capabilities and limitations of the system, before it crashes. The test showed that it was hard to ask the ‘right’ questions, we therefore added more ‘AI ques’ to simplify the interaction. We also used the principles for designing conversational agents. When talking about User-centred design of AI there are three (tentative) design principles: learning, improve and fuelled by large data sets (Følstad, 2018). The principle of learning is how the system is designed for change. Setting the expectations right, with the system's ability to perform and its ever changing nature. The principle of improve is how the system should be designed with ambiguity. The system is more than likely to make mistakes, so learning from these are an important principle to improve the system. The principle fuelled by large data sets is how the system is reliant on getting access to enough data.

7. Evaluating the chatbot

We wanted to evaluate the prototype in the right context, which for the IFI chatbot was at IFI. As mentioned before, most of the new students are more or less ‘integrated’ per now we could not test on “real potential users”. How ever we consider IFI-students as a good substitute since they have been in the situation before and a group that we easily can make contact with.

We listed a set of questions and tasks, see figure 4, wich we asked the participants to answer and preform. We also included a few control questions to investigate the participants experience with the chatbot and to find out if they had any suggestions for further improvement. The evaluation ended with a short talk about the experience, where we were open for any kind of feedback the evaluators could provide.

Due to time and capacity during this project we decided on including five participants acting as evaluators. The number of participants is also chosen on the basis that five participants can contribute to finding 80% of the usability flaws (Lazar et. al. 2017). The evaluation was formed as a formative usability test where the goal is to look at metrics that are more qualitative than quantitative (Lazar et. al. 2017). In the evaluation we wanted to combine small semi-structured interviews with the users executing tasks because this could give us more information about the experience beyond the metrics.

7. 1 The evaluation plan

Set up	Candidates: Five randomly picked evaluators, the only criteria is that they hav to be students from IFI. Context: In the Institute for informatics building
Warming up	- Have you talked with a chatbot before? If yes: What type of chatbot? - How do feel about getting information from a chatbot? Do you consider the information as more or less reliable?
Task's	**Scenario:** Imagine you are a new student. Use the chatbot and try to figure out when your next lecture starts, which room it is in and where is it located? Later you are feeling thirsty and are interested in a cup of coffee near the university. **Tasks:** Use the chatbot to find out: Where is the room named 'Normarc'? Where can you buy coffee at ifi? Have a chat with the chatbot
Control questions	- Did you feel like the chatbot gave you a good answer? - Do you think that the answer from the chatbot was trustworthy? - Do you feel a need to 'double check' the answers you got from the chatbot? - If you were to rate this chatbot from 1-6 where six is the best, what would you rate it? - If low: What improvements does it need to get a six?

Figure 4: Evaluation plan

7. 2 The evaluation

The evaluation was carried out with 5 participants at IFI, where each session took about 5 minutes. After the first session we had to make some quick changes to the chatbot because it suddenly froze. We also discovered that it was casesensitive which we changed before the next session. In general the evaluation went good and we gained a lot of insight from the participants. Bellow we have summarized the main findings from the evaluation.

7.3 Findings from the evaluation

All of our participants reported that they had interacted with chatbots before, but had very little knowledge about how they worked. They found the chatbot to be nice to interact with and enjoyed that it had a friendly and casual tone. One of the participants said that she did not want a chatbot that felt too 'human-like', and that the prototype did not feel 'human-like' at all. This became clear when the same error message appears several times during the test.

They found it hard to get the right answer but when they did they were very satisfied with the answers. *"It was a good answer when I finally got the right one.."*. It was pointed out that the chatbot was not a smart chatbot, but that it provided the most necessary information sparing them from precious time spent on 'Google'.

They also reported that they trusted the answers they got, and they all pointed out that it was good that the chatbot provided a source along with the information it gave. The gifs and the pictures were also very popular among the participants, they said that this made the chatbot fun to interact with. One of the participants said that: "*It's casual, and extra fun with GIF's"*.

One of the participants also stated: "*I liked that the chatbot was casual and cute. I don't want a formal and boring chatbot, then I could have tried to find it on the university's web-pages.*" It was also pointed out that it was preferably that the chatbot could provide diverse information, "*Usually, the information is so spread that you don't know where to look*".

8. Discussion and conclusion

When testing the last prototype we got findings suggesting that the participants did not have a problem with getting information from a chatbot instead of a human. The information that they got was not seen as less trustworthy, this could be supported by the fact that the chatbot provided a source for the information it gave. It has been interesting to investigate how the participants interacted with the chatbot and how they reported on it afterwards. Our findings have some indicators leading towards that a chatbot could be a good alternative for acting as a helpful friend for freshmans at a new school. Still we have to stress the fact that the chatbot was not very intelligent and that the evaluators had to adjust their language to match the chatbots.

Because of the scope of the project we did not have time to conduct as much user testing and re-design to the chatbot as we would have liked. This has an impact on the validity of our research. Through the project we have touched on some theory when making the chatbot, but this should also have a larger focus for higher validity. Even though the participants trusted the information given in this project we cannot say that people trusts a chatbot as much as they trust a human being. There are also biases in our project, one of them is that all the students that we included in the project already knew a lot of the answer the prototype could provide. Another bias is that the information the chatbot provides could be seen as "casual" and are not crucial and/or vital This could have had an impact on the results regarding trustworthiness.

With that being said we also think that some of our findings could give some insights into how a very small group of people think about using a chatbot to gain information in a school context. Some of the characteristics of our chatbot was viewed as appropriate for the given context, like "casualness" and links to where the information was gathered. If the IFI chatbot is to be furthered developed, this could be something to draw upon.

REFERENCES

Cummings, M., 2004. Automation bias in intelligent time critical decision support systems, in: AIAA 1st Intelligent Systems Technical Conference. p. 6313

Følstad, Asbjørn (2018), INTERACTION WITH AI – MODULE 2 - Session 1, UIO Retrieved from https://www.uio.no/studier/emner/matnat/ifi/IN5480/h18/undervisningsmateriale/interacting-with-ai---module-2---session-1---v02.pdf

Hung, V., Gonzalez, A., & DeMara, R. (2009, February). Towards a context-based dialog management layer for expert systems. In Information, Process, and Knowledge Management, 2009. eKNOW'09. International Conference on (pp. 60-65). IEEE.

Jung, M., Hinds, P., 2018. Robots in the Wild: A Time for More Robust Theories of Human-Robot Interaction. ACM Trans. Hum.-Robot Interact. 7, 2:1–2:5.

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

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

Lewicki, R. J., & Bunker, B. B. (1995). Trust in relationships. Administrative Science Quarterly, 5(1), 583-601.

Lindblom J., Andreasson R. (2016) Current Challenges for UX Evaluation of Human-Robot Interaction. In: Schlick C., Trzcieliński S. (eds) Advances in Ergonomics of Manufacturing: Managing the Enterprise of the Future. Advances in Intelligent Systems and Computing, vol 490. Springer, Cham

Luger, E., & Sellen, A. (2016, May). 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.

Schank, R. C. (1987). What is AI, anyway?. AI Magazine, 8(4), 59.

Winograd, T. (1991). Thinking machines: Can there be? Are we (Vol. 200). University of California Press, Berkeley. (p.204-210)

Schibevaag, T.A. (2017, 27. September). - Hun vil revolusjonere Kommune-Norge. NRK. Hentet fra https://www.nrk.no/rogaland/de-robotiserer-kommunene-1.13706709

Abu Shawar, B., & Atwell, E. (2007). Chatbots: Are they really useful? Journal for Language Technology and Computational Linguistics, 22(1), 29-49. Retrieved from http://www.jlcl.org/2007_Heft1/Bayan_Abu-Shawar_and_Eric_Atwell.pdf

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In I. Kompatsiaris, J. Cave, A. Satsiou, G. Carle, A. Passani, E. Kontopoulos, S. Diplaris, & D. McMillan (Eds.), Internet Science: 4th International Conference, INSCI 2017 (pp. 377-392). Cham: Springer (LIGGER UNDER RESSURSER)

Molla, D. & Vicedo, J.L. (2006). Question Answering in Restricted Domains: An Overview. https://www.mitpressjournals.org/doi/pdfplus/10.1162/coli.2007.33.1.41

Minock, M. (2005): Where are the "'Killer Applications' of Restricted Domain Question Answering. https://pdfs.semanticscholar.org/2c94/9cacd519877a8b784e14b14b9beceb8e237c.pdf

Hirschberg, J. & Manning, C, D. (2015). Advances in natural language processing. *Science 349, 261/266.*

Interaction design foundation (unknown). User centered design. *https://www.interaction-design.org/literature/topics/user-centered-design*

Appendix 1: Report on conversational interaction assignment

To make the chatbot we used the program ‘Chatfuel’, that allowed us to make a chatbot in Facebook’s messenger app. This was easy to use and we managed to actually make a chatbot within a day.

In the making of the chatbot, we thought about how the chatbot could be useful and easy to interact with. The chatbot we ended up making was a chatbot that new students could use to get simple information such as where you can get coffee, where you can find the room you are looking for and where you can get food when you are at school.

To make the interaction more enjoyable we tried to make the conversation playful and we also included some gifs to make it more fun. To make the chatbot easier to use we included a lot of trigger words so that you didn’t have to know the specific words to trigger the right answers. We also included a message that said “I’m sorry I’m not that smart yet, try google” with a link to google, for whenever the chatbot could not answer. While we built the chatbot we also tested it a lot, to make sure that it gave the answers it was supposed to do.

Appendix 2: Report on machine learning assignment

For this task, the purpose was a bit unclear. We could see that it changed when tweaking the values on Epoch. As one epoch consists of one full training cycle on the training set, we predicted that it would get smarter as we changed the number to 15. But the validity accuracy did not get higher than 0,03 and the conversation was still very abstract. Difficult to decipher which of the characters that were talking.

Each of the layers is mathematical layers, given the input we get the output. In our chatbot, we only had two layers, but if you add more layers you will get more a more complex network which then could create more patterns. The drawback is that it would take much longer time.

Appendix 3: Report on problems with AI task

To this assignment we used this video:
https://www.youtube.com/watch?v=sgJLpuprQp8

Fig X Screenshot from *'SMARTHUS | Det enkle er ofte det beste | REMA 1000'* video on youtube.

Which is a constructed video made by 'Rema 1000'. The video shows a man living in a smart house where he interacts with various technologies using his voice. The video starts smoothly, describing a simple life living in a smart home. The problems arise when he has to go to the dentist, where he gets anesthesia which makes it difficult for him to say certain words and letters. This complicates things in a smart house where everything is controlled by his voice.

Even though the story portrayed is a fictitious one we consider it to be a possible scenario in real life. Especially with the voice recognition technology we have now.

By proper testing this problem would probably have been detected early. The system should also have other interaction possibilities like text input when speech is not possible—like in the video. You could have a functionality when training the speech-recognition software where you should can talk unclearly so the software knows this. But we also think there also should be a possibility to "override" the main interaction, like with the use of text. Because it can be very difficult to predict every possible outcome.

Appendix 4: Report on human-machine partnership task

We think that an intelligent agent that will take care of recruitment and hiring of new employees should have the following functionality:

- **Screening of applications**: like CV to look for experience, education etc. that are of relevance to the company. This can reduce the time it takes to go through applications, but the relevant “keywords” must be defined by the company hiring.
- **Connected to Linkedin:** screen through profiles that can be of relevance for recruiting and send mail to people with relevant backgrounds.
- **First interview:** have a mini interview with relevant applicants through the use of a chatbot etc.

Scenario 1 level 6 - “ Computer and human generate decision options, human decides and carries out with support”: The computer does all the screening of applications and comes with recommendations and options for the human to decide which candidates they should proceed the process with and which to discard. Further the interview process will include both computer and human together where the human makes all the final decisions with help from recommendations from the computer. The advantages in this scenario is that the computer takes a lot of workload from the human so that the human can focus on the what she/he considers important for the hiring process. Some of the disadvantages are that the candidates might have something more to offer than the agent can interpret. That a human could have a bigger chance of recognizing.

Scenario 2 level 8 - “Informs the human only if asked”: When the candidate applies for a job he or she are introduced to a chatbot that asks the candidate a series of questions to check if its a good fit. For example “Are you prepared to work overtime?” and “Do you have experience with data analysis?”. If the candidate turns out to be a good fit then the robot will schedule their interview.

Unfortunately humans are inherently biased and by introducing robots to the hiring process you can remove some of that. One possible problem can be that the robot is to generic and ignores the cultural fit because the applicant does not have the pre-defined characteristics that the agent takes into account. That humans probably has defined in an algoritme beforehand. An advantage is that this can speed up the hiring process. The human recruiters that remain will need to have a slightly more different skill set that the AI has. Using AI for searching and matching, putting candidates into piles could be a good solution for solving this, and then the human recruiter can do more of the tasks that are more directed (that the AI cannot perform).