



# IN5480 - Final Report

Group 6

Rajani Shrestha - [rajanis@uio.no](mailto:rajanis@uio.no)

Mattias Ovesson - [mattiaov@uio.no](mailto:mattiaov@uio.no)

Stan Rustad - [stiaru@uio.no](mailto:stiaru@uio.no)

# Table of Content

<b>About us</b>	<b>2</b>
<b>Area of interest</b>	<b>2</b>
<b>Related Literature</b>	<b>2</b>
<b>Method</b>	<b>3</b>
Litteratur review	4
Observation	4
Ethical considerations	5
<b>Findings &amp; Analysis</b>	<b>6</b>
Observation from Human - Voice assistant	6
Observation from Human - Human assistant	7
<b>Discussion</b>	<b>8</b>
<b>Conclusion</b>	<b>9</b>
<b>Lessons learned</b>	<b>9</b>
<b>References</b>	<b>10</b>
<b>Feedback</b>	<b>12</b>
<b>Appendix 1</b>	<b>13</b>
<b>Appendix 2</b>	<b>14</b>
<b>Appendix 3</b>	<b>14</b>
Abusability template	14
Useful technology	14
Benefits	15
Vulnerabilities	15
Abuse scenario	16
Guidelines for Human-Ai interaction	16

## About us

Our group consists of Rajani Shrestha, Mattias Ovesson, and Stian Rustad. All of us are on our first year on our masters in Informatic: Design, use, and interaction, two of us took our bachelor's degree in Informatic: Design, use, and interaction at the University of Oslo whilst one of us has a bachelor in Digital design & innovation from Halmstad University.

## Area of interest

We are interested in the way humans interact with a virtual assistant (VA) such as Google assistant. Google assistant is one of many popular voice assistants on the market today. Voice assistants are virtual agents that understand human speech and are able to respond to questions (Hoy, 2018). Today these VA exist in most sold smartphones, laptops, and computers (Tulshan & Dhage, 2019). These virtual assistants have many different types of functionalities (Tulshan & Dhage, 2019). People use Siri and Google assistant quite often to perform several tasks such as setting an alarm, asking about the weather and gets an instant response.

Thus we are doing an observational research study on how people will interact with a VA in contrast to a human. Our target group is between the ages of 18 to 30 years old as most of them have grown up with technology and are comfortable with using it. We will answer the research question: How does the way humans interact with an assistant differ if the assistant is replaced by an AI-infused system?

## Related Literature

We have seen an increase in the use of voice assistants in everyday life. They are now easily accessible for everyone with a smartphone, through the use of Siri, Alexa and Google assistant. Previous work in this area has tried to explore what motivates and limits the use of VAs in everyday life and what to consider in future design interactions (Luger & Sellen 2016). Most types of use cases are relatively simple such as checking the weather or setting reminders (Luger & Sellen, 2016). The study by Luger and Sellen (2016) showed that users used natural/colloquial language at first (such as, 'should I take an umbrella today'), but changed into more simple language if the voice assistant did not understand the input.

The value of using a voice assistant rather than direct manipulation like in a graphical user interface is to enable multitasking and using the hands for other activities such as cooking or bicycling (Luger & Sellen, 2016). Another value is the

time-saving aspect since completing tasks with a graphical user interface sometimes requires multiple steps. In the process of learning how voice assistants work, users removed complex words, reduced the number of words used, and used more specific terms to accomplish their tasks (Luger & Sellen, 2016). Cowen et al. (2017) writes that users are frustrated when the VA interrupts the hands free interaction and they are too embarrassed to use the VA in public spaces.

Previous work in communication between human-robot has been done by Hill et al. (2015). Interesting insights from this article is the fact that language skills transfer easily to human-chatbot communication but there are notable differences in the content and quality of those conversations, and there is an increase of profanity used when talking to a chatbot (Hill et al., 2015). Porcheron et al. (2018) argues that the conversational interaction should be considered *embedded* in a conversation and the voice assistant is not a conversationalist.

## Method

In this study we wanted to explore how humans interact with a VA compared to how they interact with a human assistant. The human assistant will be like a travel agent helping the human in planning the trip. When looking at the interaction between the human and VA, as well as the interaction between the human and the human assistant we are going to mainly focus on speech patterns, and their choice of words.

To document the interaction between the human and the two different assistants we used observation and a recording. When using both of these we have the ability to catch the oral aspect with the recording and the physical reaction while taking notes during the observation of the interaction (Kawulich, 2005).

The main focus of the observation was to carefully observe, record, and analyze the behavior of the participants while they are communicating with Google Assistant and humans.

In addition, we used the article by Hill et al. (2015) as an inspiration to our study. Hill et al. (2015) analyzed changes in communication when people were talking with an intelligent agent as opposed to another human. In this case, they used a chatbot to represent the intelligent agent, however, we used the Google voice assistant. In Hill et al. (2016) article they analyzed the conversation in seven dimensions, such as words per message, words per conversation, and emoticons used.

In our study, we asked the participants to plan a trip to a location they had never visited before. The given task was the same for the human-human, and human-voice assistant interaction. The same task was given so we could analyze the differences

between these interactions. We recorded the entire interaction so we could transcribe the conversation. With the interview transcriptions, we were able to identify and compare the interactions. Other interesting aspects of the interaction such as body movement and gestures were noted down. We transcribed all the recordings and used this when analyzing the data. For the analysis we decided to focus on: their choice of words, their attitude (politeness and behavior), and if they managed to complete their task.

## Litteratur review

A literature review was done to create an understanding of earlier research within the subject area of conversational assistant. With the literature review, we were able to explore theories and methods to answer our research question. Some examples of the search terms we used are; Voice assistant, virtual assistant, and voice interfaces. To know if the articles were relevant in our study we read the abstract and compared it to our subject area. If they were considered relevant to our subject area we read them in more detail. We used backward and forward reference search on the article by Hill et al. (2016) as it has been central in our study. We used Google Scholar as our main way to search for literature.

## Observation

We decided to do observations with people that are used to technology from a young age. As such we decided to take people in the age group 18-30 years as most of them fall into the desired group. We originally planned to observe people from the department of informatics at the University of Oslo, but due to the covid-19 situation this proved to be more difficult than originally planned as students were studying from home. We managed to get some people to perform the given task so we could observe them, but these observations were only the pilot testing of the task. When the time came to perform the actual observation, finding people willing to let us observe them doing the task was difficult as a consequence we ended up involving friends and relatives for these tasks. Using friends and relatives and not strangers means that there is a potential for bias. This is something we had in mind when we did the analysis of these observations. The part that has the highest potential for bias is the observation of the human to human interaction as the people are talking to someone they know and are comfortable with, this can make them act differently than they would a stranger. To reduce the potential bias of them changing the way they talked to the AI to be more or less polite we did not tell the participant what part of the interaction we were interested in before after the observation.

The initial plan for the observation of the task was made simple. The person was to either use GoogleAsistance or one of us as their assistant in planning a trip to a

location of their choice. During the observation, we would take an audio recording that was to be transcribed and analyzed.

We performed a pilot test of our tasks and found out that the task was too vague and needed to be worked on to get the desired information from the participants. During the testing of the human-human interaction, it became evident that the role of the human assistant was not clear enough so this was something we tried to make more specific in future tests. In both the human-human interaction as well as the human-VA interaction there were some problems in that the interaction was short and only had a couple of questions each. We gathered that this was because the people knew a lot about the location they were going to and really only needed to know for a hotel to stay at and a plane trip to get there. To get more information out of the task we decided to make the location somewhere the person had never been before. To help people that did the task we gave them some keywords to help them if they couldn't think of anything. We chose to keep the keywords to a single word each in the hope of affecting the way they asked their questions as little as possible.

After the pilot observation, we ended up changing the task so the user used Google assistance or one of us as their personal assistance to plan a trip to a location they have never visited. In case they got stuck we also handed them a paper with a couple of keywords to help them along, the keywords were as follows: Currency, Climate, Lodging, Activities, Travel, and Safety. During the task, we would observe, take notes, and record audio. The recorded audio was then later transcribed and anonymized before getting analyzed.

## Ethical considerations

We recorded a verbal consent of all the participants before we started the observations, we explained the project's purpose and gave clear instructions on what the participants' role was. We also informed them that we would be recording what was being said during the observations, and for transcribing. It was made clear that the recordings would be deleted after the project was finished. The participants also had the opportunity to cancel the observation at any time they felt uncomfortable or didn't want to further participate. In the report, they will be completely anonymous and their name will be changed to 'participant' and a number based on when they were recruited onto the project.

## Findings & Analysis

### Observation from Human - Voice assistant

The participants often used very precise and short questions when trying to accomplish a task. Very few times did the participant ask follow-up questions. Follow-up questions refer to the supporting questions that are related to previous questions which also helped in communication building around the topic. For example, if a participant asked about a suggestion for a hotel to stay then later asked for a hotel within their budget. The only time someone asked a follow-up question was when P5 started by asking for some tourist attractions and once given some suggestions, the participant wanted more information on one of them mentioned by the VA.

Interestingly P5 used the word 'search' when requesting information about hotels. P5 also formulated this as a request for the VA to do a task rather than a question about information.

*"Hey google, can you search for hotels in Rio de Janeiro" - P5*

There were several times during the interview when the VA chose to show the information on the screen rather than replying orally the participants seemed to be unsure as to whether they should continue or stop. The participants got puzzled, and this action interrupted the conversation.

The structure of the questions asked to the VA is neither polite nor rude rather they are quite neutral and short in most of the cases. Our participants rarely used polite phrases or words such as please and could you. During the conversations with the VA short and direct questions were asked like *"Weather of Kathmandu Nepal?"*, *"What is the currency of Nepal?"*, *"How can I travel to Kathmandu Nepal?"*. These questions patterns do not resemble politeness at all but not rude as well therefore was short of neutral.

There are a couple of places in the observation where it is visible that the participants do not trust the VA to understand where they are referring to and as such they state both the city and the country.

*"Weather of Kathmandu Nepal?" - P1*

*"I want to know what's the currency in Moscow Russia" - P3*

During some of the tasks, the VA was not able to understand the participant's queries. One such instance was when P5 wanted to know about travel destinations

in Brazil and the VA started reciting parts of the Wikipedia article about Brazil. At another point, P5 wanted to know about a popular mountain located in Rio de Janeiro and the VA started reciting the Wikipedia article about Peru.

*“Hey google do you know any travel destinations for brazil” - P5*

*“Hey google can you tell me about the Corcovado mountain in Rio de Janeiro” - P5*

Even though the task we gave to participants did not include giving feedback on their interaction with the VA, some of the participants told us they were surprised to know about the ability of the Google Assistant. It also denoted that the target group is not aware of the potential of VA or on the other hand we can say VA has not been explored to its fullest extent by the users.

## Observation from Human - Human assistant

In most of the human to human assistant communication, we can notice participants asking for follow up questions after they got an answer. In the case of P2, when P2 asked about traveling from one place to another and got a reply. Then P2 asked about the availability of tickets and then the cost of the ticket consecutively. This also made it possible to divide the whole communication into different topics as participants were suggested to ask questions about climate, currency, accommodation, and so on related to their travel in the future.

The participants and human assistant communication were more informal. Those who have English as a secondary language, added English words while talking to the human assistant in their native language. In one of the interviews, P6 asked a question where English words were used though the conversation was in Norwegian. As this is quite common behavior, we can generalize this as something most humans would easily understand.

In the human to human assistant interaction, both the participant and the assistant interrupted one another to add additional information or correct each other. As both the participant and the assistant are human and capable of listening and talking at the same time this was not a big problem, but this showed the informality of the conversation.

Human assistants could accept rapid feedback about the answers they were giving to the participants and adjust their answers accordingly. Such feedback could be gestures like shaking of the head, expressions with a smile, or words such as yeah and ok.

When talking to a human assistant the humans often include humor in their requests. An example of this is when P6 is asking for hotels in Hanoi that are “survivable” as a



student. While what P6 actually wanted was a cheap hotel, the human assistant understood the request and gave recommendations.

*“What about hotels in Hanoi that are survivable as a student?” - P6*

Some responses included multiple questions that related to each other,

*“That sounds reasonable, how is it located in relation to tourist activities and things to see?” - P6*

## Discussion

Luger and Sellen (2016) bring up the fact that people used natural language in their first interaction but slowly started using more simple language as soon as the voice assistance didn't understand the input. This also correlates to our findings, where the participant used few words to complete their task of finding information. The fact that a participant used the word 'search' in their dialog when talking to the VA, could indicate that they still see the VA as a search engine rather than a personal assistant. On the other hand it could mean that the participant thinks of the VA as something humanoid and is making a request for the VA to do a search for them.

Luger and Sellen (2016) also mention that users remove complex words and reduce the number of words used. This was also very evident during our observations, the participants used short sentences and didn't use language that they were unfamiliar with. One participant even shortened one of their requests after the VA did not understand. However, during our human-to-human assistant interaction, this did not occur.

When talking to a human the participant acknowledged when enough information was acquired with “jaha ja” or “ok“. Being able to interrupt the assistant seemed to be useful when given a lot of information and ask a new question. There is currently no possibility for the VA to take oral input when reciting information thus it needs to give all the data before taking in new requests. With more flexibility and the possibility to interrupt, could make the conversation more natural.

Another theme in our observation was that the participant mixed the use of English and Norwegian words. These were understandable to the human assistant and thus there were no problems in completing the request. However, this way of speaking is something that the VA is not able to handle. If the VA is not able to understand this more natural way of speaking, it could be hard for the VA to replace a human assistant. Or rather it needs to excel at the unique features that make it a VA, such as availability (Tulshan & Dhage, 2019) and functionalities (Hoy, 2018).

We found out that participants who interacted with a VA did not trust the VA as they did with human assistants. We believed that due to the doubt towards the VA, participants decided to ask fact based questions rather than descriptive. Participants might have perceived that if they say only city names, the VA might not understand them. They used words that were more obvious instead of asking informal questions that included humour or use of different languages than english.

## Conclusion

We set out to explore the research question: How does the way humans interact with an assistant differ if the assistant is replaced by an AI-infused system?

After evaluating and analyzing the findings from the observation in addition to the literature review, we could notice that there is a huge difference in interactions between humans while interacting with VA in contrast with humans interacting with another human.

The duration of interaction with VA was shorter than interaction with human assistants. The interaction with VA was more neutral rather polite whereas participants were more polite and used phrases like “please”, “could you” and “would you suggest me” while communicating with humans. But on the other hand interactions with human assistants were more informal that included descriptive text, questions with humor.

And found that when the participants were interacting with a VA they generally used shorter and more concise questions with all the necessary information to be answered separately without context. While when they talked to a human some of the questions were very closely related to the context of the conversation and would be confusing if taken out of the context.

## Lessons learned

The hardest part might have been the analysis. Using a thematic analysis on the collected data would have helped us to easier find patterns and commonalities. We should also have performed a brief semi-structured interview after the observation was done to get their thoughts and opinions on their use of a voice assistant. This would have given us richer data and helped us answer our research question.

The method, Observation is not always appropriate as it has some limitations. It is difficult to conduct and might not provide desired data and has high chances of obtaining biased data, thus the data is less reliable. Therefore combination with other

methods or use of triangulation for data collection would have been more beneficial instead.

Getting more people to participate in the observation was something we wanted to achieve but with the covid-19 situation, limited time, and a rather small group this was not something we managed to do.

Throughout the semester we have learned about AI, chatbot, robots, and human interaction with AI. We have also gained practical experience in undergoing research studies and the use of research methods such as literature review and observation. Besides this, we have learned techniques of analysis and evaluation not only by using them in our study but also by providing feedback to fellow students and groups and working on received feedback from other students and groups.

## References

- Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., ... & Bandeira, N. (2017, September). "What can I help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1-12).
- Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. *Computers in Human Behavior*, 49, 245-250. doi:10.1016/j.chb.2015.02.026
- Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly*, 37(1), 81-88. <https://doi.org/10.1080/02763869.2018.1404391>
- Kawulich, B. B. (2005, May). Participant observation as a data collection method. In *Forum qualitative sozialforschung/forum: Qualitative social research* (Vol. 6, No. 2).
- Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA". *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/2858036.2858288

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018, April). Voice interfaces in everyday life. In proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-12).

Tulshan, A. S., & Dhage, S. N. (2019). Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa. In S. M. Thampi, O. Marques, S. Krishnan, K.-C. Li, D. Ciunzo, & M. H. Kolekar (Eds.), *Advances in Signal Processing and Intelligent Recognition Systems* (pp. 190–201). Springer.  
[https://doi.org/10.1007/978-981-13-5758-9\\_17](https://doi.org/10.1007/978-981-13-5758-9_17)

# Feedback

## Feedback on Iterasjon 2

### Star:

- Good that you conducted a pilot test and described insights from it - Interesting and relevant positioning to extent research in the background section
- It creates a good flow that you choose to present the study from Hill et al. (2015). First in the background, and then getting back to it more in detail, relating it to how you plan to conduct and analyse your own study.

### Wish:

- Maybe you could write some about research ethics related to the observation study?

Yes this is a good point and something that we now have included.

- Include different methods for triangulation

Interesting point. However due to the time limit and the fact that we are only 3 people on this project this is not something that we will pursue.

- Maybe you could refer to the articles with the page number to make it more clear where the information is obtained from?

We are using the APA format for our references. If we are quoting an author(s) we will use page numbers.

### Questions:

- Is there supposed to be a second research question?

We only have one research question

- Is there a specific target group?

Yes. This will be included in the final rapport

- Did you forget to upload Appendix 1 and 2? We can not find them on your web page.

This is something we forgot to include but will be in the final rapport

# Appendix 1

## **Process:**

First, we had a quick brainstorm session to come up with a purpose for the chatbot. Some of the areas we were considering were medical advice, hotel booking guide, school counseling guide, and travel guide.

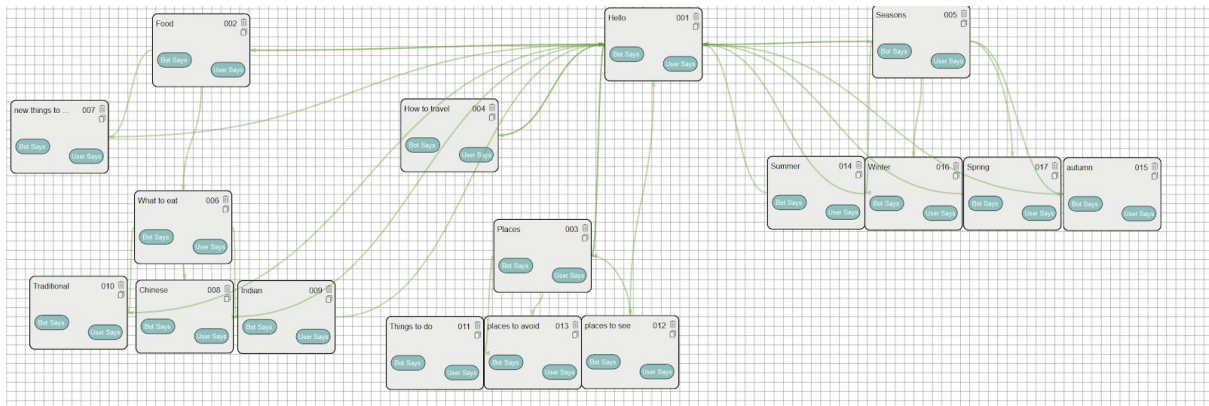
We all agreed on “Travel guide” as we all faced several problems while traveling to new places. Searching on several websites is a bit traditional and tiresome, asking someone you find on the way might not be so reliable. Thus to avoid these situations, this chatbot can be one of the solutions which will guide you during your travel period as a local guide. We based the chatbot on our own previous experiences, the problems we faced while traveling new places, we generated some questions that might be common to all the travelers. Group discussion on design flow and intents with several alternatives.

We have chosen Chatron as the main platform to develop a chatbot. We also tried out Dialogflow, however, after playing around with it for a few hours we realized that it was too complex and required more programming than we were interested in using.

## **Outcome:**

The user is presented with four options, Places, Food, Seasons, and how to travel. These are common questions a tourist has when traveling to a new place. Clicking on food brings the user to two questions, either ‘new things to try out’ or ‘what to eat’. These are different since the first one presents the user with pre-made recommendations the second ‘what to eat’ simply gives the users multiple alternatives on different cuisines to find in Oslo. The ‘how to travel’ gives advice and explanations on common ways to get around Oslo, this includes the trikk and scooters. Clicking on Places, the user is presented with typical tourist destinations and things to do in Oslo. If the user knows he or she is traveling to Oslo during a certain season they can click on ‘Seasons’ to find out information about the weather or common activities to do during their stay.

## Menu and back buttons



## Reflections and lessons learned.

Learn more about the software before you get far into the creation of the chatbot. It's not as good at text processing as we thought.

## Appendix 2

To begin with we had some problems getting into the external system we were supposed to use for the task. We also had a bit of a problem getting the hang of how to edit the code so it did what we wanted to. When we got the hang of it we started by editing the amount of layers in the neural network and det amount of neurons in each layer. We tried to go from a large amount of neurons to a smaller amount and then to a larger amount again then reduced it down to the needed output size. This gave us a higher mistake count in the compiling period than the one given to us as a template. At another round we tested going from 10240 then to 5120 then continuing to half it til we ended up close enough to the needed output this amazingly enough gave us an increase in the mistakes for each iteration which we can't explain and was not expected. During this activity we learnt that an AI based on a Neural network is really hard to teach in a way that ends with a desirable way, at least when you only have really limited knowledge about them

## Appendix 3

Abusability template

Useful technology

Chatbot from Helsenorge.no about the coronavirus

## Benefits

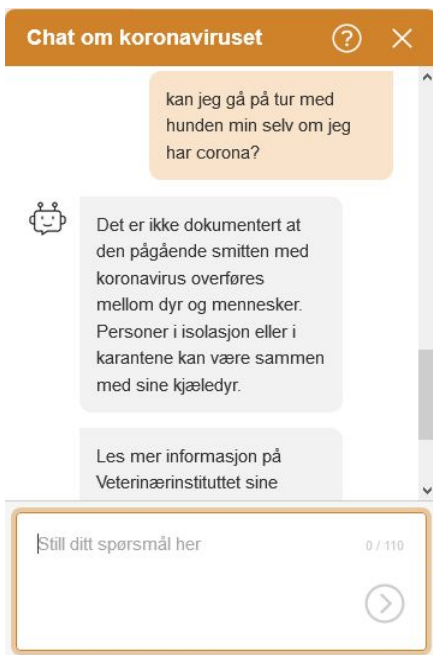
The chatbot gives an overview of the rules and regulations put in place by the government to reduce the potential risk for the people to get infected by the virus. And reduce the potential of you infecting others.

The chatbot is capable of answering questions the public might have about the coronavirus which is going to help the individuals get a better understanding of why the rules and regulations have been put in place and make them more willing to follow them.

The chatbot is able to give some information in English to people who aren't Norwegian speakers or don't have a good mastery of the language. Although this information is not as extensive and easily digestible it is a good addition added by the creators of the chatbot.

The chatbot informs you when it doesn't know the answer or doesn't understand what you want to know and asks if it should send you to a human that can help you with your question.

## Vulnerabilities



The chatbots' English capabilities are very limited and it responds in Norwegian. This makes it so that all the information people who don't understand Norwegian only get limited use out of the chatbot. The non-Norwegian speaking people that want to get information from a trusted source in the digestible way that the chatbot provides are excluded.

The chatbot has the potential of spreading miss information if it misunderstands a question and gives a wrong answer to the user which they take as a reliable answer. With this the potential of miss-information is rather big.

The chatbot has the potential of receiving feedback on the answers it gives the user and improving upon it, this can potentially make the answers less reliable if people decide to give false results.



## Abuse scenario

Imagine if the chatbot gives wrong or dangerous information about the coronavirus it does not only endanger that one individual. That one individual respects the source since it's coming from a trusted source (it's a government-owned website). Misinformation is able to spread fast and wide, this is something the whole world experienced first hand at the beginning of the coronavirus outbreak. If misinformation is spread widely enough it's possible that it gets to a point where people spread the virus to one another continually so even the ones that have had the coronavirus could potentially get infected once again and continue spreading the disease til there are so few people in the world that the virus or the human race goes extinct.

## Guidelines for Human-Ai interaction

Guideline 1. "Make clear what the system can do." - why something happened, after the fact

"Help the user understand what the AI system is capable of doing. It does make clear about its potentiality" The chatbot seems to partly fulfill this guideline. Though it's called "chat about the coronavirus", it does not indicate how the user can start a chat, for example with regards or a question or a statement but makes specific use of links and information buttons, with several themes. It also does not explain how it is getting the data that is provided, since it is operated by government level we assume that the information is from the right source.

Guideline 2. "Make clear how well the system can do what it can do."

The coronavirus chatbot from helsenorge.no clearly state when it is not capable to help the user further by writing back "Det har jeg ikke et konkret svar på. Ønskar du å bli satt over til veileder?" or "Jeg er ikke sikker på om jeg forsto alt det du spurte om, men jeg tror jeg kan hjelpe deg med:". With this response the user is able to understand the limited knowledge of the chatbot, but is also provided with alternatives to find answers outside of the AI system from either a human assistant or a link to a related subject on the webpage.

G12 Remember recent interactions. Maintain short term memory and allow the user to make efficient references to that memory.

The chatbot is not able to remember previous interactions or conversations with the user, but the chatbot does remember the last answer it gave a user and informs the user if the answer it would normally give is the same as the previous one.

G15 Encourage granular feedback. Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.

The user is able to vote on the answer given by the chatbot, letting the system know when it gave a good or bad response, according to the user. The effect of this is two folded, the first is that the user is afforded knowledge that the user can affect the outcome of the system. The second effect is that the user is helping the system getting smarter.

Guideline 11. "Make clear why the system did what it did."

It uses preset answers and does not inform why it recommended you some buttons/ links with different titles directed to information related to that title.

### **References (Appendix 3)**

Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019, May). Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).

<https://www.helsenorge.no/en/coronavirus/when-should-you-contact-a-doctor/>