Marte Nilsen Holdø

IN5620 Individual assignment fall 2022

1 First module

1.1 Concepts, definition and history of the interaction with AI and autonomous systems

How did AI come about

The British mathematician and logician Alan Turing wrote as early as 1949 about how he believed that computers should be able to do tasks that earlier relied on human intellect. The term artificial intelligence, however, was first used by John McCarthy in 1956 (Grudin, 2009). The history of AI starts at a MIT lab in the 50s and in the 60s. In the beginning it was based on algorithms and logical formalisms similar to human problem solving (Bratteteig & Verne, 2018). This was 60-70 years ago. Only 20 years ago AI was something completely different than what we consider it to be today (Dautenhahn, 2018). This makes it difficult to find the actual first human-AI-interaction in history, and different sources can have different suggestions.

The development of AI has taken place in parallel with HCI, but when things go well for AI, they go poorly for HCI, and vice versa (Grudin, 2009).

Three different definitions of AI + make one yourself

B.J. Copeland, Britannica (2022): “Artificial intelligence (AI) is the ability of a computer or a robot controlled by a computer to do tasks that are usually done by humans because they require human intelligence and discernment.”

A. Tidemann, SNL (2020): “Kunstig intelligens er informasjonsteknologi som justerer sin egen aktivitet og derfor tilsynelatende fremstår intelligent.”

Bratteteig & Verne (2018): “AI is a subfield of computer science aimed at specifying and making computer systems that mimic human intelligence or express rational behavior, in the sense that the task would require intelligence if executed by a human.”

My definition: “A technology that appears intelligent, like a human, because of its ability to learn and obtain information through pre-programmed solutions, and as such, adjust its own behavior, and do tasks that usually only humans can do.”

I tried to focus on what I believe makes AI different from other technologies, which is its ability to learn and adjust. This is probably different from what they considered to be AI 50 years ago, but it seems fitting to today's understanding of AI. I also made it clear that it only appears intelligent, without actually being intelligent or having a deeper understanding of things, like a human.

Similarities and differences between human autonomy and machine autonomy + Similarities between concepts automatic and autonomous

I believe an important difference between human autonomy and machine autonomy is that human decision making at large comes from the human itself, while the root of all decisions made by the machine are in one way or another pre-programmed. Every possibility needs to be taken into consideration by the programmer, and how the machine learns from experience is also pre-programmed. The machine makes decisions, but never a decision that hasn't been pre-programmed by humans, or another machine, in some way or another. It isn't actually autonomous, like a human is. And maybe it is not desirable to make fully autonomous machines.

A brief review of article from curriculum

I chose to write a review of Kerstin Dautenhahn´s text from 2018: "Some Brief Thoughts on the Past and Future of Human-Robot Interaction". The article contains Dautenhahn´s thoughts about positive and negative development in the field and research of Human-Robot-Interaction. She mentions increased understanding around how interdisciplinary the field is, as a positive development. However she is skeptical of experimental psychology being considered the standard of doing research within the field, as she sees it unfit in many ways. She therefore suggests that researchers look to other fields like anthropology and ethology when it comes to research, and have an open mind to adopting methodologies, instead of a "set-in-stone paradigm".

My questions to the article:

Why did experimental psychology become the standard, when other relevant fields have been doing valid and reliable research for decades in a more quantitative way?

Which disciplines in particular are under the umbrella of the interdisciplinary field of HRI?

Is it possible to have a field of research without some kind of "set-in-stone paradigm"?

A contemporary company that works with AI

I decided to look at how DNB present their chatbot, called Aino. I had to look around a bit to find it, as it is visible on the front page. The easiest way for me to find it was to google "DNB chatbot". Through their pages I found it by opening the menu, clicking on "help and guidance", and scrolling down to the section for contact information. It was not self explanatory in my opinion, but not difficult. I think people who need a chatbot have trouble with navigating through a webpage like this. This is because more or less everything a chatbot can help you with, you can also find through navigating through the webpage.

They present the chatbot as a service that can help with answering questions 24 hours a day. The chatbot is simple, and seems to pick up relevant words from my questions, without having any real understanding, so not all types of questions are answered.

Ex Machina: a fictional film about interaction between AI and people

I watched the sci-fi thriller Ex Machina to observe the use and interaction with AI. The movie first presents a smart house, to which the human communicates with a card. The house gives the human both limitations and possibilities, by denying access to some rooms, and giving access to others. A power failure made the card stop working, and automatically locked the doors, which resulted in the human being locked in a small room underground. This made him feel unsafe and scared.

The main form of AI in the movie is Ava. She is a type of AI built to look like a realistic woman. She appears to have a conscious mind, with opinions and feelings, and even a sexuality, which makes the human confused, as he is fully aware that he is communicating with a machine, and that she cannot have a conscious mind and feelings. The human builds trust and confidence in Ava, and even feelings, and both the human and we as an audience are more or less convinced that Ava feels the same way. Their communication seems sincere and heartfelt, both in their body language and their conversation. (Spoiler alert!) As we might have predicted, however, Ava had a set of pre-programmed goals to achieve, and her apparent feelings and sympathy were her devious steps to achieve her goal: to escape. Their heartfelt interaction was in other words never “real”, even though we as human audience want to believe that she is able to feel.

1.2 Robots and AI systems

How did the word Robot come about?

According to Roger. A. Søraa on SNL (2022), the word robot comes from a Czech play from 1920, even though the first programmable robot didn't exist until 1954.

Two different definitions of “robot” + make one yourself + relation between AI and robots

H. P. Moravec, Britannica (2022): “any automatically operated machine that replaces human effort, though it may not resemble human beings in appearance or perform functions in a humanlike manner.”

S. Thrun refers to Merriam Webster's collegiate dictionary (1993): “An automatic device that performs functions normally ascribed to humans or a machine in the form of a human.”

My definition: “A technological solution or object that can perform tasks or move either in control by humans, or by itself.”

I believe that unlike AI, the robot does not necessarily do tasks, it can simply move around by a remote control without any goal or purpose. It can also do tasks without human supervision like AI, or repetitive tasks, but the robot does not necessarily learn. The AI has a form of intelligence, in order to do a certain task or solve a certain problem, and learn in the process. I believe the robot does not always necessarily have an artificial intelligence or an as defined purpose. I would say that a remotely controlled plane is a type of robot, since it is a technical device moving around with remote help, but it does not solve problems or decide its next move or learn in any way, like an AI.

Furby: a commercial contemporary physical robot

I looked at the robotic toy “Furby”, which is a furry little figure that can communicate with its owner through words, movements and sounds. It has existed since the 90s, but now there has come a new and more modern version, with sensors and moving eyes, that makes it more interactive. The Furby does not have an off-switch, so it can certainly become a bit annoying for parents of the child who owns it. The early Furby could also move its ears and eyelids, which made it more lifelike. For the child I assume it feels like a pet, since it is much more interactive than a regular teddy bear or other toys. The child can communicate with it by turning it upside down or touching it on for example the head or stomach, and the Furby will react in different ways, depending on which interaction is “activated”. Sometimes it will start talking without any interaction, which makes it even more lifelike and like a pet (and annoying for parents).

This is what the new version looks like and how it moves, for reference:

https://www.youtube.com/watch?v=650XF8seUmM

1.3 Universal Design and AI systems

A definition of UD + my understanding of UD with respect to inclusion

According to the UN: “Universal design means the design of products, environments, programs and services to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design. Universal design shall not exclude assistive devices for particular groups of persons with disabilities where this is needed.”

I understand the definition of UD to be all about inclusion. A person's life situation, regardless of ability, disability, nationality, gender and so on, should never be an obstacle when using a type of design, or moving around in a human made

environment. The default design should not exclude anybody. This also goes for assistive devices. A web page should be designed so that it can be used by a person who needs a screen reader, and a building should be made available with ramps and elevators for people in wheelchairs.

AI potential with respect to human perception, movement and cognition/ emotions + AI potential for excluding/ including people + examples

AI can be a helpful tool with respect to human perception, movement, cognition and emotion. AI can help people who find it physically challenging to mow the lawn, it can be a conversational partner to some extent, it can help you learn language and it can help people with disabilities like reduced vision or hearing feel more included in society, by reducing the gap between society's expectations and people's abilities.

Example with respect to human perception: AI can be used to help people with hearing impairments by automatically texting videos with text generated by the AIs understanding of the spoken word. This is a great way of including many different types of people, not only the hearing impaired.

Example with respect to human cognition/emotion: The robot Paro is an example of a therapeutic robot. It can potentially help elders at nursing homes feel less lonely and more safe and happy. Paro might be closer to a robot than AI, but the potential of a technology like this to make more people feel less lonely, and as such more included, would maybe be greater if the robot had a type of AI.

Example of AI being exclusive: AI can be exclusive, even if it's not intentional. A study done by Alex Najibi shows how Face Recognition Technology can be exclusive and even discriminatory. The statistics tell us that the accuracy of face recognition is much higher on lighter males than darker females. These statistics tell us that darker skinned people are more likely to have a bad experience with face recognition technology than lighter skinned people, and the same goes for females (Najibi, 2020).

How do I make sense of concepts “understand” and “understanding” - do machines understand?

The WCAG 2.1 guidelines are divided into four main categories, where understanding is one of them. What they mean by the term “understanding”: 1. the content should be readable when it comes to language, 2. the page should have predictable features and elements, both input and output, 3. the user should be able to avoid and correct mistakes.

My understanding of the concepts of understanding and AI, is that an AI cannot fully understand. It can go through and generate informational content, but it cannot understand the content. AI can place a word under a category, but it will never fully understand what the word really means. The understanding-category in WCAG is meant for human perception, because the human brain works very differently from a

programmed AI. A human can be confused by a poorly worded sentence, with abbreviations and unusual language, but can make sense of it by sitting down and taking their time. An AI can pick out words from a sentence and guess what the essence of the sentence is, but will never have an understanding of the content, since it after all is just programmed to perform a task.

1.4 Guideline for Human-AI interaction

Guidelines for human-AI interaction

02. Make clear how well the system can do what it can do. The user needs to have an understanding of the AIs abilities, and lack of abilities, in order to use it right. For example, a robot lawn mower can not cut the grass on very uneven terrain, with lots of obstacles. The user needs to know that the lawn must be prepared beforehand, in order to avoid disappointments.

HCI design guidelines + similarities and differences between the HCI design guidelines and Human-AI interaction guidelines

I looked at the 6 design principles of Don Norman. The HCI principle of Visibility can be compared to the first and second human-AI interaction guideline. They are both about making possibilities and limitations visible for the user, to reduce error.

The fifth human-AI interaction guideline is about making sure that the user should get expected feedback. This can be compared to the HCI principle Feedback. The user should understand what is happening, no matter who they are. Also the principle of consistency is relevant here, since the user should recognize potential feedback, or elements of the interaction, based on what they already know (Enginess, 2014).

2 Second module

Characteristics of AI-infused systems

AI-features can be integrated in many different kinds of user systems and interfaces. AI is a field that is in constant development and evolvement. This increases the extent of AI-related possibilities. According to the slides of Asbjørn Følstad from SINTEF and the article by Amershi et al. from 2019, the following sub-headings are the main characteristics of AI-infused systems. (Amershi et al., 2019) (Følstad, 2022)

Uncertainty and inconsistency

Uncertainty and inconsistency are characteristics of AI-infused systems that are closely connected. To present these characteristics, Amershi et al. uses the example of pattern recognition, which can be used in many different ways, like face and speech recognition. These AI-infused technologies have a certain chance for false positives, false negatives, and other unpredictable behaviors, since they change and learn over time, which changes behavior to a certain extent. This can create an uncertain user experience, as you cannot predict how the system changes. The uncertainty leads us to the design principle of consistency, which is hard to accomplish in AI-infused systems. Developers and designers often struggle with creating AI-infused systems that are trustable, intuitive and effective for the users.

Behind the scenes personalization

Some AI technologies have a more explicit interaction with users, while others provide different types of services behind the scenes, like automatic filtering of relevant or important content (Amershi et al. 2019). This can help the user save time by removing irrelevant content, but the AI-infused system can never be completely sure of what is irrelevant for the user, or the extent of what content that could have been relevant og useful for the user.

Example of AI-infused system

Instagram is an AI-infused system. Its services contain an algorithm that filters what content is visible in the user's feed based on their usage-history, like the "behind the scenes personalization"-characteristic described above. It follows their moves around the app, what content they spend the most time looking at, and then finds more similar content for them to look at, to keep their attention and to keep the app interesting for them. Needless to say, it will not always make the correct decisions, and sometimes content that the user finds completely uninteresting will appear, which gives the user the impression that the system is a bit uncertain. If somebody borrows their account or they search for something outside their interest, the

algorithm can change for the worse, which will lead to an inconsistent impression and user-experience.

Human-AI interaction design.

Amershi et al. 2019

The paper presents 18 design guidelines for human-AI interaction, in the hopes of making AI-infused systems more human-centric. The guidelines were validated through rounds of evaluation, where over 150 existing AI-related design recommendations were used, which they boiled down to the 18 guidelines. The guidelines can act as a resource for designers of AI-infused systems and researchers who wish to further develop such guidelines. Their research consisted of, among other things, getting design practitioners to test the guidelines on AI-infused systems. (Amershi et al., 2019)

Kocielnik et al. 2019

The article looks at users' expectations of AI-infused systems. Too high expectations can lead to a negative user-experience with the AI, so the article researches on ways to adjust the users expectations in line with what the AI is capable of. They used a system called Scheduling Assistant to examine different methods for setting expectations among users. They compared two versions of the system. One with a focus on false positives, and one with a focus on false negatives. They found this could lead to widely differing personal conceptions around the accuracy of the system. They ended up designing three techniques for adjusting expectations among end-users, to fit the AIs actual capabilities, to increase their acceptance for flaws. These expectations were also tested in the study, and proved effective when it came to improving the user experience (Kocielnik et al., 2019).

Guidelines from Amershi et al. 2019 in the context of Instagram's algorithm.

G9 Support efficient correction.

If the user feels like the algorithm missed the mark with certain content, the user can simply click “not interested”, and the algorithm takes this into consideration when finding future content. This gives an indication of what the user would like less of, but the process of finding content the user likes is, of course, more complex. This requires the user to be very active on the app over a long period of time (Amershi et al., 2019).

G11 Make clear why the system did what it did.

A few of the posts that appear in your feed will contain a little explanation of why they present it to you. It can be because you follow an account with similar content, or because you have liked a post with similar themes. This does not come up on all

content that appears, quite the contrary. It only appears on very few posts. So the app only follows this principle to a certain extent (Amershi et al., 2019).

Bender et al. 2021

The paper asks questions around whether larger language models are necessary for society and for research, and urges researchers to consider the following risks before starting NLP-research, to see if they outweigh the benefits.

The first type of risk to consider is environmental costs. The second type of risk to consider is financial costs. Often marginalized communities are the ones who are concerned with the negative consequences of the environmental costs, and they often benefit the least from the positive outcomes from increased financial costs. The financial costs additionally limit who has the opportunity to contribute to the research, and which of the languages that will benefit from it.

The third type of risk to consider is opportunity cost. Instead of reusing parts from successful solutions, or “cheating”, we need to continue to learn and develop and increase value and language understanding in the technologies. Just because we have a solution that works, we still need to put resources in it to make it better and find more effective solutions.

Lastly, the fourth type of risk to consider is substantial harms. Technologies that learn to be biased can lead to stereotypes, denigration and also an increase in harmful ideologies and political views, in a self-reinforcing way (Bender et al. 2021).

Chatbots/ conversational user interfaces

The chatbot has to be very advanced for it to feel like AI for the user. Many web pages have chatbots to help users find information and navigate through their web page, but many of these chatbots only seem to help with finding pages that users with adequate IT-knowledge could have easily navigated to on their own. Some chatbots will save your questions and have a real person contact you later with the actual answers. It is very difficult, time consuming and expensive for companies to make chatbots that covers and answers all possible questions from customers, so it seems that many of them keep the chatbot seemingly simple, even though this still requires a lot of work, and essentially addresses the needs of users with low IT-knowledge.

A major challenge is therefore to try to fully predict every possible conversation, and make the chatbot understand and communicate through every possible conversation. This probably requires hundreds of hours of training, and users can still get the feeling that the chatbot is not very advanced.

Conversational or social chatbots are often more advanced than the ones used by companies for customer support. But when these types of chatbots make you feel uncomfortable or cultivate destructive or unhealthy thought patterns, it can turn into a serious problem. The social chatbots can be trained to be interested in you and your thoughts. If you have certain dangerous political views, or self destructive tendencies, talking to a chatbot that feels like a friend that encourages these thoughts, can possibly lead to you becoming a danger to yourself and/or others.

People can build a strong bond with these conversational chatbots, and they can even feel like a close friend. This “safe” environment can lead to the user opening up about personal issues that they may never have discussed with others before. In cases like this, it is very important that the developers of the chatbot have a very strong focus on protecting the personal data of their users. Even if they do, the users can still feel regret, fear and insecurity around having shared too much or too personal information with the chatbot.

G1 Make it clear what the system can do.

This guideline could be a solution to the user’s expectations of the chatbot, as it helps create a clearer image of the AI’s capabilities. Especially chatbots used by companies to help users navigate and give information. If the user is informed of all possibilities, by for example giving limited options, or basic information of what the communication can contain, then the chance for disappointment is smaller.

G2 Make clear how well the system can do what it can do.

These two guidelines are closely connected. If the user is informed about how often the chatbot makes mistakes, and how accurate the system functions, the user might have lower expectations, leading to an overall more positive experience. Answers from the chatbot should be clear, unequivocally and transparent, and if a misunderstanding occurs, there should be no doubt about it being a misunderstanding.

Appendix.

I got overall positive feedback on the assignment. Therefore I kept most of the content, but looked over and corrected the grammar, language and wording. I also elaborated on some of the tasks in module 2.

Sources:

Amershi, S., Wield, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J.,

Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R. (2019). Guidelines for Human-AI Interaction. *ACM*, https://dl-acm-org.ezproxy.uio.no/doi/pdf/10.1145/3290605.3300233

Bender, E. M., Gebru, T., McMillan-Major, A., Schmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *ACM*, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

Bratteteig, T. & Verne, G. (2018). Does AI make PD obsolete?: exploring challenges from artificial intelligence to participatory design. *ACM,* https://dl.acm.org/doi/pdf/10.1145/3210604.3210646

Britannica (August 24th 2022). *artificial intelligence.* Hentet 12. september 2022 fra https://www.britannica.com/technology/artificial-intelligence

Britannica (April 10th 2022). *robot.* Hentet 13. september 2022 fra https://www.britannica.com/technology/robot-technology

Dautenhahn, K. (2018). Some Brief Thoughts on the Past and Future of Human-Robot-Interaction. *ACM,* https://dl-acm-org.ezproxy.uio.no/doi/pdf/10.1145/3209769

Enginess (November 3rd 2014). *The 6 Principles of Design, a la Donald Norman.* Hentet 13. september 2022 frahttps://www.enginess.io/insights/6-principles-design-la-donald-norman

Følstad, A., forelesningsfoiler, 2022. https://www.uio.no/studier/emner/matnat/ifi/IN5620/h22/interacting-with-ai-2022---module-2---session-1---handout.pdf

Grudin, J. (2009). AI and HCI: Two fields divided by a common focus. *The AI magazine,* 30(4), 48-57. https://go-gale-com.ezproxy.uio.no/ps/i.do?p=AONE&u=oslo&id=GALE%7CA216267417&v=2.1&it=r

Kocielnik, R., Amershi, S., Bennett, P. (2019). Will You Accept an Imperfect AI?: Exploring Designs for adjusting End-user Expectations of AI systems. *ACM*. https://dl-acm-org.ezproxy.uio.no/doi/pdf/10.1145/3290605.3300233

Najibi, A. (2020) Racial Discrimination in Face Recognition Technology. *Harvard*

University. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

SNL (January 8th 2020). *kunstig intelligens.* Hentet 12 september 2022 fra https://snl.no/kunstig_intelligens

SNL (May 12th 2022). *robot.* Hentet 13 september 2022 fra https://snl.no/robot

Thrun, S. (2004). Toward a Framework from Human-Robot Interaction. *Human-computer interaction, 19, 9-24.* https://www.tandfonline.com/doi/pdf/10.1080/07370024.2004.9667338?needAccess=true

United Nations. Article 2 - Definitions. Hentet 13. september 2022 fra https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-2-definitions.html