Individual assignment

erikhma - Iteration 2

IN5480

1.1 Concepts, definition and history of AI and interaction with AI

How AI came about:

The American mathematician and logician John McCarthy is credited as the source of the first appearance of the term "artificial intelligence". While the term itself was only thought of and thereby "born" in preparation for a conference at Dartmouth College in the US in 1956 (LiveScience, 2014), the idea of intelligent machines had already been around for several years. Alan Turing wrote in the *London Times* in 1949 that "I do not see why [the computer] should not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms" (Grudin, 2009, p. 49). Isaac Asimov had also been working with similar ideas by introducing three laws of robotics through his collection of novels by the name "I, Robot" (ibid.).

Three definitions of AI & my own:

Definition of AI taken from Bratteteig & Verne, 2018 (p. 1-2) :

*"*AI is a subfield of computer science aimed at specifying and making computer systems that mimic human intelligence or express rational behaviour, in the sense that the task would require intelligence if executed by a human."

This is a relatively recent definition and shows the concern of the authors when it comes to "mimicry" of human intelligence and perhaps discretely pointing to AI's lack of true emotions and consciousness. The field of Participatory Design (PD) is generally focused on human-human interactions where sharing knowledge and mutual learning is key. While AI may be an extremely fast learner in some regards, it may lack other valuable human traits such as the ones mentioned previously.

AI defined by John McCarthy (according to Stanford University):

"The science and engineering of making intelligent machines"

This broad definition is starting to show its age a bit now as we have seen the rise of AI neural networks and the like which are able to "engineer" themselves through self-learning. The field of AI research is still very much focused on the science and engineering of making intelligent machines though, but likely in a different way than what John McCarthy would have experienced and imagined in the mid-1950s.

ICO definition of AI:

"AI is an umbrella term for a range of technologies and approaches that often attempt to mimic human thought to solve complex tasks. Things that humans have traditionally done by thinking and reasoning are increasingly being done by, or with the help of, AI."

ICO's definition is my preferred one, and it is quite similar to the definition by Bratteteig & Verne (2018). First off, it points out that AI is a broad term used to describe different technologies. This is very important to make clear as many different technological gadgets, programs, artefacts, etc. can be described by saying that they are artificially intelligent. This definition also includes a partial description of intelligence by saying that AI attempts to mimic human thought (e.g. logic, learning, problem-solving) to solve complex tasks. The latter being what we usually think of as requiring intelligence by humans, which Bratteteig & Verne (2018) also included in their definition of AI.

Based on these definitions I would like to incorporate parts of them into my own. I think the man-made aspect of John McCarthy's definition is interesting and deserves to be included. For the intelligence to be artificial it has to be created by humans - at least at some level. The ability to solve "complex" problems is also important to the definition. Such problems require

varying levels of intelligence from humans and hence machines which are able to solve these should also be considered "intelligent" to some degree, even though they merely mimic human thought and are not able to be conscious or have emotions - yet.

Brief review of The 'Problem' with Automation: Inappropriate Feedback and Interaction, not 'Over-Automation' (Norman 1990):

The article addresses the notion of "over-automation" and presents a few dramatic examples where accidents happened due to a lack of human intervention when unexpected events occured. Automated systems were involved in most of these, and the author digs deeper into what *really* caused these accidents and how they could have been avoided. Taking a stand against the stance of "automation is too powerful", the author instead claims that it is not powerful enough due to how the presented systems acted in relation to their human operators under the unexpected circumstances of the examples given.

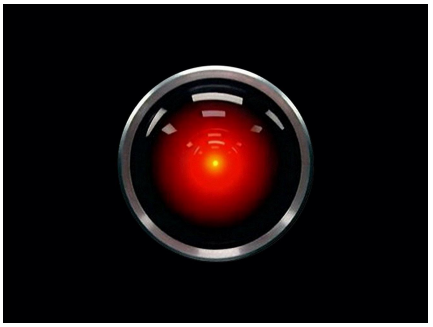
Subtle remarks such as "huh, this is weird" would be what a human pilot could say in an unexpected situation while flying, whereas the autopilot may just automatically adjust what it is doing in order to compensate for mechanical failures in the aircraft, without giving any indication to the human crew that something might be wrong. A lack of appropriate feedback from the automated systems, in addition to what the author refers to as "mental isolation" on the part of the humans involved, are the culprits of these accidents according to the author.

Description of AI used by Boston Dynamics' Spot robot:

"Out-of-the-box, Spot has an inherent sense of balance and perception that enables it to walk steadily on a wide variety of terrains. This form of AI that we call 'athletic intelligence' allows Spot to walk, climb stairs, avoid obstacles, traverse difficult terrain, and autonomously follow preset routes with little or no input from users."

The way Boston Dynamics presents their use of AI is that it "allows" their robot to do certain things through what I would call mimicry of balance and perception. It appears as though they see AI as something that enables functional features through a kind of framework.

Film about human interaction with AI:



The movie *2001: A space odyssey* depicts the fictional AI character Hal 9000, who describes himself as "the most reliable

computer ever made" who is "incapable of making mistakes or distorting information". While he may appear helpful to the human characters in the film, the viewers are more likely unsettled by the way he is presented through close-up shots of his red "eye". He communicates with the human characters with a natural male voice and slightly formal language. ((Spoiler alert!)) As many viewers could perhaps predict, he turns against the human characters when they start planning to disconnect Hal, which would effectively be killing him. Hal appears to have a state of consciousness, which slowly fades as his circuits are being disconnected by the humans.

1.2 Robots and AI systems

How the word "Robot" came about:

The word Robot has its roots in the Slavic language as robota, which means forced labour ("Robot", 2021). It first appeared as a term to describe "artificial human bodies without souls" in a 1920s play by Czech writer Karel Capek (ibid.).

Two different definitions of "robot":

The Robot Institute of America defined a robot as:

"A reprogrammable, multifunctional manipulator designed to move materials, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks" in 1979 (Thrun, 2004).

This is a very specific definition going into detail about the operations a robot should perform, but perhaps not broad enough to cover what we would consider a robot today. Thrun (2004, p. 9) also touches on this in the same article where he states that "Robotics is a field in change; the meaning of the term robot today differs substantially from the term just 1 decade ago.". This seems to ring very true, looking back at this definition from over 4 decades ago.

Definition of robot by Britannica (2021):

"any automatically operated machine that replaces human effort, though it may not resemble human beings in appearance or perform functions in a humanlike manner."

Britannica provides a much broader definition compared to the Robot Institute of America's definition. Robots in 2021 are likely very different from the robots of old from several decades ago, and this shows in the definitions given.

My own definition of robot:

A machine which operates with a high degree of autonomy and is able to sense its environment and also makes decisions and acts based on what it processes about it.

The relation between AI and Robots:

I think AI and Robots are quite connected. Especially modern robots incorporate AI to a high degree in order to mimic human actions and perform tasks which would require intelligence if done by humans. A high degree of autonomy as part of my definition of robot would likely involve some form of artificial intelligence. Processing signals from sensing an environment and deciding on actions based on them requires some form of intelligence, at least I would imagine so as it involves solving relatively complex tasks.

About a contemporary robot:

"Dyret" (Dynamic Robot for Embodied Testing) is a four-legged robot designed and made at the Institute of Informatics at UiO. Its purpose is to use AI to "teach" itself how to walk on different surfaces, such as grass and varying types of carpets. This robot has the ability to change the length of its legs in order to facilitate this. Human interaction is limited, but its "owner" and maker Tønnes Nygaard seems to have developed somewhat of an emotional bond to his creation (Torgersen, 2020).

1.3 Universal Design and AI systems

Description of Universal Design from DO.IT (2021) with an explanation:

"Universal design is the process of creating products that are accessible to people with a wide range of abilities, disabilities, and other characteristics."

This definition shows the broadness and inclusive nature of Universal Design. A core element is making products more accessible and better for everyone, which also includes people "without disabilities" (although we can all have disabilities under certain circumstances).

The potential of AI with respect to human perception, human movement and human cognition/emotions:

AI is extremely able to mimic human cognition in particular. AI vision and speech recognition/reproduction has the potential to be vastly superior to their human counterparts. Measurements based on sensors such as LIDAR could be orders of magnitude more precise than what humans are able to produce. Autonomous/self-driving cars are only showing us a glimpse of what this type of technology can do in combination with AI.

Speech and language processing has already come a long way. We are now able to automatically annotate videos with audio in order to make them more accessible, although the fidelity of this technology is not yet in a great place. Reproduction of natural language is starting to take off in the tech sphere, as AI algorithms are able to be trained on audio examples of a specific person speaking and then eventually being able to produce words and sentences with their voice which that person has never said themselves (see for example https://www.resemble.ai/). This has incredible potential for converting books, learning material and other written works into something akin to audio books, read out loud by the AI-generated voice of the author.

The potential for AI to include and exclude people:

As already mentioned, AI has great potential to include people regardless of their abilities or disabilities. The potential for AI to exclude people can be seen in face recognition where the AI has been trained using a limited data set. If a facial recognition AI is trained with pictures of white men, it will become an expert at recognizing images of exactly that. However, if presented with images of a person with a different skin color or gender, this AI will likely seem less-than intelligent as it fails to recognize the person being shown (Vox, n.d). The principle of "shit in, shit out" is very true when it comes to diversity. AI can be great at recognizing patterns in images, but if the collection of patterns they are being trained on is not diverse enough this could lead to exclusion of one or more groups of people based on their gender, skin color, etc.

Explanation of "understand" and "understanding" with regards to AI. Do machines "understand"?:

Note for iteration 2: added how I make sense of "understanding".

A key part of “understanding” something is, from my perspective, the ability to apply knowledge and experience with certain familiar concepts to new and unfamiliar ones. As far as I know, machines are not able to do this yet, but I think some AI are perhaps close to being able to “understand” certain things, as in they are able to make models of what they perceive, either digitally or physically. That being said, I think it would be near-impossible to make an AI fully understand something more abstract such as a complex theoretical concept. An AI capable of truly understanding would likely pass the Turing Test for most, if not all topics, and be at a similar level of intelligence to the “replicants” seen in the Blade Runner movies. I believe that the ability to understand concepts the way a human does requires a consciousness in order to assess and think critically about something. As far as I am aware, no AI is able to realize or “understand” that it does **not** understand something either. AI can be great at faking understanding, though, and due to this, I think AI will often appear more intelligent than it actually is.

1.4 Guideline for Human-AI interaction

Description of 1 human-AI interaction guideline from Microsoft with a different example:

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

The autopilot discussed by Norman (1990, p. 586-587) in “the case of the loss of engine power” is a great example of why this is an important guideline. The crew of the plane was not sufficiently informed of the critical failure in the aircraft by the automated system and this nearly resulted in a tragic accident, a situation which could have been entirely avoided if the system made it clear why it was severely compensating to keep the plane stable.

Comparison of the Microsoft human-AI interaction guidelines and Norman’s 7 Fundamental Design Principles:

Norman’s first principle of discoverability is directly related to Microsoft’s first 2 guidelines of making it clear what the system can do and how well it can do it. Discoverability is all about is possible, given the current state of something. It also goes without saying that Norman’s second principle of feedback is closely connected to guideline 11, which I already covered. I would say that both Norman’s principles and Microsoft’s guidelines are quite focused on usability and user experience, although Norman is perhaps more on the side of helping with the latter.

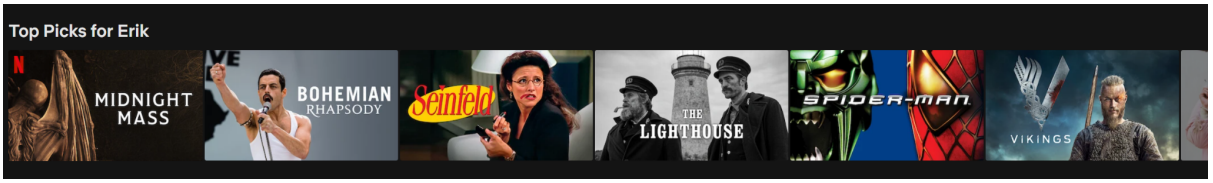
2.1 Characteristics of AI-infused systems

Identification and description of key characteristics of AI-infused systems:

AI-infused systems can be described as systems that "have features harnessing AI capabilities that are directly exposed to the end user." (Amershi et al., 2019). Introducing four characteristics of AI-infused systems, Asbjørn Følstad describes these as "Learning, Improving, Black box and Fuelled by large data sets" (Følstad, 2021). Through learning, these systems are constantly changing, which in turn should lead to improvements. Big data sets from real-world use of the system, or training sets (Goodwin, 2021) are used to "feed" the system. Many of the learning processes which can lead to improvements are not visible to the user, and can often be too complex even for the creators of the system to fully understand. This is why AI-infused systems are described as having the property of being opaque, black boxes (Følstad, 2021). In order to learn, such systems rely on making mistakes. The process of trying and failing is inherent to their design, as this is how AI-infused systems can gradually improve over time.

Key characteristics of Netflix's recommendation system and the implications of these:

Netflix has a recommendation system which presents each user with a selection of movies which have likely been picked out by an AI:



This system seems to be heavily influenced by what I, as a user, have been watching recently. It is likely to have been "trained" on data containing my watching habits, and perhaps other factors such as my age, gender, etc. These are all assumptions I am making about the AI however, as I am in no way presented with an explanation explicitly here as to how the system learns and why I am getting exactly these recommendations, which makes the system appear as a black box. I would imagine that this system is trained on a very big data set on a "generic" level. How it finds recommendations based on actors, genres, themes, etc. is likely based on aggregated data from most users of Netflix.

I can imagine this system being a nightmare for other users who share their accounts with parents/children/partners/friends without separate profiles. This system also seems to be in

charge of which categories of series and movies are shown on the page at all times, so if a user who is into horror shares a profile with their parents which are into romantic comedies, I could see the AI getting very confused and the result being a worse user experience for both.

2.2 Human-AI interaction design

Summary of Amershi et al. (2019):

Amershi et al. have created a set of 18 design guidelines for interaction between humans and AI. These are based on thorough research of previous studies on best-practices and interaction mechanics. Their goal is to empower designers who are creating AI systems which interact with humans. The guidelines are split into 4 categories, being different states: “Initially”, “During interaction”, “When wrong” and “Over time”. These guidelines can also be used as evaluation criteria for designers when looking at pre-existing systems.

Summary of Kocielnik et al. (2019):

Kocielnik et al. performed a series of experiments to explore how users would react to an AI-infused system which was imperfect in different ways. They also tried seeing how users responded to different methods of adjusting their expectations, through the use of three different techniques, prior to the use of the system. The imperfections of the system being tested were adjustable, and the researchers could skew the error rates of false-positive results against false-negative results in order to conduct experiments with users. The goal was to see whether users prefer one type of error over the other, under the condition that the system had the same overall rate of producing errors.

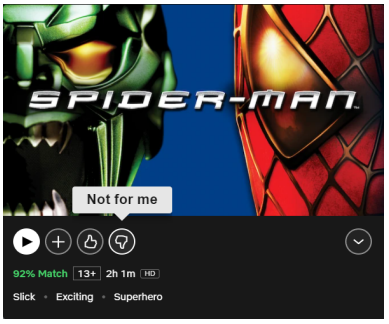
They were able to conclude that preparing the user for errors by adjusting the users’ expectations prior to use of the imperfect system would increase their acceptance of it. Their results for the two different types of errors were more inconclusive.

Summary of the main argument in Bender et al. (2021):

Bender et al (2021) present some important challenges in developing and using large-scale language models (LMs). They argue that the notion of ‘bigger is better’ does not necessarily apply for LMs, as the constant need for these to increase in size has also led to massively increased environmental and financial costs, as well as a seemingly reduced focus on data quality. These large LMs are referred to as ‘stochastic parrots’ by the authors, due to the

LMs' lack of "reference to meaning" in their output. The presented dangers of these LMs are manyfold, with perhaps the most worrying of which being the potential for automation bias in text generation based on flawed input data and lack of sufficient filtering parameters set within the LMs.

How Netflix's recommendation system relates to a few of Amershi et al. (2019)'s guidelines:



When I get recommendations which I do not like, I am thankfully able to "tell" the system that I do not like the recommendation, which follows the principles of efficient dismissal and correction in Guidelines G8 and G9 (Amershi et al. 2019) respectively. Removing the recommendation is as simple as mousing over the thumbnail of the unwanted suggestion and selecting "Not for me". This will likely also inform the system that the recommendation it made was wrong and it should then learn from this.

2.3 Chatbots / conversational user interfaces

Design of chatbots/conversational user interfaces:

One of the key challenges in designing chatbots or conversational user interfaces is the "need" for traditional GUI interaction to be converted into interactions through dialogue with chatbots. This type of conversion has also brought to light a need for a more holistic approach to this type of design. "Zooming out" and looking at the whole service being provided, instead of just the direct user interaction with a tool or a website. As an example, the chatbot HelseVenn appears to only be a part of the health services' offerings to high school students in Norway. Looking at the design of HelseVenn in isolation seems to make little sense, as it is part of a bigger service system, and it should then be treated as such.

Another challenge comes from having to design with a heterogeneous set of users in mind. The language they use and the way they perhaps ask questions can vary wildly depending on the individual, and chatbots should be designed to help them and treat them all equally. It is not hard to imagine that even more advanced AI conversational agents will likely require vast amounts of parameters and training data to accomodate all users.

With regards to Guideline G1 - *Make clear what the system can do*:

I think a key challenge related to Guideline G1 is communicating the abilities and limitations of a chatbot in an efficient manner. By making it very explicit, for example through a message containing all types of questions the bot is scripted to handle, there is a risk that users will be annoyed by an overwhelming amount of information. If the first thing I saw when I started a chat with a bot was 6 messages containing every type of question it could assist me with and all limitations of the system I can imagine I would be put off using it immediately. I think it should be made clear that the chatbot is in fact limited in its ability to help a user, but designers should be careful to not cause a cognitive overload for users.

With regards to Guideline G2 - *Make clear how well the system can do what it can do*:

The intention of Guideline G2 seems to be setting an appropriate level of expectations on the behalf of chatbot users. For example, if a chatbot presents itself and says that it **may** be able to help answer **some** questions the users have, instead of saying that it **will** be able to answer **all** questions, this will likely lower the expectations of the chatbot's performance to a realistic level. In some cases it may even be possible to present users with statistics on how sure the chatbot is in its decisions of which answers to give to certain questions.

References

LiveScience. (2014, December 4). A Brief History of Artificial Intelligence. https://www.livescience.com/49007-history-of-artificial-intelligence.html accessed 07/09/21 16:18.

Bratteteig, T., & Verne, G. (2018). Does AI make PD obsolete?: Exploring challenges from artificial intelligence to participatory design. *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2*, 1–5. https://doi.org/10.1145/3210604.3210646

Manning, C. (2020, September). Artificial Intelligence Definitions. https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf accessed 07/09/21 19:14

Boston Dynamics. (n.d). https://www.bostondynamics.com/about#Q7 accessed 07/09/21 21:45.

Norman, D. (1990). The problem of automation: Inappropirate feedback and interaction, not over-automation. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, Vol. 327, No. 1241, Human Factors in Hazardous Situations (Apr. 12, 1990), pp. 585-593 (9 pages)

Robot. (2021, September 8). In *Wikipedia*. https://en.wikipedia.org/wiki/Robot#History accessed 08/09/21 20:10.

Thrun, S (2004). Toward a Framework for Human-Robot Interaction. Human–Computer Interaction, 19:1-2, 9-24, DOI: 10.1080/07370024.2004.9667338

Moravec, H. Peter (2021, February 4). *Robot*. *Encyclopedia Britannica*. https://www.britannica.com/technology/robot-technology accessed 08/09/21 20:32

Torgersen, E. (2020, August 27). No title.

https://titan.uio.no/teknologi-informatikk/2020/det-var-et-stort-oyeblikk-da-dyret-kunne-ga-rundt-pa-laben accessed 08/09/21 23:34

DO.IT. (2021, April 9). What is Universal Design? https://www.washington.edu/doit/what-universal-design-0 accessed 09/09/21

ResembleAI. (n.d). https://www.resemble.ai/ accessed 09/09/21 13:52

Vox. (n.d). Why algorithms can be racist and sexist. https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency accessed 09/09/21 19:55

UX Collective. (2020, February 4). Don Norman's seven fundamental design principles. https://uxdesign.cc/ux-psychology-principles-seven-fundamental-design-principles-39c420a05f84

Følstad, A. (2021). *Interaction with AI.* University of Oslo. https://www.uio.no/studier/emner/matnat/ifi/IN5480/h21/interacting-with-ai-2021---module-2---session-1---handout.pdf accessed 16/10/21

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

Goodwin, M. (2021). *Interacting with Artificial Intelligence.* University of Oslo. https://www.uio.no/studier/emner/matnat/ifi/IN5480/h21/lecture-notes/presentation_uio_16_9_2021.pdf accessed 20/10/21