Innleveringsoppgave i IN5480

Ashwin Rajeswaran

29. Oktober 2021

Concepts and theory

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

The origin of AI can be traced back to Alan Turing, whom in 1949 developed a code breaking machine to decipher German code in World War II for the British government (Grudin, 2009; Haenlein & Kaplan, 2019). Following this incident, several perspectives were pursued concerning robotic intelligence and computing, leading to exploration of the three laws of robotics (Grudin, 2009). The actual term "artificial intelligence" was however coined by John McCarthy six years later for the research project DSRPAI - *Dartmouth Summer Research Project on Artificial Intelligence* – which was aimed at building machines able to simulate human intelligence (Grudin, 2009; Haenlein & Kaplan, 2019). The AI research field flourished following this research project into the next decade (Grudin, 2009; Haenlein & Kaplan, 2019). This implies that the *initial* definition for AI was leaning towards *human intelligence simulation*.

There are, however, several definitions of what is considered artificial intelligence. One of which is defined through the Turing test, which became the benchmark based on Alan Turing's article *Computing Machinery and Intelligence,* claiming that if machine behavior is indistinguishable from human behavior when a human is interacting with the system, then the machine is *intelligent* (Haenlein & Kaplan, 2019). This became a setting stone for the DSRPAI project later in 1950, for the newly founded AI community. In the article "*Does PD make AI obsolete?"* by Bratteteig and Verne (2018) a similar definition is used, the said definition being "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". Both these definitions highlight the computer's ability to *mimic* human behavior.

Another definition by Hans Moravec in his issue of Journal of Evolution and Technology from the 1970s is the following: " (…) computers as locomotives of thought, which might outperform humans

in higher mental work as prodigiously as they outperformed them in arithmetic's (…)" (Grudin, 2009). Here the focus on the AI definition concerns mathematical and logical performance, in line with Bratteteig and Verne's (2018) definition of intellectual behavior.

A slightly different perspective is also seen from Nicholas Negroponte of MIT in 1970, the implication being that [intelligent] machines should understand the context which they operate in (Grudin,2009). This presents a more social-science view on what it is to be considered intelligent and is also somewhat like *rational* behavior humans are expected to have as stated by Bratteteig and Verne (2018).

From these three definitions we see distinctions in what is considered as "intelligence". One aspect of the AI considers fully mimicking human behavior, the other considering mathematical, logical, and arithmetic intelligence (close to human crystallized intelligence), and the third highlighting specific human traits like context adaptation and rationality. These definitions are all offspring of the ambiguous nature of the word intelligence, which is challenging to define with, definite definition (Kok et al., 2009; Goertzel & Wang, 2006, p.21).

With these three definitions in mind, one could say that AI is a system that behaves in undistinguishable fashion from human behavior, is able to outperform human arithmetic and mathematical computations, and has unique human traits such as the ability to understand and adapt in its operating context. This means that the system must:

- Possess the ability to compute complex arithmetic calculations
- Possess human traits such as self-adaptation, understanding of speech and understanding of context.
- Be able to express rational behavior in line with humans.

This implies that a definition of what is *truly* AI is somewhat subjective, but that there are a few common grounds which are representative in some shape or form in all definitions. AI is therefore considered to be a system that is simulating human behavior, can adapt to its contextual situation, understand advanced arithmetic calculations and logics, but also express rational behavior.

Thoughts about Norman (1990)

The article "The problem with automation" by Norman (1990) briefly comments on that over-automatization is a downturn for AI, and that it might be because the system has been made *too intelligent*, hinting that the automatization is doing more than what meets the eye behind the scenes. Their main argument is that the *lack of feedback* is what is really causing trouble amongst automated systems; users are thrown 'out of the loop' when the automation process does not provide enough feedback on what is happening (Norman, 1990). I agree with this claim, as too automized systems *do* cause trouble for users who do not understand how the automatization works due to the lack of response. In any given man-machine interaction the man needs a 'dialogue' to orient themselves in the context and act accordingly, but when the system refrains or is unable to do so it leaves room for misunderstandings.

Google AI

Google founded 'Google AI' in 2017 and has actively attempted to evolve how AI is perceived and used by everyone. They are presenting themselves as a company that is focused on research innovative ways to incorporate AI in new domains in hopes of expanding the use of AI to more people. They introduce AI as a service (portal for research), and present products that incorporate AI capabilities to 'help users solve their problems, big or small'. In other words, their main objective is to "augment the abilities of people, to allow us to accomplish more and to allow us to spend more time on our creative endeavors" (Dean, 2021).

AI interaction in media

A videogame that tackles human-AI interaction is Nier Automata, which is about android (fully conscious robots that are as close to humans as possible) and their war against machine lifeforms. As the game progresses, we come to learn that machines have learned to communicate in a non-programmatic manner (meaning, they are able to hold dialogues and not monologues). The interaction between androids (the "humans" of the game) and machines (the "AI" of the game) shows that the more intelligent the machines are understood to be, the more respect and deserving of life they are.

1.2 Robots and AI systems

The word 'robot' may be traced back to a Czech play by author Karol Capek called Rossum's Universal Robots in 1920 and means 'forced labor' in Czech (Coiffet & Chirouze, 1983, p.17), which resonates with the strict order-obeying creatures they are in the play. The idea closely resembles that of industrial robots, which are used in factories (Thrun, 2005, p.11). The origin of the term implies that these are devices with low autonomy.

Thrun (2005, p.11) provided two different definitions in their article. First one is from the Robot Institute of America in 1979, defining 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". Merriam Webster's definition (1993, cited from Thrun, 2005, p.11) is that a robot is "An automatic device that performs functions normally ascribed to humans or a machine in the form of a human". Based on both definitions, I define a robot as *a human-resembling device which can be programmed to perform various tasks*. My definition emphasizes that a robot should be able to be programmed to do various tasks physically, but also replicate human behavior or traits.

Based on the definitions from earlier, one could imply that the difference between robots and AI is the autonomy. The key difference lies in how *intelligent* the system is; how autonomously it can act and how adaptive it is to its surroundings. However, both definitions also suggest that both robots and AI shall resemble humans in some shape or form. Both robots and AI are also expected to perform either tasks humans are incapable of doing or to release humans from having to perform redundant and "boring" activities.

Sony AIBO

One example of a contemporary robot is the Sony Aibo, a robotic puppy developed by Sony powered by AI (Thrun, 2005, p.16; Sony, n.d.). The robot is capable of voice- and gesture recognition as ways of communication between itself and the user. For the time being, local movement from the user results in global movement in the Aibo (Sony, n.d.). Its joins permit it to perform somewhat believable (but rigid) animation, which gives it an innocent and energetic personality, and can express emotions like anxiousness based on heights and tight places (Sony, n.d.). The robot is also able to sense and analyze its surroundings and has memory of its daily experiences. Aibo can also wag its tail in happiness if the user scratches its chin and perform various tricks that it learns from the user.

1.3 Universal Design and AI systems

Universal Design is "The design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (Persson et al., 2015). In other words, Universal Design is not about specialized design *for* inclusion, but designing *with* inclusion in mind. The fundamental difference between universal design and specialized design is ensuring accessibility for everyone rather than making a new system only customized to a certain user-group.

AI can potentially be used to elevate the life quality of users that have perceptive, cognitive/emotional or movement related impairments. AI can use recognition of motion or scanning to prevent potential danger that can befall people, or that can overall improve a user's interactions and reducing barriers for them to reach their goal activity. For example, a user who is blind can rely on speech in a smart home, but the AI can also be adaptive and learn patterns of the users' behavior to do certain tasks autonomously if it is a strong, daily pattern. This also serves as an example of the potential a AI system has to include physically or mentally impaired users. An example of the opposite, AI systems excluding users, is rooted in confirmation bias, dataset bias, association bias, automation bias or interaction bias (Chou et al., 2017). Joy Buolamwini (2016) provides examples of an AI excluding based on gender and skin-color due to the systems failure to use a heterogenous dataset, giving the AI dataset bias which made it unable to detect women of color in facial recognition software.

The concept of "understand" and "understanding" is used in several AI related guidelines. For me, "to understand" is the ability to either be able to learn and comprehend a situation, or to be sympathetic in relation to someone else's situation. I would say a machine's ability to "understand" is falls short in comparison to human understanding simply because as of now, machines are unable to interpret subtle hints in body language, tones, and gestures. Therefore, in my eyes, machines *do* understand when they learn and adapt from their datasets and contexts, but they don't "understand" to the same degree humans do.

1.4 Guideline for Human-AI interaction

"G8: Support efficient dismissal" is one of Microsoft's 18 guidelines for human-AI interaction. The guideline suggests that the AI services should be easily dismissible if they are undesirable for the user. An example of *not* following this guideline is Instagram's "suggested posts" feature; The feature generates new content for you to scroll through after you've seen any recent photos and videos from accounts you follow (Bonifacic, 2021), but the suggested posts cannot be dismissed unless the user stops scrolling to interact with a dialogue prompt (which also scrolls along the content and is easy to miss).

One set of HCI design guidelines is Nielsen and Molich's Heuristic evaluation (Nielsen & Molich, 1990, p.249). In comparison to the human-AI guidelines, both guidelines focus on the usability for the user. They also emphasize the need for context dependent dialogue, substantial amount of system feedback and easily dismissing the service. In contrast, the human-AI guidelines focus more on *how well* the AI provides relevant content for the user, while the HCI guidelines focus more on interchangeability and if it is 'fit for purpose'.

Characteristics of AI-infused systems

AI-infused systems are ' systems that have features harnessing AI capabilities that are directly exposed to the end user' (Amershi et al., 2019). As mentioned in the first lecture of module 2, there are several key characteristics of these types of systems; these may include learning, improving, black box (unseen processes) and that they are fuelled by large data sets (Følstad, 2021). From Amershi et al. (2019) we discover traits like uncertainty, inconsistency, and automated personalisation. In the literature, the authors elude these characteristics combined can lead to unpredictable behaviour which can be "(…) disruptive, confusing, offensive, and even dangerous" for the end users, but also that the system is more prone to errors (Amershi et al., 2019; Yang et al., 2020). Other traits for these kinds of systems may be that they are probabilistic, malleable by user interaction and have transparency issues (Kocielnik et al., 2019). As a result, they may possess natural language understanding, behaviour prediction or content recognition software. On the other hand, these services may be of lower accuracy as they are probabilistic operations.

These systems also have varying complexity that span between probabilistic operations to evolving and malleable systems based on user behaviour. Yang et al. (2020) has distinguished AI-infused systems between (1) probabilistic systems, (2) adaptive systems, (3) evolving probabilistic systems and (4) evolving adaptive system. Each of these levels present a new set of challenges that are tied to the key characteristics of AI-infused systems.

Apple Siri – voice assistant and automation

Apple's Siri voice assistant is a well-known AI-infused system that is available to the public. Siri processes user input such as voice queries and text commands to fulfil requests (Apple, n.d.). The assistant initially possessed AI-characteristics such as natural language understanding to act as a natural assistant (learning trait), but in recent times the system is also used to provide suggestions based on user behaviour and use-patterns (improving trait). This addition of personalised suggestions has given the users the ability to silence notifications based on pattern behaviour such as usual sleep cycle and typically used apps for given times of the day (black box trait).

Human-AI interaction design

Summary of Amershi et al. (2019)

The purpose of the article is to create and evaluate a set of guidelines for developing and harness AI technologies. The guidelines are derived from previous studies and other relevant guidelines. These guidelines are divided after the time of the interactions taking place to address the many challenges that AI interaction faces. To further solidify these guidelines, the designers went through evaluations in four iterations. The guidelines' main contribution to the AI field is that it serves as a starting point for further research in the field, but also as principles for creating and maintaining AI interaction.

Summary of Kocielnik et al. (2019)

This article studies the impact of methods seen considering what the users expect of the system, namely False Positive and False Negative scenarios. Their hypothesis is that there is a significant difference between how an AI-infused system is perceived based on the accuracy of the system and subjective perceptions. The designers create two versions of a 'Scheduling Assistant' to investigate their research hypothesis and see how subjective perceptions affect the perceived accuracy of the

system, and the findings of this study show us how people are more forgiving of flaws in AI-infused systems if it is admitted the system is imperfect to begin with.

Apple Siri and the human-AI interaction guidelines

Guideline G1 ('make clear what the system can do'): If the user activates Siri but does not give any voice commands, the system will return a list of possible speech-based commands that the user may perform. These commands are connected to the apps that are installed on the device by default, but it may also perform scripts with other applications if they have been defined in the helper-app Siri Shortcuts. Thus, the system makes clear what functionality the voice assistant has and what it can provide to the user. This guideline can inspire to be clearer with how the user can customize the voice assistant and their interactions; for example, by giving users information about how to add scripts to other applications that are present in the helper-app but not in the main interaction window.

Guideline G13 ('learn from user behaviour'): Siri suggests which apps the user can interact with based on how the user's current location, the time of the day or upcoming events. One can therefore say the system learns from user behaviour as it uses these metrics and ties them to certain use patterns. For example, Siri will suggest sending money to a friend in a timeslot between 11:30-12:30 if you often Vipps money to your friend for often buying you lunch. This guideline can inspire the developers to further enhance how much the user can influence these suggestions, as it is currently no easy method for the users to remove suggestions that they feel are prominent and irrelevant.

Bender et al.'s (2021) discussion

The paper raises concerns about whether large language models (LMs) are necessary and what costs one should consider in deeper learning. The authors mention several costs and risks associated with large LMs; environmental and financial costs, opportunity costs, and harms like abusive language that can lead to stereotyping, denigration, increases in extremist ideology etc. in the training data. Another argument by the authors is that quantity does not guarantee diverse content due to, for example, overrepresentation of certain user bases or the like. There is also a challenge with changing social norms in the society, cause data containing these can be misinterpreted. The main risk with these factors is that an LM that has these kinds of training data will take in these problematic conceptions, biases, and profanity language. Solutions presented in the paper is oriented towards a proactive view; handling and being cautious of the training data to mitigate the above-mentioned risks. This can include a thorough planning of the data sets and consider costs.

Chatbots / conversational user interfaces

Key challenges in design of chatbots:

A key challenge with chatbots that is mentioned in the lectures and Bender et al. (2021) is natural language. As mentioned in Kocielnik et al. (2019), natural language understanding is a core component of AI-infused systems, which chatbots operate under. Because of natural language, there are several factors that can be challenging for chatbots, including variation in language, abusiveness, and misinterpretation. People can convey the same message in different ways, which then gets interpreted during conversation based on context, behaviour, social cues and so forth. This poses as a challenge for chatbots as they have higher probability to not being able to interpret the actual meaning behind the received message.

Adherence to guideline G1 and G2

Guideline G1 ('make clear what the system can do'): A common issue with chatbots is that people ask questions that the chatbot is incapable of answering. If chatbots were to adhere to this guideline, then users would know exactly what questions the chatbot can answer. This would mean that the user would quickly understand the software's capabilities and have realistic expectations to it. By doing so, the user would become more forgiving of system errors that may be present. This is shown in a study by Luger and Sellen (2016), where it was apparent that "(...) two most frequent users who tended to be more experimental and forgiving, all of those interviewed raised issues of trust as limiting the tasks they would ask their CA to perform". This implies that setting clear boundaries on the systems performance and capabilities, the users will gain more trust to the system.

Guideline G2 ('Make clear how well the system can do what it can do'): Users usually feel frustrated with chatbots. If chatbots were to adhere to this guideline, then users would be more forgiving of the chatbot. The chatbot would give the impression that it does not have a deep enough understanding of the user (in other words, enough data to make solid suggestions and decisions), and the user could then take action to confirm or reject the suggestion. As mentioned above, the same statement by Luger and Sellen (2016) implies that the system should be transparent about not being error-free so that the users will be more forgiving whenever errors during use emerge.

Feedback from iteration 1 and 2

Important feedback I received from iteration 1 was that the text was unclear when changing topics from one to another. To mitigate this, I have added clearer headings and (hopefully) made the language sharper and clearer. For iteration 2 I received a request to elaborate a bit more about the G1 and G2 principles and add a few examples from the literature. To do so I have provided more material from Lugen and Sellen (2016) to complement.

References

- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM. (https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf)
- Apple (2021, 15 .October). Siri. https://www.apple.com/siri/
- Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM. (https://dl.acm.org/doi/pdf/10.1145/3442188.3445922)
- Bonifacic, I. (2021, 23 June). Instagram tests placing 'suggested posts' throughout your feed. Engadget. https://www.engadget.com/instagram-suggested-posts-feed-191126529.html
- Bratteteig, T., & G. Verne. (2018). Does AI make PD obsolete? Exploring Challenges from Artificial Intelligence to Participatory Design. Proceedings of PDC 2018, Belgium, 1-5. https://doi.org/10.1145/3210604.3210646
- Buolamwini, J. (2016, November). How I'm fighting bias in algorithms[Video]. TED. https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms
- Chou, J., Murillo, O., & Ibars, R. (2017). How to recognize exclusion in AI. Medium. https://medium.com/microsoft-design/how-to-recognize-exclusion-in-ai-ec2d6d89f850
- Coiffet, P., & Chirouze, M. (1983). An introduction to robot technology. Springer Science & Business Media (2012). https://books.google.no/books?id=jkvzCAAAQBAJ&lpg=PA10&dq=robot%20word%20origin&lr&hl=no&pg=PA10#v=onepage&q=robot%20word%20origin&f=false
- Dean, J. (2021). Bringing the benefits of AI to everyone. Google AI. https://ai.google/about/
- Goertzel, B., & Wang P. (2006). Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. IOS Press. https://books.google.no/books?id=t2G5srpFRhEC&lpg=PA17&dq=artificial%20intelligence%20definitions&lr&hl=no&pg=PA2#v=onepage&q&f=false
- Grudin, J. (2009). AI and HCI: Two Fields Divided by a Common Focus. *AI Magazine*, *30*(4), 48. https://doi.org/10.1609/aimag.v30i4.2271

- Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. California Management Review, 61(4), 5–14. https://doi.org/10.1177/0008125619864925
- Helm, J.M., Swiergosz, A.M., Haeberle, H.S. *et al.* (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Curr Rev Musculoskelet Med* **13,** 69–76. https://doi.org/10.1007/s12178-020-09600-8
- Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 411). ACM. (https://www.microsoft.com/en-us/research/uploads/prod/2019/01/chi19_kocielnik_et_al.pdf)
- Kok, J. N., Boers, E., Kosters W., van der Putten, P., & Poel, M. (2009). ARTIFICIAL INTELLIGENCE – Artificial Intelligence: Definition, Trends, Techniques and Cases. Encyclopedia of Life Support Systems (EOLSS), 1-5. https://www.eolss.net/Sample-Chapters/C15/E6-44.pdf
- Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. Processing of the SIGCHI Conference on Human Factors in Computing Systems (CHI'90). Association for Computing Machinery, New York, NY, USA, 249–256. DOI: https://doi.org/10.1145/97243.97281
- Persson, H., Åhman, H., Yngling, A.A. *et al.* Universal design, inclusive design, accessible design, design for all: different concepts—one goal? On the concept of accessibility—historical, methodological and philosophical aspects. Univ Access Inf Soc 14**,** 505–526 (2015). https://doi.org/10.1007/s10209-014-0358-z
- Sony (2021, 7. September). aibo – robotic puppy, powered by AI. https://us.aibo.com/
- Thrun, S. (2004) Toward a Framework for Human-Robot Interaction, Human–Computer Interaction, 19:1-2, 9-24, https://www.tandfonline.com/doi/pdf/10.1080/07370024.2004.9667338
- Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In Proceedings of the 2020 CHI conference on human factors in computing systems (Paper no. 164). (https://dl.acm.org/doi/abs/10.1145/3313831.3376301)