IN5620 - Interaction with AI and Autonomous Systems - Module 1

3. september 2024

1.1

In the summer of 1956, Marvin Minsky, John McCarthy, and others gathered at Dartmouth College to research artificial intelligence (AI). Although this event is considered the dawn of AI, and the participants were carefully selected, they did not anticipate some of the scientific advancements that would later emerge. Naturally, it is much easier to make such claims almost 70 years later than it was for the AI pioneers to spend eight weeks trying to make machines use language to solve problems once reserved for humans and to improve themselves. (McCarthy, Minsky et al. 1955)

In his 1955 proposal (McCarthy, Minsky et al. 1955), McCarthy coined the term 'Artificial Intelligence. He later defined AI as, "It is the science and engineering of making intelligent machines, especially intelligent computer programs.» (McCarthy 2007). The European Union AI Act defines AI as "... a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."(European and Council of the European 2024)

Having discussed various definitions of AI, ~~When we zoom in on the term AI we arrive at machine learning,~~ lets now zoom in on machine learning, which is the area where recent advancements have been most significant. One of the most influential AI researchers, Ilya Sutskever, has played a key role in developing AlexNet and OpenAI's GPT models. The former is considered a major leap in object recognition through computer vision, and the latter is widely recognized. His definition of the learning aspect of machine learning is to have a computer discover a function so intricate that it cannot be humanly programmed, using a large training dataset to learn the function by minimizing the difference between the function's output and a ground truth. (Sutskever 2013).

Two of these definitions come from academic figures, with Sutskever elaborating on McCarthy's definitions regarding how we create these intelligent machines (or at least one way to do it: supervised learning). The EU definition is somewhat more "political" and very broad. However, it is interesting because it addresses the ethical and security implications of these intelligent systems. Overall, all definitions describe the same concept: an intelligent program that takes an input and produces an output.

Based on these definitions, I propose my own: AI is an intelligent system without biological interference between the system's input and output, where the output achieves some stated goal. By "intelligent," I mean the system's or person's ability to learn and evolve based on complex causal observations. This definition clearly distinguishes artificial from biological intelligence and introduces the concept of "goal," which is crucial for the development of AI systems to maintain some form of alignment. The alignment of goals between artificial and human intelligence is evident when considering the lack of alignment between different forms of biological intelligence.

If we understand autonomy as a self-governing actor, I would argue that the actor's goal is paramount. Therefore, whether the actor/system is based on human or machine autonomy is largely similar. Differences arise when the actor/system acts in relation to another system/actor's goal. For humans, this is called altruism, which is arguably beneficial for both the actor and the beneficiary. However, for machine systems, altruistic acts are irrelevant since their goals are purely logical and not emotional.

A human heartbeat is automatic—we cannot control whether our heart beats (at least not biologically). This differs from a dolphin's ability to adapt and control its heartbeat, which is an autonomous, self-governing concept of heartbeat.

Laura Bainbridge discusses the automation paradox in her article "Ironies of Automation» (Bainbridge 1982). She argues that despite its aim to reduce human intervention, automation often increases complexity for human operators and that human experience and capabilities are necessary for handling unexpected situations that autonomous systems cannot manage. This is still evident today, 40 years later, with autonomous taxis and the need for remote drivers to assist in

situations where the taxis cannot navigate on their own. Bainbridge emphasizes the pitfall of operators losing their skills after long periods without use. She suggests that those responsible for taking action when an autonomous system fails should maintain their skills through periodic training. All this is to ensure a robust and effective human-computer interaction. ~~(Bainbridge 1982)~~.

DNB uses AI in products and services to help their employees be more effective and to adapt their work to reflect DNB's tone of voice. DNB therefore uses AI daily to enhance the company's operations, not just as a concept. It employs AI in their chatbots and to assess loan applications.(Giske 2024)

Max Tegmark discusses Human-AI interaction in his book *Life 3.0*. Tegmark describes how AI can serve as both an ally for human efficacy in various areas of life, such as medicine, poverty, and environmental issues, and as a potential threat to human existence. The former highlights AI's positive impact, while the latter focuses on the dangers of uncontrolled AI and autonomous systems development, emphasizing the importance of aligning AI goals to enhance human values rather than undermine them. (Tegmark 2017).

1.2

While AI encompasses a broad range of intelligent systems, robots represent a specific application of AI. The word "robot" was introduced through the Czech play *R.U.R. (Rossum's Universal Robots)*, and it originates from the Czech word "robota," which means forced labor. The play depicts artificial workers who work for human goals. The Oxford Encyclopedia defines a robot as "A machine capable of carrying out a complex series of actions automatically, especially one programmable by a computer.» (Press 2016), Britannica offers a similar definition but emphasizes that a robot replaces human effort in a human or non-human-like manner. (Moravec 2024)

None of these definitions mention autonomy, so a robot does not possess its own will to perform actions, unlike a dolphin's ability to control its heartbeat. These definitions follow Asimov's Three Laws of Robotics, with the most important law being the well-being of humans. My own definition would be something like: A robot is a machine capable of performing automatic

tasks on behalf of an intelligent will. On Earth, we can assume that humans are currently the only beings capable of creating robots that can perform intelligent tasks—so any will a robot might have in the future would have originated from humans. We can imagine an AI exerting its will over a robot, but still, a human was involved at some point in aligning the AI's goals. It's really a "chicken or egg" problem. A chatbot clearly fits this definition due to its intelligent capabilities as determined by human design.

As mentioned earlier, an AI can control a robot, but the AI might be part of the robot and thus the same physical entity. However, AI can also exist in a digital form, like Jarvis in the Marvel Universe. A robot will never be able to "live" purely digitally. While robots are pre-programmed to automatically perform a series of tasks with human safety as the highest priority, AI can evolve beyond its original intent or goal. The challenge of aligning the intent and goals of any autonomous systems, whether physical or cloud-based, is crucial in determining if these systems will serve human needs or pursue their own objectives.

When examining some of the most advanced humanoid robots, such as the 1X Technologies Eve robot, Tesla Optimus, or Figure1 from FigureAI, we see that they are still in the early stages of development. Both 1X and FigureAI have deployed their robots for various tasks in warehouses. The goal is to create robots that move and look like humans to adapt to our society, which is designed for human-like bodies and operations. Humans would use these types of robots to assist with tasks, but these robots are also equipped with AI systems to make autonomous decisions on how to maneuver and solve tasks that have not been pre-programmed. Still in their infancy, these robots require human intervention if they encounter problems, just as Brainbridge foresaw.

1.3

Universal design is the principle of designing products and environments to be inclusive and accessible to as many people as possible, regardless of disabilities. (World Wide Web 2021). My understanding is that this means designing any physical or digital environment to maximize inclusion.

AI can significantly enhance human perception, particularly for individuals with disabilities affecting one or more senses. AI can interpret the environment and "translate" between senses, allowing those with disabilities to experience the world more fully. For example, imagine a pair of

glasses worn by a hearing-impaired person that visually alerts them to the presence of ambulance sirens, or an AI that describes the visual environment in front of a blind person through voice.

AI can include blind people in events where information is conveyed visually. Conversely, it can also exclude by amplifying existing biases. Since AI systems are trained on human-labeled data, any biases in these datasets are often reflected and magnified in the AI models.

AI systems do not understand data in the same way humans do. Although AI is powerful in analyzing complex correlations in large datasets, it identifies patterns without actually experiencing the data. For instance, a newborn baby feels the love of its mother through the warmth of her body and the sound of her heartbeat. The baby understands this in a human way, while an algorithm merely recognizes patterns between heat and sound vectors.

1.4

Guideline #6: Mitigate social biases by ensuring the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.

This can be understood as an AI system that mitigates social biases by ensuring that the AI system is trained on data that is not biased and is as close to the truth as possible.

To find a set of HCI guidelines, we can refer to those mentioned earlier, such as WCAG. Both WCAG and Microsoft's Human-AI guidelines focus on user-friendliness, ensuring that computer and AI products are accessible and inclusive. However, while WCAG emphasizes web content accessibility and usability for all users, Human-AI guidelines specifically address the complexities of AI interactions, such as managing biases and providing contextually relevant feedback.

In contrast, WCAG emphasizes broad accessibility and inclusion, whereas Human-AI guidelines focus more on mitigating risks associated with intelligent systems, such as reinforcing social stereotypes and biases. Essentially, the complexity of AI systems adds an extra layer of consideration beyond the general accessibility principles of WCAG.

1.5

JOURNAL

In my experience with various journaling methods, I've noticed distinct differences in my writing style. Using Siri for dictation often results in more expressive and emotional content, as I tend to be more "floral" in my language. On the other hand, typing with a QWERTY keyboard leads to more direct and to-the-point entries with fewer emotional nuances. Handwritten journals fall somewhere in between these two extremes.

Reflecting on past experiences with handwritten journals, I've observed that my emotions can be discerned through the neatness of my handwriting. For example, when I'm feeling angry, my handwritten notes are often messier and harder to read compared to when I am calm. This suggests that the state of my emotions can influence the tidiness of my handwriting.

Modul 2

2.1

AI-infused systems, such as those described by Amershi et al. (2019), have some unique characteristics which differ from traditional user interfaces. One of the most important aspects is perhaps unpredictability (uforutsigbarhet). AI-infused systems operate under the laws of probability, which can lead to unexpected behavior. This unpredictability (uforutsigbarhet) can be confusing for users and can lead to a decline in the system's trustworthiness. Due to its probabilistic nature, these systems are inconsistent and vary their responses based on probabilistic algorithms, system software updates, or adaptive personalized learning from interactions. This complexity in user interface and experience is a challenge that needs to be addressed to gain and keep the users' trust in the systems.

Another central characteristic of such systems is adaptive learning, as mentioned above. This is the system's ability to continuously learn from data and change its behavior over a sequence of interactions. Amershi et al. (2019) emphasize that this ability to learn can lead to both enhancements and challenges. The latter can include creating unpredictable changes that users do not expect, while the former can lead to better personalization and effectiveness. Yang et al. (2020, p. 6-7) note that this adaptability in the systems is ideally positive with the right kind of interaction and data, but unfortunately, with less ideal data, it worsens the experience for the user. They also mention the probabilistic nature being a challenge for designers when prototyping. Therefore, it is of utmost importance that these systems give users insight into how they adapt so that users can themselves adapt their interactions.

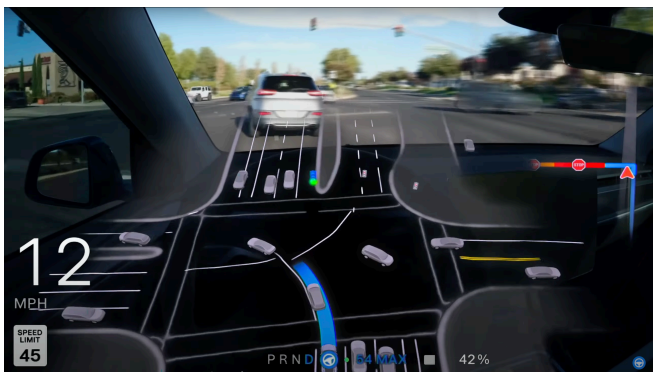
AI-infused systems can take the adaptive part a bit further by introducing automatization of decisions for the user. This is when a system has the ability to take actions on behalf of the user, such as auto name-tags on uploaded images (Yang, Steinfeld et al. 2020, p. 9). The challenges of such systems making mistakes are implicit, thus giving the user clear responses to what decisions to make, and giving the user control over them (Kocielnik, Amershi et al. 2019, p. 1). Therefore, the systems should convey their abilities and limitations to the user, including what they can do, how well they can do it, and what mistakes they may make. The 18 guidelines proposed by Amershi et

al. cover such points in the interaction between the user and the AI system. By following these steps, the system can reduce uncertainty and gain trust.

To address these and other challenges with respect to Human-AI systems, it is essential that developers and designers consider ethical considerations. This includes addressing social biases in data, as Amershi et al. describe, to mitigate discrimination and exclusion. By giving the user control over the system's behavior, one can develop intuitive and user-friendly systems that are ethical. These characteristics in AI systems are beneficial for increased effectiveness and usability for their users, while also introducing some challenges, especially with respect to ethical considerations.

One AI-infused system that I closely follow the development of and have some experience with is autonomous driving systems. The characteristic mentioned above that I find very interesting in such systems is essentially taking all the driving decisions away from the driver and executing them autonomously. The human driver always has the ability to abort the system and take control for themselves. Mistakes made by an autonomous vehicle in a complex driving environment are implicit; they can lead to casualties. There are numerous driving assistents, ranging from adaptive cruise control to more advanced systems such as Ford's BlueCruise or Tesla's FSD. Human-driver trust in the system is of utmost importance if the system is to be used. The transparency and explainability of the systems I trust most come down to how well the autonomous system shows me its interpretation of the world around it and how it plans to behave in the next few seconds given the environment around the vehicle. As a user, I am affected by these characteristics, and I have far more trust in systems that convey this information to me the best. The two aforementioned systems are visualized in the pictures to the right. I encourage the reader to reflect about how the two systems are conveying its intent to the user, and how it affects the trustworthiness of the system.





2.2

The design of the interaction in AI-infused systems are important, can either mitigate or even make use of the characteristics in AI-infused systems. One way is to adjust the end-users expectations to the systems before use. Kocielnik et al. (2019), discusses three techniques for setting the right expectations to the system by the user. The first is an an indicator for the systems accuracy, the second is an explanation to the user to help the user understand the basic principles to how the systems is reasoning, and the third is control-slider for the user to adjust the system performance. As they had hypothesized the when an accuracy indicator was given to the user the expectations to the user was lowered, and to the two other techniques the expectation were higher as we see in figure 6 (p. 8). All changes were statistically significant.

Another approach was taken by Amershi et al. (2019). They propose 18 guidelines that Human-AI or AI-infused systems should follow when designing such systems. The guidelines were derived through decades of research on the field, validation/revised rounds and studies including numerous designers. The guidelines focuses on different «time-steps» in the interaction. They discuss considerations with from various ethical and societal implications on how the models are trained, to how the system explain to the user its accuracy, much like Kocielnik et al. They also emphasize that the guidelines are not an exhaustive list in the rapid development in inclusion of AI-infused systems, and thus acknowledging value in future work in the field even though the guidelines were found to be objectively relevant, explicit and of help in most cases.

With respect to the systems discussed earlier, both of them is a clear application of guideline 7 as they require an active interaction by the user is some form to activate or deactivate the AI-infused system. When it comes to guideline 4 which is to show the user relevant contextual information I would argue that one of the system is adhering better than the other when considering displaying relevant information to the end-user. The fact that there are guidelines for such systems that emphasizes the importance of contextual relevant user information, these guidelines could inspire the developers to improve the system to convey relevant information to the users to increase trust in the system. If we consider, some time in the future, that these systems operate independently as fully autonomous vehicles without supervising human drivers, guideline 7 can be used to improve other infused services like tourist recommendations relevant to the area.

One guideline discussed in section 1.4 refers to guideline 6 - mitigating social biases. One big problem in LLM is directly coupled to this guideline. The immensity of the training data the systems are trained on includes really harmfull data. This includes such as literature on human psychology and how to manipulate us, to really social, ethically and stereo-typical challenging characteristics (Bender, Gebru et al. 2021, p. 613-614). LLMs ability to generate textual content which can reflect its training data, makes them prone to generate harmful content, without anyone responsible for generating such content which deeply understands the language- They call such systems stochastic parrots due to its probabilistic nature (s. 616-17). Such systems also have a negative impact environmentally (s. 612). They suggest that curating training data is crucial, and that guidelines like the ones we've discussed can help adjust users' expectations and awareness of the system's accuracy.

2.3

According to the participants in Skjuve et al. (2024) study of the motivations for using LLMsit is encouraging that 'productivity' is the most likely primary motivation. However, there are numerous pitfalls associated with AI-infused systems like LLMs. One significant challenge in designing such user interfaces is balancing the stochastic nature of these models while maintaining a trustworthy interface.

Referring to the Boston Consulting Group's report (François Candelon, Lisa Krayer et al., 2023) balancing this aspect is critical as average productivity performance is higher when utilizing LLMs. However it is important to consider the potential for making users over-reliant og AI-systems. Another challenge is designing the interface so users can take calculated actions considering environmental and financial costs. In other words, users need to be made aware of these challenges.

For instance, I often ask an LLM questions I could find answers to almost as quickly but in a more environmentally friendly manner. Referring to Kocielnik et al. (Kocielnik, Amershi et al. 2019) explanations and user controls positively impact user expectations. Adhering to guideline G1 positively aligns user expectations and system control, fostering a critical view of generated content. The accuracy indicator proposed by Kocielnik et al. is a clear adaptation of guideline G2, significantly reducing unconscious trust in the system.

Disclaimer

I have used both Grammarly and ChatGPT to improve sentences and correct errors in this text.

Appendix 1:

The feedback I received from the first iteration helped me to better link some of the paragraphs together and to be more specific in my references. The changes are highlighted in blue in sections 1.1 and 1.2.

Bibliography

(2023). Robot. Oxford, England, Oxford University Press.

Amershi, S., et al. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland Uk, Association for Computing Machinery**:** Paper 3.

Bainbridge, L. (1982). "Ironies of Automation." IFAC Proceedings Volumes **15**(6): 129-135.

Bender, E. M., et al. (2021). Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event, Canada, Association for Computing Machinery**:** 610–623.

European, P. and U. Council of the European (2024). "Regulation (EU) 2024/1689: Laying Down Harmonised Rules on Artificial Intelligence."

François Candelon, et al. "How People Create and Destroy Value With Gen AI." Boston Consulting Group (BCG).

Giske, M. E. (2024). Generasjon KI inntar arbeidsmarkedet. dnb.no, DNB.

Kocielnik, R., et al. (2019). Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland Uk, Association for Computing Machinery**:** Paper 411.

McCarthy, J. (2007). "What is Artificial Intelligence."

McCarthy, J., et al. (1955). "- A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence."

Moravec, H. P. (2024). Robot. Encyclopaedia Britannica. Chicago, IL, Encyclopaedia Britannica, Inc.

Press, O. U. (2016). Definition of Robot. Oxford, England, Oxford University Press.

Skjuve, M., et al. (2024). "Why do people use ChatGPT? Exploring user motivations for generative conversational AI." First Monday **29**(1).

Sutskever, I. (2013). "Training Recurrent Neural Networks." 1-101.

Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. London, Penguin Books.

World Wide Web, C. (2021). "Web Content Accessibility Guidelines (WCAG) 2.1."

Yang, Q., et al. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, HI, USA, Association for Computing Machinery**:** 1–13.