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 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)

When we zoom in on the term AI, we arrive at 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: su-

pervised 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." 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

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.

Disclaimer

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

Bibliography

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

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

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

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

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.

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."