AI and Society
Msc AI, Vrije Universiteit Amsterdam
P3 2021-2022

Human & Machine Report

Table of Contents

1 Introduction

1) Theme definition

Since the invention of the first machines, humanity has become fascinated by the nature of human-robot relationships. One big milestone was the creation of the 'Turing Test', alongside the computer in the 20th century (Copeland, 2021). They became crucial building blocks that led to the scientific field of Artificial Intelligence (AI). This field has now become an omnipresent force in society and influences economic and social areas of life. Although it is nowhere close to human-level intelligence in the sense of creating a conscious being, the remarkable progress through decades has enabled us humans to confidently label the advanced AI systems as 'intelligent machines' (Ostheimer et al., 2021). This led to the creation of the necessity to think and reason about how humans and machines should interact in the future.

2) Issue definition

In all human aspects, decision making plays a vital role and enables us to become better at what we do. The process of decision making is very important in various business sectors. The advancements of AI technologies has made it easier to integrate models in such processes (Pomerol, 2017). However, some areas, such as medicine, require high levels of accuracy and correctness of certain decisions and predictions. Therefore, accurate decision making is of vital importance and a central agenda of discussion.

In general, there are good reasons for basing decisions on predictions of a computing machine. One advantage in that regard is the lack of emotion that the machine has (Stark, 2021). Naturally, this absence of emotion is beneficial for tasks that require a logical and rational decision process. With technological advances and the rise of Industry 4.0, fast and efficient processes based on huge data sets have seen to bring a major competitive advantage for companies (ibid.). However, there are areas where the use of AI technology is questionable and those areas are usually related to jobs where an emotional bond or empathy is needed (Spinola, 2005). One needs to question the impact of AI decision making tools on the future of the economy on the job market. Are there areas that can not be exchanged through AI or is more research needed, for example, to create an emotional intelligent system?

Research of human-level intelligence development investigating artificial general intelligence (AGI) raises fundamental ethical considerations for humanity (Fitzgerald et al., 2020). While there is great potential in this type of technology, one must also consider the pitfalls it brings, especially concerning ethics and policy-making.

As it can be seen, several themes have been identified on the topic of Human and Machine. AI decision making tools can be beneficial in economic sectors where high accurate results are needed. However, by increasingly using AI tools, there is a high risk of AI replacing human jobs. Currently, jobs which require human empathy can not be replaced by AIs as they lack emotions. With rapid AI developments, more AI systems are being considered for all economic sectors. Therefore, the first identified issue of this research is the following:

Issue 1: To which extent should human decision making be outsourced to AI systems?

Due to the complexity of AGI research, only a limited number of scholars are conducting studies in that field. The field of AGI aims to create and study software or hardware systems with general intelligence. Ultimately, the goal is to develop a tool that has

general human intelligence and perhaps be even more intelligent than human beings (Goertzel, 2014). Regardless of the potential benefits this technology might bring, one must also consider the issues it may bring to society, especially concerning ethics and policy-making. Therefore, the following second issue will be investigated for this report: *Issue 2: What are the potential benefits and harms of an Artificial General Intelligence (AGI)?*

2 Related Issues

1) Scientific viewpoints

a) *Deploying human decision making into AI systems*

There are areas in life where wrong or slightly inaccurate decisions do not make a significant difference. In these areas, a high accuracy machine learning model can be used without much needed thinking (Pomerol, 2017) However, the areas where prediction accuracy and uncertainty quantification of such models are crucial, are usually areas of great risks where the results of a decision potentially cause a lot of (side-) effects (Fitzgerald et al., 2020).

Furthermore, there are areas of great uncertainty where there is not enough available information for making a well-informed decision. In such domains, Nordström (2021) shows that it is of particular importance to think about uncertainty implications for public policy and to analyze the possible outcomes that could arise when using AI algorithms. Moreover, the process of how decisions are made by AI algorithms should be arguably as transparent as possible. Since most AI algorithms are more or less black-boxes by the way they are designed, the effort to make them more transparent and understandable could outweigh the benefits of it (Nordström, 2021).

Thus, de Fine Licht (2020), proposes a framework that helps analyze how transparency in and about AI decision-making can affect the public's perception of the legitimacy of decisions and decision-makers. The reason is that a limited form of transparency that focuses on providing justifications for decisions has the potential to provide sufficient ground for perceived legitimacy without producing the harm that full transparency would bring (de Fine Licht, 2020).

In 2016, healthcare AI projects attracted more investment than AI projects in any other sector of the global economy. However, among the excitement, there is equal skepticism, with some urging caution at inflated expectations. (Buch, V. H., Ahmed, I., & Maruthappu, 2018). The scientific community has found both positive and negative aspects about the application of AI tools for human decision making tasks. To implement more AI decision making tools in society, more transparency and further research on accuracy is needed.

b) Creating human-level (general) intelligence in AI systems

Since the development of artificial general intelligence (AGI) is very demanding, both in the complexity of research and the needed computational capacities, there are only a few teams and companies in the world that have the resources to perform research on this topic (Naudé1 & Dimitri, 2019). As of now, there are two main areas of research and development (R&D) that have the potential to create an artificial general intelligence (AGI). The field of AGI is about the creation and study of software or hardware systems with general intelligence comparable to, and ultimately perhaps greater than, that of human beings (Goertzel, 2014).

The first R&D area is the development through evolution of intelligent agents. It has the potential to create intelligence during the path of evolution, just like nature did with us humans (Eiben et al., 2011)

The second one is the upscaling of artificial neural networks (ANNs) to the size of a human brain (Schaefer et al., 2021). Even though both approaches are quite promising, a lack of computational power and technological advancements impacts the research process and slows it down. As a result, it is unlikely to achieve major breakthroughs anything close to this point in time (Fjelland, 2020).

However, with rapid AI developments, an arms race for an AGI among those teams would be detrimental for and even pose an existential threat to humanity if it results in an unfriendly AGI. Therefore, establishing policies for regulating the competition among those teams and setting the right nuanced incentives is crucial for a responsible development of this technology (ibid.).

The scholar Graham (2021) argues that, to date, social scientists have dedicated little effort to the ethics of AGI or AGI researcher. As a result, it has become an even more urgent topic of discussion. Furthermore, a supposition of this is that public debate will inevitably advance to AGI ethics as the urgency of this technology becomes more apparent making AGI ethics a crucial humanitarian issue (ibid.). This shows that individuals operating in the tech industry are less focused on thinking about policy making and more on the ways how technology can solve the problems it causes.

2) Popular viewpoints

a) *Deploying human decision making into AI systems*

The article by Collins (2019), discusses the impact of AI systems on human decision making. He highlights that humans often need to make decisions under difficult circumstances which can cause bad outcomes. Humans will not take the right decision every time due to their nature (Collins, 2019). Decision making is needed in everyday life and humans need to make quick decisions when, for example, getting behind the wheel of a car to drive. However, humans also make faulty decisions and can lead to bad outcomes such as car accidents (ibid.). When they were asked what caused them to crash, they often stated that they thought they had made the correct decision at the time. This reflection may come as a surprise but in fact, holds truth because none of us humans choose to get things wrong (ibid.). As a matter of fact, what separates us humans from other mammals is our capability for thinking. However, human intelligence is not without its flaws and we often use the quote ‘To err is human’ for it (ibid.). The tendency for humans ‘to err’ during thinking can be seen when discussing details of past events with friends or family. Oftentimes, each person will have different recollections of the events and can cause conflicts between individuals (ibid.). However, with the accessibility of technology and AI systems such as Google search, arguments of who is right or wrong can be avoided and false thinking can be corrected very quickly. Collins (2019) states that 25 years ago, it was unthinkable that these sophisticated information systems would scan through all the world's knowledge and provide a useful answer.

Today, these AI systems are being taken for granted, causing humans to think less and take less decisions independently (ibid). With the fast development of AI, debates have been raised on the pros and cons of AI to humanity. The inventor Kurzweil, praised AI and all its benefits (ibid.). In his book, he described how AI will benefit and enhance humanity through human-machine symbiosis. On the contrary, the theoretical physicist Hawking, criticized AI and stated that humanity needs to be cautious of AI because the full development of it could end the human race (ibid.). Clearly, AI takes the form of super intelligence, moving beyond

the limits of human intelligence, that is, the capacity of the single human brain. To gain a better idea of the uncertainties that AI holds, one should closely evaluate the relationship between human thinking and AI (ibid.).

The book 'Thinking fast and slow' by Kahneman (2011) shows that the human brain is not purely rational due to its dual thinking modes. The two ways of human thinking are System 1 and System 2 (Kahneman, 2011). System 1 is fast thinking, requiring little to no effort and occurs automatically. System 2 is slow thinking, involving precise attention to decode complex relationships and to understand details (ibid.). System 1 is therefore intuitive, deterministic, and undoubting. System 2 is rational, probabilistic and highly aware of uncertainty and doubt (ibid.). Currently, AI does not have the capability to execute dual thinking exactly like humans. AI acts rational and analytical and fast. Nevertheless, if the future of AI allows us to do System 2 thinking at System 1 speeds, it is likely to be a powerful resource that significantly reduces, if not eliminates, the inherent biases and predispositions in human decision-making. However, even if this might seem like a perfect idea, one needs to evaluate the possible dangers of it. Collins (2019) raises concerns and states the question: "How can we be sure that this new super intelligence will not pose a future threat to humanity?". He proposes that it is dependent on whether or not our AI systems are designed as collective intelligent systems, showing that societal values should play a role in the development of AI (Collins, 2019).

b) *Creating human-level (general) intelligence in AI systems*

The TED talk by Kai-Fu Lee (2018), discusses the era of AI discovery and its social opportunity costs. The talk starts with his personal experience in which he had to make the choice of either witnessing the birth of his first child or presenting his AI discovery at Apple. Luckily, the birth took place in the morning, enabling him to attend both events. What is striking is that he prioritized his work ethics over his personal values, impacting his personal life and his relationships (Lee, 2018). Nevertheless, his presentation went brilliantly and led to further research and the rise of a new field of AI known as 'deep learning'.

Deep learning is a technology that can take a huge amount of data within one single domain and learn to predict or decide at superhuman accuracy (ibid.). This type of technology is a key component of AI and is currently employed in various domains. For instance, it is used in self-driving cars where it receives sensory data from driving on the highway, enabling it to drive a car as good as a human being (ibid.).

Nowadays, humanity is found to be in the era of implementation where execution, product quality, speed and data are of most importance. Chinese entrepreneurs in the technology sector have taken this very seriously and in return, execute rigorous work ethics where employees work from 9am to 9pm, seven days a week (ibid.). In the past decade, Chinese product quality has improved immensely because of a fiercely competitive environment. In the US, corporate competitiveness is more diplomatic and open for discussion between the parties involved (ibid.). Comparing the two nations, one can conclude that China operates in a brutal environment, leading to its rapid growth in the industry. As a matter of fact, due to its large market and implemented AIs, entrepreneurs have the chance to collect a huge amount of data, the key ingredient for AI (ibid.). Resultantly, Chinese companies are now one of the most valuable in computer vision, speech recognition, machine translation and drones. With the US leading the era of AI discovery and China leading the era of implementation, humanity is witnessing two AI superpowers driving the fastest revolution in technology (ibid.).

The AI transformation will make entrepreneurs tremendously wealthy but also poses vital challenges in terms of potential job replacements. The past industrial revolutions led to

more jobs but AI completely replaces individual jobs (ibid.). Evidently, this change has already been witnessed in factories. However, with rapid AI developments, jobs requiring social intimacy like customer service, teaching and counseling will be gradually replaced by AI in the next 15 years (ibid.). One may pose the question if this transformation will overtake mankind not only in the work environment, but also in all aspects of human life. To what extent is it ethical to create human level intelligence in AI systems and how much should humans sacrifice?

Lee (2018) later introduces his near-death experience and how he came to the realization of the human importance of love. He started to consider how AI should impact, work, and coexist with mankind. Despite the fact that AI is replacing routine jobs, work is not the reason for why humans exist on this earth (Lee, 2018). The reason why we exist is love. Only humans can uniquely receive and give love and that is what makes us different from AI (ibid). Even if AI will carry general human intelligence, it will never replace the unique hearts and brains of humans. In conclusion, AI can increase productivity, become great tools for the creatives and can work with humans as analytical tools in high-compassion jobs in the future (ibid.). Mankind will always differentiate itself from AI with uniquely capable jobs that are both compassionate and creative due their irreplaceable hearts, feelings and unique brains.

The potential danger of AI to humans has been illustrated in many science fiction novels. In fact, the writer Asimov (1942) introduced the ‘Three laws of Robotics’ in his short story ‘Runaround’ and is used as an ethical guideline for the development of robots in the real world. These laws were created to keep robots from hurting humankind and many other science fiction novels were published based on this notion. However, the majority does not focus on the more significant social effects of AI on humanity. The article by Christakis (2019) therefore investigates the ways AI could affect how humans interact with one another. The fast development of AI and machines has led to a radical transformation in society, both individually and collectively (Christakis, 2019). Those innovations have made a change to how us humans communicate and store information. Nevertheless, fundamental human behavior such as love, support and friendship have not changed through AI and remain vital in social communities (ibid.).

The constant changes of AI, especially adjusting physical and behavioral features to be more human-like may be much more disruptive to humanity. If human-like AIs become part of our daily lives, the fundamental human behaviors like love or kindness might be altered, impacting the way we interact with each other and the AIs (ibid.).

Basic human values such as love is warned to be disrupted by the current development of sex robots. On one hand, robots will be dehumanizing and lead us humans to pull away from real life intimacy (ibid.). In 2018, a Japanese man ‘married’ his AI hologram because he got so attached to it. All kinds of sexual acts between the human and AI can also be performed, posing danger to real life partners when trying to treat them the same way sexually (ibid.). On the other hand, sex robots could improve sex between humans when it takes up an assisting role. Furthermore, they cannot carry sexually transmitted diseases or result in unwanted pregnancies (ibid.). Humans are unique creatures and AI will never feel and act like humans due to their nature. The danger is not creating an almost human AI, it is that the new AI inventions have the potential to manipulate and alter human behavior (ibid.). If more people choose human-robot over human-human interaction, human beliefs like friendship, love and compassion might eventually disappear. Additionally, if humans start to develop more intimate relationships with AIs, could it be possible that humankind might go extinct and start off a new era where AI takes over?

3 Discussion

1) Deploying human decision making into AI systems

The application of AI systems for human decision making has been a topic of discussion in both the popular press and scientific community. According to Nordström (2021), the uncertainty issues of AI for policy making needs to be extensively examined for the sake of humanity and for the prevention of possible catastrophes caused by algorithms. Additionally, research suggests that more transparency is needed in the AI decision making process. The scholar de Fine Licht (2020) proposes using a framework to help analyze how transparency in and about AI decision-making can affect the public's perception of the legitimacy of decisions and decision-makers. This method can help provide legitimate justifications for decisions made by AIs without producing the harms of full transparency. New AI developments, especially healthcare AI projects, have received major investments from both the public and private sectors around the world. However, both popular and scientific publications have emphasized to view this skeptically, with some urging caution at inflated expectations. (Buch, V. H., Ahmed, I., & Maruthappu, 2018).

The popular press article by Collins (2019) highlights the natural inaccuracy of decision-making amongst humans and the relation to highly accurate AI systems. Humans need to make quick decisions, and can cause bad outcomes at times. When reflecting on poorly made decisions, the majority believed they had made the correct decision at the time. By implementing AI systems for decision-making, for instance, self-driving cars, it can prevent negative outcomes, like accidents, caused by human error (Collins, 2019). Humankind has greatly benefited from technology and AI systems such as Google search where their false thinking can be corrected very quickly by scanning through all the world's knowledge and providing a useful answer. However, the downside of superhuman accurate AI systems is that it causes humans to think less and take less decisions independently (ibid).

Throughout history, inventors and scientists have debated on the outcomes of AI decision making. On one hand, the innovator Kurzweil, praised AI and all the benefits it can bring through human-machine symbiosis. On the other hand, theoretical physicist Hawking, criticized AI and stated that humankind needs to be cautious of AI because the full development of it could end the human race (ibid.).

To closely evaluate the relationship between human thinking and AI, one needs to understand the human mechanism for making decisions. The popular book by Kahneman (2011) introduces the dual thinking modes humans have. What differentiates us from AIs is our way of thinking. Humans can think in two modes, System 1 which is fast thinking, and System 2, slow thinking, requiring precise attention to decode complex relationships and details (Kahneman, 2011). Both popular and scientific views have raised similar issues concerning the human complexity of decision-making. AI systems are accurate and analytical and therefore pose to be beneficial for rational decision making (ibid.). One problem of AI decision-making is its lack of consideration of social values. The outcomes of those systems are dependent on whether or not they are designed as collective intelligent systems, showing that societal values should play a role in the development of AI(Collins, 2019).

2) Creating human-level (general) intelligence in AI systems

Both scientific and popular press sources perceive the creation and process of creating human-level intelligence robots as a double-edged sword.

The scientists seem to look at the problems more from a technical standpoint, thus estimating the process that the research teams undergo by creating such algorithms as the most crucial (Naudé1 & Dimitri, 2019). The popular press pays more attention to the perceived problems of a possible created AGI and its interaction with humans (Christakis, 2019).

Moreover, Graham (2021) suggests a structured distributed framework to develop guidelines and ethical boundaries that can be taken into account when working on AGI.
Such a framework does have the potential to alter the perception of the public and potentially shifts the discussion to more scientific problems and fundamental algorithmic questions that arise when using such technology (Graham, 2021).

Articles of the public press shed light into the possible outcome of a more human-like interaction with AI-systems, especially for humanoid AI tools. Christakis (2019), sees good and bad sides in such a development. The bad aspect of it is that robots will be dehumanizing and lead us humans to pull away from real life intimacy (Christakis, 2019). On the positive side, sex robots could improve sex between humans when it takes up an assisting role. Furthermore, they cannot carry sexually transmitted diseases or result in unwanted pregnancies (ibid.) The issue of humanoid AI machines is not the resembling human features.
The danger of this creation is that the new AI inventions have the potential to manipulate and alter human behavior (ibid.). If more people choose human-robot over human-human interaction, human beliefs like friendship, love and compassion might eventually disappear (ibid.).

In summary, both scientific and popular press see the potential advantages and disadvantages that can result from increasingly using intelligent AI systems in society. Moreover, they try to weigh those against each other to figure out whether the pursuit of such a development is reasonable or should be questioned. There is a huge intersection and unanimity, for the creation of a guided ethical framework to establish and reason about the effects of the development of AGI systems and AI systems in general.

4 Conclusion

After the evaluation of scientific and popular press articles, relevant information was identified for the given proposed research issues:

Issue 1:To which extent should human decision making be outsourced to AI systems?
Issue 2: What are the potential benefits and harms of an Artificial General Intelligence (AGI)?

The main topics that appeared in the articles concerned the usage of AI-driven decision making in areas of high-risk. AI-based decisions are often black-boxes, making the user unable to determine the reasoning behind such a decision. This can undoubtedly lead to several problems. On the other hand, the usage of AI can bring a higher accuracy, because a lot of data can be processed and used, that a human would be unable to process.

When reasoning about AGI, it is very important not only to think about the finished intelligent machine, but to also keep into account, that nuances in the process of development can lead to extremely different behaving AGIs, ranging from a dystopian view of a bad AGI that harms humanity to a utopian perfect AGI that helps us in any way possible. Thus, many authors agree that a standardized framework for implementing and creating ethical guidelines for the development and usage of intelligent systems is very important and needs to be advanced.

References

Buch, V. H., Ahmed, I., & Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *The British journal of general practice : the journal of the Royal College of General Practitioners*, *68*(668), 143–144. https://doi.org/10.3399/bjgp18X695213

Binns R (2018) *Algorithmic accountability and public reason.* Philos Technol 31(4):543–556

Christakis, Nicholas A. (2019): *How AI Will Rewire Us* https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/ (Links to an external site.)

Collins, Rob, (2019): *How artificial intelligence will transform human thinking https://www.management-issues.com/opinion/7362/how-artificial-intelligence-will-transform-human-thinking/ (Links to an external site.)*

Copeland, B. (2021, December 14). *artificial intelligence*. *Encyclopedia Britannica*. https://www.britannica.com/technology/artificial-intelligence

de Fine Licht, K., de Fine Licht, J. *Artificial intelligence, transparency, and public decision-making. AI & Soc* **35,** 917–926 (2020). https://doi.org/10.1007/s00146-020-00960-w

Eiben AE, Kernbach S, Haasdijk E. E*mbodied artificial evolution: Artificial evolutionary systems in the 21st Century.* Evol Intell. 2012 Dec;5(4):261-272. doi: 10.1007/s12065-012-0071-x. Epub 2012 Apr 20. PMID: 23144668; PMCID: PMC3490067.

Fitzgerald, Mc Keena Aaron Boddy, and Seth D. Baum, (2020). *2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy.* Global Catastrophic Risk Institute

Fjelland, R. *Why general artificial intelligence will not be realized. Humanit Soc Sci Commun* **7,** 10 (2020). https://doi.org/10.1057/s41599-020-0494-4

Goertzel, Ben. (2014). *Artificial General Intelligence: Concept, State of the Art, and Future Prospects.* Journal of Artificial General Intelligence. 10.2478/jagi-2014-0001.

Honneth, Axel (2018): *critical essays: with a reply by Axel Honneth.* Brill Academic Publishers, Leiden, pp 207–232

Krijger, Anderson Joris(2011) *Situating Axel Honneth in the Frankfurt school traditio*n. In: Petherbridge D (ed)

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan

Lee, Kai-Fu (2019): How AI can save our humanity. Youtube, https://www.youtube.com/watch?v=ajGgd9Ld-Wc

Meng, Jingbo & Dai, Nancy. (2021). *Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not?*. Journal of Computer-Mediated Communication. 26. 10.1093/jcmc/zmab005.

Naudé, W., Dimitri, N. *The race for an artificial general intelligence: implications for public policy. AI & Soc* **35,** 367–379 (2020). https://doi.org/10.1007/s00146-019-00887-x

Ostheimer, Julia, Soumitra Chowdhury, Sarfraz Iqbal (2021): *An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles,* Technology in Society, Volume 66, (2021),

https://doi.org/10.1016/j.techsoc.2021.101647.

Pomerol, J.-C. (1997). *Artificial intelligence and human decision making. European Journal of Operational Research, 99(1), 3–25.*

Schaefer, M., Michelin, L., and Kepner, J., (2021), "*Naming Schema for a Human Brain-Scale Neural Network"*

Spinola, Jackeline & Gudwin, Ricardo & Queiroz, Joao. (2005). *Emotion in Artificial Intelligence and Artificial Life Research: Facing Problems.* 501. 10.1007/11550617_52.

Stark, Luke (2021) *The Ethics of Emotion in Artificial Intelligence Systems*