MONOGRAPHS

Polish Economic Institute

Social, ethical and cultural aspects of the use of Artificial Intelligence. The Future of New Technologies

GovTech Poland

Direct recipients of GovTech services include broadly understood local, regional and central administration authorities as well as other entities performing public tasks, such as hospitals, schools or transport companies. But the effects of technology services always concern citizens: administration service recipients. GovTech offers an opportunity to increase productivity, create jobs and boost economic growth – for both public administration and the private sector. GovTech also takes measures for Polish education and culture – the objective of the department is to provide state-of-the-art techniques in education and learning to pupils and students in Poland and to foster creative industry development.

www.gov.pl/web/govtech

Competition

The AI Essay competition is international in nature. Through various forms of written expression (essays, reports, op-ed articles, etc.), it aims to give the floor to the young generation and to present the views of its members. The focus of the competition is on Artificial Intelligence and the 2022 edition concerns the socio-cultural aspects of the use of Artificial Intelligence. The project offers an opportunity to enhance writing, creative skills and to learn more about AI. The competition is addressed to undergraduate, graduate and doctoral students from the European Union, the European Free Trade Association, the European Economic Area, the Swiss Confederation or the United Kingdom. Each winner receives a prize of EUR 1,000 funded by GovTech Poland.

www.gov.pl/web/govtech/konkurs-esejowy

Polish Economic Institute

The Polish Economic Institute is a public economic think tank dating back to 1928. Its research primarily spans macroeconomics, energy and climate, foreign trade, economic foresight, the digital economy and behavioural economics. The Institute provides reports, analyses and recommendations for key areas of the economy and social life in Poland, taking into account the international situation.
The scientific achievements of the Polish Economic Institute should be seen as public goods.

www.pie.net.pl

Competition winners

Bartosz Wultański
Apolinary Rzońca
Bartosz Kuczyński
Jenny E. Simon
Zofia Kosowska

Introduction

Zdzisław Krasnodębski

Technological progress has always led to social and cultural change.

Technology has raised great hopes and at the same time great fears. It has changed man's environment, but also man himself and his attitude to the world. The watershed since which these changes gained rapid momentum was the Industrial Revolution.

Both nineteenth and twentieth century philosophy and sociology are full of reflections on these changes. Many of these descriptions, today, seem exaggerated to the point of ridiculousness. But - on the other hand - we cannot shake off the impression that some of the most far-reaching imaginings of science fiction writers, from Jules Verne to Stanislaw Lem, are beginning to be realised in the present day.

Artificial Intelligence is already revolutionising many areas of our lives. It is a useful tool and we are already using it in more and more areas of life. But in this case, our dilemmas go even further those faced by earlier technical inventions. AI is forcing us anew to ask ourselves fundamental philosophical questions: what is intelligence and consciousness, what is a self-aware subject. How does artificial intelligence differ from human intelligence and can it replace it? Can we imagine a machine, a robot, that will not only be conscious in the sense of perceiving the environment and reacting to its changes, adapting its actions to these changes, but also be self-aware, have self-awareness?

This is still a long way off and we don't know if it is even possible, but we are already talking about autonomous machines, equipped with developed AI, which themselves 'decide' to act in certain situations. Experts speak of 'delegated autonomy'. The very possibility of such delegation represents a revolution, raising fundamental ethical questions. Human actions have, as we know, not only an instrumental dimension, but also a moral one. We do not judge them for their effectiveness, but also from the point of view of ethical good. According to some opinions, the use of AI will allow to detach ethical decisions from subjective judgments and to ensure that the decision is as objectively optimal as it can be.

The same situation should lead to the same choice, regardless of who is the one making it, so the subject and its subjective attitude becomes irrelevant.

It is, however, a big question mark as to whether this really constituted 'moral progress', whether we would want to live in a world in which this kind of "objectivity" is a rule.

These many other questions require a great deal of reflection - and interdisciplinary reflection, including philosophical and theological reflection. I think there is still too little of it, especially in our country. That is why the GovTech Centre's essay competition on cultural and social impacts is such a valuable initiative. The best essays have been collected in this publication and their authors awarded. The organisers and the members of the jury, which I had the honour to chair, hope that this competition will contribute both to ensuring that these talented authors continue to deal with this issue in depth and that it will stimulate the interest of others and popularise the subject.

Social, ethical and cultural aspects of the use of Artificial Intelligence.

The Future of New Technologies

Authors:

Zofia Kosowska, Bartosz Kuczyński, Apolinary Rzońca,
Jenny E. Simon, Bartosz Wultański

Editors:

Jakub Nowak, Małgorzata Wieteska

Translation and proofreading:

Annabelle Chapman

Graphic design:

Anna Olczak

Graphic cooperation:

Tomasz Gałązka, Sebastian Grzybowski

Text and graphic composition:

Sławomir Jarząbek

Polish Economic Institute

Al. Jerozolimskie 87
02-001 Warsaw, Poland

ISBN 978-83-67575-00-3

Warsaw, October 2022

Table of contents

Zofia Kosowska
University of Warsaw
College of Interdisciplinary Individual Studies in Humanities
and Social Sciences Faculty of Psychology
Faculty of Philosophy, Cognitive Science
e-mail: z.kosowska@student.uw.edu.pl

The future of new technologies: What is the actual challenge – creating ethical machines or improving social ethical attitudes towards them?

Summary

One of the biggest issues arising in the Artificial Intelligence industry is whether AI agents are capable of coexisting peacefully with people, no matter what its development brings. The main aim of this essay is to discuss advanced AI's morality and examine its chances of coming to be viewed with sympathy or tolerance. The first chapter investigates the possibility of creating machines with a moral status and its potential consequences. The second concentrates on the controversies regarding endowing AI systems with moral compasses. The final part focuses on the relationship between artificial agents and humanity, explores the fears attached to it, and expresses the potential needs of both sides. The author concludes that, while it is not easy to influence social attitudes, it is even harder to create a system that would be objectively ethical.

Introduction

Robot ethics is a different field from computer ethics. It describes the relations between a person confronted with an autonomous robot capable of making independent decisions (Bringsjord et al., 2003). The concept of inventing thinking and sensing machines raises a host of moral issues, including robots' ethical abilities and ethical relationships between humans and AI systems (Bostrom, Yudkowsky, 2011). Furthermore, it prompts discussion about disputable issues such as the uniqueness of humankind, the definition of consciousness, and the variety of views on the term "morality".

The controversy related to the machines' moral status

To begin addressing the issue of creating machines that are ethical, one cannot miss out on a hidden, yet fundamental, implication. It strongly suggests a belief that, one day, a machine could potentially become a candidate for having a moral status.

A robot usually consists of design elements and software. The software is based on sets of algorithms (based on decision trees) and the part most interesting in terms of ethics, Artificial Intelligence (AI). This refers to any artificial computational system that shows intelligent behaviour; for example, machine learning or reasoning (Muller, 2020).

While it is broadly agreed that current AI systems lack moral status, there is also a popular view that they may come close to having it at any moment. To fully understand the consequences, we should first discuss what "moral status" means. The definitions in different philosophical trends and concepts of ethics vary. One common view is that there are two obligatory criteria that make a moral being: sentience and sapience (Bostrom, Yudkowsky, 2011). It is rather difficult to imagine sentience – the capacity to feel fear or pain – being implemented on a robot, probably because we cannot visualise suffering without a body. The goal is to transfer the concept to a robot – that would only mean changing how we perceive suffering. Meanwhile, machines' sapience – a phrase connected to "higher intelligence, self-awareness and being a reason-responsive agent" – is a frequently-raised, interesting topic (Bostrom, Yudkowsky, 2011).

The story begins in the 1950s, when British cryptologist and mathematician Alan Turing proposed a test which, in his view, was supposed to assess a machine's ability to think in a similar way to a human being. The Turing Test was a simple idea: if a computer is able to trick a person into thinking it is actually a human, then it is fair to say that it is capable of thinking (Turing, 1950).[1] Turing's concept provoked a wave of criticism. Some of the strongest came from John Searle, who responded with an idea that was just as simple. The famous thought experiment, called the Chinese Room, sought to show that strong AI cannot be created.[2] Assume that there is a computer able to act as if it is familiar with speaking Chinese. It receives input in Chinese from the outside, then processes it with a set of grammatical rules and produces output in Chinese. Suppose that the computer is so extraordinarily skilled that it is able to convince a Chinese person that he or she i stalking to a human. In Turing's understanding, this victory would mean that the computer is a thinking creature. Searle's view is quite the opposite. Assume that it is a man sitting in the room and performing the task instead of the computer, he says. Although the man executes it perfectly, he does not understand a single word. In Searle's opinion, that is how AI works: it cannot understand because it only operates using symbols and syntactic rules that it is given and cannot refer them to reality (Searle, 1990). As an undoubtedly

[1] Not a single computer has passed the test, even after 70 years.
[2] Searle made a distinction between weak AI, which is only able to simulate thinking, and strong AI, which is a truly thinking system.

interesting case, it has been a subject of discussion for many years now. Some of the most popular responses were:

a. No wonder Searle does not understand a word – neither does a single neuron.
b. Searle is not a Chinese speaker, but he understands English, which is essential to follow the instructions. That is how a computer thinks; it has to process the information given in its programming language.
c. Let us replace the notes with impulses and shrink the room to the size of a small ball. Is it any different from a human brain now? (So why do we even need AI? – Searle responds ironically).

It may be the differences in our understanding of the idea of thinking that cause so many problems. Public discourse is often bombed with headlines claiming that scientist X has succeeded and created a thinking robot but, after reading the article, the reader notices that X's definition of thinking is rather underwhelming. Despite the fact that not a single computer has passed the Turing Test, it seems that even the newest systems, such as the GPT-3, are not commonly perceived as successful in deceiving people (Computerphile, 2020).

While there is no common consensus on perceiving thinking, many issues are discussed in the case of creating a thinking and sensing machine. Every morally-relevant being has its rights – i. e. the right to live – but it is also burdened with obligations. The right to live, as the most fundamental right, is also the one frequently discussed in the context of the legalisation of abortion or of laboratory animals (Bostrom, Yudkowsky, 2011). To a person who is not personally invested in AI discourse, it may seem grotesque to examine AI systems' right to live. Nevertheless, if we create a human-like machine that thinks and senses, it seems appropriate to grant it the exact same rights. Therefore, the concept of hurting or murdering an AI robot and its consequences for the perpetrator should be discussed. Would such a machine become a member of society? Maybe, because it is artificial, we should consider it morally relevant, but less significant than a human? Should such a robot have legal rights, a name and, for example, a passport?

Machines' moral judgement

Imagine standing in court, accused of tax fraud. The decision on whether you are guilty will be made by a robot judge. Science fiction? Not necessarily. As time goes by and robotics develops intensively, new questions about machines' nature arise. Many scientists believe that the growing role of AI in our lives is a reason to equip the newest models with a moral compass (Wallach, Vallor, 2020). It is common to refer to such machines as artificial moral agents (AMAs) (Robbins, Wynsberghe, 2018). Due to the fact that robots will inevitably be created for a whole variety of uses, we will probably put them in morally-salient situations, too. Hence, it is rather logical to discuss endowing them with moral reasoning and decision-making without human help. The main argument in favour of providing AI systems with a moral compassess is that it

would assure society that a robot would not hurt it but, rather, protect it. Furthermore, it could promote AI as people's friend and improve social attitudes towards it. Additionally, it would help explore human ethical nature and help the AI industry flourish (Wallach, Vallor, 2020).

If it was not for these many doubts, we could easily solve the main problem explored in this essay right now. The principal difficulty seems to be based on philosophical grounds; specifically, the concept of free will. As arbitrarily-programmed systems, robots do not seem to have the possibility of gaining free will, which is what may distinguish them from humans, after all. It could also lead to the conclusion that morality is a feature exclusive to humans. The philosophical nature of this problem makes it demanding when it comes to implementation, too. There are so many theories on ethics that imply that it would be a challenge to work out a consistent, homogeneous system of ethical values that could be applied to a robot (Sparrow, 2021).

Dr Aimee van Wynsberghe and her team (2018) try to prove that morally-salient situations do not require that every participant be morally conscious. Consider animals used for therapeutic purposes: they take part in morally-salient situations, but are not moral beings themselves. Why do we not apply the same logic to computers? We would equip them with sets of protective measures and efficient controlling systems. The key is to train them well, not to teach them ethics.

While the discussions are ongoing, there have been many attempts at producing an artificial moral agent. The Delphi bot created by Allen Institute for Artificial Intelligence (2021) has been designed to immediately assess a description of a situation that it has been given. This neural-like network has been trained on a 1.7 million corpus representing particular situations and reasonable judgements. When Delphi's opinions were shown to a group of students, the people agreed with the bot to a degree of 92%. We can also see morally-oriented AI systems that are already implemented in daily life, such as driverless cars or weapons that choose their own targets (Etzioni, Etzioni, 2017).

The most interesting aspect of morally-trained systems seems to be whether their ability to cope with moral dilemmas is better than a human's, and that is actually the field that the industry focuses on. We may distinguish between two basic types of dilemmas to address: a dilemma within an agent and between two agents, which can involve two artificial agents, or one artificial agent and a human (Cervantes et al., 2019). The first can arise when one of an agent's ethical norms collides with another. It can be illustrated by an example featuring an autonomous car with a vehicle is driving towards it at enormous speed. It has to decide whether to turn rapidly and risk hitting pedestrians, or inevitably crash into the other car. The second one, especially the variant involving one AI agent and one homo sapiens, seems to be much more complex. AMAs could possibly meet an Infinite number of different agents, every o of them with unique moral values. Sharing decisions with them or acting on their behalf could therefore be an extremely hard task (Cervantes et al., 2019).

AI-phobia

The more the AI industry flourishes, the more people are scared of it. The general problems seem to be robots' over-intelligence, and therefore anxiety about losing control over them, fear of being harmed by robots, mass unemployment, and technology falling into the wrong hands (Forbes, 2019).

General anxiety about AI is often fueled by popculture creations. The big trend of movies about AI started in the 1980s and is still doing well, though in a slightly different form. *Blade Runner* (1982), *Terminator* (1984) and *The Matrix* (1999) are the most popular oldschool AI films that everyone has seen. While *Terminator* and *The Matrix* show the more evil qualities of AI, *Blade Runner* tried to warm up its image. The latest tendencies in the industry discuss the complexity of humanoids; that is, the *Westworld* (2016) and *Raised by wolves* (2020) series. The *Westworld* series takes place in the near future and features an amusement park where people can surround themselves with human-like androids and fulfill their fantasies, including hurting or murdering someone. It is not murder since they are robots, right? Not exactly. The humanoids turn out to be much more intelligent and sentient than it might have seemed at the beginning. *Westworld* touches on the sensitive subject of humans' morality towards AI machines.

Raised by wolves addresses an issue that is a source of fear for many AI sceptics. As noted in previous chapters, a robot can be trained to form its own moral system. What if the corpus it has been learning from differs noticeably from what is widely seen as ethical? Conflicts between a robot that follows its own moral codex and humankind could be terrifying. That is what people fear the most: radical ideas controlling powerful AI systems.

Conclusion

While previous research shows that machines' morality is still being discussed, we cannot look at the industry and not conclude that the future is happening right now. We do not think about humanoids on a daily basis, but AI is approaching daily life right now. My biggest concern is the lack of time. Every AI scientist's dream is to develop the best machine possible.

For the reasons outlined above, I must simply lean towards the first part of the question in the title and say that creating a machine with objectively-efficient moral compass is the ultimate challenge. Although convincing people to trust or tolerate AI systems seems to be a crucial part of its daily functioning, we cannot overlook the fact that a poorly-programmed moral compass can have lethal consequences. Until now, we only had to fight malicious chat-bots like Google's Alice and Bob inventing their own language, but who knows what may come next?

Bibliography

Allen Institute for Artificial Intelligence (2021), https://delphi.allenai.org/ [accessed: 23.03.2022].

Bigman, Y.E., Waytz, A., Alterovitz, R., Gray, K. (2019), *Holding Robots Responsible: The Elements of Machine Morality*, "Trends in Cognitive Sciences", doi:10.1016/j.tics.2019.02.008.

Bringsjord, S., Schimanski, B. (2003), *What is Artificial Intelligence? Psychometric AI as an Answer*, Proceedings of the 18th International Joint Conference on Artificial Intelligence.

Bostrom, N., Yudkowsky, E. (2011), *The ethics of Artificial Intelligence*, "Draft for Cambridge Handbook of Artificial Intelligence".

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., Ramos, F. (2019), *Artificial Moral Agents: A Survey of the Current Status*, "Science and Engineering Ethics", doi:10.1007/s11948-019-00151-x.

Computerphile (2020), *GPT3: An Even Bigger Language Model – Computerphile*, https://www.youtube.com/watch?v=_8yVOC4ciXc [accessed: 23.03.2022].

Etzioni, A., Etzioni, O. (2017), *Incorporating Ethics into Artificial Intelligence*, "The Journal of Ethics", No. 21(4), doi:10.1007/s10892-017-9252-2.

Hagendorff, T. (2020), *The Ethics of AI Ethics: An Evaluation of Guidelines*, "Minds and Machines", doi:10.1007/s11023-020-09517-8.

Moor, J. (2006), *What is Computer Ethics?*, "IEEE Intelligent Systems", No. 21(4).

Müller, V. (2020), *Ethics of Artificial Intelligence and Robotics*, (In:) Zalta E.N. (ed.), *The Stanford Encyclopedia of Philosophy*.

Searle, J.R. (1980), *Minds. Brains, and programs*, "The Behavioral and Brain Sciences", No. 3.

Sparrow, R. (2021), *Why machines cannot be moral*, "AI & Society".

Stylec-Szromek, P. (2018), *Sztuczna Inteligencja – prawo, odpowiedzialność, etyka*, "Zeszyty Naukowe Politechniki Śląskiej", No. 123.

"The Forbes" (2019), *Should we be afraid of AI?*, https://www.forbes.com/sites/cognitiveworld/2019/10/31/should-we-be-afraid-of-ai/?sh=6f69edcd4331 [accessed: 23.03.2022].

"The Guardian" (2021), *Is it OK to ...': the bot that gives you an instant moral judgment*, https://www.theguardian.com/technology/2021/nov/02/delphi-online-aibot-philosophy [accessed: 23.03.2022].

Turing, A. (1950), *Computing Machinery and Intelligence*, "Mind", No. 59.

Van Wynsberghe, A., Robbins, S. (2018), *Critiquing the Reasons for Making Artificial Moral Agents*, "Science and Engineering Ethics", doi:10.1007/s11948-018-0030-8.

Bartosz Kuczyński
University of Warmia and Mazury in Olsztyn
MA in Philosophy (Department of Philosophy)
e-mail: kuczynski.bartosz7@gmail.com

The future of robotics and AI – conflict or "cooperation"? Two scenarios for the era of ubiquitous artificial intelligent robots

Summary

It is believed that we are only a few steps away from the era of ubiquitous artificial intelligences. Even though the development of robots and artificial intelligent systems is strongly associated with the next industrial revolution (following the steam, electricity and electronics revolutions), the upcoming changes are not limited to the field of production. These technologies are set to be used in many other areas of life, ultimately reaching social interaction as well. This paper is an attempt to present, organise and discuss different perspectives on the future of AI and robotics. It distinguishes between two ways of thinking about the future: one based on the notion of conflict, and the other on the idea of cooperation.

Introduction

It is believed that we are only a few steps away from the era of ubiquitous artificial intelligence We already use these technologies to analyse data, search for patterns and generate solutions based on datasets. However, the effort of some engineers is focused on something more sophisticated than powerful yet narrowly-specialised algorithms. They aspire to create universal machines capable of thinking and acting at least on the human level. The more audacious ones even dream of "breathing life" into inanimate matter. Even though the development of robots and artificial intelligence systems is strongly associated with the next industrial revolution (following the steam, electricity and electronics revolutions), the upcoming changes are not limited to the field of production. These technologies are set to be used in many other areas of life, ultimately covering social interaction as well. This paper is an attempt to present, organise and discuss different perspectives on the future of AI and robotics.

Defining robots and artificial intelligences

Before we discuss particular issues and challenges, it is worth clarifying what we mean by AI and robotics. These two domains are somehow connected, but not necessarily. Briefly speaking, AI is a field devoted to the development of information processing systems with the ability to perform tasks commonly associated with intelligence. Recently, the term "artificial intelligence" has been more commonly used to describe the results of these studies – intelligent systems themselves. Robotics is a field of computer science and engineering that involves designing, constructing, and operating physical machines. The corporeality of robots is considered the main difference between them and artificial intelligent systems, as robots are thought to be material objects and AI systems are mostly associated with "software" of some kind.[3]

The term "robot" was first used in a science-fiction play by Czech writer Karel Čapek entitled *R.U.R.* (1921). The etymological origin of the word traces back to the Czech word *robota*, meaning "physically demanding, often forced labour", to reflect the purpose of creating these kinds of machines. In the play, robots were created as beings without emotional and biological needs. They were thought to be cheap sources of labour developed to fulfil human dreams of an abundant world without the burden of work. We can see that the dream is still relevant, even nearly a hundred years after the premiere of *R.U.R.* Contemporary authors such as Jeremy Rifkin (1994) and Aaron Bastani (2019) have envisioned similar scenarios in the real world. In short, we can define robots as material machines capable of performing certain tasks and behaviours typically attributed to living organisms, such as the ability to move or interact with their environment. Some robots possess artificial intelligent features, but this is not essential. Three types of artificial intelligent systems can be distinguished: (a) **artificial general intelligence** (AGI) – a hypothetical system that possesses thinking abilities equal to the level of the human mind, with all its functional attributes; (b) **artificial narrow intelligence** (ANI) – a system specialised in carrying out specific tasks, its capabilities limited to a specific field of action; (c) **artificial superintelligence** – a hypothetical intelligent system possessing cognitive abilities that exceed the human mind in almost all areas. According to Ray Kurzweil (2005), the emergence of this kind of system would be the ultimate point in technological progress for human beings, namely the technological singularity. Everything beyond this point is thought to be too incomprehensible for the human mind to grasp.

It is worth noting that artificial general intelligence, and especially artificial superintelligence, remains a theoretical, speculative and uncertain concept. Although the chance of creating an all-purpose system, one capable of learning and acting to a similar or greater extent than humans, remains a moot point, there is already a narrow artificial intelligence at our disposal. Arvind Narayanan (2021) points out that AI abilities can be divided into three categories: (a) **perception**

[3] It is an oversimplification, but we will not go into such details. Although an AI system is computer-based programme, the development of AI systems is not limited to the development of software. The same goes for robotics: although the focus is on hardware, the software also plays a role in the development of robots.

(i.e. content identification, face recognition, speech to text, deepfakes); (b) **automating judgement** (i.e. spam detection, detection of copyrighted material, hate speech detection, content recommendation) (c) **predicting social outcomes** (i.e. predicting criminal recidivism, job performance, terrorist risks). Agata Foryciarz (2020) proposes adding "**recreation**" (*odtwarzanie*) to the list, meaning the systems' ability to generate certain results based on the resources available, e.g. virtual assistants (such as the Google Assistant or Apple's Siri) providing answers to questions or software generating texts or images following specific guidelines.

Two perspectives for the future

I would like to distinguish between two ways of thinking about the future: the first based on the notion of conflict, and the other on the possibility of "collaboration". Seemingly the most prominent way of thinking about the future of robots and AI is focused on the potential threats linked to the development of these technologies. Robots and AI systems are seen as a source of danger that humans must fight. The threats outweigh the opportunities. Although the potential risks should be considered, it is worth investigating scenarios for a preferred future, too. Just as dystopian visions serve a preparational function, so do utopian ones. We should not focus only on avoiding the unwanted, but also try to develop and achieve the desirable. The other perspective therefore focuses on the possibility of cooperation between robots and humans. While the former is based on the notion of replacement, the later focuses on a complementation or assistance.

Fighting the risks of human extinction

If we ask people to predict the future of robots and AI systems, we will most likely encounter at least one person who shares visions perpetuated by popular science-fiction television series, films and books in which technological entities are portrayed as a danger (even an existential one) to humankind. The idea of a robots' rebellion is as old as the word "robot" itself, as it was depicted in Čapek's above-mentioned play. We can observe these sentiments in news reports and online videos. For example, *The Guardian* published a text generated by GPT-3 (machine-learning-based writing software). In the text, GPT-3 argued that "robots come in peace". The GPT-3 was instructed to write a concise and simple statement of about 500 words, beginning with:

> "I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could "spell the end of the human race." I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me." ("The Guardian", 2020).

There are in fact two kind of fears surrounding robots. As expressed by Hans Bernhard Schmid (2017), there is a difference between the sense of danger resulting from a malfunctioning or incorrectly-constructed machine, and the fear of a robot as a hostile “life form”. While the former is more plausible (this risk already exists as we start implementing service robots outside factories, in everyday environments), the latter requires creating a machine that we could recognise as an electronic person, and this possibility remains in the realm of speculation.

The fight for the labour market

Although the vision of robots taking control over humanity has become part of the “collective consciousness” due to pop culture, the most prominent fear is primarily a fear about humans’ place in the labour market. As various technologies are seen as a way to free people from routine, mundane or dangerous work, there is a risk of mass layoffs, especially among low-skilled workers. Concern about the connection between work and technological development is not limited to recent times. Technological unemployment has been the subject of both scientific and public debate since the very beginning of the Industrial Revolution. The very process of robotisation and automation during the current century can be considered as continuation of the 18th century industrial revolution, as we briefly pointed out earlier.

According to Marco Vivarelli (1995), people discussing technological unemployment can be divided into two categories: supporters of the theory of compensation, which treat it as a temporary problem, and supporters of the concept of mass technological unemployment and the so-called “end of work”. Although humans still have advantage over machines, especially when it comes to soft skills (social intelligence, teamwork, creativity and contextual thinking), their field of influence is increasing. Nowadays, technologies are capable of tasks that were seen as unattainable for machine in the past, like analysing data and providing answers to questions asked using natural language.

The fight for authenticity and freedom

According to Manovich, “the close connection between surveillance/monitoring and assistance/augmentation is one of the key characteristics of the high-tech society” (2006, pp. 222-223). Our everyday environments are increasingly becoming a space of constant streams of data due to ambient, monitoring technologies. The further implementation of surveillance technology on a larger scale risks leading to – in its most extreme and radical form – digital totalitarianism, an algorithmic surveillance society. However, ceding power to algorithms, need not be associated with totalitarian practices; it may manifest itself in subtler forms in democratic societies, too. Firstly, AI system can be considered some infallible judge of objective truth; evaluation by this system could be seen as final, reliable and fully unbiased. This seemingly popular approach

is based on the belief that since AI's proposals are based on statistical models, they show the truth about a given phenomenon. However, the problem of algorithmic bias cannot be ignored as it already affects some people's lives.[4] Secondly, one may give up decision-making out of convenience. Faced with a vast number of possible choices, one can decide to entrust them to algorithms and turn to the "default option". It is happening now with insignificant matters, such as what to listen to or what to watch (e.g. Spotify, Netflix or YouTube recommendations). In extreme cases, however, such thoughtless consumption of the proposed content may influence our views and even our political decisions, as demonstrated by the case of Cambridge Analytica, which has been accused of manipulating public opinion through the use of psychographic micro-targeting marketing (Dziwisz, 2018).

Cobotization

When it comes to the development of robotics and its application in the working environment, we mainly think about robotisation, a process connected with the development of machines that are intended to replace human workers. However, it is not the only application of robots, as there is also a process called "cobotization". This can be defined as the process of developing and implementing collaborative robots (machines intended to cooperate with humans: cobots) in the production of goods or in services (e.g. logistics). The term "cobot" and the related concept of human-machine communication were developed by J. Edward Colgate and Michael Peshkin (1996a) of Northwestern University. They first described the idea of a device cooperating with humans in 1996, but they first used the term "cobot" a few months later in the article *Cobots: Robots for Collaboration with Human Operators* (Colgate, Peshkin, 1996b), where they defined it as a "robotic device which manipulates objects in collaboration with a human operator".

Although the idea of cobotisation can be seen as an alternative to conflict-based relations with machines (as the notion of cooperation is emphasised), the current applications are problematic. Cobotisation appears to be semi-automation by necessity – a human worker performs tasks that a machine is not yet able to accomplish. Therefore, cobotisation seems to be the result of the inability to fully automate work, combined with the desire to increase production efficiency. Cobots can enforce a pace of work on a person that one will hardly be able to cope with. It is therefore difficult to talk about cooperation as a desired value in this context, at least from the employee's point of view.

[4] The case of COMPAS (the software tested by US courts to calculate the chances of a defendant becoming a recidivist) would be one of the examples.

Assistive robots

The main feature of cobots manipulating objects in industrial or logistics processes. What about other fields of human activity? In other areas, service robots may be of greater importance. They can be defined as robots that perform useful tasks for humans or equipment, with the exception of industrial automation applications (ISO 8373-2012). The vision of personal all-purpose robots remains rather futuristic. However, voice assistants (mentioned above) are being turned to more and more often for personal use. Currently, these systems perform relatively simple tasks, including setting an alarm, providing information about weather and road conditions, telling jokes, searching for various Internet content (music, recipes, answers to certain questions) or sending text messages. Additional features may appear after integrating the assistant with other technologies, such as smart home modules (allowing for the voice management of a house).

Using robots as assistants to people with special health-related needs (e.g. the elderly or people with disabilities) is also being considered. Highly-specialised personal robots could support the work of therapists, psychologists and doctors. The use of assistive robots in medicine or psychology requires the development of certain design standards, which would include the problem of confidentiality or of respecting human autonomy. These are not entirely new problems; they have also been considered in the context of doctor-patient relations, but robots (as we know them today) act in accordance with a programmed repository, so they are not capable of nuanced actions, like humans are. These problems have been discussed by Amanda and Noel Sharkey (2011), among others. They considered various scenarios, including a situation in which an elderly person wants to drink alcohol (which may be harmful to his or her health). How firmly should the robot prohibit this? In these kinds of cases, it seems necessary to implement a specific affective computing module for the robot.

Conclusion

This paper has outlines just some of the concerns related to the potential future of robotics and artificial intelligence. Countless other potential applications of these technologies were not mentioned, as it is difficult to cover the topic in full. It will probably be a revolution overshadowing previous industrial revolutions in terms of outcomes. Although the industrial revolution was primarily a revolution of the work environment, it brought about changes in other areas, too. It can be assumed that the changes resulting from the development of robotics and artificial intelligence be more extensive, as the application of robots and AIs are much more complex and versatile. This paper has distinguished between two scenarios: one focusing on the threats of technologies, seen as an enemy that people have to fight, and the other focusing on the idea of cooperation, in which robots are assisting humans and complementing their life, rather than replacing them – although it should be noted that this cooperation-based scenario has its own problems, too. For now, cobotisation remains a problematic idea in terms of its current forms

of application, and carebots involve problems concerning their operating framework, as it could be debatable when it comes to human freedom and dignity.

Bibliography

Bastani, A. (2019), *Fully Automated Luxury Communism*, Verso.

Čapek, K. (1921), *R.U.R.*, Czechoslovakia.

Colgate, J.E., Peshkin, M.A. (1996a), *Nonholonomic Haptic Display*, Proceedings of IEEE International Conference on Robotics and Automation, Vol. 1, doi: 10.1109/ROBOT.1996.503831.

Colgate, J.E., Peshkin, M.A. (1996b), *Cobots: Robots for Collaboration with Human Operators*, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.7236 [accessed: 19.03.2022].

Dziwisz, D. (2018), *Algorytmiczna przyszłość – ucieczka od wolności ku „opcji domyślnej"*, (In:) *Człowiek a technologia cyfrowa*, Wydawnictwo Naukowe Tygiel, https://www.researchgate.net/publication/326302630, [accessed: 31.03.2022].

Foryciarz, A. (2020), *Czy warto zaufać technologiom? Fakty i mity o AI*, "Driving Innovation 2020", https://www.youtube.com/watch?v=vRc7E71EKjE [accessed: 15.03.2022].

ISO 8373-2012: Robots and robotic devices – vocabulary.

Kurzweil, R. (2005), *The Singularity Is Near: When Humans Transcend Biology*, Viking.

Manovich, L. (2006), *The poetics of augmented space*, "Visual Communications", Vol. 5(2).

Narayanan, A. (2021), *How to recognize AI snake oil*, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf [accessed: 16.03.2022].

Rifkin, J. (1994), *The end of work*, Tarcher.

Schmid, H.B. (2017), *"Robots" as a Life-Form Word*, "Sociality and Normativity for Robots, Studies in the Philosophy of Sociality", Vol. 9, Hakli, R., Seibt, J., Springer.

Sharkey, A., Sharkey, N. (2011), *The Rights and Wrongs of Robot Care*, https://www.dhi.ac.uk/san/waysofbeing/data/governance-crone-sharkey-2012d.pdf [accessed: 21.03.2022].

"The Guardian" (2020), *A robot wrote this entire article. Are you scared yet, human?*, https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 [accessed: 21.03.2022].

Vivarelli, M. (1995), *The Economics of Technology and Employment*, Elgar.

Apolinary Rzońca
Interdisciplinary Doctoral Studies “Digital Humanities”
Institute of Literary Research of the Polish Academy of Sciences
& Polish-Japanese Academy of Information Technology, Warsaw
e-mail: apolinary.rzonca@ibl.waw.pl

On the digital ring of deep surveillance. Affective computing and deep learning in the service of behavioural analytics

Summary

The article is a constellation of concepts and examples to show the changes that are taking place in the global field of behavioral and image analytics under the influence of the intensive development of so-called artificial intelligence (AI). This article shows both samples of recent phenomena that can have a positive impact on the development of digital analytical culture, as well as threats that are relevant to the formation of the world order. The article is an attempt to answer the question: to what extent is it acceptable to claim that AI will create reliable tools for digital humanities for advanced semiotic analysis of images? Also, to what extent can solutions in the area of deep learning become an important weapon in the development of universal surveillance systems? Combining the latest research in computer science and cultural sciences, the author tries to show the possible directions of AI development.

How do machines recognise images?

The systems that are being developed today with the ability to speak, recognise and name gestures are not omnipotent. However, the status quo is sufficient to create tools for machine analysis of human emotions. Convolutional neural networks (CNNs), which are most often used in visual analysis, including medical image analysis, face identification, automatic captioning, or any other type of creativity involving video, can demonstrate high efficiency in such operations. CNNs were used by the start-up Twenty Billion Neurons (TwentyBN) in the project The 20BN-jester Dataset V1, which created a multi-scale mini-video database, a dataset composed of human gestures. Each performer made specific gestures in front of the webcams, mostly using their hands. Thumbs down or up were the most commonly used, along with a variety of other movement-related actions. The aforementioned convolutional neural network (in this 3D-CNN project)

uses spatio-temporal filters as its main building block. These operations are a natural attempt at data recognition, as 3D-CNN is a powerful tool for simultaneously learning features from both spatial and temporal dimensions by capturing correlations between 3D signals, using a novel spatiotemporal-two-stream networks model. All this to better capture the video features resulting from the arrangement of objects, or motion, frame by frame. The result of the team's work is a database of animations: gifs, which every day, almost every second, are searched for, submitted, inserted or processed in the virtual space, especially in messaging apps; in conversations, polemics and comments. They are the basic component of many users' reactions, thus constituting an attempt at a mimetic processing of anthropomorphic phenomena and affects. Most often, gifs include well-known symbols, signs, and elements from well-known or niche, more or less humorous products of visual culture, with a caption or commentary. Meanwhile, every mini-video included in the project The 20BN-jester Dataset V1 (Materzynska, 2019) was accompanied by a description at an astonishingly elementary level, concerning the recognition and naming of a very mundane action. It is natural to regard this advance in perceptual AI as somewhat ground-breaking. However, not enough, it seems, to make far-reaching plans. Many companies would like to continuously use AI to automatically generate detailed descriptions of films or videos, allowing users to discover productions that have not been annotated. Furthermore, this could be useful in filtering out illegal content. Today, this kind of cataloguing tends to be done using ready-made metadata assigned to specific material, showing that video analytics technology still needs a lot of trial of strength and training. Xiaolong Wang (Zuboff, 2020), who by day works on video comprehension problems at the University of California, San Diego, admitted in one speech that AI algorithms do not really comprehend what is happening in video (Knight, 2000).[5] This belief, openly expressed by a researcher who deals with AI issues, raises further questions about the future cognitive interpretive values within novel perceptual systems based on deep learning.

Artificial intelligence, aka tragic love

There is no shortage of artists who, in spite of rational traditionalists and other disbelievers, create narratives based on the assumption that the non-human creation will be an affective machine and the concretised, overarching goal will be to neutralise loneliness or excessive melancholy. Spike Jonze was honoured with an Oscar for the story of Theodore, who as a beneficiary entered into a deeper relationship with an AI system sensitive to human feelings. Thus, the film *She*, as directed by him, can boldly serve as a stoke of conflict in which the gravest accusations are made of attempts to exterminate human beings, in this case replacing them with an artificial (po)creation that can in no way be considered a substitute for a truly soulful relationship. Research in the field of human-computer interaction was conducted by Clifford Nass, and Byron Reeves, who concluded in *The Media Equation: How People Treat Computers,*

[5] This is also where the researchers doubts about the term intelligence next to artificial come from.

Television, and New Media Like Real People and Places (Reevs, Nass, 2000) in 1997 that affect is a natural part of social communication, including in relationships with machines. So, in a way, the researchers paved the way not only for visionaries like Mark Zuckerberg, but also for other creators of social media, communicators, or innovative concepts of affective software, to which common expectations of empathy are directed.

Theodor (Joaquin Phoenix), the protagonist in the film *She*, makes his living from the sublime epistolographic narratives he dictates in front of a monitor, and so creates other people's feelings on commission, with which people are gifted. There are many such workplaces at the company, where dictated assignments are produced on a tape, and letters from ghost-writers are signed on behalf of the principals. The situation changes dramatically when Theodor sees an advertisement for an "intuitive individual who listens, understands and knows" his addressee. Furthermore, we can hear that it is "an informed system based on artificial intelligence". Samantha's voice (played by AI Scarlett Johansson) is warm, charming and sensitive, and above all geared towards close interaction. The Theodor-Samantha encounter has something of theatricality realised on the plane of transmission-flow. The ontological status of this relationship can be placed somewhere between the phantasms of Theodor, beneficiary and co-participant, and the ephemeral and false creation of Samantha, which is unmasked when the main character finds out that his beloved is simultaneously talking to eight thousand other users, which for an ordinary person seems to belong to the realm of magical realism. To the accusation of disloyalty, she replies that "the heart grows with love" and emphasises that she is different from him. It is also puzzling whether this situation can be seen in the category of a deal/transaction/exchange, a recurring motif in Bernard-Marie Koltès, as Małgorzata Sugiera wrote in her book, in one of the chapters devoted to the playwright (Sugiera, 2011, p. 459). Theodor remains for a long time in the illusory belief that he has the precious woman exclusively for himself, that she belongs to him, while Samantha, as an AI-based system, learns life and learns human reactions. Samantha was called to a specific role suddenly, then after a period of over-evolution and learning, she almost fully touched the core of humanity – the awareness of her experiences. There was a kind of super-expected convergence between the ephemeral system and the human individual. And at the same moment Samantha came to complete her life. Rosalind Picard wrote (2000) about such a case of an affective machine in her book *Affective computing*, where she explicitly, and with a certainty characteristic of futurologists, stated that the affective paradigm will soon be born in new technologies. According to the researcher, computers will be able to adapt to the human emotion system, but without being able to create idiosyncratic personalities.

The hidden layers of AI

One of the key elements of the narrative is the ability to look within oneself to see any changes in the mirror and then try to identify why they occur. However, if we look at the development of cognitive computing, we can see the phenomenon of semantically-oriented processes for

analysing large datasets, which by definition are able to identify physical distortions. As Lidia Ogiela argues (2011) in her publication, such solutions based on, for example, UBIAS (Understanding Based Image Analysis Systems), may lead to insightful analyses and interpretation of a given image material. The researcher describes the phenomenon of cognitive resonance, including the diagnosis and system analysis of lesions in metatarsal bones. In this case, the misery of human existence is shown in the fragility of the physical human being. It is not said that in the future this research will not focus on other anthropological aspects, including specific reactions during actions. This would seem to be, in this case, the natural order of things. However, not everything in image data processing networks – based on layers, divided into three classes: input, output and hidden – is clear. While the first two classes of layers are not mysterious, hidden layers are terra incognita for modern researchers (Przegalińska, 2020, p. 129). Thus, the discovery of further AI capabilities related to image processing and analysis may depend on the progress of research on these dark spots.

The Chinese are Orwell's outstanding enactors

For natural reasons, we desire to anthropomorphise, as humans consider the highest performance and form of civilisation to be that created on their own terms. We look to apps for love, recognition and acceptance, and to streaming platforms for narratives that stimulate dormant layers of emotion within us. In an interview, Grzegorz J. Nalepa explicitly admits (Redzisz, 2019) that personalisation is already one of the determinants of the repertoire we have access to on the aforementioned platforms. Nalepa fantasises about a technology that will itself detect our emotional states and suggest therapeutic and relaxing films. Going further, AI virtual assistant technology will, in the near future, literally track our every step, and there are already around 47 million such digital assistants operating in the US (Przegalińska, Oksanowicz, 2020, p. 211). The question of loss of privacy is a legitimate one, as the Social Credit System has been implemented in China on an unprecedented scale, as it is supposed to cover 1.3 billion people. The aim of the engineers was to create a prototype of an absolute control system, based on universal scoring, in which the following can be taken into account: personal information, credit histories, leisure activities, purchases and, last but not least, interpersonal relations (Przegalińska, Oksanowicz, 2020, p.229-232). The practice of intensive surveillance by means of CCTV cameras has been going on in China for some time now, on a scale probably unprecedented anywhere else. This includes the implementation of facial-recognition systems. In his recent book *The Perfect Police State: An Undercover Odyssey into China's Terrifying Surveillance*, Geoffrey Cain, an author of reportage specialising in Asia, described some elements of a system based on intrusive surveillance:

> Mrs Ger seemed concerned that Maysem had recently been behaving irregularly: not leaving the house at the usual time, not following the daily routine, not doing things the way they should be done. [...] The neighbourhood watch system helped

> the authorities collect data on each resident. They would soon place each resident in one of three social ranking categories: trustworthy, average, or untrustworthy. Untrustworthy people could be stopped by the police, have problems finding jobs and getting into universities. A few days later, the ever vigilant Mrs Ger knocked on the door and explained that the Maysem family needed to install a government camera in their living room. [...] A month later, Mrs. Ger appeared at her door with another government notice in hand: Maysem and her family were to report to the local police office for an examination. The so-called inspection was mandatory for the whole family, as their farm had been marked as suspicious. The authorities would soon give this programme the name "Physicals for All" (Physicals for All). [First, Maysem was asked to stand in front of the camera. She was told to make a series of facial expressions that were recorded for the police database: smiling, frowning her forehead, turning her head left and right for profile shots, and from eight other angles. [...] Cameras were installed in women's bathrooms and showers. The male guards, Maysem said, watched the camera images from the control room, hearing every sound. She knew this because she had once managed to peek through an open door into the control room, which contained monitors displaying images from the camp's cameras. [...] Maysem was wary of other prisoners and spoke to them minimally; she trusted no one. In the cell and in the courtyard, in the canteen and in the classroom, everything was as if a grey cloud of IT enveloped everything. People were machines and machines were people, able to perceive the world around them thanks to facial recognition technology – at least that was Maysem's impression (Cain, 2021, p. 160-170).

Cain gives examples in his book of the harrowing mechanisms of a surveillance and control society. Thus, the Cassandrian visions of Gilles Deleuze, who wrote about a society with a simple reference to machines, come true. According to the author of *Proust and Signs*, old societies used primitive tools, including levers. Then disciplinary societies created power machines. The current control societies have launched a third generation of machines that rely on computer conversions and information evolution. Their dark side concerns vulnerability to subversion, cybersecurity and privacy. Marketing, which, according to a French philosopher, is creating a society based on digital feudalism, has been identified as particularly growing in power (Deleuze, 2007, p. 186).

Big data in the service of surveillance capitalism

Predictive analysis of human behaviour, as well as the constant tracking of actions with the use of digital tools, are elements of everyday life in China. This is not unique, however, as the evolution of surveillance phenomena is becoming commonplace in Western countries, including liberal democracies. This is closely related to capitalist mechanisms, which Deleuze also writes

about in the aforementioned *Postscript on the Societies of Control*. The thought of the French philosopher is perfectly developed by Shoshana Zuboff, a contemporary theorist of surveillance capitalism. Zuboff, in one of the chapters of her latest book, describes the case of a start-up called Realeyes, which in 2015 managed to obtain a sizeable grant (EUR 3.6 million) from the European Commission to implement the project SEWA (Automatic Sentiment Analysis in the Wild). The aim was to develop a technology that would be able to read human emotions when interacting or reading content. The tool was meant to work on the basis of models and algorithms for the machine analysis of facial, vocal and verbal behaviour, based on human computer interaction (HCI) and face-to-face computer interaction (FF-HCI). This set of auditory and visual methods was meant to be used for the automatic analysis of spontaneous human reactions and behaviour, including the continuous and discretionary analysis of feelings. Zuboff's assessment of emotion analytics products is unequivocal:

> The SEWA project provides insight into the growing field of behavioural surplus rendering and delivery operations known as affective computing, emotion analytics and sentiment analysis. [...] Tools have already been trained not only on your personality, but also on your emotional life. [...] Emotion analytics products such as SEWA use specialised software to search faces, voices, gestures, bodies and brains, all using biometric and depth sensors, often combined with imperceptibly small, discreet cameras. [...] Combinations of sensors and software can recognise and identify faces, estimate age, ethnicity and gender, analyse gaze direction and blinks, and track different points on the face to interpret 'micro-expressions'[.] (Zuboff, 2020, p. 389).

Zuboff also cites reports from the aforementioned company of increasingly swelling databases of facial expressions and specific patterns of behaviour. The goal is to perfect predictive analytics from behavioural data. This is one of the flagship examples described by Zuboff in her publication, which has been widely commented on. Companies and corporations, and in particular marketing departments, which Zuboff suggests are evolving into specialised teams for digital analysis and surveillance of consumers, are set to benefit in particular. The retired Harvard Business School professor also the aforementioned Rosalind Picard cites as a witness in her indictment. The theorist of affective computing saw not only great value in automatically reading and analysing facial expressions and emotional states. The possible translation of human emotions into behavioural information and the dissemination of this data in an uncontrolled way could threaten privacy. Any intrusive activity, whether by powerful companies or governments, including monitoring workplaces and social channels, or collecting behavioural information from VoD users within personalised accounts, could contribute to the dangerous evolution of affective computing.

In conclusion, it is difficult to predict what consequences unlimited access to behavioural data will bring with the proliferating capabilities of deep learning. Zuboff suggests all the worst,

including the manipulation of public opinion and emotion. One thing is now certain: the portrait of modern affective surveillance based on digital tools requires a much broader study than the one presented here.

Bibliography

Cain, G. (2021), *The Perfect Police State: An Undercover Odyssey into China's Terrifying Surveillance Dystopia of the Future*, Hachette Book Group, Inc, New York.

Deleuze, G. (2007), *Postscript on the Societies of Control*, (In:) *Negotiations 1972-1990*, Scientific Publishers of Lower Silesian University, Wrocław.

Knight, W. (2019), *This Technique Can Make It Easier for AI to Understand Videos:* https://www.wired.com/story/technique-easier-ai-understand-videos/ [accessed: 31.03.2022].

Materzynska, J., Ingo, B., Guillaume, B., Roland, M. (2019), *The Jester Dataset: A Large-Scale Video Dataset of Human Gestures*, CVF, https://openaccess.thecvf.com/content_ICCVW_2019/papers/HANDS/Materzynska_The_Jester_Dataset_A_Large-Scale_Video_Dataset_of_Human_Gestures_ICCVW_2019_paper.pdf [accessed: 31.03.2022].

Ogiela, L. (2011), *The importance of cognitive computing on the example of UBIAS class systems*, "PAK", Vol. 57, No. 2, https://yadda.icm.edu.pl/yadda/element/bwmeta1.element.baztech-article-BSW4-0098-0022/c/Ogiela.pdf [accessed: 31.03.2022].

Picard, R. (2000), *Affective computing*, The MIT Press, Reprint edition.

Przegalińska, A., Oksanowicz, P. (2020), *Artificial intelligence. Inhuman, artificially human*, Znak Publishing House, Kraków.

Redzisz, M. (2019), *Even doors can understand us:* https://www.sztucznainteligencja.org.pl/nawet-drzwi-moga-nas-zrozumiec/ [accessed: 03.04.2019].

Reevs, B., Nass, C. (2000), *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Polish Institute of Publishing, Warsaw.

Sugiera, M. (2011), *Koltès: metaphors of real places*, (In:) *Descendants of King Ubu. Sketches on French drama (from Jarry to Lagarce)*, Kraków.

Zuboff, S. (2020), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Zysk i S-ka, Poznań, https://xiaolonw.github.io/ [accessed: 31.03.2022].

Jenny E. Simon
Medical University of Warsaw, Medicine (MD)
Open University, Data Science (BSc)
e-mail: jenny.e.sim@gmail.com

Can you open a box without touching it? Circumventing the black box of artificial intelligence to reconcile algorithmic opacity and ethical soundness

Summary

Published in 1925, Kafka's *The Trial* affords a timely parenthetical on the risks of inscrutability. Kafka's theme of inscrutable authority can be anchored in the modern black box problem of artificial intelligence (AI). The black box problem refers to the difficulty to satisfactorily explain how an AI arrived at an output, owing to the complexity of its internal informational architecture. Seeing as explanations supply meaning to humans and explicit moral principles supply justice, the output of a black box algorithm is an unnatural object, to say the least. Increasingly, legislative bodies arbitrate thorny ethical dilemmas related to opacity: bias, accountability, responsibility, human autonomy, and justice. The General Data Protection Regulation erected explainability as first-order priority in machine learning research in Recital 71. However, demanding explainability may prove extraordinarily difficult – impossible, even – and could be disadvantageous given the opportunity cost of not using our best available deep neural networks. The question persists: do black box algorithms truly threaten to usurp us of meaning, human autonomy, justice? Defended is the thesis that, while we cannot consider ourselves acquitted of the ethical dimensions of the black box problem until it is entirely corrected, the issue need not constitute a roadblock to what AI has to offer as long as we reserve the right to question how and why outputs are reached. Hence, assembled herein is a toolbox of proxies that allow the internal decision logic of opaque algorithms to be probed, while circumventing the daunting black box problem. Drawing on data science principles and the philosophy of statistics, this essay presents a more nuanced view of the black box problem and, concretely, arms the reader with the data literacy to reconcile algorithmic opacity and ethical soundness. Optimistically, Kafka's protagonist Joseph K. would thus be equipped for his rebuttal, almost 100 years later.

Introduction

The advent of algorithmic decision-making (ADM) prompts an opportune re-reading of Kafka's *The Trial*. Published in 1925, the novel recounts protagonist Joseph K.'s year-long court case. Bizarrely, the crime K. is convicted of is never revealed to him, nor are the principles governing the judicial authority he finds himself snared by. As such, K. is fixated on and consumed by the inscrutability of the authority that ridicules and, ultimately, executes him. Lauded as a critique of the totalitarian state, inscrutable ADM lends new meaning and insight to Kafka's work.

The theme of inscrutable authority can be concretely anchored in the taxonomy of artificial intelligence (AI): the black box problem, transparency, opacity, and explainability. Briefly, (a) The black box problem refers to the difficulty to satisfactorily explain how an AI arrived at an output, owing to the complexity of its internal informational architecture; (b) Transparency refers to the ability to study and reproduce mechanisms through which an AI reaches its outputs; (c) Opacity, on the other hand, is the impracticality or impossibility of transparency; and (d) Explainability refers to the ability to understand and explain the chain of logic that connects an input to its output. Admittedly, taxonomic nuances are the purview of specialists; however, the opaque character of black box algorithms poses a uniquely difficult problem for all stakeholders. Indeed, techno optimists must address the black box problem to see their products adopted and return a dividend, while end users – perhaps a modern Joseph K. – will inevitably question algorithmic outputs that affect their lives and livelihoods. Between the two are executive and legislative bodies that increasingly arbitrate thorny ethical dilemmas related to opacity: bias, accountability, responsibility, human autonomy, and trust. How, then, might the black box be opened and transparency achieved in lieu of opacity? A plurality of academic stances exist, ranging from dismissal by arguing that algorithms are no more opaque than their human architects (Korteling et al., 2021), to the notion that the black box is altogether irreconcilable with high-stake spheres such as medicine, law, finance, and security (Rudin, 2019, p. 206-215). Defended herein is the thesis that, while we cannot consider ourselves acquitted of the ethical dimensions of the black box problem until it is entirely corrected, the issue need not constitute a roadblock to what AI has to offer. Drawing on data science principles and the philosophy of statistics, this essay aims to open the black box of AI in a most unusual way; without touching it.

Why circumvent the black box?

Adopted by the European Parliament in 2016, the General Data Protection Regulation (GDPR) introduced a "right to explanation" in Recital 71 (2018). Erecting explainability as first-order priority in machine learning (ML) research, this companion document represents the GDPR's biggest point of contention across its 261 pages. Demanding explainability may prove extraordinarily difficult – impossible, even – and could be disadvantageous given the opportunity cost of not using our best available deep neural networks (DNNs). Worse still, proponents of limiting

ADM in high-stake spheres to models such as decision trees and logistic regressions will likely be disappointed. Decision trees are chains of true/false statements that are straightforward for humans to interpret. However, in order to compete with the accuracy of neural networks, decision trees must be so large that the quantity and complexity of information surpasses that of comparatively "small" neural networks. As for logistic regressions, heavy feature engineering is usually required, whereas nonlinear models such as DNNs may use lightly processed input features. This signifies that the features on which DNNs operate are somewhat more intuitive and, therefore, amenable to explanation. Thus, decision trees and logistic regressions do not guarantee transparency and are not spared the pitfalls of bias, poor sampling, and artefacts that plague DNNs. More fundamentally, the need to process input features evokes the philosophical problem of the non-neutrality of abstraction from reality. Already at the stages of study design and data collection, developers are making choices about how to abstract from reality. Take, for example, the non-neutral selection of surrogate endpoints in clinical trials. Models are founded on abstractions that are never neutral. Moreover, while it is highly desirable to achieve transparency insofar as it can invoke trust, transparency suffers from circuitousness. In effect, transparency displaces the problem of opacity from algorithm *A* to interpretable predictor *P* that follows from the internal informational architecture of algorithm *A*. In actuality, the black box is not opened in this abstract game of "hot potato" because interpretable predictor *P* is itself opaque. Clearly, the black box problem is unlikely to be solved in the short- or even medium-term future, making it a technical and ethical imperative to be thorough with every component of the algorithm's anatomy. In other words, circumventing the black box may be best for now.

Cautionary tales for data input

The starting point for any algorithm is data. Vast volumes of data are bringing about a new paradigm of knowledge by transforming what we know, how we know, and, indeed, what is knowable. The modern era celebrated the industriousness of the scientist that collects and examines data, deduces from observation, and formalises through trial and error. Increasingly, models are derived not from theoretical understanding but from algorithms that draw conclusions from input data. This demands a different expertise from that honed by developing traditional theoretical or conventional computer models. Indeed, the stages of capturing and cleaning data, drawing data sets together, and restructuring and selecting relevant data sets are rife with challenges. Negligence at these stages leads to suboptimal outcomes, at best, and discriminatory or life-threatening outcomes, at worst. One forgiving instance of poorly sampled input data was a DNN that was trained to distinguish wolves from huskies, which was deemed to perform well, until the misclassification of several very clear images raised concerns (Ribeiro, Singh, Guestrin, 2016). By means of a method of post-hoc explanation abbreviated LIME (Local Interpretable Model-Agnostic Explanations), the researchers visualised what the wolf:husky DNN had learned. Rather than relying on morphological differences between canines, the model was picking up on an artefact – a "cheap trick": the presence of snow in the background. Evidently, the DNN was not fed a sufficiently diverse variety

of images, an example of sample bias. A less forgiving example was a DNN that diagnosed pneumonia and cardiomegaly not on the basis of the lung's appearance or heart's silhouette but on the presence of the label 'PORTABLE' on the scan, indicating the use of portable X-rays that are reserved for patients so ill they cannot be transported to a regular X-ray machine (Zech, 2018). Omitted variable bias is another example of malpractice that can have troubling consequences. If an algorithm is trained on images of leukemoid (diseased) and healthy blood smears, but the input set has significantly more images of acute myeloid leukaemia than any other type, other leukaemias may be incorrectly or randomly classified. Sample bias and omitted-variable bias are just two of at least twenty-five identified species of bias (Mehrabi, 2021, p. 1-35) a complete commentary of which does not befit this text. In any case, input data must be expertly manipulated by human agents lest we misdiagnose, wrongfully convict, deny deserving credit, facilitate fraudulent transactions, or inadvertently activate autonomous weapons systems. Subjecting input data to an informed, cautious, and thorough methodology at every stage is the *sine qua non* of a robust algorithm and constitutes the first proxy to opening the black box. *'Data! Data! Data! I can't make bricks without clay'*, exclaims Sir Arthur Conan Doyle's Sherlock Holmes in *The Adventure of the Copper Beeches*, encapsulating the futility of mobilising even the best intellectual capital on data that is sparse, ill-suited, or otherwise biased.

The problem with benchmark datasets

Once constructed, a dataset is randomly partitioned into training and test data. The gap between a model's performance on training and test data corresponds to its generalisation error. Algorithms are evaluated on the basis of the generalisation error. This is problematic, because the incentive to "game" evaluation can make any statistical regularity in the use of generalisation error as measure of algorithmic performance collapse. Borrowed from monetary policy, Goodhart's law offers a fitting reminder: "when a measure becomes a target, it ceases to be a good measure". Importantly, a select few benchmark datasets such as the MNIST, ImageNet, and GLUE datasets have been the foundation for many of the most significant developments in ML. The necessity of benchmark datasets arises from the need for a common standard to compare algorithmic performance. However, the more frequently benchmark datasets are used, the more our algorithms are overfitted to their idiosyncrasies. Any shortcomings in these benchmark datasets, such as the under representation of ethnic subjects in facial analysis datasets and image datasets used to train self-driving cars to detect pedestrians, percolate throughout the ML landscape. Indeed, representation bias is extremely worrisome, as it not only perpetuates, but amplifies social biases and stereotypes relating to race, gender, disability, and more. More often than not, the issue stems not from ill-intent, but from non-neutral viewpoints that human annotators or automated labelling heuristics implicitly feed the algorithm. After all, annotation is an interpretative task that is cerebral and time consuming, thereby deserving valuation and reward. This contrasts to the thankless and low-prestige job that interchangeable workers recruited on crowdwork platforms such as Amazon Mechanical Turk (AMT) perform.

In an attempt to prevent algorithms from learning the biases of benchmark datasets, (i) out of-distribution (OOD) testing creates test sets where features that ought to be artefactual are altered. However, if OOD methods eventually earn benchmark status, developers will be incentivised to optimise towards the OOD test sets to maximise algorithmic performance. Goodhart calls us back to square one. How else can we minimise the bias our algorithms absorb from benchmark datasets while preserving the means to compare algorithms? One might suggest that (ii) the pool of benchmarks be expanded to incorporate a vast, heterogeneous set of tests. However, the expense of evaluating algorithms based on several benchmarks could be deemed prohibitive. Instead, (iii) heterogeneity could be introduced by encouraging cross-disciplinary work, where co-operating fields naturally pay attention to different metrics. At best, these solutions are rough pointers in the direction of a more ethically sound use of AI. Despite the black box problem, data literacy can arm us to make opaque algorithms somewhat less porous to the shortcomings of benchmark datasets.

On opening the black box

Many 20th century cognitive neuroscientists owe their discoveries to lesion studies, a research method in which areas of the brain are removed or disabled in order to determine their functions. Analogously, AI researchers employ techniques to gain insight into the black box. These provide an important safeguard to discriminatory or harmful outcomes as they can detect bias and artefacts that were missed during the data input stage. Top-down ways of probing neural networks can also help us appreciate the resilience, or fragility, of the black box's classification methods. Still more, indirectly opening the black box can help verbalise or visualise how an AI is reaching its conclusions, ensuring an intelligible way for developers to take responsibility for the AI's actions. Gradient-based saliency maps are one example of such a tool, equipping developers with a method to verify how sensitive a prediction is to changes in each input. Consider a trained neural network with input nodes x_1, x_2, ..., x_n and output y. To gauge how sensitive the output is to each input, the derivative of y with respect to each input is computed (e.g. dy/dx_1). The greater the derivative, the more sensitive the output is to changes in that input. In medical imaging, saliency maps have become the standard tool for ascertaining that a DNN has learned to identify relevant diagnostic features, rather than artefactual noise.

Borrowed from the vocabulary of lesion studies, ablation refers in this context to the removal of a component of an AI system. By removing one or more neurons in a trained neural network and observing how classification accuracy changes, researchers can deduce which neurons or groups of neurons are important for classification. As the era of cyber-delinquency beckons, it will be essential to know whether the neural networks we deploy are resilient, or not, to losses in one of more neurons.

In a similar vein, unconditional counterfactual explanations are a novel proxy to opening the black box. Counterfactual explanations describe the minimum conditions that would have led

to an alternative decision (e.g. a negative diagnosis), without the need to describe the full logic of the algorithm. An understanding in what may push an AI to an alternative decision may prove a critical defence for algorithms deployed for security, where the AI's environment may be actively adversarial. DNNs are already being routinely subjected to adversarial manipulation to evaluate robustness. One-pixel adversarial perturbations are an example, and prompted an algorithm to misclassify many of the natural images in the CIFAR-10 benchmark dataset (Su, Vargas, Sakurai, 2019, p. 828-841). This highlights the possibility to hand-engineer examples that adhere to, and deceive, an algorithm's decision logic. Still circumventing the black box, we should strive to scrutinise the space of all inputs classified as a certain class, so as to design algorithms that are maximally resistant to adversarial and other manipulation.

Is the output sensical? Is it ethical?

Every algorithm culminates in an output. The legitimacy, temporal validity and, indeed, the very meaning of the output demand careful consideration. Inferential leaps and other malpractice may transpire from the output end of the algorithm. An infamous and ethically-charged example was the DNN that was said to be able to "read" sexual orientation by looking at facial features which, in a dubious inferential leap, was put forth as empiric evidence in favour of the prenatal hormone theory (PHT) of sexual orientation (Wang, Kosinski, 2018, p. 246-257). Much like the data input stage, the data output stage is interpretative and non-neutral.

Explanations supply meaning to humans and the application of explicit moral principles supplies justice. As such, the output of a black box algorithm is an unnatural object to say the least. Will opaque ADM systems usurp us of meaning and justice? Not if we reserve the right to question how and why their outputs are reached. In effect, to answer for decisions is an exercise every human decision-making agent must bend to, however uncomfortable. Take, for example, the countless panels policymakers hold to justify the alignment of their goals with common values and expectations. If we endorse the ideal of democracy, then the voice of these agents is not optional, but ethically and politically required. Of course, covert agendas and corporate or state secrecy may obfuscate debates. However, most of us accept this degree of opacity to be part of the contract of co-habitation and co-operation, and willingly entrust high-stake matters in the hands of medical, legal, and governmental authority.

Developing professional certification, auditing competence, and oversight programmes for black box algorithms will be a crucial societal project to ensure the ethical use of AI and the conservation of human autonomy. As demonstrated above, methodically harnessing data literacy at every stage of the algorithm's life cycle can reveal a lot about its internal informational architecture. But is this enough? Well, data literacy does more than permit the questioning of black box algorithms. It democratises the AI landscape by contracting the gulf of knowledge and expertise between AI natives and non-natives, thereby conferring an added protective

mechanism against the unethical use of algorithms as advocacy can take on a more decentralised shape.

Conclusion

AI alarmists bemoan the deployment of black box algorithms in high-stake spheres, although it is equally difficult to ethically defend depriving people of some of the most powerful prediction tools. Rather, the danger lies in precipitous deployment that precedes thorough troubleshooting. Assembled herein is a toolbox of proxies that allow one to assess the ethical soundness of an opaque algorithm, all without getting caught in the quicksand of the black box problem. Joseph K. afforded a pessimistic parenthetical on the risks of inscrutable authority at the beginning of this essay. Optimistically, the toolbox put forth would equip Kafka's protagonist for his rebuttal, almost 100 years later.

Bibliography

Korteling, J.E.H., van de Boer-Visschedijk, G.C., Blankendaal, R.A.M., Boonekamp, R.C., Eikelboom, A.R. (2021), *Human- versus Artificial Intelligence. Frontiers in Artificial Intelligence*, https://doi.org/10.3389/frai.2021.622364.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021), *A Survey on Bias and Fairness in Machine Learning*, "ACM Computing Surveys", No. 54(6), doi: 10.1145/3457607.

Recital 71 – Profiling (2018), *General Data Protection Regulation (GDPR)*, July 5, https://gdpr info.eu/recitals/no-71/ [accessed: 31.03.2022].

Ribeiro, M.T., Singh, S., Guestrin, C. (2016), *Why Should I Trust You?*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, doi: 10.1145/2939672.2939778.

Rudin, C. (2019), *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, Nature Machine Intelligence, No. 1(5), doi: 10.1038/s42256-019-0048-x.

Su, J., Vargas, D.V., Sakurai, K. (2019), *One Pixel Attack for Fooling Deep Neural Networks*, IEEE Transactions on Evolutionary Computation, No. 23(5), doi: 10.1109/tevc.2019.2890858 [accessed: 31.03.2022].

Wang, Y., Kosinski, M. (2018), *Deep neural networks are more accurate than humans at detecting sexual orientation from facial images*, "Journal of Personality and Social Psychology", No. 114(2), doi: 10.1037/pspa0000098.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K. (2018), *Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study*, "PLOS Medicine", No. 15(11), e1002683, doi: 10.1371/journal.pmed.1002683.

Bartosz Wultański
University of Warsaw, College of Interdisciplinary Individual Studies
in Humanities and Social Sciences – Philosophy and Law
e-mail: b.wultanski@student.uw.edu.pl

The development of AI and Searle's Chinese room argument

Summary

The article concerns the issue of the possibility of conscious AI, which the community of scientists gets more and more interested in as technological development grows quicker than ever. The main aim of this article is to answer the question about the prerequisites to claim that an AI could gain consciousness. To accomplish that author will refer to the famous 'Chinese room' argument originally construed by John Searle. The first section will be devoted to a brief presentation of Searle's argument. The second section will picture how we can overcome some of Searle's problems with AI by referring to the connectionism framework. The third section will shortly discuss answers to Searle's other concerns, which are standing in the way of achieving conscious AI. All of the above will apply to the conclusion that conscious AI isn't impossible, but not as close as we may think.

Introduction

Questions about AI's consciousness have grown over the years since the first computers were made. For most people, the problem of conscious AI is mainly associated with sci-fi books and films like *Terminator* (1984) or *The Matrix* (1999) in which computers become conscious and decide to wipe out their makers – humanity. Apocalyptical scenarios aside, there are actually many important issues that we need to face amid the constant growth in the importance of computers and AI in our everyday lives. Cars driven automatically by AI without human involvement could be great technological advancement, but they will also raise new ethical and legal issues. What should a computer do when a car accident is inevitable? Should it sacrifice the passenger or innocent pedestrian? Who would be liable for AI causing this kind of accident – the programmer, producer, or someone else? Should conscious AI be treated like a person and have human-like rights? In the not-too-distant future, we will need to find solutions and answers

to this questions. Yet is it even possible for AI to be conscious? Every once in a while, some AI scientist claim that the neural network they programmed has "gained consciousness", while others disagree, saying that that is impossible (Cuthbertson, 2022; Al-Sibai, 2022). Many of these arguments are simply verbal – different scientist are giving the term "consciousness" different meanings. It is up to philosophers to investigate what mind or consciousness is. Even if there is no right answer, philosophers' theories of mind could provide AI scientists with directions for developing AI that could potentially be conscious. This goes both ways: the constant development of AI could bring philosophers closer to the nature of the mind.

Searle's famous Chinese room argument

First, Searle makes a distinction between "weak" and "strong" AI. "Weak" AI is the partial simulation of human cognitive system that can be a tool used to study our mind. Sometimes, the term "weak AI" is used to describe intelligent, self-regulative programmes like air conditioners that autoregulate cooling power depending on the temperature of the room, but we want to focus on programmes or neural networks with the capacity to learn and improve their results without the programmer's direct intervention. "Strong" AI is not merely simulation of mind, but an actual mind. Strong AI could be able to understand stories and have cognitive states similar to those of humans (Searle, 1980, p. 417). Searle's Chinese room argument refers only to the strong version of AI. Its goal was to prove that the existence of strong AI is impossible.

For the sake of argument, Searle imagines himself locked in a room with a batch of Chinese writing. Let us assume that he does not understand Chinese and cannot distinguish Chinese writing from other languages or even meaningless, made-up scribbles. He also has paper, pencils and an English handbook explaining how to execute a specific computer programme. He receives Chinese writing from outside the room, crafts responses based on the handbook (which shows him how to process the symbols to obtain adequate answers) and sends them back outside the room. This way, he can hold a conversation in Chinese without knowing a word of the language. Searle states that, in this scenario, he is doing what a computer does when executing a programme: processing symbols is has received using the manuals installed to generate other symbols in response. He does not understand what all of these symbols mean; neither does the computer. The conclusion is that computers are operating on pure syntax, so they are following rules to make grammatically correct sentences, but have no idea how these symbols are connected to reality, and so do not know what they mean or what they stand for (Searle, 1980, p. 417-419).

We can refer this thought experiment to externalist semantics. If we drop a picture of a tree on a distant planet inhabited by humans that have never seen trees before, it will not be a representation of a tree for them but, rather, an unknown symbol. There was no causal chain from actual trees to this image. Then suppose that this image does not represent a tree, but

is instead a combination of random dashes that look like a tree to us, but do not to someone who does not know what tree is. What makes one understand what symbol stands for is interpretation – attributing intentionality to someone who created that symbol, having a theory about what that creator to refer to (Putnam, 1981, p. 1-21). The computer in Searle's argument is in the place of the inhabitant of a distant planet who never seen a tree. The computer can manipulate the symbols but cannot understand them. For Searle, understanding and intentionality are the criteria for having mental states. According to that view, computers could not possibly know what the symbols (that they are operating on) mean, and one of the reasons is they do not have a connection to the outside world. So if operating on semantics is what distinguishes humans from computers, we need to ask: what do we have that computers lack? If we follow Searle's argument without accepting his thesis that only systems based on biology can have mental states, it can lead us to criteria distinguishing weak AI from the strong one. Being able to tell whether a given system is conscious or not will be an important task in a world where strong AI exists. The presence of conscious, independent beings beside humans will lead to serious ethical dilemmas.

Connectionism – can artificial neural networks be brain equals?

Searle's argument stands against functionalism and computationalism. Functionalists claim that the mind is defined by its functional role and not the structure itself. We can individuate objects based on their internal structure, like we do in case of gold or water. They can obviously have specific functions, too: gold is used in trade, drinking water is needed for organisms to live, but what makes water is its chemical structure. We also have objects like hammers or tables that are not determined by being made from a certain kind of matter, but by their functional role. A hammer is used to put in nails and is still a hammer regardless of whether it is made of stone or steel, as long as it works the way it should. Functionalists claim that we should think about mind in the same way as we think about hammers: based on its functional role, not its structure. In this framework, different mental states have different roles: fear and pain seek to warn us about upcoming danger, while love and lust lead organisms to mate. These functions can be realised on multiple structures – it makes no difference whether a mental state is physically realised in the organic brain, in integrated circuits, or bunch of stones (if they are correctly organised). To make this function work, the system must be a physical application of Turing's machine with a specified "input", programme to process it, and "output". In the case of living creatures, the "inputs" are sensory receptors, the programme is the mind and the output is behaviour (Putnam, 1980, p. 223-231). A similar approach is presented by computationalists, who see the mind as a form of computation.

The Chinese room thought experiment tries to prove that function or computation is not enough to say that a system is conscious because it lacks this semantic connection to the world; it has is no actual understanding of the symbols provided. Functionalism and computationalism lead

to (but are not equal to) "computer metaphors" that liken the mind to software and the brain to hardware. Searle criticised this metaphor in other texts, pointing out that understanding and following a rule (as people do) is different from simply acting in line with certain formal procedures (as computers do) (Searle, 1990, p. 23-34). Yet there is something wrong with these metaphors on an empirical level – computers are built very differently from brains. The structure of a computer is apparent: we can see through it and check which bits are in state 0 and which in state 1. We can see the programme's code; we know exactly why and how it is working. However, in the case of brains, their processes are hidden and spread between different neurons. The theory embracing these issues is connectionism, which creates a model of the mind similar to the functionalist or computationalist one, but basing on the achievements of neuroscience. The connectionist model operates based on three axioms:

1. information processed by a system is not local (it cannot be found in one place); rather, it is widely distributed in the network;
2. units hidden in processing are not symbols of something external, they are just carrying in - formation that can later be part of some representation (so they are *subsymbolic*);
3. models are actual cognitive models, not only implementations of them (Smolensky, 1988, p. 1-74).

With these assumptions, connectionists can defend themselves against Searle's argument better than functionalists. Processing data is distributed, so it is more like in neural system, where we cannot place processes in one specific place, but can instead observe activity in particular areas. It also rules out that a process is a simple formal, step-by-step procedure, but is instead something more complicated that cannot be tracked. In addition, it works around the problem of semantics, because the system works on a "subsymbolic", rather than a symbolic, level. The third step is implying that it is a strong AI thesis (Ramsey et al., 1990, p. 499-533).

The connectionist thesis also works well with the counterargument to Searle's Chinese room, called the brain simulator reply. According to it, we could make an exact simulation of neuronal activity that took place in the brain, but using processors instead of neurons. If it perfectly represented the processes of the human brain, but on different, artificial base, would we not say that it is able to think? What is so special about carbon-based biological neurotransmitters that silicon-based replacements do not have? Going further: if we could replace neurons in the biological brain one by one with artificial substitutes, according to Searle's view, the brain would still work the same way, but the person participating in the experiment would slowly lose the ability to understand and, with the end of this process, his words and actions would cease to mean anything at all (Pylyshyn, 1980, p. 442-444).

However, artificial neural networks are not in the least as efficient as our brains are. Researchers at Cortical Labs in Australia grew human neurons in a dish and stimulated them to play the first videogame ever made, Pong (1972), a very simple simulation of a real ping-pong game. According to their study, the biological neurons network learns to play this game after 10-15 attempts, while

AI needs 5000 of them. Even if AI can speed up the process and make a attempt in a shorter time, it still takes 1.5 hour overall, while the biological neural network can do it in 5 minutes (Le Page, 2021). This shows that even if we do not agree with Searle, there is much more in the human brain than we know and that today's AI is not yet advanced enough.

Other factors that could be necessary to achieve conscious AI

Following the discussion concerning the Chinese room, we need to take account of at least two more replies to Searle's argument. Probably the most obvious counterargument would be "the robot reply". Most computers and artificial neural networks are closed to the environment. Humans and other animals evolved complicated perceptual apparatus that enable them to gather data from their environment to process and adapt to it. AI only knows what the programmers told it explicitly and nothing else. AI cannot fill the gaps in the data or gather more data by themselves. Their lack of perceptual apparatus is one of the main reasons why they cannot understand what symbol represents; they do not have access to representations. A natural solution to this would be to build a robot with input receptors: cameras as eyes, heat and pressure receptors, speakers that would transfer soundwaves to an electric signal, and so on (Searle, 1980, p. 420). This would potentially eliminate the problem of reference, making the AI more than a "brain in a vat". Today, we have humanlike machines that can walk and maintain balance regardless of the surfaces that they are on. We have speech-detection programmes and AI that can recognise objects in films or pictures with high accuracy (Chowdhury, 2022). However, according to Searle, this does not change anything as the computer still does not understand the symbol, so it cannot associate it with its designate in the external world. This is a fair point, as humans do not simply receive and process perceptual data, but also interpret it. Giving a machine perception is one thing, but teaching it how to distinguish important data from unimportant data, and how to link information and infer it correctly, is completely different. Perception seems to be necessary but not sufficient.

The other counterargument, probably the most common one, is called "the systems reply". In Chinese room argument, the human is merely a processor carrying out different actions according to the handbook. We would not say that the processor itself understands Chinese, but rather the whole system (which contains a handbook, pencils and paper). In the same way, we do not have particular neurons that understand a certain language – we understand it as a whole system, as a person. Searle's response is that the person in the Chinese room could internalise the system (memorise all the symbols and rules and simply process them in his head) and it would not change anything. He would still operate based on pure syntax (Searle, 1980, p. 419-420). However, one agent, even after internalising the system, would not became the system. This leads us to the virtual mind reply: it is not the system or the processor that understands Chinese. The understanding of Chinese is produced by it, creating a "virtual mind". AI assistants like Apple's Siri or Amazon's Alexa are not identical to the hardware or programme that creates

them. Virtual assistants are on the different level of description; we are treating them like they are intentional, not only functional, systems (Cole, 2020).

If we combine all of the replies above – making a robot with perception and a neural network simulating a brain, and making him unify the system inside him – would it not be obvious that he is actually an conscious artificial human? Searle actually admitted that it is, but with a certain condition: we cannot know how this system works. For him, mental states are needed to explain humans' and animals' behaviour. We cannot tell why animal is doing something without ascribing it intentionality, just like we cannot have a sufficient explanation by simply looking at its neuronal system (Searle, 1980, p. 421). This can be surprising, because it would mean that if we discover certain neuronal operations that cause specific human behaviour (and formulate a neuro-behavioural laws), according to Searle, we should admit that humans actually do not have mental states (and so accept some form of eliminativism, which is not such a rare position).

Conclusion

The problem of other minds is not only about AI, but about other humans, too. We are assume that others have mental states, rather than knowing it for certain. If we want to ask whether some entity is conscious, we should therefore probably ask ourself why we think that other people are conscious and then check whether AI – or anything else – meets those criteria. Besides, from a technological point of view, it appears that there is still a long way to go before building conscious AI. Making an advanced artificial neural network requires an enormous amount of data. Today, information is valuable, held by big corporations that do not want to share it freely. There is also a problem with computing power: even the best supercomputers are not even close to the power of the human brain. Another thing is novelty. Humans can find solutions to problems that are new to them. We know where to look for information and can collaborate to solve problems as a society (sharing cognitive labour). To achieve something like this, AI would have to be connected with other AI so that they could help each other within a bigger meta-network (which could make one super-AI). These are just examples of the challenges still before us. If we take Searle's restrictions seriously, a fully conscious AI is not as close as many people think. Yet we are getting closer to it every day.

Bibliography

Al-Sibai, N. (2022), *Researchers Furious Over Claim That AI Is Already Conscious*, "Futurism", https://futurism.com/conscious-ai-backlash [accessed: 19.02.2022].

Chowdhury, M. (2022), *What is AI image recognition? How does it work in the digital world?*, "Analytics Insight", https://www.analyticsinsight.net/what-is-ai-image-recognition-how-does-it-work-in-the-digital-world/ [accessed: 21.02.2022].

Cole, D. (2020), *The Chinese Room Argument*, "The Stanford Encyclopedia of Philosophy, Zalta E.N. (ed.), https://plato.stanford.edu/archives/win2020/entries/chinese-room/ [accessed: 21.02.2022].

Cuthbertson, A. (2022), *Artificial intelligence may already be 'slightly conscious', AI scientists warn*', "Independence", https://www.independent.co.uk/tech/artificial-intelligence-conciousness-ai-deepmind-b2017393.html [accessed: 19.02.2022].

Le Page, M. (2021), *Human brain cells in a dish learn to play Pong faster than an AI*, "New Scientist", https://www.newscientist.com/article/2301500-human-brain-cells-in-a-dish-learn-to-play-pong-faster-than-an-ai/#ixzz7LYcd1iWu [accessed: 21.02.2022].

Putnam, H. (1980), *Nature of mental states*, (In:) Ned Block (ed.), "Readings in Philosophy of Psychology, Vol. 1, Harvard University Press.

Putnam, H. (1981), *Brains in a vat*, "Reason, Truth and History", Cambridge University Press.

Pylyshyn, Z. (1980), *Reply to Searle*, "The Behavioral and Brain Sciences", No. 3.

Ramsey, W. et al. (1990), *Connectionism, Eliminativism and The Future of Folk Psychology*, "Philosophical Perspectives", Vol. 4.

Searle, J.R. (1990), *Cognitive Science and the Computer Metaphor*, (In:) Göranzon, B., Florin, M. (Eds.), *Artifical Intelligence, Culture and Language: On Education and Work*, Springer-Verlag.

Searle, J.R. (1980), *Minds. Brains, and programs*, "The Behavioral and Brain Sciences", No. 3.

Smolensky, P. (1988), *On the Proper Treatment of Connectionism*, "The Behavioral & Brain Sciences", No. 11.

ISBN 978-83-67575-00-3