ZNANSTVENA KNJIŽNICA 98e

Vojko Strahovnik, Jonas Miklavčič (ur.)

BEYOND ALGORITHMS

DISENTANGLING THE PHILOSOPHICAL AND ETHICAL COMPLEXITIES OF AI AND ITS IMPLEMENTATION

TEOF
UNIVERZA V LJUBLJANI
Teološka fakulteta

UNIVERSITY OF LJUBLJANA

TEOF

Faculty of Theology

Vojko Strahovnik, Jonas Miklavčič (ur.)

BEYOND ALGORITHMS: DISENTANGLING THE PHILOSOPHICAL AND ETHICAL COMPLEXITIES OF AI AND ITS IMPLEMENTATION

Teološka fakulteta
Univerze v Ljubljani
Ljubljana 2025

Znanstvena knjižnica 98e

Collection editor / Urednik zbirke:	doc. dr. Miran Špelič OFM
Reviewers / Recenzirala:	izr. prof. dr. Roman Globokar
	prof. dr. Anton Jamnik
Editorial board / Uredniški odbor:	doc. dr. Matjaž Ambrožič
	doc. dr. Igor Bahovec
	prof. dr. Branko Klun
	prof. dr. Maria Carmela Palmisano
Proofreading / Lektoriranje:	Richard H. Smith
Cover / Naslovnica:	Umetna inteligenca
Layout / Prelom:	asist. dr. Jonas Miklavčič
Published by / Izdala in založila:	Teološka fakulteta Univerze v Ljubljani
	Poljanska 4, 1000 Ljubljana
For the Publisher / Za založnika:	doc. dr. Tadej Stegu, dekan TEOF
Type of issue / Vrsta izdaje:	Elektronska izdaja
Available on / Dostopno na:	https://https://www.teof.uni-lj.si/uploads/
	ZNK_98e_Strahovnik_Miklavcic_Beyond.pdf
Price / Cena:	free / brezplačno

Publikacija je nastala v okviru raziskovalnega projekta *J6-4626 Teologija, digitalna kultura in izzivi na človeka osrediščene umetne inteligence* (ARIS), zagonske programske skupine *Presečišče kreposti, izkustva in digitalne kulture: etični in teološki uvidi* (ARIS, Univerza v Ljubljani) in raziskovalnega projekta *Theological, Philosophical and Ethical Challenges of Artificial Intelligence and New Technologies* (Templeton Foundation; University of Oxford)

This publication was produced within the framework of the research project J6-4626 *Theology, Digital Culture and the Challenges of Human-Centred Artificial Intelligence* (ARIS), the start-up research programme *The Intersection of Virtues, Experience and Digital Culture: Ethical and Theological Insights* (ARIS, University of Ljubljana), and the research project *Theological, Philosophical and Ethical Challenges of Artificial Intelligence and New Technologies* (Templeton Foundation; University of Oxford).

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

COBISS.SI-ID 263471107
ISBN 978-961-7167-39-9 (PDF)

TABLE OF CONTENTS

PERSPECTIVES ON ARTIFICIAL INTELLIGENCE

THE TRANSPARENCY PROBLEM IN ARTIFICIAL INTELLIGENCE

ARTIFICIAL INTELLIGENCE AND HEALTHCARE

INTRODUCTION

As the digital age deepens its roots in the fabric of contemporary society, the interplay between technology and ethics becomes increasingly intricate and indispensable. This collection of papers seeks to explore the multifaceted ethical dimensions that arise in the wake of advancing artificial intelligence (AI) and digital technologies. Through an interdisciplinary effort, it addresses a spectrum of critical issues at the intersection of technology, philosophy, ethics, and human values.

The contributions herein delve into the ethical quandaries engendered by the digital replication of real-world systems and beings, scrutinize the implications of AI's opacity on trust and accountability, and examine the moral obligations of human agents within the framework of machine learning and digital decision-making. This dialogue is not confined to theoretical musings but extends to practical considerations and applications, particularly in fields as sensitive and impactful as healthcare.

At the core of these discussions are the challenges of ensuring transparency and explainability in AI systems, a task complicated by the inherent complexity and often inscrutable nature of machine learning models. These concerns are further magnified by the pressing need to reconcile human intuitive judgment with AI's data-driven decisions, highlighting the ethical imperative for a balanced approach that respects human intuition and rationality alike.

This book represents a collective endeavour to navigate the ethical labyrinth that technology weaves around our personal and social lives. It calls for a critical examination of how we, as a society, conceive of and interact with digital technologies. The discussions put forth challenge us to reconsider our values, our responsibilities, and our visions for a future where technology and ethics coalesce toward the betterment of humanity.

In bringing together these diverse, yet interconnected, explorations, the book aims to contribute to a deeper understanding of the ethical implications of digital and AI technologies. Its aim is to challenge scholars, practitioners, policymakers, and the wider public to engage in a meaningful dialogue on how to navigate the digital frontier ethically. As we stand at the crossroads of technological advancement and ethical responsibility, the insights offered serve as both a reflection and a guide for the thoughtful and conscientious integration of technology into the fabric of human life.

Vojko Strahovnik and Jonas Miklavčič

Bojan Žalec

THE (IM)PROBABILITY OF HUMANLIKE ARTIFICIAL INTELLIGENCE

Introduction

The main question of this chapter is: Is the creation of humanlike AI probable? In the following, I will refer to the thesis that humanlike AI will be created as the HI thesis. When I speak of humanlike AI, I mean AI that would be capable of everything human intelligence can do. However, the question of the truth of the HI thesis is closely connected to the probability of superintelligence, an intelligence that would vastly surpass human intelligence. Therefore, I will also discuss the probability of the emergence of superintelligence, which some transhumanists promise, such as Ray Kurzweil, who predicts the onset of singularity. (Kurzweil 2005) I will henceforth refer to the thesis of the emergence of superintelligence/ singularity as the SI thesis, and use the abbreviation SI for artificial superintelligence. The HI thesis and the SI thesis are intertwined. On one hand, the emergence of SI is not probable if humanlike AI is not probable; on the other hand, the creation of humanlike AI would provide substantial support for the SI thesis. Thus, the theses are connected, and arguments for and against the HI thesis are indirectly also arguments for and against the SI thesis. Therefore, considering both theses together makes sense.

In what follows I will argue that referring to the achievements in AI so far, which are truly impressive in many respects, does not carry much weight in terms of the main question of this article. Citing them as convincing evidence for AI and HI theses is based on a misunderstanding of human intelligence. Therefore, in this article, I will critically reflect on the understandings on which predictions about the probability of the AI and HI theses are based.

In the first part of the chapter, I will present arguments against the HI thesis, as articulated by Eric J. Larson in his book *The Myth of Artificial Intelligence* (2021), and in the second part, arguments against the probability of the SI thesis, as presented by François Chollet in his influential article *The Implausibility of Intelligence Explosion* (2017).

Larson focuses on four fundamental and essential characteristics of human intelligence: generality, intuition, common sense, and the ability of abduction. He argues that existing AI lacks these capabilities and that within the current (Turing) paradigm of AI development, we will not be able to create AI possessing them. It is also unlikely that AI would develop itself to a level where it possesses them. Chollet argues that the onset of SI is unlikely. In his argumentation, he emphasizes three characteristics of intelligence: 1. Non-generality; 2. Situationality and contextuality; 3. Externalism. I refer to the first two as the particularity of intelligence, although the third also actually implies the particularity of intelligence.

Overall, this chapter can be understood as a contribution to the defense of an anti-Cartesian paradigm of understanding (human) intelligence and mind. Various predicates and names are used for understandings and approaches within this paradigm: Aristotelian and interactive (Miščević 1988), enactivism (Gallagher 2017), ecologism (Gibson 1979; Potrč 1993; 1996, 194-197; 2004, 55-61), externalism (McCulloch 1995, 184ff; Rowlands, Lau and Deutsch 2020; Potrč 1993), theory of contact (Dreyfus & Taylor), embodied humanism and anthropology of embodiment (Fuchs 2021), and others. I like the term 'being-in-the-world paradigm' for the version of anti-Cartesianism which I prefer. Therefore, it is not surprising that in the chapter, I also mention Hubert Dreyfus' understanding, but here I do not delve into it in detail.

In this chapter, I focus on humanlike intelligence. I argue that there is no such thing as humanlike AI and that its emergence is not probable. However, this does not mean that I believe there is no AI that is intelligent in the literal sense of the word. On the contrary, I believe that existing deep learning systems are intelligent in the true sense (Žalec 2023a). We must differentiate between humanlike intelligence and intelligence as such. Humanlike intelligence is not the only type of intelligence (op. cit.).

I also touch upon the social dimension of the misunderstanding of AI potentials. One of its consequences is the ideology of big data science, which is dangerous and harmful, among other things as a negative factor in nurturing (human) creativity, in the field of science and technology, as well as others. The harmfulness of this ideology indicates that the discussion about the nature and potentials of AI development is not only of narrow academic significance but also very important from a broader societal and moral perspective.

In the literature, various other arguments for and against the possibility of humanlike AI and SI can be found, which I do not mention or only briefly mention (for instance argumentation that creating an adequate mathematical model of the human mind exceeds human capabilities (Landgrebe and Smith 2023)). This is certainly a limitation of this chapter. On the other hand, I believe that the arguments presented in it alone form a good basis for a reasonable rejection of the probability of the HI thesis and SI thesis.

Essential Characteristics of Human Intelligence

Let us start the discussion with some general reasons why, at least in the near future, and given the current scientific and technical knowledge and approach, we cannot expect anyone to create AI that possesses the essential characteristics of human intelligence. As such characteristics, we can mention abilities that are closely intertwined (Larson 2021): generality, as opposed to narrowness and specialization in a specific narrow domain of problems and tasks, the ability to understand, intuition, learning ability, choosing the problems to solve, common sense, and the ability of a particular form of reasoning called abduction.

Larson argues that currently, no-one has a clue how to create humanlike AI. No one has a proper scientific and technical idea of it. Instead of scientific arguments in favour of the possibility of creating humanlike AI, many resort to various (scientifically) poorly grounded claims that spread the belief that we will create humanlike AI. Time and again, there are those who claim that we have already reached this goal (in principle), that we will soon achieve it, that we will (soon) witness the onset of AI that will greatly surpass human intelligence, and so on. Such 'false prophets' or promoters of AI create a myth about AI (Larson 2023). Here, 'myth' is meant in a negative sense, as a false belief that something exists or is probable, although it actually does not exist and is not probable. What do not exist, and are not probable, are humanlike AI and SI. Similarly, we can talk about the ideology of AI in the sense mentioned, that is, about a false belief that some, consciously or unconsciously, deliberately and intentionally, create, due to certain interests, among which financial ones are anything but negligible.

Big Data and Big (Hive) Science

An integral part of today's ideology is the ideology of big data and big science. Advocates of this ideology argue that simply integrating a large amount of data using current knowledge and theories will lead to the emergence of AI, which will first reach human level, and then quickly surpass it infinitely, reaching a level that Kurzweil calls singularity. Alongside big data science, an essential part of the ideology of big science is the concept of hive or swarm science. Its proponents claim that further development of the theory is unnecessary, or even impossible, and that the integration of large amounts of data, with the help of high-performance computers, will take care of the missing pieces to create AI. We no longer need highly intelligent and creative individuals with original ideas and insights. In short, we no longer need Einsteins. Such a stance implies the degradation of humans and human intellect merely to assistants of a large computer, mere servants and caretakers of a big computer that is supplied with necessary data and services to function. We no longer need scientists who think, who come up with new ideas, but only highly skilled technicians, service personnel, and operators. Thus, humans increasingly adapt to the machine and become more and more like machines. In terms of typically mechanical tasks, of course, the machine surpasses humans, so within such a conception, it is understandable that humans submit to the machine and become its servants. One advocate of such a concept stated that the time has come for us to set aside our ego and perform our role in the hive to which we are assigned. Such a perception implies a reduction in the significance of the individual and promotes a culture that no longer invests in creating conditions for the development of individuals with original ideas and theories, which enable true scientific and technological breakthroughs and progress. Therefore, it is not surprising that in environments and periods where the idea of big science prevails, there is no real progress in science and technology, although advocates of big science try to show that the situation is different. The situation is even worse. If the model of big science becomes more established in the future, we can expect not only stagnation in the level of scientific and technological progress but regression and ultimately its decline. Many scientists are aware of this and warn against the harmfulness of the concept of big science, actively organizing and directing opposition to its funding. Typically, enormous sums are at stake for projects operating on the model of big science. A case in point is the petition signed by more than 500 scientists addressed to the European Commission, suggesting major changes of The Human Brain Project (Larson 2021, 267), which officially began in 2013 under the leadership of Dr. Henry Markram, a neuroscientist from the Swiss Federal Institute of Technology Lausanne. The project initially involved more than 150 institutions worldwide. (243ff)

The Problem of Generality and The Trap of Narrowness

One prediction is that once we create humanlike AI, it will soon create intelligence far surpassing human capabilities, leading to the development of SI, whose capabilities we cannot imagine. However, there are several problems with this idea. First: we already have human-level intelligence. It has existed for a long time but has not evolved into SI surpassing human intelligence. Why would we reasonably expect humanlike AI to do so? Second: nobody knows how this humanlike AI could achieve this. Nobody has even scientifically described this process. Yet some predict it will happen.

Characteristics of human intelligence include intuition and independent problem selection. Because of this, human intelligence is not narrow but general. It is not limited to solving only certain problems or tasks. Larson observes that nobody has any idea how to create an AI system capable of intuition and independent problem selection, necessary for surpassing narrowness. Without these, we can hardly speak of humanlike AI.

One way to argue in favour of the SI thesis is through reference to evolution: SI will surely emerge in evolution, which produces increasingly higher and more complex systems. Even if we do not know how, we can argue, based on evolutionary grounds, that SI will eventually appear. However, such a prediction cannot be considered scientifically substantiated. One sign is that it cannot be falsified. But this is a criterion of scientificity. Moreover, this prediction is not affected by errors in prediction. Kurzweil predicted the advent of humanlike AI as early as 2029 and singularity by 2045. But even if this does not happen, proponents of SI can still insist on their prediction; it will just happen a few decades later, and if it does not happen then, it will just happen a little later.[1]

[1] In a similar non-falsifiable manner, some representatives of the New Age have predicted the onset of a new, "higher" consciousness. (Sire 2002, 164ff)

Some argue for the SI thesis based on the principle of a (necessary) leap from quantity to quality. Continuous enhancement of the ‘quantitative’ capabilities of AI must eventually lead to a leap in quality, to the appearance of a qualitatively higher level of AI, and so on to SI. The principle of the necessary leap from quantity to quality has been advocated in the past but experience does not confirm it. Friedrich Engels (1975, 118-119, 351-352, 482, 510-511, 516-517, 552) advocated it as one of the main principles of his ‘philosophy’ of nature or dialectics of nature.

Claims about the advent of humanlike AI and SI have been made in the past. So far, none of these predictions has come true. Moreover, as mentioned earlier, nobody currently has a scientific idea of how to create a system that exhibits the characteristics of humanlike intelligence listed above. Without some fundamentally new theoretical insight, breakthrough, or revolution, there is no possibility of achieving this, as proponents of the possibility of creating AI within the currently available theoretical framework actually do not offer scientific arguments for their claim, which indicates that this is an ideology.

In order to understand the challenges and obstacles to the development of humanlike AI, we must free ourselves from the narrow concept of intelligence, which reduces all intelligence to (independent) problem-solving. Independent problem-solving is a sufficient condition for attributing intelligence, but this does not mean it is sufficient for every kind of intelligence. It is merely a minimal, necessary, and sufficient condition for intelligence, which is not enough for human intelligence. Human intelligence is more than minimal intelligence. Human intelligence includes not only independent problem-solving but also independent problem selection. "The master need only know how to order that which the slave must know how to execute," says Aristotle (1995, 1992). In this sense, human intelligence is the intelligence of the master, while AI is the intelligence of the slave. It cannot command itself, it cannot set its own problems (Peirce 1887, 165; Larson 2021, 233); it needs a master. Therefore, it cannot be creative in terms of posing a new problem. Defining intelligence as independent problem-solving, let us call it the minimal definition, is very clear and therefore engineeringly useful. According to it, we can already speak of AI as intelligent. However, the problem arises if we do not realize that this definition is insufficient for defining human intelligence since it does not indicate its distinctive characteristics such as generality. In this case, it obscures the specificity of human intelligence and consequently leads to a misconception about the potentials of AI. However, abandoning the minimal definition as sufficient for the definition of any intelligence, not just minimal intelligence, is probably quite a difficult task among programmers and engineers since such understanding is part of the Turing legacy, and Turing is the founder of the paradigm within which AI research and development still take place today.

If a computer successfully or even better solves problems that previously required human intelligence, that is certainly progress in AI development. The more complex these tasks are, the more capable a human must be to solve them, and the better AI's capabilities for executing them, the more it represents progress in AI development. This we can also agree with. However, it is another matter to use them as evidence for the HI thesis or the SI thesis. Playing games like chess and go, correctly answering quiz questions, medical diagnoses, etc., performed by AI, some proclaim as evidence in favor of the HI thesis and SI thesis. I disagree with this, and the purpose of this article is to show that it is not true.

Treating intelligence as problem-solving gives rise to narrow applications. If the machine could learn to transcend its limitation to narrow tasks and be able to solve problems in general, then this would signify a transition to higher or (more) humanlike intelligence. But for now, at least, we are far from any general AI. In order to achieve its goals, every machine learning system must learn something specific. Researchers call this biasing of the system. Bias in machine learning means that the system is designed and tuned to learn something. This is precisely the production of applications for solving narrow problems. That's why the deep learning systems used by Facebook for recognizing human faces have not simultaneously learned how to calculate our taxes. The situation is actually worse: researchers have found that if a machine learning system is biased and specialized to learn a specific application or task, then the system performs worse on other tasks. There is a reverse correlation between the system's success in learning a certain thing and the success of its learning for another task. This applies even to very similar tasks. A machine learning system that learns to play chess at a high level will not learn to play go at a high level and vice versa. The go system was specifically designed with a particular bias for learning the rules of go. The problem is that we cannot get rid of biases because they are an integral part of machine learning. The proverb "there is no such thing as a free lunch" also applies to machine learning. In this case, it means that any machine learning system that is not biased will not perform any better than random chance when applied to arbitrary problems. A truly unbiased system, a system not biased by programmers, is useless. But the biased system learns only what its designers want it to. Thus, by biasing a system we make it narrow in the sense that it will not then generalize to other areas. Narrowness is thus inseparable from the success of the system; narrowness and success are two sides of the same coin in machine learning systems. (Larson 2021, 28-30)

What we now know, different from the initial enthusiasm, is that machine learning is just a type of problem-solving that can only be achieved by introducing bias into the learning system. While this allows learning of a specific application, it also reduces the performance of other applications. Even learning AI systems are just narrow problem-solving systems. No scientific or technical breakthrough is known from such narrow systems to general intelligence, as exhibited by humans. In this regard, the development of general AI has found itself in a deadlock, a standstill: understanding intelligence only as (independent) problem-solving gradually, but inevitably, leads to a theoretical dead end, to the "trap of narrowness,"[2] as Larson calls it, right at the heart of AI research. This indicates that understanding intelligence, which reduces it to (independent) problem-solving, is too narrow and inadequate to achieve broader goals than producing narrow applications. (Larson 2021, 30-32)

The central problem for the development of humanlike AI can be described with the concept of intuition. If we wanted AI to be humanlike, it would have to be capable of intuition. If we wanted it to have intuition like humans and researchers do, we would have to describe this intuition scientifically so that this description would be useful for programmers and engineers. However, nobody has any idea how to do it. Without intuition, AI cannot surpass the 'curse' of narrowness, which would be necessary for it to learn in such a way as to become similar to human intelligence. Such learning requires the system to choose problems itself, and intuition is required for this. We know that AI system designers use their own intuition to instruct AI systems on which specific problems to solve or learn to solve. But for AI to be truly intelligent, it would have to have its own intuition. It does not have that, and, as already mentioned, nobody has a scientific idea of how to create AI that would have it. (Larson 2021, 31-32)

Let us summarize the main reasons so far for rejecting predictions of the likely emergence of humanlike AI or SI, especially in the near future, as justified. SI could evolve from human or humanlike AI 'on its own', like evolution, without us technically understanding it. But this has not happened on its own since the advent of human intelligence, and that is a very long time. Why would this happen in the near future? Of course, we now have systems that perform certain activities much faster and more accurately than humans. However, for humanlike intelligence, much more than speed and accuracy is required. Intuition and the ability to choose problems are needed. But how to create an AI capable of this, Larson notes, no one has a clue.

[2] One of the first who clearly pointed out the original initiative and narrowness problems was Peirce: "Every reasoning machine, that is to say, every machine, has two inherent impotencies. In the first place, it is destitute of all originality, of all initiative. It cannot find its own problems; it cannot feed itself. It cannot direct itself between different possible procedures. /…/ [T]he machine would be utterly devoid of original initiative, and would only do the special kind of thing it had been calculated to do. This, however, is no defect in a machine; we do not want it to do its own business, but ours. /…/ In the second place, the capacity of a machine has absolute limitations; it has been contrived to do a certain thing, and it can do nothing else." (Peirce 1887, 168-169; Larson 2021, 233) Turing was aware of the initiative problem: "In the next century, Turing proposed that we take up the challenge of infusing machines with 'original initiative,' by first programming them to talk to us. Turing was aware of Peirce's objection, which he attributed to Lady Lovelace in his 1950 paper. He also had played with simple learning algorithms, and in the decade of the 1950s single-layer neural networks appeared (called a perceptron). Understandably, Turing thought perhaps we could escape Peirce's and Lovelace's objections by creating learning machines modelled on the human brain. Reading 'Computing Machinery and Intelligence,' one gets the impression that learning represented the only real escape from the inherent limitations of machines, and the only real hope for passing the Turing test. It did not – it has not happened. Believing that it will, that it must, has consequences for society that now have become all too apparent. In this book's final part, we look at some of the consequences of the inevitability myth – particularly its deleterious effect on science itself. (233-234)

Turing contemplated formalizing intuition so that a computer could use it, but he did not know how, and even now, nobody knows because current science and technological understanding of AI are heirs to Turing's understanding and approach and proceed within the framework of his paradigm and understanding of intelligence. A radical change in the understanding of intelligence and research paradigm would be required for a breakthrough, but nobody knows what those should be. Therefore, nobody has a scientific idea of how to create humanlike AI, that is, intelligence that would have intuition and would therefore be able to choose problems itself and (thus) would not be limited and could learn to be general. Similarly, nobody has even remotely described how humanlike intelligence, artificial or natural, could develop itself into SI. Therefore, we can justifiably assume that nobody has a scientific idea of how this could happen. Because of all that has been said, we can argue that humanlike AI is (for now) just a 'myth' (Larson 2021).

Abduction, Common Sense, and Understanding

We can distinguish three basic forms of reasoning that cannot be reduced or converted into one another: deduction, induction, and abduction.

In the symbols of propositional logic, they can be represented as follows (Larson 2021, 172):

Deduction:
P->Q
P

_____
Q

Induction:
P
Q

_____
P->Q

Abduction:
P->Q
Q

_____
P

Deep learning systems are based on statistics. Statistics is essentially induction. Abduction is similar to guessing, but it is not mere guessing. It is more than just association or correlation. However, no one knows how to describe this ‘more’ in a way that would enable the engineering of human intelligence. Abduction, common sense, and intuition presuppose understanding, which no-one knows how to describe in a technically useful way. Understanding is the core of human intelligence. Until we can adequately mathematically describe understanding, we will not be able to create humanlike intelligence.

The only knowledge that can be supplied to a machine learning system is what the system can recover from data in a purely syntactic way. This means that there is a blind spot in the system, resulting in incorrect predictions, as what the system cannot observe in the data, it does not know. (173) This blind spot and error in reasoning can be illustrated by the example of a turkey (121-124), who is an inductivist and arrives at a fatally incorrect conclusion through induction:

"This turkey found that, on his first morning at the turkey farm, he was fed at 9 am. However, being a good inductivist, he did not jump to conclusions. He waited until he had collected a large number of observations of the fact that he was fed at 9 am, and he made these observations under a wide variety of circumstances, on Wednesdays and Thursdays, on warm days and cold days, on rainy days and dry days. Each day, he added another observation statement to his list. Finally, his inductivist conscience was satisfied and he carried out an inductive inference to conclude, ‘I am always fed at 9:00 am.’ Alas, this conclusion was shown to be false in no uncertain manner when, on Christmas Eve, instead of being fed, he had his throat cut. An inductive inference with true premises has led to a false conclusion." (Chalmers 1982, 41-42)

This example nicely illustrates "the folly of forming 'habits of association' without deeper knowledge" (Larson 2021, 123). Machine ‘learning’ is not based on understanding and is not the learning of understanding, but is based solely on association or correlation. Therefore, machine learning is not humanlike learning, as human learning is based on understanding and is the learning of understanding. Progress in learning for humans means understanding more. In deep learning systems, we cannot speak of humanlike learning because we cannot attribute understanding to them. The progress in the knowledge of AI systems is not progress in understanding but in the quantity of information or ‘facts’ they have and in (simulation of) adaptability (Fuchs 2021, 33).

The difference between human intelligence and AI is clearly visible in the recognition of faces. While AI systems for facial recognition require a vast number of views, a baby requires only a few views to recognize a face. Of course, a baby cannot explain why it recognized the face. This knowledge is *knowledge how*. Hubert Dreyfus (1972) argued two things relevant to this knowledge (Oettinger 1972, xii-xiii): 1. Humans first perceive the whole (Gestalt) and then, if necessary, analyze it into parts; 2. A baby's recognition/knowledge is based on the human body. Therefore, robots should have sufficiently humanlike bodies to have humanlike intelligence. Dreyfus's rejection of the possibility of humanlike AI was based on the insights of phenomenological philosophers, especially Merleau-Ponty’s and Heidegger’s. Using Heidegger's terminology, we can say that for a computer, entities are (at best) only present-at-hand or occurrent (Ger.

vorhanden), not ready-to-hand or available (Ger. *zuhanden*) (Dreyfus 1991, xi). Therefore, the computer is a 'stranger' in the (human) world or even better, it cannot be in the human world because the human world already presupposes human being-in-the-world (Dreyfus 1997; Dreyfus and Dreyfus 2004, 403; Žalec 2023b, 36-37, 108-110). As long as AI remains a stranger among present objects, given to it only in a theoretical approach and not as a way of availability as given to a human, then AI that would be humanlike will not be possible. For an entity to be capable of humanlike intelligence, it must first be capable of human being-in-the-world. The essential component and foundation of this being-in-the-world is human understanding of entities as ready-to-hand. This is pre-theoretical, practical, skillful coping with the world, observing entities as merely available, not 'theoretically' (Dreyfus 1991). In practical orientation, humans use entities as tools. Being a tool is a basic way of being available, ready-to-hand. Common sense, lacking in computers, is based on grasping entities as available from the human perspective. This grasping, as it seems, is based on the human (lived) body.[3] Therefore, we come to the same conclusion again that AI should have a human lived body to have humanlike intelligence.

I believe Dreyfus was right. However, in this paper, I will not delve into a discussion of his understanding. His thoughts, though, on the importance of the purpose of intelligence and its way of being-in-the world provide a meaningful basis for presenting arguments against the likelihood of SI, as given by François Chollet (2017).

Particularity and Externalism of Intelligence

Chollet's article on the improbability of the so-called intelligence explosion has sparked a lot of interest. So far, it has already received over eighteen million views. The term 'intelligence explosion' was used in the 1960s by the British mathematician Irving John Good to claim the advent of AI which will develop to an unimaginable degree. This development will proceed with tremendous growth. In the article, Chollet proves that such an explosion is not probable. His argumentation is relevant to the present time when claims of an intelligence explosion formulated as predictions of the advent of singularity and similar phenomena are reappearing.

Chollet's approach in the article is empirical. He refers to data on the evolution and historical development of intelligence and various other empirical data, e.g. on the success and achievements of people with above-average IQs and so forth.

[3] The distinction between the lived body (Ger. Leib) and the object-body (Ger. Körper) has a rich history in phenomenology (E. Husserl, E. Stein, M. Scheler, M. Merleau-Ponty, M. Henry, H. Schmitz, et al.). For its explanation, see Gallagher 1986; Schmitz 2011; Zahavi 2019; Žalec 2023b; Ottinger 2021. Zahavi (2019, 145) defines both terms as follows: "*Körper*: the physical and biological body; the body considered as a physical object that belongs to nature. *Leib*: The lived and experienced body; the body as subjectively lived through."

Chollet says that it is very important first to define the concept of intelligence so that it will be clear what we are discussing and what we are proving. For the purposes of the article, Chollet initially defines intelligence as problem-solving. He says this is an initial definition, so to speak, a working definition, to start the discussion. Throughout the article, he further develops and supplements this definition.

The core of Chollet's argumentation in the article consists of the following three claims:

1. There is no general intelligence.
2. Intelligence is situational and contextual.
3. Intelligence is externalist.

All three characteristics are interconnected and intertwined. The first thesis claims that every intelligence is adapted to the entity that has that intelligence, to its needs, its way of life, etc. This applies to the intelligence of animals, to human intelligence, as well as to AI, which is adapted to its purposes. There is no general intelligence in the sense that it would be suitable for all tasks and all purposes. Human intelligence is useless for certain animal needs, and vice versa. Much is already innate. The second thesis says that the use, utilization, and development of our intelligence are influenced by the situation in which we find ourselves. Chollet cites the example of feral children who grew up among animals, for example, a man who lived among monkeys. He, in a way, became a ‘monkey’ and could never humanize again. He never learned language, jumped on tables and elsewhere, rejected cooked food, etc. The same applies to other feral children. Some managed to humanize more, others less, but in all cases, it was very deficient. That intelligence is externalist means that it is not only contained in our brains, but also in our civilizations (Larson 2021, 27).

Very few people with above-average IQ achieve anything outstanding in life. Most of these people work in ‘banal’ professions, performing quite ordinary and ‘average’ functions and tasks. What and how much someone will achieve with their intelligence, how they will utilize it and develop it, etc., depends on numerous factors: their historical and social position, upbringing, education, chance encounters with the right people, being in the right place at the right time, appropriate motivation and aspirations, etc. Chollet cites examples of scientists who achieved significant results, yet their IQ was nothing special. For example, physicist Feynman and Watson, co-discoverer of DNA. They had the same IQ as many other scientists who will never discover something like them. And how many people with a higher IQ than them will achieve something comparable to their achievements? He also cites the words of biologist Gould, who is less concerned with the weight and size of Einstein's brain than with the thought of how many people with Einstein's IQ have toiled away in various factories, mills, etc., where they don't even have the opportunity to develop their potential. Or if we go back in human history 10,000 years, how did people live? how could they develop their intelligence? They spoke a simple language; most of them could not read or write, etc. It also depends on other historical aspects what someone will discover. Based on various data from the past and present, Chollet argues that the growth of intelligence will be gradual, moderate, linear in the future, and not explosive, exponential, or even unimaginably steep. There will be no singularity in the sense as predicted by Ray Kurzweil and similar 'prophets' of SI.

Chollet's thesis is surprising in some respects. Take the development of science and technology, which strongly influences development in all other areas. If we consider what powerful tools we have available now, whose capabilities are constantly increasing – the sophistication of communications, the possibility of exchanging knowledge and networking, the increasing number of people engaged in science and technology development, etc. – then we would expect rapid progress. However, Chollet points out that we must look at the whole and consider the effects that such development brings. Along with development, the complexity of scientific problems increases. With the increase in knowledge, the time needed to master that knowledge, to educate scientists, and to monitor achievements in science increases, so that one can stay up to date and develop new things. If we consider the above, it is not surprising that, according to all methodologically measurable indicators, scientific knowledge in the field of physics in the second half of the 20th century did not increase more than in the first half of the same century, and one could list more examples. On the one hand, it is true that the more educated you are, the faster your knowledge will increase, as you will be able to use complex tools, mathematical knowledge, and notation, etc. But it is necessary to consider the whole and understand intelligence and knowledge growth contextually and externalistically so that we can see that things are related, that we need to take an integral and relational perspective to understand why intelligence and knowledge growth are linear. Similarly, in the field of economics; on the one hand, it is true that the more money you have, the more you can additionally gain. But if we look at the whole, with the growth of wealth and investments, other negative factors of growth also increase, and ultimately empirical data show that the growth of wealth is seen as linear, gradual, and not 'explosive'.

Chollet's conclusion is that the growth of intelligence will not be explosive, but that intelligence will grow, as it does now, gradually, linearly. Civilization, which means intelligence, will continue to develop: scientific knowledge will increase, the number of technological tools and their capabilities, etc. However, there will be no development of superbrain, SI, or anything similar, such that that would mean the end and abolition of humanity, etc. Nor will AI take over all tasks or authority, as AI has no own needs, purposes, motives, desires, intentions, etc. In all these meanings of the term, 'intelligence explosion', according to Chollet's conviction, is not probable.

Conclusion

I have presented four of Larson's arguments against the HI thesis, which can be called the argument of generality, the argument of intuition, the argument of abduction, and the argument of common sense, which, in my opinion, are good reasons for rejecting the HI thesis. In addition, I presented Chollet's argumentation against the SI thesis. This is based on empirical analysis and takes into account the particularity and externalism of intelligence. Chollet's arguments, in conjunction with the aforementioned Larson's arguments, provide a solid justification for rejecting the SI thesis. I have shown how important it is to understand and define intelligence, as incorrect understanding can have significant harmful effects that go far beyond the academic sphere. For example, such understanding provides a basis for obtaining unjustified financial and other benefits, for harmful and dangerous ideologies, such as radical transhumanism, and for instrumentalist, dehumanized and anti-personalistic views on the character and significance of human creativity and the nature of science, which jeopardize its actual progress and the welfare of humanity in general.

Funding acknowledgement
This chapter is the result of work within the research program *P6-0269: Religion, ethics, education, and challenges of modern society* and the research projects *J6-4626 (B): Theology, digital culture and the challenges of human-centered artificial intelligence* and *J6-60105: Theology and digitalisation: anthropological and ehical challenges*, all financed by the Slovenian Research and Innovation Agency (ARIS).

References

Aristotle. 1995. Politics. Trans. by Benjamin Jowett. In: Jonathan Barnes, ed. *The Complete Works of Aristotle*, vol. 2, 1986-2129. Princeton: Princeton University Press

Bostrom, Nick. 2020. *Superintelligence. Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Chalmers, Alan. 1982. *What Is This Thing Called Science.* St. Lucia, AU: University of Queensland Press.

Chollet, François. 2017. The impossibility of intelligence explosion. Medium, 27. november. https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec (accessed 22. 8. 2023).

Dreyfus, Hubert L. 1972. *What Computers Can't Do: A Critique of Artificial Intelligence.* New York: Harper and Row.

- - -. 1991. *Being-in-the-World. A Commentary on Heidegger's Being in Time, Division I.* Cambridge, MA: The MIT Press.

- - -. 1997. From Micro-Worlds to Knowledge Representation: AI at an Impasse. In: John Haugeland, ed. *Mind Design II: Philosophy, Psychology, Artificial Intelligence,* 143-182. Cambridge, MA: A Bradford Book, The MIT Press.

Dreyfus, Hubert L., and Stuart E. Dreyfus. 2004. Why Computers May Never Think Like People. In: David M. Kaplan, ed. *Readings in the Philosophy of Technology*, 397-413. Lanham: Rowman &

Littlefield Publishers, INC.

Dreyfus, Hubert, and Charles Taylor. 2015. *Retrieving Realism.* Cambridge, MA: Harvard University Press.

Engels, Friedrich. 1975. Dialektik der Natur. In: Karl Marx, and Friedrich Engels. *Werke*, vol. 20, 307-570. Berlin: Dietz Verlag.

Fuchs, Thomas. 2021. *In Defense of Human Being. Foundational Questions of an Embodied Anthropologie.* Oxford: Oxford University Press.

Gallagher, Shaun. 1986. Lived Body and Environment. *Research in Phenomenology* 16:139-170

- - -. 2017. *Enactivist Interventions. Rethinking the Mind.* Oxford: Oxford University Press.

Gibson, James Jerome. 1979. *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Good, Irving John. 1965. Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers* 6:31-88.

Kurzweil, Ray. 2005. *The Singularity is Near: When Humans Transcend Biology.* London: Duckworth Overlook.

Landgrebe, Jobst, and Barry Smith. 2023. *Why Machines will Never Rule the World: Artificial Intelligence without Fear.* London: Routledge.

Larson, Erik J. 2021. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do.* Cambridge, MA: The Belknap Press of Harvard University Press.

McCulloch, Gregory. 1995. *The Mind and Its World.* London: Routledge.

Miščević, Nenad. 1988. *Radnja i objašnjenje* [Action and explanation]. Zagreb: Hrvatsko filozofsko društvo.

Oettinger, Anthony G. 1972. Preface. In: Hubert Dreyfus, *What Computers Can't Do: A Critique of Artificial Intelligence*, xi-xiii. New York: Harper and Row.

Ottinger, Richard. 2021. Körperliche Leiblichkeit als Bedingung der Erfahrungsmöglichkeit von Authentizität: Walter Benjamins Begriff der Aura, (Neue) Phänomenologie und digitale Mediatisierung. *Zeitschrift für Theologie und Philosophie* 143, no. 3:388-404.

Peirce, Charles Sanders. 1887. Logical Machines. *The American Journal of Psychology* 1, no. 1:165-170.

Potrč, Matjaž. 1993. *Phenomenology and Cognitive Science*. Dettelbach: Verlag J. H. Röll.

- - -. 1996. Phenomenology and organic unity. In: Elisabeth Baumgartner, Wilhelm Baumgartner, Bojan Borstner, Matjaž Potrč, John Shawe-Taylor, Elisabeth Valentine, eds. *Handbook Phenomenology and Cognitive Science*, 185-197. Dettelbach: Verlag J. H. Röll.

- - -. 2004. *Dinamična filozofija* [Dynamical Philosophy]. Ljubljana: Filozofska fakulteta.

Rowlands, Mark, Joe Lau, and Max Deutsch. 2020. Externalism About the Mind. In: Edward N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition). https://plato.stanford.edu/archives/win2020/entries/content-externalism/ (February 1, 2024)

Schmitz, Hermann. 2011. *Der Leib.* Berlin: Walter de Gruyter.

Sire, James W. 2002. *Izazov svjetonazora: pregled temeljnih svjetonazora* [orig. The Universe Next Door, 1997]. Zagreb: STEPress.

Turing, Alan. 1950. Computing Machinery and Intelligence. *Mind* 59, no. 236:433-460.

Zahavi, Dan. 2019. *Phenomenology: The Basics*. New York: Routledge.

Žalec, Bojan. 2023a. Ali je umetna inteligenca inteligenca v pravem pomenu besede? Vprašanje psihičnih značilnosti in splošnosti [Is Artificial Intelligence an Intelligence in the True Sense of the Word? The Issue of Mental Characteristics and Generality]. *Bogoslovni vestnik* 83, no. 4:813-823.

- - -. 2023b. *Človečnost v digitalni dobi. Izzivi umetne inteligence, transhumanizma in genetike* [Humanity in the Digital Age: The Challenges of Artificial Intelligence, Transhumanism, and Genetics]. Ljubljana: Teološka fakulteta. https://www.teof.uni-lj.si/uploads/Zalozba/ZnK86-Zalec-clovecnost_elektronska.pdf (February 1, 2024)

Octavian-Mihai Machidon

WE SHAPE AI, AND AI SHAPES US: PHILOSOPHICAL AND THEOLOGICAL CONSIDERATIONS ON AI'S ALGORITHMIC DETERMINISM

Introduction

Today, artificial Intelligence (AI) is a pervasive, transformative force impacting many domains. We are witnessing AI-powered systems bringing the promises of improved efficiency, increased productivity, reduction of costs (Aly 2020, 2-5), and, in general, higher and faster computing capabilities for any given computing task (Zhang and Lu 2021, 2-4). AI is automating and speeding up processes while also influencing decision-making. AI has become so embedded in the social fabric that we interact with it regularly, and most of the time seamlessly: in our smartphones, cars, homes, and work environments. Consequently, because of its data-driven nature, there is a high chance that many of our actions end up as input data for AI-powered systems. However, AI's pervasive and obfuscated nature and its tight coupling with people's lives may also lead to a significant social transformation potential.

Considering this potential to transform society and people, theology and philosophy can bring consistent contributions to the ongoing interdisciplinary debate on the role and dangers of AI in today's society. Theology can answer fundamental questions regarding the relationship between Imago Dei, human creativity, and the limits of AI evolution (Dorobantu 2019, 14). The link between theology and AI is also visible in the tendency of some AI researchers and advocates to resort to theological terminology when describing AI. Among other things, they assert that from a religious evolution perspective, AI is the ultimate step, playing a crucial role in the salvation of humanity (Oeming 2022, 354-355), and doing so creates a somewhat 'mystical' aura around artificial intelligence.

Another societal implication of AI's widespread use is AI-induced fear and anxiety (Li 2020, 1). Studies show that 34% of people fear AI, with 24% thinking AI will harm society (Maddox 2015). In addition, people hesitate to put their personal lives in the hands of an AI assistant, especially when that assistant makes decisions without providing transparent reasoning for choosing one solution over a set of alternatives (Polonski 2016). Like always, in the case of AI, too, people fear the unknown – which is understandable since we do not know how AI will impact society.

AI may generate vast societal impacts, aligning it with past transformative technological changes such as the industrial or agricultural revolutions (Parson et al. 2019). However, these impacts are surrounded by a veil of uncertainty. AI is credited for generating both good, positive societal changes and detrimental effects, especially since its technological building blocks are also diffuse, labile, and uncertain.

While there is some agreement on specific issues, like AI's impact on labour markets, which is almost unanimously expected to be disruptive, causing a potential increase in unemployment in specific sectors, there is significant uncertainty on other issues. How will AI impact the concentration or distribution of economic and political power on the world stage? Will it help society as a whole, and human lives in particular, flourish in diversity or make them more uniform? What impact will it have on individual liberty? Will human capabilities be enhanced or degraded because of AI? On all these points, the range of present speculation is vast.

The Relation Between AI and Society

The societal impacts of AI can be analysed within the broader context of how technology impacts society. Three philosophical perspectives on the relationship between technology and society can be identified (van de Poel 2020, 500):

- Technology as an autonomous force that determines society;
- Technology as a human construct that human values can shape;
- A co-evolutionary perspective on technology and society where neither determines the other.

The first perspective was established in the 20th century by philosophers such as Martin Heidegger, Jacques Ellul, Marshal McLuhan, and Langdon Winner. Applied to AI, this perspective is shared not only by modern techno-pessimists like Stephen Hawking and Nick Bostrom but also by techno-optimists and AI supporters, such as Frank Tipler and Ray Kurzweil (van de Poel 2020, 506).

In his work *The Technological Society*, Jacques Ellul introduced the concept of autonomous technology, i.e. technology is a closed system, "a reality in itself (…) with its special laws and its own determinations" that ultimately conquers every aspect of human society (Ellul 1967, 134). One can also say this to be the case for AI, given its widespread use across all areas of human life. For Ellul, technology and its effects on society cannot be seen as good or evil. All technology is a disruptive, self-augmenting force that engineers the world on its terms, which in the case of AI would translate to it shaping our world one way or another simply by existing. Ellul concludes that the world technology creates is "the universal concentration camp" (Ellul 1967, 100), a dark image very similar to what today's AI's harshest critics warn: that humanity will end up enslaved in a world ruled by AI (Bostrom 2002, 15-16).

Can we say about AI that it is augmenting itself, according to Ellul's theory? To a certain extent, something along these lines is currently happening in today's IT industry: AI and machine learning are hot topics, and companies are in a quest for more machine-learning solutions and AI products even if they do not fully understand them or even need them (Johansson 2019). The demand for new machine learning tools reached unprecedented heights leading to increasing requirements on behalf of the companies to have more employees dedicated to monitoring and guiding neural networks, writing scripts for chatbots, and maintaining other AI-based services. In short, AI is a 'brand' that 'sells': adding AI to the title of a product or service will make it more popular and sought after, increasing profit. This is also the case in academia, where including AI in your scientific paper will increase the chances of it getting published while blending AI into your thesis or project will contribute to getting a better grade.

This might lead to unwanted consequences, such as novel projects that can be even more impressive than those using AI getting pushed aside. Also, talented and skilled engineers and computer scientists could end up chasing the positions that offer the most money, while the next genuinely original computer science breakthrough could be pushed back to ensure we ride this current AI wave to its fullest. Meanwhile, the demand for lower-level IT experts will shrink, making it harder to find entry-level positions and unlikely to maintain a career at a high level indefinitely unless one has some niche skill-set (Johansson 2019).

In just a matter of months, we have witnessed the emergence of AI-based image generators like DALL-E 2, Midjourney, and Stable Diffusion that make it possible for anyone to create unique, hyper-realistic images just by typing a few words into a text box (Roose 2022). The Google-acquired British AI Company DeepMind Technologies Ltd. designed a program that "mimics any human voice." Along the same lines, in just two hours, an artificial intelligence algorithm called GPT-3 wrote an academic thesis on itself. The researcher who directed the AI to write the paper submitted it to a journal with the bot's consent, stating, "We just hope we did not open Pandora's box" (Getahun, 2022). This is AI self-augmentation at its finest since new and more advanced AI solutions will be required to determine if the content is human- or AI-created.

AI as an Extension of Man

Sharing with Ellul the same perspective on technology and media, Marshall McLuhan introduces a more developed vocabulary and defines any technology as an "extension of man" that ultimately and inevitably causes unforeseen cultural implications (McLuhan 1994, 7-16). People create new technologies (new 'media') to fulfil a particular intent or need. However, it was only after that technology became mainstream and widely used (often decades later) did its cultural implications (what McLuhan called its "message") become visible (McLuhan 1962, 110-111). In McLuhan's words, "the medium is the message," and thus, it can change us and our society without us being aware.

The mainstream view is that AI represents a new, enhanced form of intelligence that can improve our society. Applying McLuhan's model to AI, however, we are faced with the question: is AI a different type of intelligence, or is it extending human intelligence (Braga and Logan 2017, 2)? McLuhan states that “all media are extensions” of some human faculty – psychological or physical (McLuhan 1994, 21). These extensions are connected closely to our senses, to the human faculties they extend, and tend to shift our sensory balance outwards, from the human sensor or faculty towards the extension, leading to a form of ‘discarnation’. According to McLuhan, “when these [sensory] ratios change, men change” (McLuhan and Fiore 2005, 41).

Technology (media) extends, and consequently changes, humans through another concept McLuhan introduced: “amputation”. The ultimate unintended consequence of an extension is the numbing – going as far as an amputation – of the faculty it extended (McLuhan 1994, 42). If AI extends human intelligence, will it contribute to its decline to some extent? For example, will it cause us to lose some of our cognitive autonomy to AI, ultimately altering our perspective on the nature of the human spirit (Braga and Logan 2017, 6)?

McLuhan states, “by continuously embracing technologies, we relate ourselves to them as servomechanisms. To use them at all, we must serve these objects, these extensions of ourselves, as gods or minor religions” (McLuhan 1994, 46). He gives the example of Narcissus, who fell in love with his image reflected in the water as an analogy for people seeing a reflection of themselves in the technology they are using and ending up serving or worshipping that technology as if they were worshipping themselves.

We can see AI as the pool Narcissus looked into and fell in love with his image. AI supporters seem mesmerised by the beauty of logic and rationality to such an extent that they end up dismissing (or amputating) the remaining dimensions of the human intellect, such as the emotional, moral, or spiritual ones (Braga and Logan 2017, 6-7). AI is limited and oversimplifies the concept of intelligence. It can be viewed as a unicameral brain with a left-brain bias, missing the dynamics of emotional chemistry present in a human brain (Braga and Logan 2017, 7)

McLuhan's view on technology can be summarised as “We become what we behold. We shape our tools, and then our tools shape us” (Culkin 1967, 70). We devise AI algorithms, systems, and agents that interact with us (they ‘watch’ how we move and how we act and ‘learn’ from this, i.e. the data used for training AI systems is ‘produced’ by humans). At the same time, AI also ‘designs’ us by recommending (and thus influencing) what videos we see, what products we buy, what content we read, and so on. AI is thus converging us into our bubbles and feeding us constantly with content of their choice, shaping us in this process without us noticing it.

AI's Algorithmic Determinism

Imagine a typical day in an AI-augmented world (Polonski 2016): your AI assistant greets you with a friendly greeting before preparing your favourite breakfast. During the morning workout, it plays songs that perfectly match your taste. For the commute to work, it recommends articles for reading based on the trip duration and past reading history. At some point, a notification pops up, reminding you that elections are closing in. Next, your AI assistant recommends which candidate to vote for based on a prediction model that considers your previously expressed views and data on other voters that match your profile. It then asks through a pop-up message whether you want it to cast the vote on your behalf. You tap 'agree' and get on with your life.

Personal recommendation systems tend to "steer the user towards the content, thus ghettoising the user in a prescribed category of demographically classified content." (Polonski 2016) Going back to McLuhan, extension is followed by amputation. Hence, as AI gets to decide for us, our decision skills might get 'amputated' or at least 'numbed', leading to our personal development getting hindered.

There is an increasing tendency to rely more and more on personalised AI recommendation systems in everyday life. The more we use such systems, the more our data gets fed to them; consequently, they make better decisions for us. However, this tendency also leads such systems increasingly to shape our decisions, preferences, actions, and, ultimately, our way of living (Polonski 2016). Nevertheless, there is also the thick end of the stick; any bias in such personalised recommendation algorithms may induce or amplify our biases and deepen social divisions (Polonski 2016). Moreover, such AI algorithms use past data regarding our actions to predict, suggest or anticipate our future needs or decisions. This form of algorithmic determinism is troublesome since it reproduces established behaviour patterns, providing old answers to new questions while also impeding our natural need for experimentation and exploration to the detriment of our identity's diversity (Polonski 2016).

We can therefore ask ourselves: what role do humans play in the design of algorithms – are a creator's subconscious beliefs and biases encoded into the algorithms that make decisions about us? Most of the time, algorithmic bias originates from the data used to train such algorithms. The biased world we live in can result in biased datasets and, in turn, biased artificial intelligence frameworks. Moreover, the massive AI training data is often gathered through participatory sensing: our movements, pictures, and thoughts expressed in posts on social media all become training data for AI algorithms which, in turn, will make decisions and provide recommendations. A relevant example is Tay, the AI bot that Microsoft released on Twitter in 2016. By observing the content and interactions on Twitter and mimicking it, the bot quickly learned to be a misogynist and a racist. Microsoft had to pull the bot offline hastily (Srinivasan 2018, 107).

AI builds an algorithmic identity for its users, encompassing several dimensions, such as use patterns, tastes, preferences, personality traits, and the structure of their social graph. This digital identity is not directly based on users' personhood or sense of self but on a collection of measurable data points and the machine's interpretation thereof. The embodied user identity is replaced by an imperfect, simplified digital representation of itself in the eyes of the AI. In turn, based on the digital representation of a human user, AI is then interacting with that user: providing content, recommendations, and suggestions. Consequently, we can state that AI and humans are becoming paradigms of each other.

'Mutual Paradigms' Principle in Light of Theological Anthropology

Becoming what one beholds is not something new – in fact, it is a millennia-old principle dating back to the Old Testament times. In Psalm 115:8, the Psalmist warns those who trust idols, "Those who make them become like them; so do all who trust in them" (Ps 115:8, English Standard Version). In the book of Jeremiah, God asks Israel regarding its pursuit of idols: "What wrong did your fathers find in me that they went far from me, and went after worthlessness, and became worthless?" (Jer 2:5, English Standard Version). Christianity gave this principle an even higher understanding. "God became man that we might become God" states St. Athanasius of Alexandria in his work *On the Incarnation*. Along the same lines, three centuries later the byzantine monk and theologian St. Maximus the Confessor will affirm (in Ambigua 10) that "God and man are paradigms of each other" (Maximus the Confessor 2014a, 165), God taking bodily form in man to the extent that man deifies himself through the cultivation of virtue.

A theological analysis of the implications of Artificial Intelligence has to consider the broader discussion on the meaning of technology in the context of theological anthropology. St. Maximus the Confessor in Ambigua 45 discusses three different understandings of technology as an anthropological reality following the Fall of Man (Maximus the Confessor 2014b, 193):

- A close relationship between technology and pathos, linking man's prelapsarian apatheia (dispassion) with the lack of needing artifacts: the first man lived "a life devoid of artifice";
- Before the Fall, man was not just in harmony with the environment but also had a single need: "the unconditioned motion of the whole power of his love for what was above him, by which I mean God" and thus having no intellectual curiosities and being "wholly undistracted by any of the things that were beneath him, or around him, or orientated to him";
- The original man was perfectly and naturally virtuous and had "no need to rely on ideas discursively drawn from sensible objects to understand divine realities."

Commenting on that excerpt from Maximus, Orthodox Neo-patristic writer Fr. Dumitru Stăniloae argues that three layers are standing between man and God, which are pulling man towards those things beneath him, hindering his ascent (Maximus the Confessor 2006, 450):

- The irrational fantasies of passions;
- The principles of technical skills;
- The natural principles derived from the law of nature.

Adam, before the Fall, did not have to face these three layers, having a direct, unmediated experience of God. We now must proceed through and beyond these layers to re-establish the prelapsarian, Adamic state and relationship with God. To achieve this, Stăniloae argues that we must first recognise the irrational fantasies of passions for what they are (inconsistent mirages) and consequently dismiss them. The principles of technical skills, according to Stăniloae, "are made by man, who in turn to make them use the natural principles" (Maximus the Confessor 2006, 451). However, these 'natural principles' must become known to man "not only for the help they provide in making technological principles" but also because through them, man satisfies "his natural thirst for knowledge," which includes the knowledge of God (ibid.). Stăniloae concludes that "technology must not develop beyond the real needs of man and should not be used to harm him. Man must remain its master, and he should not be impeded by it in his ascend towards God" (ibid.).

According to Maximus, the postlapsarian world is implicitly technical: humans are bound to create and use technology, make tools that have a practical use, and "mediate and transform their experience and knowledge of the rest of creation" (Delicata 2018, 42). Based on Maximus and the interpretation of Stăniloae, the "natural principles and the principles of technical skills" are necessary until the eschaton. Humans must get to know, learn to master, and rightfully use them in their ascend towards God to fulfil their destiny – returning to the same level of closeness to God as before the Fall. However, as Stăniloae warns, a correct understanding and use of technology are mandatory, so it will meet its purpose and not become an obstacle in man's spiritual ascent. Hence, technology should mediate our relationship with the divine without separating us further from God (by discarnation and amputation of our senses, intellect, and emotions).

Technology should contribute to a better ordering of human society; however, it is a human product, and human creativity is closely linked with human freedom. Because of that, the "sphere of creative activity is so susceptible to being corrupted by sin" (Bulgakov 2002, 331). If on the personal level, the sin could be characterised by a refusal of communion with God and his divine grace, on the communitarian one, it can be described as "a broad development of creativity in its name, by a deluge of anthropotheism, in the form of a luciferian creative intoxication, and by an immersion in dull sensual paganism, still so present in our society characterised by a secular culture, atheism, and sensuality" (Ibid., 332).

According to Orthodox theologian Sergei Bulgakov, such developments cannot be overcome by mere rejection; they can be overcome only by unfolding a positive Christian doctrine of the world and creative activity and by manifesting its power (loc. cit.). Nevertheless, how to achieve this doctrine? As Maximus writes in Ambigua 45, to correctly "per¬ceive in all things the ray of true knowledge," one must first remove "all the dark fluid of passions and every material at¬tachment from their intellective eyes (Maximus the Confessor 2014b, 193). Only purified from all passions can our intellective eyes correctly relate to the principles of technical skills and contemplate "the meanings of all things encountered" (Maximus the Confessor 2006, 447). This way, we will see and use things for what they are, without fear and anxiety towards them. AI and technology would transcend from tools that can deterministically shape humans and society into means by which humans participate as co-creators in the world, fulfilling God's commandment.

Conclusion

We have seen that AI is tightly coupled with human existence, both in terms of its ubiquitous and pervasive nature in today's society and its ability to resonate with and mimic human conceptions and attitudes embedded into its training data. We shape AI through our lives since all our actions, thoughts, and emotions might be training data for AI algorithms and services. AI shapes us in turn – its algorithmic determinism and data-driven nature make recommender systems, personal assistants, and virtually any AI service influence how we think, what content we follow, and what choices we make. At the same time, AI's creative potential (its ability to generate human-like content) challenges us to reflect on the purpose and limits of human creativity. In addition, the above generates anxiety and fear about what the AI-augmented future will bring.

Considering all this, the Christian patristic tradition offers us a positive and liberating perspective to relate correctly to AI and technology and to use them for mediating our relationship with the divine. This perspective excludes fear towards AI since we are meant to master and use technology as co-creators with God and through technology to satisfy our thirst for knowledge – including knowledge of God. It offers a correct understanding of AI and human creative activity. Moreover, such a Christian understanding can thus help regulate and avoid AI usage 'in its own name'. It can lead to limiting AI's algorithmic determinism by following the Christian exhortation of committing ourselves to the path of deification by grace and the practice of virtues. This is important since human nature – its passions and biases – are reflected in the data we produce and, ultimately, in how AI algorithms behave. Finally, such a perspective on AI can lead us to a more in-depth understanding of our anthropology and relation to God in light of the 'mutual paradigms' of Man/AI and God/Man.

References

Aly, Heidi. 2020. “Digital transformation, development and productivity in developing countries: is artificial intelligence a curse or a blessing?” *Review of Economics and Political Science.*

Bostrom, Nick. 2002. “Existential risks: analyzing human extinction scenarios and related hazards.” *Journal of Evolution and Technology* 9.

Braga, Adriana, and Robert K. Logan. 2017. “The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence.” *Information* 8 (4). https://doi.org/10.3390/info8040156.

Bulgakov, Sergius. 2002. *The Bridge of the Lamb.* Translated by Boris Jakim. N.p.: Eerdmans Publishing Company.

Culkin, John M. 1967. “A schoolman's guide to Marshall McLuhan.” *Saturday review* (New York), March 18, 1967.

Delicata, Nadia. 2018. “Homo technologicus and the Recovery of a Universal Ethic: Maximus the Confessor and Romano Guardini.” *Scientia et Fides* 6 (2): 33-53. http://dx.doi.org/10.12775/SetF.2018.020.

Ellul, Jacques. 1967. *The Technological Society.* New York: Vintage Books.

Dorobantu, Marius. 2019. “Recent advances in Artificial Intelligence (AI) and some of the issues in the theology & AI dialogue.” *ESSSAT News & Reviews* 29 (2): 4-17.

Getahun, Hannah. 2022. “Artificial Intelligence Bot Wrote Scientific Paper in 2 Hours.” *Insider*. https://www.insider.com/artificial-intelligence-bot-wrote-scientific-paper-on-itself-2-hours-2022-7.

Johansson, Anna. 2019. “How AI and Machine Learning Are Affecting the Computer Science Industry.” *IEEE Computer Society.* https://www.computer.org/publications/tech-news/trends/how-ai-and-machine-learning-are-affecting-the-computer-science-industry.

Li, Jian. 2020. “Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory.” *Technology in Society* 63 (November). https://doi.org/10.1016/j.techsoc.2020.101410.

Maddox, Teena. 2015. “Research: 63% say business will benefit from AI.” *TechRepublic.* https://www.techrepublic.com/article/research-63-say-business-will-benefit-from-ai/.

Maximus the Confessor. 2006. *Ambigua*. Romania: Editura Institutului Biblic şi de Misiune al Bisericii Ortodoxe Române.

– – –. 2014a. *On Difficulties in the Church Fathers: The Ambigua*. Edited by Maximos Constas and Nicholas Constas. Translated by Nicholas Constas and Maximos Constas. N.p.: Harvard University Press.

– – –. 2014b. *On Difficulties in the Church Fathers: Ambigua to John, 23-71*. Edited by Maximos Constas and Nicholas Constas. Translated by Maximos Constas and Nicholas Constas. N.p.: Harvard University Press.

McLuhan, Marshall. 1962. *The Gutenberg galaxy; the making of typographic man.* Edited by University of Toronto. N.p.: University of Toronto Press.

– – –. 1994. *Understanding Media: The Extensions of Man.* N.p.: Cambridge, Mass.

McLuhan, Marshall, and Quentin Fiore. 2005. *The Medium is the Massage: An Inventory of Effects.* Edited by Jerome Agel and Quentin Fiore. N.p.: Gingko Press.

Oeming, Manfred. 2022. "Intelligentia Dei: Artificial Intelligence, Human Reason and Divine Wisdom." In *Intelligence - Theories and Applications*, edited by Rainer M. Holm-Hadulla, Joachim Funke, and Michael Wink. N.p.: Springer International Publishing. https://doi.org/10.1007/978-3-031-04198-3_21.

Parson, Edward, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant, and Nick Novelli. 2019. "Could AI drive transformative social progress? What would this require?" *AI Pulse.* https://aipulse.org/could-ai-drive-transformative-social-progress-what-would-this-require/

Polonski, Slava. 2016. "Algorithmic determinism and the limits of artificial intelligence." *Medium*. https://medium.com/@slavaxyz/algorithmic-determinism-and-the-limits-of-artificial-intelligence-d32397b8f618.

Polonski, Vyacheslav W. 2016. "Algorithmic determinism and the limits of artificial intelligence." *OII*. https://www.oii.ox.ac.uk/news-events/news/algorithmic-determinism-and-the-limits-of-artificial-intelligence/.

Roose, Kevin. 2022. "A.I.-Generated Art Is Already Transforming Creative Work." *The New York Times*. https://www.seattletimes.com/business/technology/art-generated-by-artificial-intelligence-is-transforming-creative-work/.

Singh, Amita. 2018. "In search of regulations and legal personhood." In *Human decisions: thoughts on AI*, 124-129. N.p.: United Nations Educational, Scientific and Cultural Organization.

Srinivasan, Rajeev. 2018. "The ethical dilemmas of Artificial Intelligence." *In Human decisions: thoughts on AI,* 104-107. N.p.: United Nations Educational, Scientific and Cultural Organization.

van de Poel, Ibo. 2020. "Three philosophical perspectives on the relation between technology and society, and how they affect the current debate about artificial intelligence." *Human Affairs* 30 (4): 499-511. 10.1515/humaff-2020-0042.

Zhang, Caiming, and Yang Lu. 2021. "Study on artificial intelligence: The state of the art and future prospects." *Journal of Industrial Information Integration* 23 (September). https://doi.org/10.1016/j.jii.2021.100224..

Vojko Strahovnik

TRANSPARENCY OF AI SYSTEMS AND HUMAN JUDGEMENT: RESPONDING TO THE DOUBLE-STANDARD ARGUMENT

Introduction

The transparency issue in artificial intelligence (AI) and machine learning (ML) revolves around the issue that we have a limited understanding of how specific AI algorithms make decisions or generate recommendations. This fundamental problem results in humans being unable to track, comprehend, or scrutinize the decisions made by such systems. This aspect of AI systems has evoked various reactions, one of which suggests that these decision-making systems should adhere to the same criteria, norms, and standards that apply to human decision-makers, with transparency being a crucial factor. Therefore, the options are either to enhance the transparency of these systems or to prohibit their use. Conversely, there is a counterargument known as the double standard argument, asserting that demanding extensive transparency for AI-based decision-making systems is unfair or unjustified because human judgment also lacks transparency. This article presents a proposed response to the latter argument. The paper begins by first outlining the transparency problem. Next, it elaborates on the double standard argument (Section 3), while Section 4 deals with models of human transparency and elaborates on an alternative view of it. In conclusion, the relevance of the model of transparency called chromatic transparency is highlighted.

Outlining the Transparency Problem

In machine learning, the ‘transparency’ or ‘opacity’ problem concerns the fact that we supposedly have no significant insight into the functioning and, consequently, the decision-making processes of algorithms. Such algorithms are given a certain amount of input data and use it to produce an output, i.e. some concrete classification. More complex algorithms are opaque in the sense that when we receive their output (the classification decision), we do not know how and why a certain classification was produced or selected. A concrete example could be the following: it is practically impossible to understand an algorithm’s decision-making process for rejecting a request for a bank loan. We often cannot know what weight was attributed to individual features or attributes on the basis of which a decision was made and, consequently, why a machine has decided as it did. This is particularly true if the recommendation-making system is provided with a broad range of data about the loan applicant and was also previously trained on a large amount of data. In order to understand the broader ethical and legal aspects of such non-transparency, we must first consider the following question: What are the main reasons why (at least some such) algorithmic classifications are non-transparent?

As multiple reasons can be identified, we also have multiple types of opacity. Jenna Burrell introduces three basic forms, but it seems relevant that they all refer to situations where we're dealing with the opacity of classification mechanisms that have concrete social consequences (spam filters, search engines, insurance or loan qualification, credit scoring, security threat detection, etc.) and that, needless to say, rely on computational algorithms and machine learning algorithms. The mentioned three forms are:

1. Opacity as intentional corporate or state secrecy.
2. Opacity as a consequence of technical illiteracy.
3. Opacity emerging on the characteristics of machine learning algorithms and the scale required to apply them usefully (Burell 2016).

The first type of algorithmic opacity is entirely intentional and can serve, for instance, as self-protection by the corporation's intent on maintaining its trade secrets and competitive advantage (e.g. search engines and advertisement) or the state's interest in pursuing other aims and values (e.g. national security and the safety of its citizens). In addition, opacity is essential in some fields, for instance, in a network security application of machine learning which deals with spam, scams, and fraud, since such systems must remain opaque to a certain extent if they are to perform their task. One of the problems is that opacity not only gives corporations a competitive advantage in the market but also provides them with a means possibly to circumvent regulation and manipulate consumers.

Opacity as technical illiteracy occurs because code writing and the design of algorithms are specialised skills inaccessible to the majority of the population. Codes are implemented in various programming languages and systems, whose specialised task is a precise description of data manipulation within a mathematically rigorous interpretation of data. This strictness is vastly different from the human language that is used in less formalised everyday settings. Thus, for data models of sufficient complexity, even an educated expert cannot verify all the aspects of the code without investing effort comparable to, or even exceeding, the effort of developing the code. A code is good if it can be interpreted by both humans and computers, but this linguistic mediation requires a specific education. Consequently, even if codes and algorithms are openly accessible to the public, only a few can make sense of them.

The last type of opacity stems from the fact that machine-learning algorithms often prove difficult, even for those who program them. What causes these difficulties is not only the length of codes, the number of people writing a code, and the multitude of interlinkages between modules and subordinates but also the continuous changes in the logic of an algorithm's decision-making process as it learns from training data, which makes tracking and interpretation extremely difficult. An algorithm, therefore, becomes a 'black box' that we cannot simply open to see what is inside. "While datasets may be extremely large but possible to comprehend and code may be written with clarity, the interplay between the two in the mechanism of the algorithm is what yields the complexity (and thus opacity)." (Burell 2016, 5). Deep learning models frequently operate as 'black boxes', which means that the mechanisms through which they make decisions are not easily understood by humans. The models discover patterns and relationships from enormous amounts of data, but extracting a clear explanation of how or why a certain choice was reached can be difficult. Many AI models operate on high-dimensional data, making identifying the individual qualities or elements that drove a single choice difficult. Also, AI models rely significantly on the data on which they are trained. If the training data is biased or incomplete, the AI system may make decisions that are prejudiced or fail to account for specific conditions. Identifying and correcting biases is challenging because of the lack of openness in the decision-making process. Lastly, AI models can be extremely complex, with millions or billions of parameters. Understanding how each parameter influences the decision-making process can be difficult, making it difficult to explain the model's functioning.

In the paper, I focus on the issues of transparency or opacity of AI in this latter meaning and highlight some of the ethical conundrums that this aspect of AI raises. I particularly focus on the comparison between AI and human judgment and decision-making, investigating whether the AI decision-making process really lacks transparency and whether a double standard is involved in the calls for more comprehensive transparency in AI systems. (Zerilli et al. 2018) One background presupposition is that such AI-based systems should be transparent. Bostrom and Yudkowsky outline broader ethical concerns related to AI and deep learning as related to social norms in the following way: "Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to replace human judgment of social functions; all criteria that may not appear in a journal of machine learning considering how an algorithm scales up to more computers. This list of criteria is by no means exhaustive, but it serves as a small sample of what an increasingly computerised society should be thinking about." (Bostrom & Yudkowski 2014, 2).

There are many ethical problems regarding the opacity of algorithmic decision-making. One major ethical problem is, for example, the problem of trust. Opacity can, in fact, disguise discrimination, as would happen if an algorithm for loan application assessment assigned unjustified significance to attributes such as race, ethnicity, or religious belief, and we actually would not have any knowledge of it. Proposals on how to address the problems with the first two types of opacity have already been put forward. When it comes to intentional opacity, they usually entail oversight and regulation of codes by a 'trusted auditor' that would maintain the algorithm's secrecy on the one hand and act in the public interest on the other. As for opacity as technical illiteracy, a possible solution given by Burrell would be to incorporate the development of computational thinking into all levels of education. This would give the public the ability to more easily assess and criticise mechanisms that directly affect their real-life opportunities and possibilities.

No tenable solutions have yet been proposed for the third type of opacity, even though many of the AI community have been exploring the methods of Explainable AI (XAI), which would replace or amend the 'black box' algorithms currently in use. Simpler (and consequently more transparent) models of machine learning do exist, but there is a significant correlation between the accuracy or success of classificatory decision-making and the complexity of an algorithm. If a machine-learning model is to be useful and practical for accurate classification, it will inevitably be highly complex. In our desire to reach the highest possible precision of high-stakes decision-making, which has a direct impact on our lives, and to simultaneously preserve some degree of transparency, which makes it possible to avoid many forms of discrimination, we must, for the time being, search for the right balance between the two extremes. This alone opens up many ethical and legal questions we can only face if we first try to understand the models of decision-making processes themselves – both in humans and in machines.

In what follows, I predominantly use the term transparency for the central problems we are discussing. There are several other terms in its vicinity; terms such as opacity, which more or less point to the same phenomenon and which I use interchangeably with the first. Next, one can also discern other dimensions that either cross-cut or further elaborate on the initial notion, e.g. explainability, interpretability, intelligibility, understandability, clarity, traceability, legibility, surveyability, auditability, improbability, responsibility, etc., which raise more specific problems.

The Double-Standard Argument

In this section, I will first briefly consider the 'double-standard' argument that is being made in defence of AI algorithms (Zerilli et al. 2018.). This argument is based on the premise that we should not set higher standards for AI decision-making than we do for human decision-making since that would be a double standard that is (in the absence of other relevant considerations) indefensible. If we combine with this the social intuitionist model of human judgment and decision-making, the conclusion is clearly that we should not set the bar for AI so high as to include full transparency, explicability, or even absence of bias.

The double-standard argument in defence of AI

> *Premise 1.* Human decision-making is not transparent or lacks proper transparency.
> *Premise 2.* We should not set higher standards for AI decision-making than we do for human decision-making since that would be a double standard, that is, in the absence of other relevant considerations, unjustified.
> *Premise 3.* There are no relevant considerations that would support a double standard.
> *Conclusion.* We should not expect full transparency from AI or worry excessively about the transparency problem in AI.

I agree with the double-standard presupposition (P2 and P3) but disagree with the social intuitionist model as the benchmark for transparency of judgment and decision-making. The model is too pessimistic. Besides the classical rationalist model and social intuitionist model, there is a third alternative view on human judgment and decision-making that we see as more feasible than both of the mentioned alternatives. I now turn to these issues.

Models of Human Transparency

Let me begin this section with a clarificatory remark. For the purposes of this paper, I will understand decision-making in a very broad sense, including, for instance, the formation of belief, the formation of judgments (including moral judgments and other normative judgments), and the formation of practical judgments that are related to our intentions and actions. Although one can argue that some of these do not include a fully fledged process of decision-making or are not agentive at all, I think that there are good reasons to view them as part of such a broader class of decision (e.g. it is certainly true that I do not 'decide' what to believe (for example, that there is a cat sleeping on the chair next to me), but it is also true that I base my beliefs on reasons or what I regard as evidence for my belief; therefore my belief is based on reasons and is not merely a result of some involuntary causal process. I do not just find myself believing that there is a cat on the chair near me out of the blue, so to speak) (Horgan, Potrč and Strahovnik 2018). Transparency and explainability will thus be positioned in terms of reasons and rationality.

Traditional Rationalism

The traditional model of judgment and decision-making is a rationalist one (Audi 2006; 2013). According to this model, decision-making is closely connected with the process of reasoning or deliberation, which is understood as an inferential process that proceeds from one set of propositions (data) to another by means of deductive or (broadly) inductive steps and results in judgment. A judgment or a decision could thus be understood as an outcome of such a reasoning or deliberation process in which one explicitly represents general considerations (e.g. one's goals) or principles and facts about the particular case at issue. What is crucial for this model is the stress on reasoning, and thus the model offers a similar view regarding the justification of judgments or decisions, which is understood as the giving or stating of reasons that were operative in the formation of the initial judgment. Transparency, according to this model, is thus a kind of two-fold transparency; first, in making the judgment or decisions, the reasons are explicitly represented in the very process of a decision-making system, and second, in providing the explanation, these reasons can be, at least in principle, clearly laid out for others to survey (at least reasons that were most significant).

Social intuitionism, modularity, and hybrid models

This rationalist model came under heavy attack in recent decades, especially from the so-called social intuitionism model of judgment (as formulated for the domain of moral judgment and moral decision-making by Jonathan Haidt (2001; cf. Haidt and Bjorklund 2008) and by proponents of the so-called modular conception of mind (Carruthers 2006; Mercier and Sperber 2017). According to this model, a substantive part of our judgment and decision-making is intuitive or affect-laden, which means either based on our emotions or otherwise a result of cognitive operations of various mental modules, heuristics, etc. The resulting judgment is thus not a result of the process of reasoning; on the contrary, 'reasoning' in the sense of providing reasons occurs only *post hoc* when one is trying to provide an explanation or justification for one's judgment or decisions to others. The latter process is thus essentially a *post hoc* confabulation or fabrication of reasons that did not play any role in the formation of a judgment or decision. Transparency, according to this model, therefore, cannot even be framed in terms of reasons; it can only be framed in terms of our understanding of the underlying causal processes that lead to a judgment or decision.

Consider, for example, the following quote from Mercier and Sperber that nicely illustrates the spirit of the mentioned models. "Whether or not it would be better to be guided by reasons, the fact is that in order to believe or decide something, we do not need to pay any attention to reasons. Purely intuitive inference, which generates so many of our beliefs and decisions, operates in a way that is opaque to us." (Mercier and Sperber 2017, 114). The social intuitionist model of moral judgment includes the following fundamental tenets: Most moral judgments are the consequence of moral intuition at work. Moral thinking is not typically part of the process of developing moral judgments (at least not in circumstances where moral judgments result from the operation of moral intuition). Individuals who attempt to give reasons for moral judgments they make tend to follow a 'makes-sense' moral script, offering what they believe are the considerations that led them to their judgment but which actually played no role in producing the moral judgment. They confabulate that, "moral reasoning does not cause moral judgment; rather, moral reasoning is usually a post hoc construction, generated after a judgment has been reached" (Haidt 2001, 814; cf. Horgan and Timmons 2007).

The two models of reasoning have their counterparts in the approaches to AI decision-making. Both can be illustrated by their application of credit scoring, which also opens fundamental questions for each model of decision-making. The reasoning model can be paralleled by exact algorithms of Turing machines and rule inference of symbolic computation. There, a set of predetermined rules is applied to a symbolic representation of inputs to derive new data consistent with the rules and the input data. In this model, transparency can be achieved on two levels: algorithmic transparency is achieved by providing a set of operative rules. Explanatory transparency can be achieved by stating the complete list of steps the algorithm has performed, thus enabling everyone to examine the correctness of each of them. The intuitionistic model can be seen as parallel to statistical and probability-based AI algorithms. They can be illustrated by classification using support vector machines and more advanced neural networks. The key method in such algorithms is to represent known data instances as points in a highly-dimensional space and then use a function of many variables in this dimensional space to separate the points of one class from the points of other class(es). In the crediting example, each input data attribute is represented by one dimension of the representation space, and each past credit recipient is represented by one point corresponding to the combination of values of its attributes. Neural networks are obtained by combining such classifiers, thus producing new attributes and feeding them into new layers of classifiers, which are fed into further layers and so on, until a complex network of classifiers is obtained. Due to its high dimensionality, the whole space in which creditors are representable can hardly be displayed efficiently. Even less can these be explained when neural networks are used, as the number of dimensions of such space is beyond human comprehension.

Setting the Standards in Line with the Transparency of Human Decision-making

As stated above, I agree with Premise 2 and Premise 3 of the double-standard argument but disagree with the social intuitionist model as the benchmark for transparency of judgment and decision-making. The model is too pessimistic. Besides the classical rationalist model and social intuitionist model, there is a third alternative view on human judgment and decision-making that is more feasible than both of the mentioned alternatives. It was initially developed in the field of epistemology and based on debates arising out of connectionism (Henderson, Horgan, and Potrč 2019) and the lessons of the frame problem (Fodor 1976; 1983; 2000). I will use the label 'chromatic rationalism' for this model. Chromatic rationalism claims that judgment and decision-making follow a dynamical model of reasons, according to which (some) reasons are situated in an agent's structured or morphological cognitive background, illuminating the judgment from this background. This background is often characterised in terms of morphological content. "Morphological content is embodied in a cognitive system's persisting structure, rather than being occurrently represented; … morphological content is best understood not in terms of the physical structure of the physical network, but in terms of the persisting structure of the high-dimensional dynamical system that the network subserves – i.e., the activation landscape" (Horgan and Tienson 1996, 165; cf. Horgan and Potrč 2010). In forming a judgment or making a decision, we are sensitive to a rich body of background information, not all of which is represented in consciousness at the time of making a decision (pace classical rationalism), but which nonetheless figures in judgment (pace social intuitionism).

The latter fact is reflected in our conscious experience as a phenomenon of chromatic illumination, i.e. an aspect of phenomenology that grounds judgment or decision and its subsequent understanding. The space does not permit me to elaborate the model in full, but its core is that while human agents are sensitive to reasons, not all of these reasons are represented in consciousness at the time of judgment or decision and thus also cannot be a part of an explicit and tractable reasoning process which would be fully transparent. The process of giving reasons thus cannot be reduced to mere *post hoc* confabulation or fabrications of reasons, as the social intuitionist model would claim. It is more of a process of spelling out the reasons that were operative at the time of making a judgment or decision.

In relation to AI recommendation-making systems, implementation would be along the following lines: Suppose a bank would use an artificial neural network to decide about credit applications. To test it, an analyst could produce a sample of random creditors with believable combinations of input attributes, observe the decisions of the neural network, and feed them into a rule inference algorithm that would convert the neural network decisions into an observable form. This would be a *post hoc* fabrication of reasons. The spelling out of the actual reasons would look into the topology of the last few neurons of the network, observe which trigger each other, and, in a similar manner as proposed earlier, explain each relevant neuron being triggered. This would allow one to understand and inspect the neural network decision process and infer the rules from the topology and the data it used.

Conclusion

In answering the challenge of the double-standard argument, I have proposed a model of human judgment called chromatic rationalism as the new standard for AI transparency. Human agents are sensitive to reasons, but not all of these reasons are overtly represented in consciousness at the time of judgment or decision and thus also cannot be a part of an explicit and tractable reasoning process that would be fully transparent. The process of giving reasons thus cannot be reduced to mere post hoc confabulation or fabrications of reasons, as the social intuitionist model would claim. It is more a process of spelling out the reasons that were operative at the time of making a judgment or decision. Transparency, according to this model, is thus possible, but less than full transparency in the sense that one would be able to follow or subsequently audit all considerations that were operative in a given judgment or decision. Some of these reasons are not part of the explicit occurrent content of a cognitive system but are forming the cognitive background. They illuminate the judgment or decision in terms of enabling it and sustaining it (phenomenology) as well as being the basis for answering (disposition) the probe questions that pertain to the case at hand. Nevertheless, even with probe questions, some of the relevant considerations (and the way they combine) cannot be explicitly brought to light. If the model just described is the right one to go with, we have also established a plausible alternative standard for transparency of AI decision-making and one that avoids the 'permissive' conclusion of the double-standard argument. Under this approach, intuitive and opaque AI methods would be used to generate decision proposals. These proposals would then be subjected to explainable AI techniques to provide explanations, verifiable validation, and adherence to transparency standards. If the two approaches yield conflicting decisions, the disparity can be researched, and the explainable algorithm can be updated to implement the updated decision. Alternatively, the explainable, transparent, and agreed-upon decision can override the opaque decision of the intuitionistic AI. This proposal aims to combine the strengths of both approaches, leveraging the statistical power of intuitionistic AI while making their results comprehensible and reasoned by humans. In future research, we will explore the potential implications of this approach, particularly regarding the implementation of anti-discrimination rules within this regulatory model.

Funding acknowledgement

The publication was prepared as part of a research program *P6-0269 Religion, ethics, education, and challenges of modern society*, research program *The intersection of virtue, experience, and digital culture: Ethical and theological insights*, and research project *P6-24684 Theology, digital culture and challenges of human-centred artificial intelligence*, which are all financially supported by the Slovenian Research and Innovation Agency, and as part of research project *Epistemic identity and epistemic virtue: Human mind and artificial intelligence*, which is supported by the John Templeton Foundation (New Horizons for Science and Religion in Central and Eastern Europe project).

References

Audi, Robert. 2006. *Practical Reasoning and Ethical Decision*. London: Routledge.

– – –. 2013. *Moral Perception*. New Jersey: Princeton University Press.

Bostrom, Nick and Yudkowsky, Eliezer. 2014. The Ethics of Artificial Intelligence. In: *Cambridge Handbook of Artificial Intelligence* / Frankish, K., Ramsey, W. (ed). New York: Cambridge University Press, pp. 316–334.

Burrell, Jenna. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society,* January–June 2016: 1–12. https://doi.org/10.1177/2053951715622512

Carruthers, Peter. 2006. *The Architecture of the Mind. Massive Modularity and the Flexibility of Thought.* New York: Oxford University Press.

Fodor, Jerry. 1976. *The Language of Thought.* Sussex: Harvester Press.

– – –. 1983. *The Modularity of Mind: An Essay on Faculty Psychology. Cambridge*, MA: MIT Press.

– – –. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology.* Cambridge, MA: MIT Press.

Haidt, Jonathan. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108: 814–834. https://doi.org/10.1037//0033-295X.108.4.814

Haidt, Jonathan and Bjorklund, Fredrik. 2008. Social Intuitionists Answer Six Questions About Morality. In: W. Sinnott-Armstrong, ed., *Moral Psychology*, Vol. 2. The cognitive science of morality: Intuition and diversity. Oxford University Press, pp. 181-217.

Henderson, David, Horgan Terry and Potrč, Matjaž. 2019. Morphological Content and Chromatic Illumination in Belief Fixation. In: Chan, T. & Nes, A. *Inference and Consciousness.* New York: Routledge, pp. 229–252.

Horgan, Terry and Potrč, Matjaž. 2010. The Epistemic Relevance of Morphological Content. *Acta Analytica* 25: 155–173.

Horgan, Terry, Potrč, Matjaž and Vojko Strahovnik. 2018. Core and Ancillary Epistemic Virtues. *Acta Analytica* 33(3): 295–309.

Horgan, Terry and Tienson, John. 1996. *Connectionism and the Philosophy of Psychology.* Cambridge: MIT Press.

Horgan, Terry and Timmons, Mark. 2007. Morphological Rationalism and the Psychology of Moral Judgment. *Ethical Theory and Moral Practice*, 10: 279–295. https://doi.org/10.1007/s10677-007-9068-4

Mercier, Hugo and Sperber, Dan. 2017. *The Enigma of Reason.* Cambridge: Harvard University Press.

Zerilli, John, Knott, Alastair, Maclaurin, James and Gavaghan, Colin. 2018. Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology,* 32: 661–683. https://doi.org/10.1007/s13347-018-0330-6

Martin Justin

IS EXPLAINABILITY NECESSARY FOR TRUSTWORTHINESS?

Introduction

In recent years, we have witnessed a leap forward in the advancement of artificial intelligence (AI) technologies.[1] These technologies have proven useful in solving complex problems in various domains, both in science and society. For example, AI systems have proven to be very good at predicting the risk of lung cancer (Ardila et al. 2019) and pneumonia from CT scans (Harmon et al. 2020). Similar systems have also been used to predict successfully future health complications based on patient records (Rajpurkar et al. 2022). AI systems have significantly accelerated otherwise very time-consuming experimental approaches to the discovery of novel protein structures (Jumper et al. 2021). Beyond science, in the United Kingdom and the United States, AI systems are already being used in the judicial system (Burgess 2018). In the United States, some universities are using similar technology to identify potential prospective students (Selingo 2017).

However, the use of these technologies can sometimes be fraught with danger. For example, a high-profile investigation by a journalism team at ProPublica found that COMPAS, a system used in some US states to predict the likelihood of recidivism, systematically discriminated against black defendants (Angwin et al. 2016). More generally, many researchers, philosophers and scholars point to the fact that AI systems are often so complex that even their designers do not fully understand how they work. The use of such systems to make decisions that ultimately affect individuals thus runs counter to our institutional and ethical standards of transparency and accountability (Strahovnik, Miklavčič, and Centa 2020; von Eschenbach 2021; Chatila et al. 2021; Mittelstadt et al. 2016). Strahovnik, Miklavčič, and Centa (2020) clearly explicate some central ethical issues in the use of AI systems in society: Who is responsible when an AI system makes a mistake? Can we trust an opaque system? Should citizens have a right to reject a decision made by an opaque system? As they point out, these debates also raise an important conceptual question: What is the relationship between the notions of opaqueness, trust, and accountability (Strahovnik, Miklavčič, and Centa 2020, 323–25)?

In this essay, I will focus on the problem of trust when using opaque AI systems. Several authors have argued that the use AI systems cannot be

[1] By AI systems, I mean computational models capable of performing tasks that require intelligence in humans. Much of the current progress in artificial intelligence is based on advances in machine learning methods, in particular deep neural networks (Bringsjord and Govindarajulu 2022).

transparent and trustworthy unless we understand how these systems work (Strahovnik, Miklavčič, and Centa 2020; Balasubramaniam et al. 2022). In other words, they argue that the explainability of AI systems is a necessary condition for the transparency and trustworthiness of the decision-making processes that involve these systems. In contrast to this position, I will try to show that explainability is not a necessary condition for a transparent and trustworthy use of AI systems.

The essay is organised as follows: In the next section, I will explain in more detail the notions of transparency, opacity, explainability and trust. In the third section, I will present a reconstruction of the argument that transparency requires explainability. In section 4, I will present two counterarguments to this, one existing and one novel. The novel counterargument highlights ways of achieving transparency in decision-making processes that do not require that we explain the underlying AI systems themselves.

Transparency and the Problem of Trust

AI systems can be opaque in many ways (Strahovnik, Miklavčič, and Centa 2020, 323). The underlying software code may be protected as intellectual property and thus unavailable to the public. Alternatively, since the developers frequently rely on existing libraries and modules when writing software code, AI systems may also be opaque in the sense that no single individual involved in their development has detailed knowledge of all the software (Durán and Jongsma 2021). In this essay, I will leave these external sources of opacity aside and focus on the fact that these systems can also be inherently opaque to human users. As researchers in the field of explainable AI have noted, current AI models are so complex and use such large data sets that even their developers cannot explain the precise mechanisms of their operation (Linardatos, Papastefanopoulos, and Kotsiantis 2020). More specifically, this inability of understanding expresses itself primarily as an inability to answer two questions: (1) Why does this input lead to this output? (2) What information does the system contain? (Gilpin et al. 2018). By opacity of AI systems, I mean the fact that people can fruitfully develop and use these systems without having a precise understanding of how they work.

Opacity is problematic for multiple, both normative and epistemological, reasons. In what follows, I will focus on one of them, namely the relationship between opacity and trustworthiness. Although different authors agree about the existence of this problem, they often disagree even about the basic terms. Vereschak, Bailly and Caramiaux (2021), in their review of existing empirical research on trust in AI systems, attempt to develop a definition of trust that will summarise all the different aspects of this phenomenon. They arrive at a definition of trust as "an attitude that an agent [trustee] will achieve an individual's [trustor's] goal in a situation characterized by uncertainty and vulnerability" (Vereschak, Bailly, and Caramiaux 2021, 10). One the other hand, Chatila et al. (2021) take a completely different approach and define a new concept of "technological trust."

According to their definition, trustworthy technological solutions should be accessible, reliable, secure, maintainable, should protect data and maintain integrity. Hatherley (2020) takes a third approach, arguing that AI systems are not moral agents and thus cannot be objects of trust.

In this chapter, I will adopt the classical analysis of trust as reliance plus some additional elements (McLoad 2022). This account is motivated by an intuitive difference between relying on someone and trusting him. Let us look at an example.[2] Imagine I am graduate student writing my thesis. I rely on my supervisor to answer my emails regularly and, say, suggest me some relevant literature. If she does not meet these expectations, I might be disappointed, but I would not see it as an injustice. At the same time, I trust my mentor not to plagiarise my work and publish it under her own name. If she were to do that, I would rightly feel as if an injustice had been done. This intuitive distinction is also apparent in other contexts. I rely on Google Maps to guide me to my destination accurately, but I trust my friends to give me honest advice. I rely on my doctor to prescribe me the right medicine, but I trust the European Medicines Agency to make sure that these medicines are safe and effective, etc.

There are several accounts of trust that attempt to conceptualise and clarify this distinction. They can be roughly divided into two groups: (1) motives-based theories and (2) non-motives-based theories (McLoad 2020). Motives-based theories say that someone is trustworthy only if his both willing to help us and has the right motives to do so. One such motive, frequently discussed in the literature, is ‘good-will’: a trustee must be motivated by the good-will toward the trustor to be deemed trustworthy. In the non-motives-based theories, on the other hand, it is essential for trust that the trustor takes a particular stance or holds certain expectations that bind the trustee in a normative way. An example of such theory is the theory of normative expectation. This theory argues that trust differs from reliance in the fact that the trustor takes a particular stance towards the trustee, i.e. has some normative expectations of himm.

In this essay, I will not argue for any specific theory of trust. Since I am interested in the relationship between trust and transparency, I only need to show that transparency is a requirement for trustworthiness. I will show that is the case both under motives-based and non-motives-based theories. Let us focus first on the motives-based theories. Here, the link between transparency and trust is obvious: to know whether someone is trustworthy, I need to know his motives. The situation is a bit more complex for non-motives-based theories. But here too the relationship is clear enough. The possibility of trust here is essentially conditioned by “what the trustor believes he ought to be able to expect from [the trustee]” (McLoad 2020). So, if I want to trust someone, I need to be able to determine whether he can meet my expectations in an appropriate way. An example can help us understand what this means. Let us say I trust a judge to give a fair judgment. Under the non-motives-based accounts, I do not need to know or approve of her

[2] This example is based on Townley and Gardield (2013).

motives to think that she is trustworthy. I might even find her motives objectionable. Perhaps I think that she is motivated to give her judgement primarily by a desire to take a lunch break. Nevertheless, I can trust her if I believe that I ought to be able to expect her to give a fair judgement. This belief can only be based on the fact that she is a qualified judge and that judges are required to make such fair judgements. Alternatively, it can be based on my knowledge of her track-record in judging similar cases. In any case, some things about the judge – either her qualifications or her track-record – must be transparent to me.

Argument from Opacity to Untrustworthiness

Some degree of transparency is therefore a necessary condition for trust. This, of course, poses a problem for the use of opaque AI systems to make decisions that affect people's lives. Take, for example, the use of AI in the legal system. Say a judge uses an AI model that predicts the likelihood of recidivism in defendants. The system takes the defendant's record as input and returns an estimate of the likelihood of re-offending as output. The judge can take this score into account in her final judgement. We know that the input data of the model includes race and place of residency of the defendant. We also know that due to unjust societal conditions, this data correlates with the likelihood of re-offending. It is possible than that the AI model uses the information about race and place of residency in determining the likelihood of re-offending. But making the decision in this way would be unfair. To be able to trust the AI model, we thus need to know that it does not rely on race or any other such personal circumstances. But if the model is opaque, we cannot know that. Therefore, we cannot trust such a model and, consequently, the judicial process that relies on it.

It seems, therefore, that decision-making processes involving AI systems are necessarily untrustworthy if these systems are opaque. This line of reasoning can be summarised by the following argument:

- If a decision-making process relies on an opaque AI system, then this decision-making process is itself opaque, i.e. not transparent.
- Some AI systems, embedded in decision-making processes, are opaque.
- If a process is opaque, i.e. not transparent, it is not trustworthy.
- Thus: The decision-making processes that rely on opaque AI systems are not trustworthy.

For now, I will assume that premise (1) is correct; I will say more about it below. Premise (2) is an empirical fact which I assume is true. Premise (3) is the results of the above analysis of the concept of trust. The conclusion follows deductively from the premises. If this argument is correct, then explaining AI systems embedded in decision-making processes is a necessary condition for making these processes transparent and, consequently, a necessary condition for making them trustworthy. Notice that it does not show that it is a sufficient condition since premises (1) and (3) are implications, not equivalences. Nevertheless, I take this to be a strong argument in support of the view that explaining AI systems is an important step in ensuring that these systems are used in an ethical way.

Counterarguments

An Existing Counterargument

The above argument seems quite convincing. However, it is not universally accepted. In this section, I will present one existing counterargument, presented by Zerilli et al. (2019). This counterargument denies that premise (2) is correct in a meaningful way. In essence, Zerilli et al. (2019) argue that the reasoning of human decision-makers is also largely opaque. Since we consider some human decision-makers as trustworthy, requiring that we explain in detail the workings of AI systems as a necessary condition for trustworthiness constitutes a double standard. They present this argument in two steps. First, (a) they argue that some everyday human decisions, including bureaucratic and administrative ones, do not need detailed explanations at the level of mechanisms of action to be deemed transparent. In (b), the second step, they argue that requiring such explanations for AI systems in comparable situations would constitute a double standard.

Let's look first at point (a). Zerilli et al. (2019) argue for it in the following way: First, they argue that a given decision can be explained and justified at different levels of abstraction. The level of explanation required to justify a decision depends on the nature of the decision. Everyday practical decisions, such as choosing between whether to cook or eat out, do not require a detailed explanation at the level of mechanisms of action, but can be justified by a brief explanation, expressed in the language of intentions, reasons, and beliefs. Some administrative decisions are similarly mundane and practical. So, some administrative decisions do not require a detailed explanation at the level of the mechanisms of action.

Their argument for (b) goes as follows: First, they argue that requiring detailed explanations at the level of operating mechanisms for AI systems but not for human decision-makers, would constitute a double standard in the absence of further arguments for the need for such explanations. They then present a possible such additional argument: AI systems, as we have seen in the above example with the legal system, can be biased and make unexpected mistakes. But they argue that human decision-makers are also biased and can make unexpected mistakes. Therefore, this is not a good argument to support the demand for more detailed explanations for AI systems. So, this request constitutes is a double standard.

I do not find this argument especially convincing. First, it seems questionable to me whether we can really compare practical decisions, such as the decision where to eat dinner, with administrative decisions, such as granting people parole. The authors argue that these decisions are "formally identical" and differ mainly “in their content” (Zerilli et al. 2019, 667). But the content seems crucial here. My decision to eat out rather than cook dinner can hardly have negative consequences for other people. On the other hand, granting or denying parole directly affects someone’s ability to exercise his most fundamental human rights. It therefore seems reasonable to require a more precise and detailed explanation to justify practical administrative decisions.

Second, I think there is a significant difference between the opacity of human decision-makers and the opacity of AI systems. As we have seen above, AI systems are opaque in the way that we do not even know what information they include. This is, of course, not the case for humans. Consider the following example. To put it simply, large language models are trained to predict the next word in a sentence. A trivial example: based on the training data, the model “learns” that the sentence “The glass stands on ...” should most probably be completed with the word “table”. We humans are also very good at this. But there is a fundamental difference between how we humans do it and how the large language models do it. We understand what a glass is and what it is for, we know that it can stand on a stable flat surface, we know that table is such a surface, etc. But we don’t really know what is happening in the large language models. We know that they can approximate very complex mathematical functions with thousands of parameters and that they are very good at detecting statistical patterns. What we do not know is which patterns they detect and what information about words these patterns actually convey (Gilpin et al. 2018). So, in this sense, at least some AI systems are more opaque than humans.

A New Counterargument

In this section, I will present a new argument against the idea that explaining AI systems embedded in decision-making processes is a necessary condition for making these processes trustworthy. I will focus on premise (1), which states that if a decision-making process relies on an opaque AI system, then this decision-making process is itself opaque, i.e. not transparent. In what follows, I will therefore try to show that decision-making processes can be considered transparent despite containing opaque AI systems.

Let us look again at the court system example. There, the opaque AI system proved problematic because there was a possibility that it could discriminate based on race or place of residence. As a solution, it was suggested that the AI system needs to be explained or made transparent. But there are other ways of checking whether the system discriminates. ProPublica journalists did just that: they analysed system’s past predictions, which showed that the system predicts different levels of risk for comparable defendants of different skin colours (Angwin et al. 2016). Thus, without making the system itself any less opaque, they revealed a relevant aspect of its operation and made the whole decision-making process more transparent (in this case, more transparently bad). While learning about its flaws and biases would probably be easier and faster if the system were transparent, the work of ProPublica journalists show that explainability is not the only way to transparency.

Testing the reliability and robustness of AI systems is, therefore, one way in which the decision-making processes involving these systems can be made more transparent (see Durán and Jongsma 2021; Kawamleh 2022 for a similar proposal and details). One might immediately object here that this merely reduces trustworthiness to reliability. It seems to follow from the above argument that reliable AI systems can already be trustworthy, whereas above I have made an explicit distinction between trustworthiness and reliability. But this objection misses an important distinction in my position. I argue that we can have transparent decision-making processes, even if the AI systems involved in those processes are opaque. Thus, in the case of the judicial system, we see that we can have a transparent (albeit transparently unfair) decision-making process that involves an inherently opaque AI system. Transparency and the resulting potential trustworthiness are therefore properties of decision-making processes, while the AI system itself can only be reliable.

Nevertheless, that does not mean that designing transparent decision-making processes with embedded opaque AI systems is easy or unproblematic. I think that besides reliability of embedded AI systems, there are two additional broad requirements that need to be met to achieve this. First, it must be specified exactly what role an AI system plays in the decision-making process. Decision-makers should be able to answer questions such as: How did the output of an AI system influence the final decision? Under what circumstances would the decision-maker ignore the AI system output? Second, it should also be clear why an AI system was included in the decision-making process in the first place. What do I mean by this? AI systems provide technological solutions to specific problems. In the justice system, for example, better predicting the likelihood of recidivism can help reduce the number of incarcerated people by helping to determine who can get a reduced prison sentence. But the use of technological tools such as AI models is only one possible answer to this problem. For example, the problem of the increasing prison population could also be addressed by decriminalising certain drugs, reducing prison sentences, etc. Implementing AI systems is therefore not a neutral solution but a political decision that should be made transparently.

Conclusion

In this essay, I argued that the explainability of AI systems is not necessary for the transparency, and thus trustworthiness, of decision-making processes relying on the outputs of these systems. In the second section, I showed that that transparency is a necessary condition for trust. Then, I explicated an argument concluding that explaining AI systems embedded in decision-making processes is a necessary condition for making these processes trustworthy. In the following section, I presented an existing counterargument, first presented by Zerilli et al. (2019). In the fourth section, I present a new argument against the explanation condition. I argued that we need to distinguish between the transparency of the AI system and the transparency of the decision-making process, and that the former is, in fact, not necessary for the latter. I also presented three alternative requirements for transparent AI decision-making processes: (1) the AI systems involved in the decision-making process must be reliable, (2) the way the AI system is implemented within the decision-making process must be transparent, (3) the reasons for implementing the AI system within the decision-making process must be transparent.

References

Angwin, Julia, Jeff Larson, Surya Mattu, Lauren Kirchner, and ProPublica. 2016. Machine Bias. *ProPublica*, 23th May, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 10. 10. 2022).

Ardila, Diego, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* 25, no. 6: 954–961.

Balasubramaniam, Nagadivya, Marjo Kauppinen, Kari Hiekkanen, and Sari Kujala. 2022. Transparency and Explainability of AI Systems: Ethical Guidelines in Practice. In: *Requirements Engineering: Foundation for Software Quality*, 3–18. Eds. Vincenzo Gervasi and Andreas Vogelsang.

Bringsjord, Selmer and Naveen Sundar Govindarajulu. 2022. Artificial Intelligence. *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/artificial-intelligence/ (accessed 3. 10. 2022).

Burgess, Matt. 2018. UK police are using AI to inform custodial decisions – but it could be discriminating against the poor. *The Wired*, 1. 3. 2018. https://www.wired.co.uk/article/police-ai-uk-durham-hart-checkpoint-algorithm-edit (accesed 3. 10. 2022).

Chatila, Raja, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. 2021. Trustworthy AI. In: *Reflections on Artificial Intelligence for Humanity*, 13–39. Eds. Bertrand Braunschweig and Malik Ghallab.

Durán, Juan Manuel, and Karin Rolanda Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47: 329–335.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA),* 80–89.

Harmon, Stephanie A., Thomas H. Sanford, Sheng Xu, Evrim B. Turkbey, Holger Roth, Ziyue Xu, Dong Yang, et al. 2020. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nature Communications* 11, no. 1: 4080.

Hatherley, Joshua James. 2020. Limits of trust in medical AI. *Journal of Medical Ethics* 46, no. 7: 478–481.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, no. 7873: 583–589.

Kawamleh, Suzanne. 2022. Against explainability requirements for ethical artificial intelligence in health care. AI and Ethics 3: 901–916.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, no. 1: 18.

McLoad, Carolyn. 2020. Trust. *The Stanford Encyclopedia of Philosophy.* 2020. https://plato.stanford.edu/entries/trust/ (accessed 28. 9. 2022).

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, no. 2: 1–21.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. AI in health and medicine. *Nature Medicine* 28, no. 1: 31–38.

Selingo, Jeffrey. 2017. Colleges' Endless Pursuit of Students. *The Atlantic*, 10th April 2017. https://www.theatlantic.com/education/archive/2017/04/the-business-of-college-marketing/522399/ (accessed 15. 10. 2022).

Strahovnik, Vojko, Jonas Miklavčič, and Mateja Centa. 2020. Etični vidiki uporabe algoritemskega odločanja in ostalih sistemov UI v času pandemij oz. izrednih razmer. *Bogoslovni vestnik* 80, no. 2: 321–334.

Townley, Cynthia, and Jay L. Gardield. 2013. Public Trust. In: *Trust: Analytic and Applied Perspectives*, 95–107. Eds. Pekka Mäkelä and Cynthia Townley.

Vereschak, Oleksandra, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5: 1–39.

von Eschenbach, Warren J. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology* 34, no. 4: 1607–1622.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology* 32, no. 4: 661–83.

Jonas Miklavčič

TRANSPARENCY AS A PRINCIPLE AND A REQUIREMENT

Introduction

For a long time, artificial intelligence (AI) was thought of as a field of 'fun and games'. Both in the literal sense and in the figurative sense. On the one hand, computer scientists have tested AI on games in a literal sense, building computers that play chess or robots that can play football with other robots (Wooldridge 2021). But even in a figurative sense, the field has had a tinge of 'fun and games', with experts wondering what it really means to 'think', whether we can artificially simulate the workings of the brain, and whether machines will ever be able to feel emotions and understand language in the true sense of the word 'understand' (Copeland 1993).

But with the advent of machine learning a decade or so ago, suddenly everyone realised that the whole thing was no longer just 'fun and games' (Bryson 2020). Today, machine learning systems make decisions in banks, courts, make medical diagnoses, predict the spread of viruses, help with security and surveillance and many other things (Taulli 2019). Generative AI systems have also made it possible for computers to 'create' text, images and music (Strahovnik 2023; Centa Strahovnik 2023). In many areas, they have taken on the roles that humans used to have, and in many others, they have started to perform tasks that we never thought could be done. But machine learning algorithms have also come into the everyday lives of the people. On the web, they filter the flow of information we see, they filter unwanted emails, enable face recognition on smartphone cameras, YouTube videos have automated subtitling, and Facebook ads are personalised to the individual user (Alpaydin 2016). With machine learning, which has enabled the truly massive implementation of artificial intelligence systems in various areas of people's daily lives, came also a strong concern about how to use these systems in an ethical and safe way (Coeckelbergh 2020).

The fact that machine learning has led to a 'boom' in the implementation of systems has also encouraged a parallel 'boom' in the ethics of AI. If the first 20 years of the history of AI (from McCarthy's Conference in 1956 until around 1975)[1] are considered to be 'the golden age of AI' (Wooldridge 2021), it is quite

[1] Most often, authors see the beginning of AI history in a 10-week conference organised by John McCarthy at Dartmouth College in Hanover, USA, in 1956. McCarthy is also the father of the term 'artificial intelligence', which he coined precisely for the purpose of raising funds for this conference (McCarthy et al. 1955).

possible that the period in which we are currently living will go down in history as the 'golden age of AI ethics'.

Today we are witnessing an incredible increase in the number of documents that are published every year suggesting their own ethical guidelines for the safe use of AI systems, or for what is sometimes called 'ethical' or 'trustworthy' AI. These guidelines are issued by non-profit organisations, educational institutions, government organisations and even many private companies themselves (e.g. Google, IBM, Nvidia). The documents differ in many ways (depending on the organisation that issued the document, who the document is aimed at, what values are represented in it and more). But in many ways, they are similar. Almost all documents include the ethical principles of human-centredness, accountability, privacy, fairness, reliability, safety and security – and, of course, transparency (Miklavčič 2021).

Transparency as a Principle and a Requirement

Transparency is mentioned as an important value and principle in almost all documents that propose ethical guidelines. Below we can see a summary of the analysis of the documents that have been published worldwide up to 2019.[2]

Ethical principle	**Number of documents**	**Included codes**
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non)bias, (non)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access, and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom & autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)

[2] Despite being several years old, the relevance of the study is undiminished. For the present paper, it is important to establish a broader picture of how global organisations have been trying to regulate AI in recent years, and what values have guided the recommendations. We refer to this research for its comprehensiveness, precision, and clarity, but we stress that trends are virtually unchanged since its publication, as is evident in much more recent research (Bhandari 2021; Hagendorff 2020).

Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Table 1: Overview of the 84 documents with ethical guidelines, and the ethical principles they propose (Jobin, Ienca, and Vayena 2019, 7).

In the research, 84 documents proposing ethical guidelines for the use of AI had been analysed, to see which ethical principles emerge most frequently. 'Transparency' is at the top of the list of ethical principles that appear in these documents (by frequency) and is listed as an ethical requirement in 73 of these 84 documents.

In the right-hand column we can see that several different terms have been included in the sum total of the appearance of the transparency principle. These include transparency, explainability, explicability, understandability, interpretability, communication, disclosure and showing. There may be nothing wrong with that, at least in principle, but we must point out that these are not just different terms – they are different phenomena. While the difference between the phenomena denoted by some of these terms may not be obvious (for example 'explainability' vs. 'explicability'), it is probably quite intuitive that 'understandability' (ability to understand or to be understood) and 'disclosure' (openly making information known) are not the same thing.

But it is no coincidence that this research has combined many phenomena into one term – 'transparency'. This is perhaps not surprising, given the breadth of the term 'transparency', which derives from its metaphoricity, abstractness and consequent ambiguity (Miklavčič 2023). Even more so, many of these documents with ethical guidelines merge these phenomena into a single term themselves. They may distinguish perhaps between two of these terms, but not many more.

If several different values or principles are embedded in the broader principle of transparency, there may be no problem at all. But when we start 'demanding' transparency, issues arise. The problem is that if the demand encompasses many different phenomena, the demand for transparency, or rather the 'transparency requirement', in fact involves many different requirements, and if we call all these requirements, in most cases, simply a 'transparency requirement', we don't actually know exactly what we are demanding when we demand transparency. This makes it very difficult for us to try to make ethical guidelines useful. Let us look at some examples.

OECD – "Recommendation of the Council on Artificial Intelligence" (2022)

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation (IGO) with 38 member countries. It was founded in 1948 to stimulate economic progress and world trade. Recently, it has been devoting a lot of time to the issues of AI. That is why it has issued a document with recommendations on the ethical use of AI, and in its latest update (2022) suggests the following five key principles for the safe use of AI systems (OECD 7–8):

1. Inclusive growth, sustainable development and well-being
2. Human-centred values and fairness
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability

The third principle is called ‘transparency and explainability’. We can see that this document distinguishes between (at least) two terms, two phenomena, two requirements, when referring to this third principle. Let us point out that the explainability requirement is usually referring to the fact that AI systems are required to be built in such a way that we can understand (explain) how they work. This requirement exists because machine learning systems are often not explainable by human beings (Zednik 2021). This is called ‘the black-box problem’ (Espindola 2018). The operation of some machine learning systems is so complex that even the programmers who wrote the algorithm and trained the model cannot understand the reasons behind the decision-making of the algorithm in a meaningful way.[3] (Burrell 2016) This third principle (transparency and explainability) of the five principles proposed by the OECD is explained in more detail later in the document. It is written that AI actors should provide meaningful information (OECD 8):

i. to foster a general understanding of AI systems,
ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,
iii. to enable those affected by an AI system to understand the outcome,
iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

If we look closely at what compliance with the transparency principle requires, we see that the principle includes several different requirements. The first point is referring to the requirement of at least *explainability* (which they themselves have detected in the naming of the principle) and *understandability* (implied in the phrase ‘general understanding’), but also *interpretability* (because people should be able to understand ‘the meaning of the explanation’). The second point concerns, for example, *disclosure of information* and *communication with*

[3] To put it very concretely: if the algorithmic system used by a bank to decide who should (or should not) be granted a loan makes a decision not to grant loan to a given customer, no-one, not even the programmers who wrote the algorithm and trained the model, has any insight into why the algorithmic system decided not to grant a loan to that particular customer.

users. The third point seems to shift the understanding from the level of 'understanding in principle', which probably primarily includes programmers, to the *users* of such systems. The fourth point involves the fact that it must be possible to challenge the result, which means that the possibility of questioning the decision and demanding answers as to why the decision was made must be enabled. This already includes the requirement for *intelligibility* (and all other requirements related to the provision of reasons behind decision) and even the requirement to be able to define responsibilities clearly and who should be held accountable when things go wrong.

So, at this point, we would like to point out, that the transparency requirement always requires many different things. This ambiguity is perhaps even more evident in the document issued by UNESCO in 2021.

UNESCO – "The Recommendation on the Ethics of Artificial Intelligence" (2021)

The United Nations Educational, Scientific and Cultural Organization (UNESCO) has also published its document with ethical guidelines. UNESCO's document "The Recommendation on the Ethics of Artificial Intelligence" (2021) sets out the following ethical principles (UNESCO 2021, 20–23):

1. Proportionality and Do No Harm
2. Safety and security
3. Fairness and non-discrimination
4. Sustainability
5. Right to Privacy, and Data Protection
6. Human oversight and determination
7. Transparency and explainability
8. Responsibility and accountability
9. Awareness and literacy
10. Multi-stakeholder and adaptive governance and collaboration

UNESCO explains what is actually required by Principle 7, which is relevant to our paper:

> People should be fully informed when a decision is informed by or is made on the basis of AI algorithms, including when it affects their safety or human rights, and in those circumstances should have the opportunity to request explanatory information from the relevant AI actor or public sector institutions. (22)

UNESCO then adds that individuals should have access to the reasons for a decision that affects them and the possibility to contact someone who can check and, if necessary, correct the system's decision (22). It is also added that AI actors should inform users in an appropriate and timely manner when a product or service is provided by or through an AI system.

With the 'transparency principle', UNESCO is therefore proposing the requirement for the people to be *informed* in a timely manner when an AI system is deciding (or assisting in a decision) about them – that people should be able to *request further explanation*, that people should be able *to see the reasons* for a decision that affects their rights and freedoms, and that they should be able *to object* to a chosen decision.

We can tell that 'requiring transparency' actually includes (at least) four separate requirements:

1. the requirement to inform people that they are in contact with the AI system (e.g. that it is making decisions about them);
2. the requirement to provide additional information and explanations at the request of those involved
3. the requirement to provide the reasons for the decision which affects those involved
4. the requirement to provide the possibility to exercise the right to object to the decision

We propose to call the first requirement *open communication* (as it concerns the real-time information of individuals), the second *disclosure* (additional information and explanations must be disclosed at the request of the individual), the third *explainability*, and the fourth *the right to object* to the decision.[4]

We hope it is obvious why we find this problematic. The importance of distinguishing these phenomena with clear terminology lies in the fact that these different phenomena require different solutions. The fact that people need to be informed that an algorithm is deciding on their fate when loan is granted at a bank (or not) is a relatively easy problem to solve; people just need to be informed. The fact that all documentation must be openly accessible, available for inspection, already becomes a slightly more problematic issue; what about the right of companies not to disclose the workings of their algorithms publicly to maintain an advantage in the market? Or the right of the state not to disclose the workings of algorithms that help with cyber-security and fight hacker attacks? (Burrell 2016) These scenarios require us to think more carefully. And then there is the 'explainability problem' or the 'black-box problem', which does not look very solvable at the moment.[5] (Knight 2017) The fact that all these issues are grouped under the term 'transparency', does not help in addressing these practical problems in the way that is needed, and it should be emphasised that we have highlighted these two documents (OECD's and UNESCO's) because they are among the better ones.

[4] Although the fourth perhaps falls more within the reasons for the need for explainability rather than its own requirement.

[5] Although there is a whole field of AI that deals specifically with trying to explain AI systems – Explainable Artificial Intelligence (XAI), due to all the challenges that this field often encounters (Arrieta et al. 2019), deep machine learning systems may never be completely explainable.

Transparency as a principle and a requirement

Ethical principles and ethical requirements are central to ethical systems but have different characteristics and objectives.

Ethical principles are very general fundamental ideas that guide moral behaviour and decision-making. Principles are usually quite abstract and provide a broad general framework for ethical thought (Iserson 1999). Although they may vary from culture to culture, their generality and abstractness make a number of principles almost universal, such as autonomy (or respect for individual autonomy), justice (treating people equally and fairly), beneficence (acting for the good of others), non-maleficence (not causing harm to others) and integrity (acting responsibly and credibly in accordance with codes) (Beauchamp and Childress 2013). Precisely because of their breadth and considerable degree of abstractness, the principles are easily transferable from one field to another but they do not provide any clear guidance on how to apply them in practice (Gert et al. 1997).

Whereas ethical principles are general and relatively abstract, and therefore often independent of their fields of application and the particularity of situations, ethical rules are much more specific directives that can clearly guide an individual's behaviour in particular situations (Iserson 1999). They communicate much more clearly what is required of people and what is forbidden to them, which is why it is often said that ethical rules can be formulated in a positive form (what one should do) or in a negative form (what one should not do). In the context of transparency, we are primarily interested in ethical requirements (the positive form of ethical rules), as the main object of our research is precisely the requirement to achieve transparency. Ethical requirements are derived from principles and offer 'instructions' on how to apply the principles in real-life scenarios (High Level Expert Group on Artificial Intelligence 2019). Requirements can also be more or less specific. 'Respect the elderly' and 'Get to work on time' are requirements with different degrees of abstraction, as the former allows for a much wider variety of interpretations and applications than the latter, but what all requirements have in common is that they serve as very practical guidance on how to behave in accordance with the underlying ethical principles. Ethical rules (both requirements and prohibitions) can remain at the level of morality (as 'unwritten rules'), or they can be formalised in the form of specific codes of ethics or even laws and regulations.

The relationship between ethical principles and requirements could be defined as ethical principles being more general ideas that provide the basis for moral action, while ethical requirements are specific rules that help individuals to put these principles into practice in concrete situations by clearly defining their obligations. If principles enable requirements to be implemented, requirements enable principles to be followed in practice. For example, general principles such as 'respect for others', 'respect for private property' and 'fairness' lead in society to the development of an ethical rule (requirement) of 'thou shalt not steal' and then a formal rule that theft is forbidden and punishable.[6]

If a principle guides our behaviour, a requirement imposes certain concrete obligations on us. Following this understanding, we can clearly see that the principle of transparency and the requirement for transparency are two different, and at least partly separate, things. The principle of transparency serves as a general guideline that we must keep in mind when using AI and tailor specific requirements, constraints and laws to follow this principle; the transparency requirement asks those involved (more or less formally) to achieve transparency in specific situations in a much more concrete way. Whereas the transparency principle emphasises the importance of openness, honesty, public communication and traceability, the transparency requirement imposes a duty on those involved (or some of them) to achieve all of these. A company may remind staff of the importance of the principle of transparency, thereby actively promoting openness and honesty as a matter of principle, or it may write into its code that each staff member must provide his or her superiors once a week with relevant documentation showing his or her work during the previous week – a very clear requirement which, in its concreteness, imposes a specific duty, going well beyond the general framework of transparency as a bare guiding principle.

We have seen (in OECD's and UNESCO's documents) that often when we refer to the principle of transparency, we have, in fact, many different requirements hidden in it, which are not clearly separated in terms of terminology, but all fall under the transparency requirement. Perhaps it is not really a major problem when we are talking about guidelines because, although guidelines are important, they are really just that – guidelines. A much more serious problem arises when the requirement for transparency appears in a formal context and is required at the level of regulation. The legal documents that are in the making at EU level at this very moment (and often based on NGO guidelines) do not make it any clearer what is actually required when they call for transparency of AI. Often, it is much less well defined. And these are the bills that will regulate what is, and is not, allowed in the EU in relation to the use of AI in the coming decades. We do not have much room for error here.

European Commission – "AI Act" (2021)[7]

The AI Act by the European Commission is a regulatory framework proposed to govern the use and deployment of AI within the European Union. The

[6] In this example of the relationship between a principle and a rule, we cite the example of a prohibition because it is much more illustrative in referring to the creation of legal constraints than is perhaps the case with requirements, but of course the description of the relationship applies to prohibitions as well as requirements.

[7] The full name of this proposal for a regulation by the European Commission is "Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts" (2021).

proposal has been adopted but has not yet entered into force. In the AI Act, the term 'transparency' appears in all its forms twenty-seven times, which is actually quite high, as the other terms already mentioned that are used in relation to transparency appear much less frequently: 'disclosure' twelve times, 'provision of information' eight times, 'traceability' five times and 'explainability' once.

Here again, 'transparency' in its twenty-seven occurrences means different things. In some places it means *provision of information*: "For other, non-high-risk AI systems, only very limited transparency obligations are imposed, for example in terms of the provision of information to flag the use of an AI system when interacting with humans." (European Commission 2021, 7) For the sake of additional terminological clarity, we call this real-time information *open communication*. Elsewhere again, transparency and information provision are separate things – or so the conjunction 'and' seems to suggest: "The requirements will concern data, documentation and traceability, provision of information and transparency, human oversight and robustness and accuracy and would be mandatory for high-risk AI systems." (9) While the first example clarifies what is meant by 'transparency', in the second example it is no longer obvious.

Furthermore, 'transparency' in the AI Act also refers to the requirement to disclose certain things, such as the algorithmic code used by the system. In this sense, it is stated that the increased transparency obligations will not disproportionately affect the right to protection of intellectual property (e.g. algorithmic code), because only the minimum necessary information that users need to exercise their rights will be required to be disclosed (11). Transparency is thus also linked to *disclosure* in this document, although apparently to a somewhat 'limited disclosure'.

In relation to transparency, more general public information is also required, for example in the form of user manuals: "High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate." (30) Also mentioned are 'technical documentation' and 'record-keeping' requirements, which are often subsumed under the transparency requirement through the term *traceability* (Jobin, Ienca, Vayena 2019), although they are presented separately from 'transparency' in the AI Act.

In places, therefore, the AI Act refers to phenomena that are often (semantically) subsumed under 'transparency', even if other terms are used for them. The term 'transparency' itself, however, is used to mean different things in the document – sometimes it is defined what is meant by 'transparency', sometimes it can be interpreted from the context, and in about half of the cases it is not entirely clear to which meaning the term refers.[8] The problem is that if this is the case, the

[8] Example: "It is therefore appropriate to classify as high-risk a number of AI systems intended to be used in the law enforcement context where accuracy, reliability and transparency is particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress." (European Commission 2021, 27)

wording loses its practical usefulness. If we do not know what is required when transparency is required, in an important sense, nothing is actually required.

The Problem of Terminology

Since a detailed analysis of the legal use of the term ‘transparency’ is beyond the scope of this paper, which is primarily focused on the analysis of the ethical requirement for transparency, we will conclude the review here by noting the key findings:

1. All documents that seek to ensure the safe use of AI (both *soft law* and *hard law* documents) also require transparency.
2. The term ‘transparency’ can refer to many different phenomena, including *communication* (to the public), *information provision* (to individuals interacting with the system), *disclosure* (of algorithmic code), *traceability* (of the use of the system through archives and documentation), and *explainability* (of the reasons for the decisions made by the system).
3. It is very often not clear from the documents themselves what is required when ‘transparency’ is requested or demanded.
4. The vagueness of the definition of the term ‘transparency’ makes it difficult for guidelines or regulations to be useful in practice.
5. The field needs clearer terminology in the context of the requirement for transparent AI. The individual requirements currently included in the transparency requirement should be referred to as individual requirements for the sake of clarity – this should be addressed by broadening the terminology, which currently relies primarily on the term ‘transparency’.

So, we have an obvious problem: How to differentiate better between the issues that now all fall under the ‘transparency requirement’ and clearly define the appropriate terminology for them?

Proposal for a More Appropriate Terminological Division

In order to make (ethical and legal) transparency requirements as useful as possible, we propose a more precise terminology to provide a clearer overview of what we are asking for when we ask for ‘transparent AI’. Having reviewed the documents described above and some others,[9] we identify the following requirements in the requirement for transparency (in the field of AI):

[9] Other documents of particular relevance reviewed for this research include: “White Paper on Artificial Intelligence: a European approach to excellence and trust” (European Commission 2020), “Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems” (European Group on Ethics in Science and New Technologies 2018), “Artificial Intelligence: From Ethics to Policy” (van Wynsberghe 2020), “Algorithms and Human Rights: Study on the Human Rights Dimension of Automated Data Processing Techniques and Possible Regulatory Implications” (Council of Europe 2018), “Everyday Ethics for Artificial Intelligence” (IBM 2022), “Google AI Policy Perspectives” (Google 2019), “Microsoft Responsible AI Standard, v2” (Microsoft 2022).

1. **Provision of information**
   a) *Public information* (e.g. instructions for the use of AI)
   b) *Open communication* (providing individuals with real-time information about their interactions with the AI system)
2. **Traceability**
   a) *Technical documentation* (e.g. who is the system manufacturer, who is the supplier, who is the importer, etc.)
   b) *Record-keeping* (e.g. automated logging of system performance and archiving of activity information)
3. **Disclosure**
   a) *Technical information* (e.g. possible disclosure of the code used in machine learning at the request of users)
   b) *Explanatory information* (e.g. further explanations on decisions made must be provided which are appropriate to the individual concerned)
4. **Explainability** (e.g. the ability to understand how AI systems work)

In our view, this list contains everything that is covered by the 'transparency requirement' today. Although not all of them or always, the above-mentioned requirements, hidden under the term 'transparency', are, in fact, at work when we demand transparent AI. John Zerilli, in his book *A Citizen's Guide to Artificial Intelligence* (2020), presents his own division of the notion of the term 'transparency', which may be similar in some respects to our own, and while very useful in showing the breadth of the use of the term 'transparency' and its various meanings, is, in our view, not precise enough for the actual *requirement*, and therefore not very useful in our context. Zerilli identifies the political notion of accountability as the basic meaning of 'transparency', and this meaning of the term then he divides into 'responsibility', which can be moral or legal; 'inspectability', which can refer to the process or technical aspects of a system (which can further be general or particular and include explainability, intelligibility and justifiability); and 'accessibility' (2021, 25). As this division is much more related to the understanding of the concept of transparency than to the practical requirement for it, we stick to our division of the requirement for transparency into more specific individual requirements.

Conclusion

The use of the term 'transparency' in the context of AI has been shown to be very inconsistent and not clear enough. It is *inconsistent* in the sense that different documents that talk about the principle or requirement of transparency use the term to mean different things. It is *not clear enough* in the sense that individual documents, when they refer to a requirement for transparency, are in fact requiring many different things by that requirement.

The first important conclusion is that we need to ensure that we have appropriate terminology in the use of the term 'transparency' in the context of AI, so that it is clearer. Accordingly, we argue that the use of the term 'transparency' is perfectly appropriate to denote a specific ethical principle, but not to denote an ethical requirement. Transparency is a rather broad and vague concept, but this may not be too much of an obstacle to using the term in the context of the 'principle of transparency', which is deliberately set broadly and somewhat abstractly, as this makes it more easily transferable to many different fields and more easily leaves open the possibility of different approaches to its application. In our view, such transparency can be quite legitimately, appropriately, and reasonably referred to as an ethical principle.

What we have seen with requirements is that they need to be as clear and specific as possible if they are actually to tell the individual how to act in practice in a particular environment or situation. In other words, if we are asking for something, we need to be very clear about what we are asking for. In the case of a requirement, the term 'transparency' is no longer appropriate because of its breadth and abstractness.

The claim is therefore that 'transparency' can be reasonably a principle, but not a requirement. In practice, our suggestion is that documents with ethical guidelines should still refer to transparency as an important ethical principle that, in combination with other important principles, should guide the entire life-cycle of AI systems. However, when it comes to making requirements, both in the form of perhaps less binding codes of conduct and in the form of legal laws, we suggest avoiding the term 'transparency' and, for the sake of clarity, identifying which specific requirements the transparency requirement has actually included so far. These requirements should then be also distinguished terminologically and understood as different requirements. In line with our suggestion for more appropriate terminology for a transparency requirement we propose that in the case of requirements, instead of 'transparency', we should require *public information*, *open communication*, *technical documentation*, *record-keeping*, *disclosure of technical and explanatory informatio*n, and *explainability*.

Funding acknowledgement

The publication was prepared as part of a research programme *The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights*, financed by the University of Ljubljana, and research projects *society* and the research projects *J6-4626 (B): Theology, digital culture and the challenges of human-centered artificial intelligence* and *J6-60105: Theology and digitalisation: anthropological and ehical challenges*, both financed by the Slovenian Research and Innovation Agency (ARIS).

References

Alpaydin, Ethem. 2016. *Machine Learning: the new AI.* Cambridge: MIT Press.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI. *Information Fusion*. https://doi.org/10.48550/arXiv.1910.10045 (accesses August 12, 2021).

Beauchamp, Tom L. in James F. Childress. 2013. *Principles of biomedical ethics*. New York: Oxford University Press.

Bhandari, Avantika. 2021. Artificial Intelligence: the global landscape of ethics guidelines. *Montreal AI Ethics Institute*. October 8, 2021. https://montrealethics.ai/artificial-intelligence-the-global-landscape-of-ethics-guidelines/ (accessed March 31, 2022).

Bryson, Joanna J. 2020. The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In: Markus D. Dubber, Frank Pasquale in Sunit Das, eds. *The Oxford Handbook of Ethics of AI*. New York: Oxford University Press.

Burrell, Jenna. 2016. How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*. January–June 2016: 1–12.

Centa Strahovnik, Mateja. 2023. Identiteta in pogovorni sistemi umetne inteligence. *Bogoslovni vestnik*, 83, no. 4: 853–864.

Coeckelbergh, Mark. 2020. *AI Ethics*. Cambridge: The MIT Press.

Copeland, Brian Jack. 1993. *Artificial intelligence: a philosophical introduction*. New Jersey: Wiley-Blackwell.

Council of Europe. 2018. *Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications.* March, 2018. Brussels: Council of Europe.

Espindola, David. 2018. The Black Box Problem – When AI Makes Decisions That No Human Can Explain. *Intercepting Horizons*. 29. oktober 2018. https://www.interceptinghorizons.com/post/the-black-box-problem-when-ai-makes-decisions-that-no-human-can-explain (accesses 10. marca 2022).

European Commission. 2020. *White Paper on Artificial Intelligence: a European approach to excellence and trust*. February 19, 2020. Brussels, Belgium. COM(2020), 65 final.

– – –. 2021. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* April 21, 2021. Brussels, Belgium. COM/2021/206 final.

European Group on Ethics and New Technologies. 2018. *Statement on Artificial Intelligence, Robotics and ‚Autonomous‘ Systems.* Brussels: European Commission.

Gert, Bernard, Charles M. Culver in Danner Clouser. 1997. *Bioethics: A Return to Fundamentals.* New York: Oxford University Press.

Google. 2019. Perspectives on Issues in AI Governance. *AI.Google.* https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf (accessed July 18, 2023).

Hagendorff, Thilo. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30: 99–120.

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI.* April 8, 2019. Brussels: European Commission.

IBM. 2022. Everyday Ethics for Artificial Intelligence. *IBM.* https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf (accessed July 16, 2023).

Iserson, Kenneth V. 1999. Principles of biomedical ethics. *Emergency Medicine Clinics of North America*, 17, no. 2: 283–306.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. „Artificial Intelligence: the global landscape of ethics guidelines“. *Document before publication in Nature Machine Intelligence*. https://doi.org/10.48550/arXiv.1906.11668

Knight, Will. 2017. The Dark Secret at the Heart of AI. *MIT Technology Review*. April 11, 2017. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/ (accessed March 4, 2022).

McCarthy, John, Marvin Lee Minsky, Nathaniel Rochester in Claude Elwood Shannon. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* August 31, 1955.

Microsoft. 2022. Microsoft Responsible AI Standard, v2: General Requirements. *Blogs.Microsoft.* https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf (accessed July 18, 2023).

Miklavčič, Jonas. 2021. Zaupanje in uspešnost umetne inteligence v medicine. *Bogoslovni vestnik*, 81, no. 4: 935-946.

– – –. 2023. Ideal transparentnosti v digitalin dobi. *Bogoslovni vestnik*, 83, no. 4: 825-838.

OECD. 2022. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449.

Strahovnik, Vojko. 2023. Etični in teološki izzivi velikih jezikovnih modelov. *Bogoslovni vestnik*, 83, no. 4: 839–852.

Taulli, Tom. 2019. Artificial Intelligence Basics: A Non-Technical Introduction. New York: Apress.

UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence*. SHS/BIO/PI/2021/1.

van Wynsberghe, Aimee. 2020. *Artificial intelligence: From ethics to policy*. Brussels: European Union.

Wooldridge, Michael. 2021. *A brief history of artificial intelligence: what it is, where we are, and where we are going*. New York: Flatiron Books.

Zednik, Carlos. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34: 265–288.

Zerilli, John, John Danaher, James Maclaurin, Colin Gavaghan, Alistair Knott, Joy Liddicoat, and Merel Noorman. 2020. A citizen's guide to artificial intelligence. Cambridge: The MIT Press.erlini, Sanam 2011. In: Marshall, K., Haywardwith, S., Zambra, C., Breger, E in Jackson, S. 2011. *Women in Religious Peacebuilding*. Berkeley: United States Institute of Peace. http://www.usip.org/sites/default/files/PW71-Women_Religious_Peace-building.pdf (accessed 20. januarja 2017).

Saša Horvat

WHAT DO AI, MEDICINE AND CHESS HAVE IN COMMON? ISSUES IN DECISION-MAKING PROCESS

Introduction

Artificial intelligence (AI) is present in many areas of human activity, and one of them is medicine. Many ethical issues arise in the context of the broader application of AI in medicine. In our work, we focus on the issue of the application of AI tools in the decision-making process in clinical practice. We approach the issue of decision making also from a cognitive perspective, highlighting challenges faced by physicians. The relationship between AI and physicians is considered from the perspective of the ethical principle in medicine – respect for autonomy. The broader relationship between AI, physician and patient (Bracanović 2021) will be addressed on another occasion. To help us think about these complex issues, we consider the relationship between artificial and human intelligence in the field of chess, where it has been developing intensively for more than seven decades.

In the first part of the paper, we present the expected benefits of AI in medicine. Then we analyse the decision-making process, before looking at the state of AI tools specifically developed for the decision-making process in medicine. After that we will highlight the main findings from the field of chess and, in the conclusion, offer certain guidelines for further reflection on the issue of the relationship between AI and physician.

Concerning the terminology, we will follow the study prepared for the European Parliament by the Panel for the Future of Science and Technology (Panel): *artificial intelligence* (AI) is when a machine is able to perform tasks that mimic human intelligence (e.g. medical prognosis); *machine learning* (ML) stands for methods of learning and is a subset of AI that performs a task by learning from data and improving through experience; *deep learning* is a subset of ML that uses neural networks and big data to solve complex problems; *AI algorithm* is used to develop *AI models* for specific tasks; and *AI tool* (a term we use frequently in this article) is an AI model that has been developed to the point where it can be used by end users (e.g. physicians) (Panel 2022, 2–3).

Benefits of AI in Medicine

In this chapter, we will briefly familiarise ourselves with the (possible) benefits of applying AI in medicine, as well as with the great expectations associated with it. According to the Panel it is expected that "AI can benefit future healthcare, in particular by increasing the efficiency of clinicians, improving medical diagnosis and treatment, and optimising the allocation of human and technical resources" (2022, 69).

The European trade association for the medical technology industry (The MedTech Europe) has presented figures on the potential impact of AI on European healthcare systems. According to the report, AI applications have the potential to free up 1.6 million to 1.9 million working hours of healthcare professionals each year. This is an important factor considering that there is expected to be a shortage of 4.1 million healthcare workers in the EU by 2030 (WHO, 2016). MedTech Europe also stated that AI applications could potentially save up to 403,000 lives per year and €212 billion annually (Deloitte and MedTech Europe, 2020). Big and well-known companies are deeply immersed in the field of AI in medicine joining forces with health services, like Google (*The DeepMind*), Intel (*Lumiata*), IBM (*Watson*), Microsoft (*Hanover Project*), etc. (Lidströmer, Aresu, and Ashrafian 2022, 64). Briganti and Le Moine (2020) claim that AI in medicine could enable the development of a 4P model of medicine (predictive, preventive, personalised and participatory), which could also promote patient autonomy.

Our focus in the paper is on the potential of AI tools in clinical practice and this "ranges from the automation of diagnostic processes to therapeutic decision making and clinical research. The data necessary for diagnosis and treatment comes from many sources, including clinical notes, laboratory tests, pharmacy data, medical imaging, and genomic information" (Ibid, 5).

We will now give some examples of the application of AI tools in different areas of medicine (not only in clinical settings). We also provide some test studies that show potential benefits.

Cardiology is considered to be at the forefront of AI in medicine (Lopez-Jimenez et al. 2020), and AI tools are seen as a great help to cardiologists. For example, AI tools can save time when analysing data and enable faster patient assessment (Ibid). An example of AI solutions accessible in everyday life can be found in the Apple Watch 4 that has an electrocardiogram with the ECG app (Apple, 2022) and can help with early detection of atrial fibrillation.

In pulmonary medicine, AI tools are being tested for the interpretation of pulmonary function tests. In a recent study, 120 pulmonologists made the correct diagnosis in 44.6% (out of 6,000 evaluations), while the AI-based software made a correct diagnosis in 82% of cases. The authors conclude that "the AI-based software has superior performance and may provide a powerful decision support tool for clinicians" (Topalovic et al. 2019, 9).

In gastroenterology, we find promising studies on a wide range of AI applications in the clinical setting. AI has been developed and tested for diagnosis, prognosis and image analysis (Yang and Bang 2019). Convolutional neural networks (CNNs) from Deep Learning, for example, show promise in pattern recognition (Valueva et al. 2020) and are being tested for a real-time polyp detection system, for example (Qadir 2021). A 3-dimensional convolutional neural network can detect colonic polyps with an accuracy of 76.5% (Misawa at al. 2018). The authors conclude that “artificial intelligence has the potential to provide automated detection of colorectal polyps. Further machine learning and prospective evaluation are mandatory; however, such CADe systems are expected to fill the gap between endoscopists with different levels of experience” (Ibid, 2028). Another recent study by Shaukat et al. evaluated the utility and safety of using a CADe device during colonoscopy and concluded that “for experienced endoscopists performing screening and surveillance colonoscopies in the United States, the CADe device statistically improved overall adenoma detection (APC) without a concomitant increase in resection of non-neoplastic lesions (THR)” (Shaukat et al. 2022, 732). These findings are important because the misdetection rate for adenomas is 26% (Zhao 2019) and adenomas, if left untreated, can become malignant and dangerous (Arnold et al. 2017). AI has also brought significant developments and potential benefits to many other medical fields (see Lidströmer, Aresu, and Ashrafian 2022).

Nevertheless, some caution is warranted here. We find that the results of AI tools are increasingly being compared with those of physicians. Going through the studies on this issue, Liu and his team found “the diagnostic performance of deep learning models to be equivalent to that of health-care professionals” (Liu et al. 2019, e271). However, the same study also found that “few studies presented externally validated results or compared the performance of deep learning models and health-care professionals using the same sample. Additionally, poor reporting is prevalent in deep learning studies, which limits reliable interpretation of the reported diagnostic accuracy” (Ibid). Therefore, we are still far away from all-encompassing scope of comparison between AI and physicians.

Other promising AI tools include Natural Language Processing (NLP) algorithms that help physicians create medical records (Locke et al. 2021). In addition, Ambient Clinical Intelligence (ACI) (Acampora 2013) is eagerly awaited as a digital environment that enables better workflow between physician and patient, but also addresses the problem of physician staff shortages (Giovanni and Olivier 2020). Great successes with AI are also expected in the field of pharmacology (Wallis 2019), where AI tools are used for “drug discovery and development, drug repurposing, improving pharmaceutical productivity, and clinical trials” (Paul et al. 2021, 80).

Despite all the encouraging results, however, there is still some reluctance to see AI applied more widely in clinical medicine. Briganti and Le Moine state that there are four reasons that need to be addressed and are the cause of physician’s caution: a) lack of training in digital medicine; b) electronic health records (EHR) have imposed a large administrative burden on physicians (leading to burnout); c) will AI replace physicians (the prevailing view is that AI will complement human intelligence); d) lack of legal framework – who should be held responsible in case of acceptance or rejection of algorithm recommendations (2020, 2).

In this paper, we will address, to some extent, all four issues raised by Briganti and Le Moine, but with a narrower focus on c) and d), as they are directly related to the decision-making process in clinical practice.

The Decision-Making Process in Medicine

Why do so many like the medical series *House, M.D.*? Most people are impressed by Gregory House's ability to make accurate diagnostic decisions that save lives, despite all the factors surrounding him. As on the screen, so in reality.

Knowledge and successful decision-making are extremely valued qualities in the profession (Croskerry 2014). But knowledge is something that is acquired, that is accessible and that is not so problematic. It seems that the way physicians make decisions involves more problems than what they know (Croskerry 2020a, 165), or: “by far the greatest number of errors we make in medicine are in the ways through which our thoughts and feelings impact our decision making” (Croskerry 2020b, 1).

The percentage of diagnostic errors in medicine ranges from under 5 % up to 15 %, depending on the speciality (Berner and Graber 2008) and errors in diagnosis are recognised and highlighted by the World Health Organization (Cresswell et al. 2013). Croskerry states that “optimal diagnosis depends upon optimal decision making which, in turn, depends upon optimal rationality. In fact, we can refine this further and say that clinical prediction is the main challenge” (2020a, 165). Rationality and the decision-making process are influenced by 40 contextual factors, which is an important point when we talk about AI tools in medicine. The spectrum of factors ranges from individual cognitive characteristics to health systems, culture, etc. (Croskerry 2020a, 165).

Decision-making in medicine is considered within the framework of dual process theory (Croskerry 2005, 2009, 2017a; Croskerry et al. 2017). The theory states that our cognition can be described by System 1 and System 2. System 1 is intuitive, fast, automatic, natural and acquired through experience. System 2 is slower and involves reflective thinking. System 2 can influence system 1, but also vice versa.

It is important to note that there are numerous biases that influence the human decision-making process through System 1. In the diagnostic process, for example, we can expect to see ‘classic’ cognitive biases such as the availability bias, which “occurs when a clinician judges the likelihood of a diagnosis based on how easily similar examples come to mind (whether because the diagnosis is seen frequently, a rare diagnosis was seen recently, or a specific case had a significant emotional impact, making it easier to recall)” (Morgenstern 2022). Another possible bias is confirmation bias, or “the tendency to look for, and favour, evidence that supports our prior beliefs and to discredit evidence that refutes them” (Kosmidis 2021, 83). There are many other possible biases, such as anchoring, framing, ascertainment bias, unpacking failure, etc. (Croskerry 2020a, 167). Given that System 1 is fast and autonomous, it is more energy efficient for the body. In the situations where cognitive overload occurs (also in clinical practice), the human mind prefers System 1 and this opens the door to the biases that could lead to making erroneous decisions.

We see the possible issues during the decision-making process in clinical practice. Therefore, the question arises: can AI contribute to the decision-making process in clinical practice?

AI and The Decision-Making Process in Medicine

Since the mid-20th century, decision making in medicine has been heavily influenced by mathematics, symbolic logic, and probability, which provided "statistical approaches to perform diagnosis" (Buchard and Richens 2022, 160). In the 1980s, AI came into play with logic-based systems, "which, to this day, are among the most successful and finalized clinical decision support systems" (Ibid). However, these systems were mainly used as "hospital EHR subsystems, such as drug dosage and interaction tools, or for simple automated detection, such as rule-based alarms for continuous monitoring in critical care units or ECG machines able to perform diagnosis" (Ibid). As computers become more powerful, more medical data available, and the development of ML accelerated (especially deep learning methods), research into AI tools to support clinical decision named *Clinical Decision Support Systems* (CDSS) is again gaining a momentum. Buchard and Richens rightly note that the slow development of automated decision tools points to the complexity of the decision-making process in medicine and the difficulty of implementing AI tools in a set environment (Ibid).

Given the complexity and vastness of currently available digitalized medical data, CDSS are "designed to assist the physician with medical decision support, especially very complex clinical cases. In this way these systems may provide a bridge between clinical observations with medical science and have an impact, depending on background algorithms, to affect the ultimate choices made by medical doctors in a sharp medical setting" (Lidströmer, Aresu, and Ashrafian 2022, 67). According to Lidströmer and his team the aim of the system is to support the physician, offering suggestions or reminders "not to miss to consider a diagnosis in a complex case, i.e. the clinician uses both the system and own knowledge to evaluate the cases" (Ibid, 68). Sutton and his colleagues describe traditional CDSS as a software with "which the characteristics of an individual patient are matched to a computerized clinical knowledge base and patient-specific assessments or recommendations are then presented to the clinician for a decision" (Sutton et al. 2020, 1). CDSS are used for various functions, "including diagnostics, alarm systems, disease management, prescription (Rx), drug control, and much more. They can manifest as computerized alerts and reminders, computerized guidelines, order sets, patient data reports, documentation templates, and clinical workflow tools" (Ibid.).

We are not going to divide or direct our argumentation depending on the methods for clinical decision-making, like logic-based, learning-based, Bayesian or combination (Buchard and Richens 2022, 160). The same goes for 'standard' classification of CDSS as knowledge-based (rules IF-THEN plus data) or non-knowledge based (data plus AI/ML). Although, given that non-knowledge based CDSS are black-boxes when they are based on deep learning, our argumentation will have in mind more this kind of CDSS, that are still not wide-spread in medicine (Sutton et al. 2020, 1). In terms of healthcare decision domains (basic science and medical research, clinical, logistics, and policy-making domain), we will mainly discuss the clinical domain, which includes tasks such as triage, diagnosis, prognosis, therapeutic decisions, medical imaging, etc. (Buchard and Richens 2022, 161). CDSS developed for diagnosis are known as *diagnostic decision support system* (DDSS) and still do not have as much success as other kinds of CDSS. Sutton and his team state that the reasons include "negative physician perceptions and biases, poor accuracy (often due to gaps in data availability), and poor system integration requiring manual data entry" (Sutton et al. 2020, 5).

The CDSS are going through various testing phases and are "showing promising performance in pre-clinical, in silico, evaluation, but few have yet demonstrated real benefit to patient care" (Vasey et al. 2022, 924). For early, small-scale and direct clinical evaluation, Vasey and colleagues have proposed a reporting guideline called DECIDE-AI. Their reports address four important aspects: "proof of clinical utility at small scale, safety, human factor evaluation, and preparation for larger scale summative trials" (Ibid). Furthermore, the authors emphasise that these types of AI-based clinical decision support systems must support, not replace, human intelligence. However, in their reporting guide, they have not considered the notions of interpretability and trust, which are important factors in the development of the human-machine relationship. In terms of trust, for example, two extremes are possible: over-reliance (when experts trust or rely on the CDSS too much) and the 'carry-over effect' (after a certain period of using the CDSS, a training effect leads one to think that it is no longer necessary to use it) (Sutton et al. 2020, 7).

A recent meta-study on the concrete application of AI in perioperative clinical decision-making concluded that AI can be useful at every step of the decision-making process (Giordano et al. 2021). Based on their own knowledge and current discussions between experts, the authors identified five topics related to the application of AI in clinical decision-making: risk stratification, optimisation of patient outcome, early warning of acute decompensation, potential bias in AI and future medical education. For four of these topics, AI could play an important role and improve medical and healthcare options. Nevertheless, caution is needed regarding potential biases.

Since AI is a product of the human mind, and we have seen how bias can affect people's decisions, it is not surprising that bias is also present in the field of AI. Numerous forms of bias have been identified at different stages of algorithm development and application: from data to algorithm (most common), from algorithm to user, and from user to data (Mehrabi 2022). Safety measures against bias are developed, such as: human-in-the-loop approaches, logic-based constraints and safe reinforcement learning (Buchard and Richens 2022, 169). This is of particular relevance for CDSS based on deep learning (black box), especially given that these kinds of CDSS could enable more accurate diagnosis (Sutton et al. 2020, 5).

Another problem for CDSS is that real-life context enters the clinical decision-making process and it is not easy to translate contextual factors into data for AI. Moreover, there can be crucial changes in context that lead to a dataset shift – a mismatch between the data on which the AI tool has been trained and the data with which the AI tool is working (Quiñonero-Candela et al. 2009; Subbaswamy and Saria 2020). The most recent example of context change is pandemics. AI tools may not recognise when the context changes overnight and this has an impact on the data the AI tool is working with (Finlayson et al. 2021).

On the other hand, AI tools can be of great help in medical education, as it has been postulated that it will take 73 days for medical knowledge to double in 2020 (Densen 2011). Another side of the educational perspective is the urgent need to train: a) future physicians trained in the field of AI (Giordano et al. 2021); b) future specialists in computing who have biomedical knowledge (Kaushal and Altman 2019); c) experts who have a deeper understanding of the social and ethical implications of the use of technology (Ibid.); the general population, in order to overcome possible biases towards AI solutions in medicine.

Having presented only parts of the complex topic of possible applications of CDSS (for an overview of benefits, harms and solutions to mitigate harm, see Sutton et al. 2020), let us take a brief look at how humans and artificial intelligence have been making decisions together in the field of chess for many decades.

Chess and AI

Chess is not a sport that usually fills the front pages of news websites and newspapers. Yet the portals were recently flooded with news of a scandal after the long-time world chess champion Magnus Carlsen lost a game to the young player Hans Niemann during the Sinquefield Cup tournament in September 2022.

The day after the game, Carlsen announced that he was withdrawing from the tournament, which is very unusual, and that he was not allowed to talk about the reasons for his withdrawal. After a few weeks, after a lot of information had been released to the public, it became clear that the World Champion thought that Niemann had cheated in the game with the help of AI. Indeed, the golden rule in the classical chess game is that artificial intelligence (or any kind of human help) must not be used in any way. There are also strict controls where players are scanned before the game to look for devices that could be used to signal moves. Recently, these measures are becoming even more restrictive (Hudoon 2022).

Nevertheless, even if you are not allowed to use AI during a game (unless otherwise agreed), AI chess programmes are a routine tool for most chess players.

So what is the nature of this relationship between AI and humans and what can we learn from it?

Chess programmes began to show their face in the 1950s. Chess was one of the first areas to be conquered by AI. After the world champion, Garry Kasparov, won the first duel with the IBM supercomputer Deep Blue in 1996, he lost the return match 3.5 to 2.5. It was the first defeat of a world champion against an AI machine. Since then, chess programmes have flourished, are widely available and help chess players improve their game. Interest in chess has not waned; classical tournaments with only human players are still incomparably more interesting for spectators than tournaments with AI programmes. Moreover, the programmes have helped many players in their development, so that there are fewer and fewer differences in playing strength between professional players. On the other hand, there are more draws or fewer decided games at the highest level because of the good preparation with the help of AI.

AlphaZero – a Game Changer

A major step forward was taken in 2017 with the development of DeepMind's AI programme AlphaZero based on the deep reinforcement learning. AlphaZero teaches itself how to play chess. At the beginning, it only knows the rules for moving pieces (and some other rules), plays with itself and learns from defeats, victories and draws. Then it learns to checkmate, develops the branches of the candidate's moves and the ability to evaluate the position. It thus creates value for itself through reinforcement learning (Tomašev et al. 2022). After AlphaZero has played 200,000 games in a few hours, it already plays like a superhuman. Interestingly, human intelligence is still more efficient because it needs less experience and fewer games to learn the rules of chess and become an expert player over several years. AlphaZero does the same in a few hours, but with countless more games. Interestingly, AlphaZero shows a certain style of play that is amazingly aggressive and often relies on activity. The DeepMind team has also defeated champions the game of Go and Shogi.

AI and Humans – a New Relationship

Chess champion Carlsen actively uses AI and learns from it. Grandmasters acknowledge that AI helps them imagine new, unthinkable possibilities in positions and elevates them to a new level of understanding the game. Often, top grandmasters do not know how to explain why a move made by the AI programme is good. Before computers existed, human intuition about certain positions was crucial when making decisions about complex chess positions where it is difficult to calculate all possible directions of play.

Chess players use AI mainly for preparation. AI helps them evaluate positions, find new surprise moves and prepare for opponents. The player or his team, which is preparing, guides the AI to the positions to be studied, and ultimately the human decides which position to try to achieve on the board, taking into account the opponent's characteristics. So, the human takes into account the whole context of a particular game. But when the game is played over the board, the use of AI is considered cheating. In a way, you could say: the AI helps in making decisions about certain positions, but over the board, the responsibility lies solely with the player. Today, AlphaZero is also used to develop and study different variants of the game (Tomašev et al. 2022).

Another important insight into the use of AI in the decision-making process in chess is the so-called Kasparov's Law. In 2005, Kasparov observed an online freestyle tournament where players could use AI. It was expected that strong grandmasters with great programmes would win the tournament. But that was not the case. Two amateur players using three different computers won the tournament. The way they used AI prevailed over strong grandmasters plus AI or strong AI only. Kasparov concluded: "It was a triumph of process. A clever process beat superior knowledge and superior technology. It didn't render knowledge and technology obsolete, of course, but it illustrated the power of efficiency and coordination to dramatically improve results. I represented my conclusion like this: *weak human + machine + better process* was superior to a strong computer alone and, more remarkably, superior to a *strong human + machine + inferior process*" (Kasparov 2017, 236). Today this is referenced as Kasparov's law. The conclusion from this insight is that AI was not efficient as AI plus humans. The synergy between humans and AI is more efficient when humans understand what they are doing and understand the AI they are using.

It is not just that medicine could use insights from chess about AI. There are ideas about how understanding the relationship between chess players and AI can help in the development of future war strategies (Phillips-Levine et al. 2022). Senior computer scientist Andrew Lohn (2020) predicts that AI will be the first to enter the world of war defence strategies thanks to knowledge gained from chess and Kasparov's Law. AI will help train soldiers/military personnel and help them spot mistakes before they happen. As AI progresses, it will gradually enter direct combat as an equal partner to humans (as in chess) and shape the fight. Interestingly, Lohn believes that AI will not replace humans in combat, just as it has not displaced humans from chess.

The Maia Program – a New Understanding of the Relationship

Who likes to lose all the time from the computer machine? One of the biggest challenges today is to develop a programme that plays like a human and makes decisions that adapt to the strength of each player. While AlphaZero trains itself, the Maia programme is trained on human games to play as humanly as possible (McIlroy-Young et al. 2020a; McIlroy-Young et al. 2020b). Maia can successfully predict human decisions at the individual level (McIlroy-Young et al. 2022a). When looking at selected games from a dataset, Maia can identify the playing style and the chess player in question with high accuracy (McIlroy-Young et al.

2022b). A more fine-tuned Maia can also detect and predict the mistakes that an individual player is likely to play in certain positions. Thus, if Maia points out persistent mistakes made in a player's decision-making process, players can learn from Maia and avoid repeating them.

This is a step that will not only help chess players, but also provide new opportunities to develop the relationship between humans and AI in general. For example, can AI tools like Maia be used in medicine for detecting/predicting wrong decisions?

Conclusion

AI is, and will be, transforming medicine, opening up new, previously unimagined possibilities. The time is ripe to develop ethical principles for the process of development, evaluation, and application of CDSS, especially non-knowledge based CDSS. We strongly believe that ethical principles should be essential in shaping the collaboration between clinicians and AI. However, this is easier said than done.

First of all, it must be understood that each field of medicine has its own particular conditions. As far as the decision-making process in clinical practice and the co-operation with CDSS is concerned, the most important insight in determining the future direction seems to be that a standalone or autonomous AI tool is not an option we should be looking for. Given the complexity of the data, the great potential for bias (from the creation of algorithms to their use), and the inevitable element of clinical context in which the AI tool is applied and the final decision is made, we can say with a high degree of confidence that the goal is an algorithm that collaborates with, or supports, clinicians. Therefore, AI tools that satisfy the principles of explainability and interpretability are desired, allowing physicians to review and explain the incorporation of AI suggestions into their decisions. This is especially true for deep learning algorithms. Another important fact is the development of physicians' confidence in AI predictions which could be fostered by integrating 'uncertainty estimation', with the aim of providing physicians with useful indications (Panel 2022, V).

An important application of AI tools can be expected in the field of medical education. Using ideas like the Maia chess programmes, this is a step forward in bridging the communication gap between the human and artificial, which is obvious when we compare artificial and human ways of looking at and solving a particular problem. Also, we still lack experts educated in the combined field of biomedicine and informatics.

We can agree with Human-Centred AI researcher Pontus Wärnestål (2019) when he claims that success "is not only about data and algorithms. Value is not in the algorithm itself. Value is instead derived from algorithms that have been designed for the context where they should operate". To create human-centred AI, Kasparov's law is a great help in setting the framework for human-AI collaboration.

We need to keep in mind that the process of developing human-centred AI is aimed at providing better healthcare for the patient. Given the complexity of applying AI tools and, in particular, understanding algorithms based on deep learning, the moment of informing the patient about treatment steps will also be an ethical challenge.

To conclude: As a society, we have a responsibility to mandate the principles of explainability and interpretability in AI tools, especially for those products that aim to contribute to the decision-making process. This will contribute to the empowerment of physicians and help mitigate possible negative effects on user skill (for example, over-reliance and 'carry-over effect'). Furthermore, as a society, we have a responsibility to help preserve the freedom and autonomy of human intelligence, in our case by allowing physicians to say 'no' to the use of CDSS and its recommended treatment options after a carefully reasoned decision, but also to 'turn off' AI tools already in use because of identified malfunctions (e.g. bias).

References

Acampora, Giovanni, Diane J. Cook, Parisa Rashidi, and Athanasios V. Vasilakos. 2013. A survey on ambient intelligence in health care. *Proceedings of the IEEE* 101, no. 12:2470-2494. https://doi.org/10.1109/JPROC.2013.2262913

Apple. 2022. *Take an ECG with the ECG app on Apple Watch*, 12th September. https://support.apple.com/en-us/HT208955 (accessed 23. 9. 2022).

Arnold, Melina, Mónica S. Sierra, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2017. Global patterns and trends in colorectal cancer incidence and mortality, *Gut.* 66, 4:683-691. https://doi.org/10.1136/gutjnl-2015-310912

Berner, Eta S., and Mark L Graber. 2008. Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine* 121, 5: S2-23. https://doi.org/10.1016/j.amjmed.2008.01.001

Bracanović, Tomislav. 2021. Umjetna inteligencija, medicina i autonomija. *Nova prisutnost* XIX, no. 1:63-75. https://doi.org/10.31192/np.19.1.5

Buchard, Albert and Jonathan G. Richens. 2022. Artificial Intelligence for Medical Decisions. In: *Artificial Intelligence in Medicine*, 159-181. Eds. Niklas Lidströmer and Hutan Ashrafian, Cham: Springer.

Cresswell, Kathrin M., Sukhmeet S. Panesar, Sarah A. Salvilla, Andrew Carson-Stevens, Itziar Larizgoitia, Liam J. Donaldson, David Bates, and Aziz Sheikh - on behalf of the World Health Organization's (WHO) Safer Primary Care Expert Working Group. 2013. Global Research Priorities to Better Understand the Burden of Iatrogenic Harm in Primary Care: An International Delphi Exercise. *PLoS Med* 10, 11: e1001554. https://doi.org/10.1371/journal.pmed.1001554

Croskerry, Pat. 2005. The theory and practice of clinical decision making. *Canadian Journal of Anesthesia* 52, 1: R1–8. https://doi.org/10.1007/BF03023077

– – –. 2009. A universal model for diagnostic reasoning. *Academic medicine: journal of the Association of American Medical Colleges* 84, 8: 1022-8. https://doi.org/10.1097/ACM.0b013e3181ace703

– – –. 2014. *How Doctors Think*. YouTube. 17th November. https://www.youtube.com/watch?v=GFE6D5460oE

– – –. 2017. The rational diagnostician. In: *Diagnosis: interpreting the shadows*, 113-129. Eds. Pat Croskerry, Karen Cosby, Mark L. Graber, and Hardeep Singh. Boca Raton: CRC Press.

– – –. 2020a. Sapere aude in the diagnostic process. *Diagnosis* (Berlin, Germany) 7, 3: 165–168. https://doi.org/10.1515/dx-2020-0079 165

– – –. 2020b. *The Cognitive Autopsy: A Root Cause Analysis of Medical Decision Making*. New York: Oxford University Press.

Croskerry, Pat, Karen Cosby, Mark L. Graber, and Hardeep Singh (Eds.). 2017. *Diagnosis: interpreting the shadows*. Boca Raton: CRC Press.

Deloitte and MedTech Europe. 2020. *The socio-economic impact of AI in healthcare*. https://www.medtecheurope.org/wp-content/uploads/2020/10/mte-ai_impact-in-healthcare_oct2020_report.pdf (accessed 21. 9. 2022).

Densen, Peter. 2011. Challenges and opportunities facing medical education. *Trans Am Clin Climatol Assoc*., 122: 48–58.

Giovanni, Briganti, and Le Moine Olivier. 2020. Artificial Intelligence in Medicine: Today and Tomorrow. *Frontiers in Medicine* 7, 4. https://doi.org/10.3389/fmed.2020.00027

Green, Brian Patrick. 2018. Ethical Reflections on Artificial Intelligence. *Scientia Et Fides* 6, 2:9-31. https://apcz.umk.pl/SetF/article/view/SetF.2018.015

Kasparov, Garry. 2017. *Deep Thinking - where machine intelligence ends and human creativity begins*. New York: Public Affairs.

Kaushal, Amit and Russ B. Altman. 2019. Wiring Minds. *Nature* 576, 7787:S62–S63, 18th December. https://doi.org/10.1038/d41586-019-03849-x

Kosmidis, Michail. 2021. Confirmation Bias. In: *Decision Making in Emergency Medicine*, 83-88. Eds. Manda Raz and Pourya Pouryahya. Singapore: Springer. https://doi.org/10.1007/978-981-16-0143-9

Finlayson, Samuel G., Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. 2021. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med*, 385:283-6. https://doi.org/10.1056/NEJMc2104626

Giordano, Chris, Meghan Brennan, Basma Mohamed, Parisa Rashidi, François Modave, and Patrick Tighe. 2021. Accessing Artificial Intelligence for Clinical Decision-Making. *Front. Digit. Health* 3:645232. https://doi.org/10.3389/fdgth.2021.645232

Horvat, Saša, Piotr Roszak, and Brian J. Taylor. 2022. Is It Harmful? A Thomistic Perspective on Risk Science in Social Welfare. *J Relig Health* 61, 3302–3316. https://doi.org/10.1007/s10943-021-01452-x

Hudoon, Fatima. 2022. Chess: Niemann-Carlsen scandal prompts new security measures. *DW*, 25th October. https://www.dw.com/en/chess-niemann-carlsen-scandal-prompts-extra-security-measures/a-63547994 (accessed 29. 10. 2022).

Lidströmer, Niklas Federica Aresu, and Hutan Ashrafian. 2022. Introductory Approaches for Applying Artificial Intelligence in Clinical Medicine. In: *Artificial Intelligence in Medicine*, 57-75. Eds. Niklas Lidströmer and Hutan Ashrafian, Cham: Springer.

Liu, Xiaoxuan, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 1, 6: e271-e297. https://doi.org/10.1016/S2589-7500(19)30123-2.

Lohn, Andrew. 2020. What chess can teach us about the future of ai and war. *War on the Rocks*, 3 January. https://warontherocks.com/2020/01/what-chess-can-teach-us-about-the-future-of-ai-and-war (accessed 29. 10. 2022).

Lopez-Jimenez, Francisco, Zachi Attia, Adelaide M Arruda-Olson, Rickey Carter, Panithaya Chareonthaitawee, Hayan Jouni, Suraj Kapa, Amir Lerman, Christina Luong, Jose R Medina-Inojosa, Peter A Noseworthy, Patricia A Pellikka, Margaret M Redfield, Veronique L Roger, Gurpreet S Sandhu, Conor Senecal, and Paul A Friedman. 2020. Artificial Intelligence in Cardiology: Present and Future. *Mayo Clin Proc*. 95, 5:1015–39. https://doi.org/10.1016/j.mayocp.2020.01.038

McIlroy-Young, Reid, Ashton Anderson, Jon Kleinberg, and Siddhartha Sen. 2020a. The human side of AI for chess. *Microsoft Research Blog*, 30th November. https://www.microsoft.com/en-us/research/blog/the-human-side-of-ai-for-chess/ (accessed 25. 10. 2022).

– – –. 2020b. Aligning Superhuman AI with Human Behavior: Chess as a Model System, arXiv:2006.01855.

McIlroy-Young, Reid, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2022a. Learning Models of Individual Behavior in Chess, arXiv:2008.10086.

– – –. 2022b. Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess, arXiv:2208.01366.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2020. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635v3 [cs.LG]. https://doi.org/10.48550/arXiv.1908.09635

Misawa, Masashi et al. 2018. Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience. *Gastroenterology* 154, 8: 2027–2029.e3. https://doi.org/10.1053/j.gastro.2018.04.003 /

Morgenstern, Justin. 2022. Availability Bias. In: *Decision Making in Emergency Medicine*, 47-53. Eds. Manda Raz and Pourya Pouryahya. Singapore: Springer. https://doi.org/10.51684/FIRS.125778

Paul, Debleena, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, and Rakesh K. Tekade. 2021. Artificial intelligence in drug discovery and development. *Drug discovery today* 26, 1:80–93. https://doi.org/10.1016/j.drudis.2020.10.010

Phillips-Levine, Trevor, Michael Kanaan, Dylan Phillips-Levine, and Noah “Spool” Spataro. 2022. Weak Human, Strong Force: Applying Advanced Chess to Military AI, *War on the Rocks*, 7th July. https://warontherocks.com/2022/07/weak-human-strong-force-applying-advanced-chess-to-military-ai/ (accessed 04. 11. 2022).

Qadir, Hemin Ali, Younghak Shin, Johannes Solhusvik, Jacob Bergsland, Lars Aabakken, and Ilangko Balasingham. 2021. Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction, *Medical Image Analysis* 68, 101897. https://doi.org/10.1016/j.media.2020.101897

Quiñonero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer and Neil D. Lawrence. 2009. *Dataset Shift in Machine Learning*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/9780262170055.001.0001

Shaukat, Aasma, David R. Lichtenstein, Samuel C. Somers, Daniel C. Chung, David G. Perdue, Murali Gopal, Daniel R. Colucci, Sloane A. Phillips, Nicholas A. Marka, Timothy R. Church, and William R. Brugge, for the SKOUTTM Registration Study Team. 2022. Computer-Aided Detection Improves Adenomas per Colonoscopy for Screening and Surveillance Colonoscopy: A Randomized Trial. *Gastroenterology* 163, 3:732–741. https://doi.org/10.1053/j.gastro.2022.05.028

Subbaswamy, Adarsh, and Suchi Saria. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21:345-52. https://doi.org/10.1093/biostatistics/kxz041

Sutton, Reed T., David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *Npj Digit. Med.* 3, 17:1-10. https://doi.org/10.1038/s41746-020-0221-y

Tomašev, Nenad, Ulrich Paquet, Demis Hassabis, and Vladimir Kramnik. 2022. Reimagining chess with AlphaZero, *Communications of the ACM* 65, 2:60-66. https://doi.org/10.1145/3460349

Topalovic, Marko et al. 2019. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *The European respiratory journal* 53, 4:1801660. https://doi.org/10.1183/13993003.01660-2018

Valueva, M.V., N.N. Nagornov, P.A. Lyakhov, G.V. Valuev, and N.I. Chervyakov. 2020. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation* 177, 232-243. https://doi.org/10.1016/j.matcom.2020.04.031

Vasey, Baptiste et al. 2022. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, *Nature Medicine* 28, 924–933, 1-12. https://doi.org/10.1136/bmj-2022-070904

Wallis, Claudia. 2019. How Artificial Intelligence Will Change Medicine. *Nature* 576, 7787:S48. Https://doi.org/10.1038/d41586-019-03845-1

Wärnestål, Pontus. 2019. Why Human-Centered Design is Critical to AI-Driven Services. *Inuse*. 9th September. https://www.inuse.se/read/why-human-centered-design-critical-ai-driven-services/ (accessed 10.10.2022).

Yang, Young Joo and Chang Seok Bang. 2019. Application of artificial intelligence in gastroenterology. *World journal of gastroenterology* 25, 14:1666–1683. Https://doi.org/10.3748/wjg.v25.i14.1666

Zhao, Shengbing, Shuling Wang, Peng Pan, Tian Xia, Xin Chang, Xia Yang, Liliangzi Guo, Qianqian Meng, Fan Yang, Wei Qian, Zhichao Xu, Yuanqiong Wang, Zhijie Wang, Lun Gu, Rundong Wang, Fangzhou Jia, Jun Yao, Zhaoshen Li, and Yu Bai. 2019. Magnitude, Risk Factors, and Factors Associated With Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-analysis. *Gastroenterology* 156, 6:1661–1674.e11. https://doi.org/10.1053/j.gastro.2019.01.260

World Health Organization. 2016. *Global strategy on human resources for health: workforce 2030*, Geneva. https://www.who.int/hrh/resources/pub_globstrathrh-2030/en/ (accessed 08. 9. 2022).

Stjepan Štivić

DIGITAL TWIN AND ETHICAL DILEMMAS

Introduction

The intention of this text is to present briefly the idea of the digital twin technology. The text covers the consuetudinary definition, history of the concept, application review with an emphasis on healthcare and ethical dilemmas when used in healthcare. In the most general terms, when speaking of digital twin technology we are talking about a ubiquitous process of digitisation and virtualisation of things, processes, situations and living beings (Lv & Xie 2021, 3–14).

The topic of digital twin technology in recent years has drawn attention in various fields, particularly in the field of ethics. On first glance, our initial association with the term 'digital twin' is childbirth and two offspring born from the same pregnancy. It is obvious that this is not the case; that it is a metaphorical term. The term digital twin emphasises a very solid and strong relationship or bond between two very close entities. In that trail, we can define the digital twin technology or a digital twin as a "high-precision simulation that maps and models events or objects in real time" (Braun & Krutzinna 2022, 1). It can be described in different terms or associations: copy, clone, avatar, replica, representation etc.

Digital twin technology represents an improvement and the next level in the modelling of reality. However, it is currently an emerging technological idea. It is not entirely a finished product, but a model, which is being developed and its future development depends on different areas of application – architecture, geography, manufacturing, healthcare, logistics, applications to traffic and smart cities etc. Digital twins of human bodies for use in healthcare have also been developed. "They are faced with measurement and big data challenges, extreme levels of complexity, and ethical issues" (Helbing & Sánchez-Vaquerizo 2022, 19)

Definition of Digital Twin Technology

As with all emergent technologies, in the example of the digital twin, the scientific language is intertwined with the language of futurology or even science fiction. One such narrative comes from the Gartner Report, which named the digital twin one of its "Top 10 strategic technology trends for 2017". Considering the previously stated, as well as different areas of application, there are various ways to attempt to define of a digital twin.

There are two ways to define the digital twin technology [DT]. First one is a strict scientific definition of DT, and the other one is a popular and descriptive definition of DT. The last one is covered by enthusiastic and futuristic benefits and use of DT in everyday life or science. The first is a scientific one with an emphasis on the technical dimension of DT functions in different fields of applications.

One could define DT as “a virtual representation of specific products, systems or machines that accompany their physical counterparts analogous to the real product lifecycle – over their entire life” (Lindow 2018, 7). However, this definition remains quite general. The following one remains incomplete or fallacious because it is not fully applicable to, for example, healthcare systems: DT “is a digital replica or a representation of a physical or an intangible system that can be examined, altered and tested without interacting with it in the real world and avoiding negative consequences” (Miskinis 2018). A much more precise definition can be found at the IBM website: “A digital twin is a virtual representation of an object or system that spans its lifecycle, is updated from real-time data, and uses simulation, machine learning and reasoning to help decision-making.” (IBM)

In general, there are several acknowledged benefits of DT that are mentioned in the upper definition. The basic advantage over previous forms of simulation is that DT has inherent exploratory and prospective power, as it can provide an advanced insight into ‘what-if’(hypothetical) scenarios (Braun 2021, 394; Helbing & Sánchez-Vaquerizo 2022, 2). Thus, a digital twin allows more accurately performing different procedures or otherwise unfeasible and very complex experiments. This opens up space for moral reflection in areas where it would otherwise be inapplicable due to speed, lack of manoeuvres or sheer complexity. However, implementation of ethical principles and safety precautions through hypothetical scenario enables procedure in which it is possible – in a much more accurate way – to avoid unethical experiments that are incompatible with the principles of responsible innovation and engineering (Huang, Kim & Schermer 2022).

Development of the Digital Twin Technology

The procedure of representing objects or situations has a long tradition. It could be said that we already have the first sketches in the Lascaux caves, or in more contemporary forms of the interplay of arts and science in the age of Leonardo da Vinci, Verancsic’s Faustus etc. In the qualitative sense the next progress appears in the age of information technology; later the procedure was improved with software tools used in architecture, engineering, geography, manufacturing, healthcare, logistics, applications to traffic and smart cities etc.

Computer Assisted Design (CAD) programs had a strong impact on development of the idea of DT. These are used widely and have enabled ever more detailed three-dimensional visualisations of planned buildings, allowing for advanced improvements and modifications before they were built (Helbing & Sánchez-Vaquerizo 2022, 2). In numerous articles, the theory is proposed that the idea of DT was strongly developed at NASA, and it is explained that some full-scale mockups of early space capsules were used on the ground to mirror and diagnose problems in orbit. These physical replicas gave way to fully digital simulations. Furthermore, opinion prevails that the concept of DT was originally presented for use in the aerospace field at NASA during the Apollo 13 mission. After that pioneer work, DT was used in the maintenance and quality assurance of flight process and aerospace flight machines through simulation (Garg 2021, 33; Bruynseels, Santoni de Sio & Hoven 2018).

This thesis is not completely true. "The origin of the DT is attributed to Michael Grieves and his work with John Vickers of NASA, with Grieves presenting the concept in a lecture on product life-cycle management in 2003. […] The initial description defines a Digital Twin as a virtual representation of a physical product containing information about said product, with its origins in the field of product life-cycle management." (David et. al 2020, 36) Grieves himself confirms these facts and clarifies the difference between the physical mockups from previous periods and the concept of DT (Grieves 2022a, 2). The DT concept presupposes three components of a physical entity: a digital representation of that entity, the mutually orientated data relations that feed data from the physical to the digital representation, and the same back from the digital representation to the physical (David et. al 2020, 36). This does not mean that the physical entity must precede the digital representation (Grieves 2022b, 3).

Today, we are talking about DT as a part of the fourth, or even fifth, industrial revolution. DT technology relies on artificial intelligence, internet-of-things, big data, big data analytics etc. DT is based on modern technologies and therefore constructed so that it can receive input from sensors gathering data from a real-world counterpart. This allows the twin to simulate the physical object in real time, while offering insights into performance and potential problems. A particular DT can be complex or simple depending on the amount (quality, analysis etc.) of data that is used to build and update it. That will determine how accurately a real-world counterpart will be simulated (Helbing & Sánchez-Vaquerizo 2022, 1).

Application of the Digital Twin Technology

DT is attributed a revolutionary character in the scientific, technological and socio-cultural sense. DT technology is present in many fields*: in the aerospace field* – flight simulation, driverless models, error detection; *in research of driverless cars* – virtual simulation testing, environment simulation, safe driving warnings; *in the intelligent manufacturing field* – virtual workshops, error early warnings, carbon emission forecasts; *in the concept of the smart city* – intelligent transportation, traffic accident tracking, real time traffic monitoring; *in robotics* – e.g. the costs go down by about 90% because there are no lab fees or equipment setup charges; *in economics* – to test and simulate, in order to avoid any mistakes on physical prototypes, or, in other words, saving time and money from costly errors that could have occurred through experimentation on materials or manufacturing processes (Helbing & Sánchez-Vaquerizo 2022, 1–15).

For instance, the McLaren and Red Bull teams use DT technology for their racing cars in Formula One racing. The DHL delivery company is making more efficient models of supply chains with a digital map. Shanghai and Singapore are both being replicated in the digital world with DT technology with an idea of improving the design and operations of buildings, transport systems and streets. Amongst others, in Singapore DT helps to find new routes for citizens to navigate, avoiding areas of pollution (Wakefield 2022).

The category which we did not directly mention above, and which is becoming one of the most receptive areas for DT technology, is healthcare and medicine. Within this category, DT is developing in areas of immunology, cardiology, transplantation medicine, diagnostics, radiology etc. (Reinhard Laubenbacher et al.) to outline a roadmap for meeting challenges and building a prototype of an immune DT (Laubenbacher, Niarakis & Helikar 2022). The 'Living Heart Project' was started by Dassault Systèmes in order to create an accurate virtual model of the human heart. This model opens the possibility of hypothetical scenarios through which it would be possible to analyse and test various procedures and devices. Boston Children's Hospital is now using this technology to map out real patients' heart conditions, while at Great Ormond Street Hospital in London, a team of engineers is working with clinicians to test devices that may help children with rare and difficult-to-treat heart conditions (Wakefield 2022).

Ethics and Digital Twin in Healthcare

Digital twin technology has, and will have, multiple use and benefits in healthcare and medicine. The prevailing opinion that causes enthusiastic acceptance in the public space is that DT technology has great potential to transform the existing healthcare system and to modify medical praxis. The central concept of that transformation is personalisation of healthcare and medicine in general (Bruynseels, Santoni de Sio & Hoven 2018; Garg 2021; Braun 2021; Braun & Krutzinna 2022). More precisely, it is about moving from a universal approach to an individualised approach. A digital twin should, via a virtual model of human cells, organs, systems, or entire bodies, collect very detailed biophysical and lifestyle information from a person over a long period of time. A virtual model, as such, should ensure early diagnosis, exact cure/treatment and accurate prediction of the health status of an individual (Braun 2021, 395).

Pei-hua Huang et al., in a recent article, offered the definition of DT applied to healthcare. The definition has two key parts: its technical description, and its definition of purpose. The technical description emphasises that DT is a computer model, which is guided by collected data and which is in constant interaction with the source of data and new upgrades. The purpose of DT is to get an individual's health picture as faithfully as possible, so that it can be modified.

A digital twin for a personalised healthcare service is a data-driven interactive computerised model that aims to offer health-related information which properly simulates or predicts the health conditions of a particular person. (Huang, Kim & Schermer 2022)

From our perspective, ethical dilemmas related to DT in healthcare could be divided into two categories: a concrete set of technical issues, and speculative questions that are no less real. By the first category we mean a set of dilemmas that are part of concrete processes and, based on that, must be solved without delay. It is about setting a boundary between the need for the development of the process and a possible threat to an individual's rights. The issue of individual autonomy is one of the crucial questions regarding hyper-collection of data, data ownership, data accessibility, decontextualisation of disease, epistemic injustice (patient testimony vs. DT information). Furthermore, the issue of the right to privacy related to data collection, data brokerage, hacking possibility, as well as the issue of distortion of the understanding of health regarding data quality, biased algorithms, a biased training dataset, epistemic injustice, and overdiagnosis. Pei-hua Huang et al. state the mentioned issues and add ethical issues of the right to bodily integrity (e.g. overdiagnosis) and doctor-patient relationship (e.g. epistemic injustice) (Huang, Kim & Schermer 2022).

Speculative ethical questions concern the direction of development, goals and the very purpose of DT in healthcare. They should indicate the possible consequences and their possible perniciousness for man. In our perspective, the remark would refer to the methodology. The methodology for exploring the possibilities of the digital twin and its use should have a corrective character, or the digital twin should place humans at the centre, if we want it to be an ethically acceptable technology; in other words, DT should be a human-centred technology. That means that the direction of development and goals of DT in healthcare and medicine should not have any ideological, economical, or simply inhumane motives. There should be no fundamental ethical conundrums in the use of the digital twin, as long as we understand and acknowledge that the digital twin does not have the possibility, methodologically speaking, of absolute representation and final solutions. Otherwise, attempts to realise an absolute representation of the body could be disastrous. For example, if DT technology were developed as a perfect and absolute representation of the human body, we would soon be faced with reductionism, in which a human could be seen merely as a machine. Furthermore, DT technology may find itself on the slippery slope of steering clinical medicine toward what is known as healthism. This idea presents the health as a process of 'perfectisation', which is basically unattainable. The result of that is the creation of a new lifestyle. This new lifestyle will surely impact what is traditionally considered to be therapy or treatment (Bruynseels, Santoni de Sio & Hoven 2018). In this respect, acceptance of normal ageing and deterioration of the organism could become a problem, and a space would open up for anti-ageing medicine as a dominant framework. In this context, many other questions arise such as, could DT technology replace the patient-physician relationship in the long term, not only in a technical perspective but in the ethical perspective also. Could it happen that a physician in the new paradigm loses his compassion in the long run, and becomes a mere technician?

It should be noted that there are clear positive potentials of DT technology. DT can assist, improve, and speed up operations, a patient's experience could be improved, it could reduce the costs of infrastructure; in general, it could provide a better service for patients. The use of simulation of various operative interventions will reduce the percentage of errors in vivo. DT technology could solve one of the key bioethical controversies: the testing of procedures and preparations on animals could be completely removed, or significantly reduced, because the simulation of the effects and consequences of the procedures and preparations will be done by a digital replica.

Conclusion

DT technology is part of the ubiquitous process of digitisation and virtualisation of things, processes, situations and living beings. It aims to produce highly realistic models of real systems. DT technology differs from digital copies or animated models in interaction with reality and with their physical counterparts. As with all emergent technologies, in the example of the DT, the scientific language is intertwined with the language of futurology or even SF. DT technology is already present in many scientific and industrial fields. However, DT technology has, and will have, multiple uses and benefits in healthcare and medicine. A great potential to transform the existing healthcare system and to modify medical praxis is attributed to DT. The central concept of that transformation is personalisation of healthcare and medicine in general. There are futurological predictions and ideological tendencies that predict a complete simulation of the human body. However, DT technology offers opportunities as well as challenges, particularly in the fields of ethics. In healthcare, the idea of DT raises new challenges regarding a interaction between person and his digital simulation. This technology raises the question of what consequences the development of such a representation of a person may have.

References

Braun, Matthias. 2021. Represent me: please! Towards an ethics of digital twins in medicine. *Journal of Medical Ethics*; 47:394–400.

Braun, Matthias & Jenny Krutzinna. 2022. Digital twins and the ethics of health decision-making concerning children. *Patterns*, 3, 4:1–7.

Bruynseels, Koen, Filippo Santoni de Sio & Jeroen van den Hoven. 2018. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Frontiers in Genetics*, 9, 31:1–11.

Garg, Harita. 2021. Digital twin technology: Revolutionary to improve personalized healthcare. *Science Progress and Research*, 1,1: 32-34.

Grieves, Michael. 2022a. Intelligent digital twins and the development and management of complex systems. *Digital Twin*, 2, 8:1–24.

– – –. 2022b. Physical Twins, Digital Twins, and the Apollo Myth. *LinkedIn*, https://www.researchgate.net/publication/365872057_Physical_Twins_Digital_Twins_and_the_Apollo_Myth (accessed 21. 12. 2022)

Helbig, Dirk & Javier Argota Sánchez-Vaquerizo. 2022. Digital Twins: Potentials, Ethical Issues, and Limitations. In: *Handbook on the Politics and Governance of Big Data and Artificial Intelligence*. Edward Elgar, Forthcoming, Available at SSRN: https://ssrn.com/abstract=4167963.

Huang, Pei-hua, Ki-hun Kim & Maartje Schermer. 2022. Ethical Issues of Digital Twins for Personalized Health Care Service: Preliminary Mapping Study. *JMIR*, 24, 1.

IBM. What is digital twin? *IBM*. https://www.ibm.com/topics/what-is-a-digital-twin# (accessed 13. 12. 2022)

Jones, David & Chris Snider, Aydin Nassehi, Jason Yon, Ben Hicks. 2020. Characterising the Digital Twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52.

Laubenbacher. Reinhard & A. Niarakis, T. Helikar et al. 2022. Building digital twins of the human immune system: toward a roadmap. *Npj Digital Medicine*, 5, 64.

Lindow, Kai. 2018. When decency meets temptation. *Best practice*, 3: 6-11. T-Systems International Gmb.

Lv, Zhihan & Shuxuan Xie. 2021. Artificial intelligence in the digital twins: State of the art, challenges, and future research topics. *Digital Twin*, 1, 12:1–23.

Miskins, Carlos. 2018. *Explaining the definition of digital twin and how it works*. https://www.challenge.org/insights/what-is-digital-twin/ (accessed 24. 11. 2022)

Panetta, Kasey. 2016. Gartners Top 10 Technology Trends 2017. *Gartner* https://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017 (accessed 14. 12. 2022)

Wakefield, Jane. 2022. Why you may have a thinking digital twin within a decade. *BBC News*. https://www.bbc.com/news/business-61742884 (accessed 20. 12. 2022).

POVZETKI

Bojan Žalec

(NE)VERJETNOST ČLOVEKU PODOBNE UMETNE INTELIGENCE

Povzetek: Avtor se ukvarja z vprašanjem verjetnosti nastanka človeku podobne umetne inteligence in nastanka superinteligence. Trdi, da ni verjeten noben pojav. Verjetnost prvega zavrača sledeč argumentaciji Erika J. Larsona, ki temelji na bistvenih značilnostih človeške inteligence. Te vključujejo splošnost, intuicijo, zdravo pamet in abdukcijo. Ugotovljeno je, da nihče nima znanstvene zasnove, kako bi lahko takšno umetno inteligenco ustvarili ali kako bi se lahko razvila. Verjetnost pojava superinteligence je zavrnjena na podlagi argumentov, ki jih je podal François Chollet, pri čemer poudarja nesplošnost, situacijsko in kontekstualno naravnanost ter eksternalizem inteligence. Avtor dokazuje pomen pravilnega razumevanja, koncepta in opredelitve inteligence. Napačno razumevanje ima lahko negativne učinke, ki daleč presegajo akademsko sfero, vključno z neupravičenimi finančnimi dobički in škodo v smislu organizacije znanstvenega dela (*big science* ali *hive science*) ter spodbujanja pogojev za ustvarjalnost.

Ključne besede: verjetnost človeku podobne umetne inteligence, verjetnost superinteligence, splošnost, intuicija, abdukcija, zdrava pamet, partikularnost in eksternalizem inteligence.

Octavian-Mihai Machidon

MI OBLIKUJEMO UI IN UI OBLIKUJE NAS: FILOZOFSKI IN TEOLOŠKI RAZMISLEK O ALGORITEMSKEM DETERMINIZMU UI

Povzetek: Umetna inteligenca (UI) vse bolj postaja vseprisotna in avtonomna sila, ki spreminja našo družbo in način, kako ljudje komunicirajo s svetom okoli sebe in drug z drugim. Zaradi njenega osupljivega razvoja in široke uporabe se pojavljajo etične razprave in pomisleki glede mnogih družbenih vplivov UI, zlasti glede na impresiven potencial za družbeno preobrazbo, ki ga je UI že pokazala. Ta prispevek obravnava algoritemski determinizem UI kot silo za družbeno preobrazbo in zadeva razmišljanje dveh filozofov 20. stoletja, ki sta ugledna znanstvenika na področju medijske ekologije: Jacquesa Ellula in Marshalla McLuhana. Analiziran je transformacijski potencial umetne inteligence na družbeni in individualni ravni, da bi ugotovili, v kolikšni meri lahko na umetno inteligenco gledamo kot na samouresničujočo se, deterministično silo, ki načrtuje svet pod svojimi pogoji (kot je menil Ellul), in ali smo kot posamezniki izpostavljeni temu, da umetna inteligenca spremeni naše kognitivne in intelektualne sposobnosti (po McLuhanovi teoriji tehnologije kot podaljška človeka). Nazadnje se to delo obrača h krščanski antropologiji v iskanju osvobajajočega pogleda na človekov odnos do tehnologije na splošno in še posebej do UI, ki omogoča individualno odgovornost in krepi človeški nadzor, hkrati pa zmanjšuje deterministični potencial UI.

Ključne besede: umetna inteligenca; algoritemski determinizem; družbena preobrazba; krščanska antropologija; patristika.

Vojko Strahovnik

TRANSPARENTNOST SISTEMOV AI IN ČLOVEŠKEGA PRESOJANJA: ODGOVOR NA ARGUMENT DVOJNEGA STANDARDA

Povzetek: Problem transparentnosti umetne inteligence (UI) in strojnega učenja (ML) zadeva trditev, da domnevno nimamo uporabnega vpogleda v postopek odločanja ali oblikovanja priporočil nekaterih algoritmov UI. To je načelna težava, ki vodi do tega, da ljudje ne moremo spremljati, razumeti ali revidirati odločitev, ki jih sprejemajo takšni sistemi. Ta vidik sistemov umetne inteligence je sprožil različne odzive. Eden od njih je, da takšni sistemi odločanja podedujejo in zanje veljajo merila, norme in standardi, ki veljajo za človeške odločevalce, med katerimi je tudi transparentnost. Zato bi morali te sisteme bodisi narediti bolj transparentne bodisi prepovedati njihovo uporabo. Na nasprotni strani pa se pojavlja odgovor v obliki argumenta o dvojnih standardih. Osrednja trditev je, da gre pri pozivih k celovitejši transparentnosti sistemov odločanja na podlagi umetne inteligence za dvojna merila, saj tudi človeška presoja ni transparentna. V prispevku je predlagan odgovor na slednji argument.

Ključne besede: preglednost, umetna inteligenca, človeška presoja, dvojni standard, racionalizem, intuicionizem.

Martin Justin

ALI JE RAZLOŽLJIVOST NUJNA ZA ZANESLJIVOST?

Povzetek: Nekateri sistemi umetne inteligence so tako netransparentni, da niti njihovi oblikovalci ne razumejo natančno, kako delujejo. To se izkaže za težavo pri uporabi teh sistemov v dejanskih procesih odločanja, saj se zdi, da krši naše etične in institucionalne norme, kot sta transparentnost in prevzemanje odgovornost, poleg tega pa zmanjšuje našo zmožnost zaupanja v te procese. V literaturi je pogosto predlagano, da lahko ta problem rešimo tako, da razložimo sisteme umetne inteligence. V tem eseju nasprotujem tej zamisli. Čeprav je transparentnost res potrebna za zanesljivost, argument, da potrebujemo razložljivo UI, ne razlikuje med transparentnimi sistemi UI in transparentnimi procesi odločanja, ki temeljijo na rezultatih teh sistemov. Teza tega prispevka je, da prvo ni potrebno za drugo.

Ključne besede: umetna inteligenca, zaupanje, razložljivost, etika umetne inteligence, razložljiva umetna inteligenca, zanesljivost.

Jonas Miklavčič

TRANSPARENTNOST KOT NAČELO IN ZAHTEVA

Povzetek: Prispevek raziskuje večplastni pojem transparentnosti na področju umetne inteligence (UI) in poudarja njegovo ključno vlogo kot etično načelo in regulativno zahtevo v različnih dokumentih nevladnih organizacij in vladnih organov. Z analizo etičnih smernic in pravnih okvirov, vključno s tistimi, ki sta jih predlagala UNESCO in OECD, ter akta Evropske komisije o umetni inteligenci (2021) razmejujemo nedoslednosti in dvoumnosti v zvezi s pojmom „transparentnost“. S primerjalno analizo trdimo, da zahteva po transparentnosti pogosto zajema raznoliko paleto pojavov, kot so razložljivost, razkritje in interpretabilnost, kar vodi v pomanjkanje jasnosti pri njenem praktičnem izvajanju. Predlagamo bolj niansirano terminološko razlikovanje, da bi bolje izrazili posebne vidike transparentnosti, ki so potrebni za etično in varno uporabo UI. Besedilo ponuja prispevek k trenutni razpravi o etiki UI s pozivom k jasnejši in bolj diferencirani terminologiji za obravnavo zapletenih zahtev po transparentnosti, s čimer bi lahko povečali učinkovitost etičnih smernic in rekulacije UI v Evropski uniji.

Ključne besede: umetna inteligenca, preglednost, etične smernice, pravna ureditev, etika.

Saša Horvat

KAJ IMAJO SKUPNEGA AI, MEDICINA IN ŠAH? PROBLEMI V PROCESU ODLOČANJA

Povzetek: Širša uporaba rešitev umetne inteligence (UI) v medicini prinaša številne izzive, eden od njih je razvoj in uporaba orodij umetne inteligence v procesu kliničnega odločanja (*sistemi za podporo kliničnemu odločanju*), razvitih z metodami strojnega učenja. V tem prispevku obravnavamo okoliščine procesa odločanja v klinični praksi in trenutno stanje na področju razvoja rešitev UI. Da bi dobili širšo sliko odnosa med človeško in umetno inteligenco pri odločanju, kritično analiziramo spoznanja s področja šaha, kjer se ta odnos razvija že sedem desetletij. Na koncu izpeljemo določene sklepe glede odnosa med zdravniki in umetno inteligenco v procesu odločanja v klinični praksi.

Ključne besede: umetna inteligenca, medicina, odločanje, CDSS, DDSS, šah.

Stjepan Štivić

DIGITALNI DVOJČEK IN ETIČNE DILEME

Povzetek: Digitalni dvojček je del vseprisotnega procesa digitalizacije in virtualizacije stvari, procesov, situacij in živih bitij. Njegov cilj je izdelava zelo realističnih modelov resničnih sistemov. Tehnologija digitalnega dvojčka se od digitalnih kopij ali animiranih modelov razlikuje po interakciji z resničnostjo in s fizičnimi dvojniki. Pojem digitalnega dvojčka je bil razvit v industriji, pozneje pa se je razširil na druga področja, kot sta ekonomija in zdravstvo. V zdravstvu je zamisel o digitalnem dvojčku sprožila nove izzive glede interakcije med osebo in njeno digitalno simulacijo. Poleg pozitivnih plati ta tehnologija sproža vprašanje, kakšne posledice ima lahko razvoj takšne upodobitve osebe. Namen tega prispevka je predstaviti tehnologijo digitalnega dvojčka in etične dileme, ki jih odpira v zdravstvu.

Ključne besede: digitalni dvojček, zdravstvo, etične dileme, izzivi, posledice.

ABSTRACTS

Bojan Žalec

THE (IM)PROBABILITY OF HUMANLIKE ARTIFICIAL INTELLIGENCE

Abstract: The author deals with the question of the probability of the emergence of humanlike artificial intelligence and emergence of super-intelligence. He argues that no emergence is probable. The probability of the first is rejected, following Erik J. Larson's argumentation based on the essential characteristics of human intelligence. These include generality, intuition, common sense, and abduction. It is noted that no-one has a scientific conception of how such artificial intelligence could be created or how it could develop. The probability of the onset of superintelligence is rejected based on arguments provided by François Chollet, emphasizing the non-generality, situationality and contextuality, and externalism of intelligence. The author demonstrates the importance of correct understanding, concept and definition of intelligence. Misunderstanding can have negative effects far beyond the academic realm, including unjustified financial gains and damage in terms of organizing scientific work (big science or hive science), and fostering conditions for creativity.
Keywords: probability of human-like artificial intelligence, probability of super-intelligence, generality, intuition, abduction, common sense, particularity and externalism of intelligence.

Octavian-Mihai Machidon

WE SHAPE AI, AND AI SHAPES US: PHILOSOPHICAL AND THEOLOGICAL CONSIDERATIONS ON AI'S ALGORITHMIC DETERMINISM

Abstract: Artificial Intelligence (AI) is increasingly becoming a ubiquitous and autonomous force transforming our society and how people interact with the world around them and each other. Its staggering development and widespread use raise ethical debates and concerns over AI's broad social impacts, especially given the impressive social transformation potential that AI has already shown. This work discusses AI's algorithmic determinism as a force for social transformation concerning the thinking of two 20th-century philosophers, both prominent scholars in the field of media ecology: Jacques Ellul and Marshall McLuhan. AI's transformative potential on both social and individual levels is analysed to determine to what extent AI can be viewed as a self-augmenting, deterministic force engineering the world on its terms (as implied by Ellul) and if we, as individuals, are exposed to having our cognitive and intellectual faculties altered by AI (following McLuhan's theory of technology as extensions of man). Finally, this work turns to Christian anthropology in search of a liberating perspective on man's relationship with technology in general, and AI in particular, that enables individual responsibility and empowers human control while minimising artificial intelligence's deterministic potential.
Keywords: artificial intelligence; algorithmic determinism; social transformation; Christian anthropology; patristics.

Vojko Strahovnik

TRANSPARENCY OF AI SYSTEMS AND HUMAN JUDGEMENT: RESPONDING TO THE DOUBLE-STANDARD ARGUMENT

Abstract: The transparency problem for artificial intelligence (AI) and machine learning (ML) concerns the contention that we supposedly have no usable insight into some AI algorithms' decision-making or recommendation-producing process. This is an in-principle problem leading up to humans not being able to track, understand, or audit the decisions made by such systems. This aspect of AI systems has elicited various responses. One of them is that such decision-making systems inherit and are subject to criteria, norms, and standards that apply to human decision-makers, transparency being one of them. This is why we should either make these systems more transparent or prohibit their use. On the opposite side, there is a common response in the form of the double standard argument. The central claim is that a double standard is involved in the calls for more comprehensive transparency of AI-based decision-making systems since human judgement also lacks transparency. This paper proposes a response to the latter argument.

Keywords: transparency, artificial intelligence, human judgement, double standard, rationalism, intuitionism.

Martin Justin

IS EXPLAINABILITY NECESSARY FOR TRUSTWORTHINESS?

Abstract: Some AI systems are so opaque that even their designers do not understand exactly how they work. This proves a problem for using these systems in real-life decision-making processes since it seems to violate our ethical and institutional norms like transparency and accountability, in addition to diminishing our ability to trust these processes. It has been suggested in literature that we can solve this problem by explaining the AI systems involved. In this essay, I argue against this idea. While transparency is indeed necessary for trustworthiness, the argument that we need explainable AI fails to make a distinction between transparent AI systems and transparent decision-making processes that rely on the outputs of these systems. The thesis of this essay is that the first is not necessary for the latter.

Keywords: artificial intelligence, trust, explainability, AI ethics, explainable AI, trustworthiness.

Jonas Miklavčič

TRANSPARENCY AS A PRINCIPLE AND A REQUIREMENT

Abstract: This paper explores the multifaceted concept of transparency within the domain of artificial intelligence (AI), emphasising its critical role as both an ethical principle and a regulatory requirement across various documents by NGOs, governmental bodies, and international organisations. By analysing ethical guidelines and legal frameworks, including those proposed by UNESCO and the OECD, and the European Commission's AI Act (2021), we delineate the inconsistencies and ambiguities surrounding the term 'transparency'. Through a comparative analysis, we argue that the transparency requirement often encompasses a diverse array of phenomena such as explainability, disclosure, and interpretability, leading to a lack of clarity in its practical enforcement. We propose a more nuanced terminological distinction to articulate better the specific aspects of transparency required for the ethical and safe deployment of AI technologies. This paper contributes to the ongoing debate on AI ethics by calling for clearer, more differentiated terminology to address the complex transparency requirements, thereby enhancing the efficacy of ethical guidelines and legal regulation in AI governance.

Keywords: artificial intelligence, transparency, ethical guidelines, legal regulation, ethics.

Saša Horvat

WHAT DO AI, MEDICINE AND CHESS HAVE IN COMMON? ISSUES IN DECISION-MAKING PROCESS

Abstract: The wider application of AI solutions in medicine brings with it a number of challenges, one of which is the development and application of AI tools to support the clinical decision-making process (*Clinical Decision Support Systems*) developed with the machine learning methods. In this article, we consider the circumstances of the decision-making process in clinical practice, and the current situation regarding the development of AI solutions. In order to get a broader picture of the relationship between human and artificial intelligence in decision-making, we critically analyse insights from the field of chess, where this relationship has been developing for seven decades. Ultimately, we draw certain conclusions regarding the relationship between physicians and AI in the decision-making process in clinical practice.

Keywords: artificial intelligence, medicine, decision making, CDSS, DDSS, chess.

Stjepan Štivić

DIGITAL TWIN AND ETHICAL DILEMMAS

Abstract: Digital twin is part of the ubiquitous process of digitisation and virtualisation of things, processes, situations and living beings. It aims to produce highly realistic models of real systems. Digital twin technology differs from digital copies or animated models in interaction with reality and with their physical counterparts. The concept of a digital twin has been developed in industry and later it was extended to other areas such as economics and healthcare. In healthcare the idea of a digital twin raised up new challenges regarding interaction between a person and his digital simulation. In addition to its positive sides, this technology raises the question, of what consequences the development of such a representation of a person may have. The aim of this presentation is to present the digital twin technology and the ethical dilemmas that it opens up in healthcare.

Keywords: digital twin, healthcare, ethical dilemmas, challenges, consequences.

AUTHORS (AVTORJI)

Assoc. Prof. **Saša Horvat**
University of Rijeka, Faculty of Medicine
sasa.horvat@medri.uniri.hr

Martin Justin
University of Maribor, Faculty of Arts
martin.justin1@um.si

Assist. Prof. **Octavian-Mihai Machidon**
University of Ljubljana, Faculty of Computer and Information Science
octavian.machidon@fri.uni-lj.si

Jonas Miklavčič, PhD
University of Ljubljana, Faculty of Theology
jonas.miklavcic@teof.uni-lj.si

Professor **Vojko Strahovnik**
University of Ljubljana, Faculty of Arts and Faculty of Theology
vojko.strahovnik@ff.uni-lj.si

Stjepan Štivić, PhD
University of Ljubljana, Faculty of Theology
stjepan.stivic@teof.uni-lj.si

Professor **Bojan Žalec**
University of Ljubljana, Faculty of Theology
bojan.zalec@teof.uni-lj.si

REVIEWS (RECENZIJI)

Review of the scientific monograph Vojko Strahovnik and Jonas Miklavčič (eds.), *Beyond Algorithms: Disentangling the Philosophical and Ethical Complexities of AI and Its Implementation*

The scientific monography, a collection of scientific papers, entitled Beyond Algorithms: Disentangling the Philosophical and Ethical Complexities of AI and Its Implementation, edited by Vojko Strahovnik and Jonas Miklavčič, with seven scientific papers, is an original scientific work that analyses and evaluates the issue of artificial intelligence from various perspectives, mainly from an anthropological and ethical point of view. The originality of the scientific collection of papers lies in the fact that it deals with the issue of AI in an interdisciplinary way, from the perspective of philosophy, theology, computer science, art, and medicine. This makes the work very holistic and, at the same time, applied. The work is structured in a very clear and transparent way, divided into three thematic sections: Perspectives on Artificial Intelligence, thee Transparency Problem in Artificial Intelligence, and Artificial Intelligence and Healthcare.

Bojan Žalec in his paper The (Im)Possibility of Humanlike Artificial Intelligence deals with the question of the probability of the emergence of human-like artificial intelligence and the emergence of super-intelligence. The author demonstrates the importance of correct understanding, concept, and definition of intelligence. Misunderstanding can have negative effects far beyond the academic realm, including unjustified financial gains and damage in terms of organising scientific work (big science or hive science), and fostering conditions for creativity. Octavian-Mihai Machidon in his article We Shape AI, and AI Shapes Us - Philosophical and Theological Considerations on AI's Algorithmic Determinism emphasise how Artificial Intelligence (AI) is increasingly becoming a ubiquitous and autonomous force, transforming our society and how people interact with the world around them and each other.

In the article Transparency of AI Systems and Human Judgment: Responding to the Double-Standard Argument, Vojko Strahovnik very originally analyses how the transparency problem for artificial intelligence (AI) and machine learning (ML) concerns the contention that we supposedly have no usable insight into some AI algorithms' decision-making or recommendation-producing processes. The central claim is that a double standard is involved in the calls for more comprehensive transparency of AI-based decision-making systems since human judgment also lacks transparency, emphasises Strahovnik. Martin Justin in his paper Is Explainability Necessary for Trustworthiness? opens the problem of how some

AI systems are so opaque that even their designers do not understand exactly how they work.

Jonas Miklavčič in his article Transparency as a Principle and a Requirement explores the multifaceted concept of transparency. Through a comparative analysis, he argues that the transparency requirement often encompasses a diverse array of phenomena such as explainability, disclosure, and interpretability, leading to a lack of clarity in its practical enforcement. The paper of Jonas Miklavčič is a very original scientific contribution to the ongoing debate on AI ethics by calling for clearer, more differentiated terminology to address the complex transparency requirements, thereby enhancing the efficacy of ethical guidelines and legal regulations.

In his article What do AI, Medicine and Chess Have in Common? Issues in the Decision-Making Process, Saša Horvat emphasises considering the circumstances of the decision-making process in clinical practice, and the current situation regarding the development of AI solutions.

Stjepan Štivić in his paper Digital Twin and Ethical Dilemmas analyses how digital twin technology differs from digital copies or animated models in interaction with reality and with their physical counterparts. The concept of digital twin has been developed in industry and later it was extended to other areas such as economics and healthcare. In healthcare the idea of a digital twin raises new challenges regarding the interaction between a person and his digital simulation.

The scientific papers collected in this monograph are a great contribution to the debate on artificial intelligence, both in Slovenia and internationally. The authors analyse the current topic with clear arguments and from the perspective of different sciences, and the special value of the scientific debates is their applicability. This scientific work will be a valuable handbook for researchers, professors, students, and persons facing the many challenges of artificial intelligence. As Pope Francis underlined in his message for the 57th World Day of Peace in January 2024: "'Intelligent' machines may perform the tasks assigned to them with ever greater efficiency, but the purpose and the meaning of their operations will continue to be determined or enabled by human beings possessed of their own universe of values. There is a risk that the criteria behind certain decisions will become less clear, responsibility for those decisions concealed, and producers enabled to evade their obligation to act for the benefit of the community".

Professor Anton Jamnik
University of Ljubljana
Faculty of Theology

Review of the scientific monograph Vojko Strahovnik and Jonas Miklavčič (eds.), *Beyond Algorithms: Disentangling the Philosophical and Ethical Complexities of AI and Its Implementation*

The scientific monography *Beyond Algorithms: Disentangling the Philosophical and Ethical Complexities of AI and Its Implementation*, edited by Vojko Strahovnik and Jonas Miklavčič is a profound exploration of the ethical, philosophical, and social implications of artificial intelligence. It is a collection of various articles written by authors with different professional profiles, but all are aware of the need for an interdisciplinary approach to find the ethical implications of AI in very different fields, from healthcare to autonomous systems. This thought-provoking anthology is divided into three parts: Perspectives on Artificial Intelligence, the Transparency Problem in Artificial Intelligence, and Artificial Intelligence and Healthcare.

The first part is devoted to fundamental philosophical reflections on the development of artificial intelligence and its implications for humanity. *Bojan Žalec* points to the need to understand the concept of intelligence correctly, as the debate on AI and, consequently, decisions on the direction of development, are often not based on real assumptions. He is convinced that the creation of human-like AI is not possible and provides several convincing arguments to support this view. Žalec highlights the unique characteristics of human cognition –generality, intuition, common sense, and abduction – that current AI lacks. He also raises objections to over-optimistic expectations about so-called super-intelligence. The very first article invites us into a depth of critical thinking that runs like a thread through all the contributions in the volume.

A very interesting in-depth philosophical and theological reflection on the social implications of algorithmic determinism has been written by the computer scientist *Octavian-Mihai Machidon*. Drawing on the reflections of philosophers Ellul and McLuhan, Machidon warns of the dangers that the uncritical use of AI brings with it for the individual and society. The computer scientist is convinced that theology and philosophy have an important role to play in the interdisciplinary debate on the role and limits of AI for today's society. When AI makes decisions for us, there is a danger that our ability to make independent decisions will slowly stagnate. Machidon's final theological reflection, inspired by the Orthodox tradition, is very valuable and original. We must learn how to use technology properly to fulfil our transcendent purpose in life.

The second part of the scientific monograph is devoted to the issue of transparency, which is one of the key topics in the ethical consideration of the use of AI. On the one hand, transparency is demanded, but on the other hand, complete transparency cannot be guaranteed.

Vojko Strahovnik is convinced that the demand for greater transparency of AI in decision-making is a double standard, since even in human decision-making, complete transparency cannot be guaranteed. After presenting the problem of transparency in using AI in a very understandable and systematic way, Strahovnik develops an original model of chromatic transparency that goes beyond classical rationalism and social intuitionism.

Strahovnik's reflections are continued by *Martin Justin* and *Jonas Miklavčič*. *Martin Justin* questions the assumption that explainability is a prerequisite for trust. He proposes that while explainability can enhance trust, it is not the sole factor; other elements, such as the system's reliability and the context in which it operates, are equally important. This nuanced view suggests that the relationship between transparency, explainability, and trust is more complex than commonly assumed. Justin argues that decision-making processes can be considered transparent despite containing opaque AI systems. *Jonas Miklavčič* critically examines the requirement for transparency in several documents published worldwide, highlighting in particular the inconsistencies and ambiguities in the use of the term transparency. His article is an important contribution to the terminological clarification of the transparency requirement, which contributes to the effectiveness of ethical and legal guidance on the use of AI. He advocates avoiding the notion of transparency in the case of ethical requirements and using terms that define the content more precisely.

The third part of the monograph focuses on the use of AI in a specific area of healthcare. *Saša Horvat* focuses on a critical assessment of the use of AI tools for decision-making in clinical practice. The author draws an intriguing parallel between AI's role in medicine and its applications in games like chess. Horvat discusses the potential benefits of AI in enhancing diagnostic accuracy and treatment personalisation while also cautioning against over-reliance on AI systems that may lack the nuanced judgment of physicians. He emphasises the importance of preserving the human element in medical decisions, particularly in cases where ethical considerations and patient values play a crucial role.

In the last paper of the anthology, *Stjepan Štivić* critically explores the concept of digital twins in healthcare – digital simulation of patients used for testing treatments and predicting health outcomes. He is convinced that digital twin technology will have widespread use and benefits in healthcare and medicine, as it will enhance the possibility of personalising healthcare. Despite the positive prospects, Štivić highlights ethical dilemmas, particularly concerning respect for individual autonomy and privacy.

In the face of the extremely rapid technological advances in artificial intelligence, the scientific monograph *Beyond Algorithms* provides a critical and in-depth reflection on how to use these new advances for the benefit of the individual, human society and the natural environment as a whole. The authors promote the many positive effects of the new technologies, but warn in particular of the anthropological and social consequences of their uncritical use, which could lead to the erosion of individual rights and to greater inequality in society. All contributions advocate the human-centred use of AI and the need to follow ethical guidelines in its development. The contributions are characterised by their topicality, familiarity with the latest literature and developments in the field of AI, scientific rigour and a strong sense of responsibility towards the future development of humanity.

Despite the complexity of the subject matter, the contributions are written in accessible language, making them valuable for students, researchers, and policy-makers. In reading the contributions, the reader is encouraged to take a critical stance towards the further development of modern technologies in the context of respect for fundamental ethical values and the dignity of each individual, which will ensure social equality and prosperity for all. Artificial intelligence as a human product can be of great help in this endeavour.

Assoc. Prof. Roman Globokar
University of Ljubljana
Faculty of Theology