UDK 004.8:159.9
Review Article
https://doi.org/10.46630/gpsi.20.2023.09

Exploring The Horizon of Science: A Brief Introduction to AI Ethics[1]*

Miloš Stojadinović**
Department of Psychology, Faculty of Philosophy, University of Niš, Serbia

Abstract

In recent years, a modern field of artificial intelligence (AI) ethics has been emerging. Today, there is virtually no sphere of social functioning untouched by AI in one form or another. Furthermore, the functional autonomy possessed by these intelligent systems is rapidly increasing. All indications suggest that this trend will continue and likely intensify in the near future. In this process, it is natural for important questions to arise that warrant thorough philosophical and psychological analysis. Timely engagement with these issues could prevent potential disagreements and unwanted outcomes. Therefore, the aim of this paper is to provide a brief introduction to the emerging field of AI ethics, highlighting the problems and questions that the contemporary field of AI ethics addresses. It is based on the recognition of the inseparability of any ethical discussion from its psychological antecedents and consequences. The paper will first delve into the definition of artificial intelligence, as well as the definition of AI ethics and its subject of inquiry. It will then explore the most significant questions and issues in AI ethics (in terms of urgency), including autonomous systems, machine bias, the problem of opacity (i.e., the "black box" problem), machine consciousness, technological singularity, and other important topics addressed by AI ethics. Finally, the paper will discuss the researchers and professionals engaged in AI ethics, the interest of psychologists in AI ethics, and their potentially critical role in this emerging scientific field.

Keywords: artificial intelligence (AI) ethics, machine bias, the problem of opacity (the black box problem), machine consciousness, technological singularity

[1] Corresponding author: milos.stojadinovic@filfak.ni.ac.rs
Acknowledgement: This study was supported by the Ministry of Science, Technological Development and Innovations of the Republic of Serbia (Contract No. 451-03-47/2023-01/ 200165).
*Please cite as: Stojadinović, M. (2023). Exploring The Horizon of Science: A Brief Introduction to AI Ethics, *Godišnjak za psihologiju, 20*, 157-175. https://doi.org/ 10.46630/gpsi.20.2023.09.
** https://orcid.org/0000-0001-6535-2945

Exploring The Horizon of Science: A Brief Introduction to AI Ethics

One, a robot may not injure a human being, or, through inaction, allow a human being to come to harm. [...] Two, […] a robot must obey the orders given it by human beings except where such orders would conflict with the First Law. [...] And three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

In 1942, Isaac Asimov introduced these "Three Laws of Robotics" in his renowned short story, "Runaround," which has since become a seminal work in science fiction literature. About 80 years later, at the intersection of scientific advancement and speculative fiction, the field of AI ethics emerged, seeking to extend the discussion from its previous realm of science fiction into academic domains. Asimov's laws have sparked discussions spanning several decades, engaging not only science fiction experts but also professionals in technical sciences and computer technologies. Undeniably, Asimov's laws have played a pivotal role in the evolution of AI ethics, prompting Asimov himself to delve into the ethical implications, both positive and negative, that may arise from these laws in many of his subsequent works.

Over the past few decades, AI has permeated nearly every aspect of human society, influencing everything from everyday consumer choices to medical diagnoses and crucial state and international decision-making processes. The extent to which AI technology is reshaping our social interactions, primarily through the algorithms driving numerous social networks, is widely acknowledged and evident. AI is now a ubiquitous presence in virtually all spheres of social functioning. Tristan Harris, a renowned technology ethicist and former design ethicist at Google, famously stated that "AI already runs the world" (Orlowski, 2020). The level of functional autonomy exhibited by these intelligent systems is rapidly advancing. The undeniable fact is that AI significantly impacts our lives, decisions, and interpersonal relationships, and all signs indicate that this influence will continue and likely escalate in the near future. Modern historian Yuval Noah Harari (2018) places AI among the three paramount challenges that humanity must seriously confront in the coming decades, alongside nuclear war and ecological collapse. This process raises crucial questions that warrant thorough examination from both philosophical and psychological perspectives, as timely engagement with these issues has the potential to preempt numerous conflicts. Consequently, the objective of this paper is to provide a concise introduction to the emerging field of AI ethics, highlighting the problems and inquiries addressed within the contemporary realm of AI ethics (with ethical discussions inherently intertwined with the field of psychological science).

What is artificial intelligence (AI)?

Artificial Intelligence (AI) is, simply put, the programming of machines to perform tasks and processes that typically require human intelligence. Just like human intelligence, there are numerous definitions of AI, each emphasizing different aspects, largely based on the perspective and stance of the definer. One

widely accepted definition of AI in academic circles was proposed by Copeland (2020), describing AI as the ability of a computer or computer-controlled robot to perform tasks characteristic of intelligent beings, such as reasoning, symbolic thinking, generalizing conclusions, or learning from experience. The term "artificial intelligence" was coined in 1955 by a group of researchers who organized a two-month summer school at Dartmouth College (McCarthy et al., 1955). This event is often regarded as a pioneering endeavour in the study of AI.

Today, in the early years of the 21st century, when engineers, IT professionals, social science experts, and AI enthusiasts hear the term "AI," they often envision a robust intelligent system closely resembling human intelligence, albeit with a machine origin. However, for several decades, there has been a debate about the feasibility and achievability of this notion. This debate has been significantly shaped by the emergence of large language models such as GPT-4 (https://openai.com/gpt-4). The discussion was initiated, in part, by the influential philosopher John Searle (1980), who contends that *strong or general AI* is fundamentally unattainable. Present achievements in the field of AI can be categorized as *weak AI* since current intelligent systems do not surpass the ability to solve seemingly complex tasks that can ultimately be decomposed into numerous simpler steps. Searle's central argument posits that, regardless of a machine's complexity and sophistication, it lacks *consciousness and/or a mind*, which he deems essential for genuine understanding – an attribute beyond the mere capability of performing highly intricate computational operations (Searle, 1980).

However, Searle (1980) momentarily overlooks the fact that consciousness and the mind remain unresolved inquiries in the realms of psychology, neuroscience, and philosophy. Many psychological and neurological phenomena can be adequately elucidated without invoking the concepts of consciousness or mind. Consequently, we lack a definitive argument against the proposition that human intelligence is not solely a complex system of biological algorithms, implying that it is only a matter of time before we begin to deconstruct human intellect into a greater number of simpler mental processes that can be mathematically represented or encoded as computer instructions (a belief shared by the organizers of the aforementioned Dartmouth College summer school).

Some authors (for example: MacClellan, 2023; Torrance, 2013), taking a broader perspective on the concept of intelligence rather than a *strictly biocentric* one, challenge the idea that intelligence is limited solely to living organisms. They propose that intelligence can also emerge in fully mechanical environments, provided that the system attains a sufficient level of complexity (Chalmers, 1996). For instance, xenobots, which are bioengineered robots created using stem cells from the African frog (Xenopus laevis), have already astounded researchers with their ability to move, self-heal, and even autonomously gather scattered debris (Kriegman et al., 2021). However, when the original synthetic particles were substituted with individual stem cells, these miniature living bots exhibited a remarkable behaviour – they self-assembled, bringing the cells together to construct entirely new xenobots. The very assertion that intelligence must necessarily have organic origins lacks scientific evidence or thorough investigation, thereby lacking a convincing scientific argument for why intelligence as a capability should be restricted solely to beings composed of organic-carbon chemistry. It fails to

account for the potential development of similar abilities in systems based on alternative materials, such as silicon, as exemplified by modern computer chips.

Therefore, it is evident that there is a broad spectrum of interpretations regarding the true nature of AI. The inquiry into what constitutes AI will likely remain unresolved until we gain a more precise understanding of long-standing concepts such as intelligence, consciousness, mind, and similar constructs that have accompanied us throughout millennia – a pursuit in which psychologists can undoubtedly play a pivotal role. Moreover, the progress in AI ethics and the advancements in comprehending intelligent systems offer psychologists a fresh perspective to explore the human mind – by conducting a comparative analysis of the human and artificial minds.

What is AI ethics and what does it deal with?

AI ethics, as a nascent scientific discipline, heavily draws on the slightly older field of machine ethics as its foundational framework. Anderson and Anderson (2011), distinguished pioneers in machine ethics, define its objective as the endeavour to create machines that adhere to ideal ethical principles or a set thereof during decision-making processes. In essence, machine ethics, as the name implies, aims to infuse an ethical dimension into the realm of machines. Moreover, machine ethics addresses concerns surrounding the moral status of intelligent machines, contemplating whether these machines should be attributed moral and legal rights. Machine ethics resides within the interdisciplinary and multidisciplinary domain of technology ethics, which is itself a subdivision of applied ethics. Presently, distinguishing AI ethics from machine ethics with precision remains challenging, and there seems to be no immediate necessity for such demarcation.

In the early 21st century, multiple approaches have emerged for integrating ethics into the realm of intelligent machines and systems. Within this context, we can identify at least three distinct types of approaches (Gordon & Nyholm, n.d.): (1) the bottom-up approach, (2) the top-down approach, and (3) the mixed or hybrid approach.

A *bottom-up approach* is exemplified by the systems discussed by Guarini (2006), which are rooted in casuistry. *Casuistry* is a process of reasoning aimed at resolving moral dilemmas by applying theoretical rules that were previously used to address other, often similar, moral dilemmas ("Casuistry", 2021; Schmidt, 2014). These systems employ artificial neural networks to learn how to navigate specific ethical dilemmas that already have predetermined answers. Following a designated learning period (referred to as the "training phase" in AI systems), the system should possess the capability to autonomously address new ethical dilemmas. However, AI systems based on casuistry face challenges of reclassification and reflection (essentially, reconsidering a made decision). Guarini himself acknowledges that casuistry alone is an inadequate principle if the goal is to construct a comprehensive ethical AI system (which aligns with the aspirations of AI ethics).

The *top-down approach* combines two fundamental ethical theories, utilitarianism and deontology[2], with analogical reasoning (Dehghani et al., 2011).

[2] For introductory information on utilitarianism, please refer to Duignan & West (2021); for a basic

In these systems, the utilitarian mode prevails until "sacred values" are challenged, prompting a switch to a deontological mode that places less emphasis on utility and the consequences of actions (which are central to utilitarian ethics). To align such a system with the process of moral decision-making in humans, researchers rely on empirical findings from psychological studies on human decision-making in diverse ethical situations. One significant advantage of the top-down approach is its integration of the two prominent ethical theories, deontology and utilitarianism, in a coherent manner. However, the use of empirical psychological studies on human moral decision-making can pose challenges as it implies that the majority's decision-making becomes the yardstick for ethical correctness in decision-making.

The *hybrid/mixed approach*, as its name implies, seeks to integrate the characteristics of the two aforementioned approaches. These systems are highly complex and are still in the early stages of development.

AI ethics is a burgeoning scientific discipline, which poses challenges in providing a concrete and precise definition of its scope. Furthermore, no matter how we attempt to delineate the interests of AI ethics, any enumeration will inevitably fall short as the field continually uncovers new topics to explore – an inherent characteristic of a growing scientific discipline. Gordon and Nyholm (n.d.) have presented a comprehensive overview of the primary areas that currently command the attention of AI ethics. They have also provisionally categorized the questions and concerns within AI ethics based on their urgency and their present relevance to individuals and societies. Table 1 presents their categorization. (AI ethics literature often carries a tone of urgency and a "call to action," with authors underscoring the imperative of promptly addressing AI ethics issues.)

Table 1

The most important questions and issues in AI ethics by urgency

Short-term questions (early 21st century)	Mid-term questions (from the 2040s to the end of the 21st century)	Long-term questions (end of the 21st century and beginning of the 22nd century)
• autonomous AI systems (in transportation and weaponry) • machine bias in legislation • privacy and surveillance • black box problem • decision-making in intelligent systems	• AI in government administrations • moral and legal status of intelligent machines • human-machine interaction • mass automation	• technological singularity • mass unemployment • space colonization

understanding of deontology, consult Encyclopaedia Britannica (2021).

Müller (2020) has also presented a similar categorization of the key questions and challenges in the field of AI ethics. He identifies the following critical debates: privacy and surveillance, behaviour manipulation, opacity and lack of transparency in AI systems, bias in intelligent decision-making systems, human-machine interaction, automation and its impact on the job market, autonomous systems, machine ethics, moral status of intelligent machines, and technological singularity. While this list is not exhaustive, it provides a comprehensive overview of the topics that are likely to gain significant importance in the years and decades ahead. The field of AI ethics is highly dynamic, and new questions and issues continue to emerge. As Gordon and Nyholm (n.d.) point out, this is arguably the fastest-growing area within the realm of ethics and moral philosophy, and undoubtedly a field in which psychologists can contribute significantly. Subsequent sections will delve into a more detailed examination of some of these themes.

Autonomous systems

When it comes to AI ethics, discussions about autonomous systems often focus on two phenomena: (a) autonomous weapons systems and (b) autonomous vehicles. This is understandable given the significant potential consequences that could arise from the direct interaction between these technologies and individuals.

Opinions regarding autonomous weapons systems are, expectedly, highly divided. Some authors argue that such systems could serve as a beneficial substitute for human armies (Müller & Simpson, 2014). For instance, if warfare were delegated to machines, it might potentially reduce the occurrence of war crimes, provided that the machines are equipped with ethical algorithms ensuring consistent adherence to internationally prescribed rules of warfare (Arkin, 2010). Conversely, there are those who believe that the widespread use of autonomous weapons systems could lead to a more casual engagement in warfare, raising numerous concerns about the practicality of developing algorithms capable of accommodating all rules of combat (Gordon & Nyholm, n.d.).

Autonomous vehicles are another important aspect to consider. With the growing popularity of the electric vehicle industry, which often incorporates autonomous driving capabilities (a trend spearheaded by Tesla, Inc. and its founder Elon Musk), it is only natural that autonomous vehicles are a focal point in AI ethics. The key concept revolves around the necessity of equipping autonomous vehicles with ethical algorithms that dictate their response in situations where human safety is jeopardized (Gordon & Nyholm, n.d.). This is crucial to accelerate the realization of the numerous benefits autonomous vehicles bring, including enhanced traffic safety, more efficient fuel utilization, and improved traffic management (Harari, 2018). Undoubtedly, this is one of the most pressing topics in AI ethics, given the gradual global deployment of autonomous transportation systems (TEDx Talks, 2018).

The urgency of this issue has already been recognized, as several individuals have lost their lives in accidents involving autonomous vehicles (National Highway Traffic

Safety Administration, 2022; Salter, 2023). One of the earliest incidents occurred in March 2018 when an experimental vehicle operated by Uber struck a pedestrian who was crossing the road, resulting in a fatal outcome (Wakabayashi, 2018). It was discovered that the vehicle encountered difficulty in classifying the object (in this case, the pedestrian) that suddenly appeared in its path. It initially identified the object as unknown, then as a vehicle, and finally as a bicycle. The vehicle began braking a few moments before the collision, but it was already too late to avoid the impact. Hence, the accuracy of AI systems within autonomous vehicles can be a matter of life or death. However, one may question why a pedestrian was on the road without any regulations and how, from an ethical perspective, an accident of this nature involving a vehicle with a human driver would be handled. As we will see, such double standards are a common occurrence in AI ethics, where people impose demands on machines that even the most morally upright individuals of the human species would struggle to meet.

Whether it is autonomous weapons systems or autonomous vehicles, the central question in AI ethics within these fields remains: What ethical principles should govern the decision-making processes in these systems when there is a possibility of endangering human life? While fatal outcomes related to autonomous vehicles are mostly seen as unfortunate and hopefully infrequent side effects of their use, deaths caused by autonomous weapons are generally regarded as something that can and should be avoided (there is even an active campaign on this topic called "Stop Killer Robots"; https://www.stopkillerrobots.org/). In both cases, it is widely believed that some level of control should remain in human hands (Santoni de Sio & van den Hoven, 2018). However, as some authors argue (for example: Königs, 2022; Santoni de Sio & Mecacci, 2021), this can lead to *responsibility gaps*, where it becomes unclear whether the autonomous machine[3] or the human retaining some level of control over it is accountable for a particular outcome. By nature, Homo sapiens tends to attribute credit (i.e., responsibility) to themselves in cases of positive outcomes resulting from machine decisions, while assigning and shifting responsibility to the machine in cases of negative outcomes caused by its decisions. This phenomenon, known as *attribution bias*, has been a recognized concept in psychology since the mid-twentieth century (Heider, 1958).

Machine bias

There is a widespread belief that the utilization of smart technologies will eradicate human bias, thanks to the (wrongly) presumed "ethical neutrality" of machines. However, empirical evidence from the past few decades has demonstrated that machines can perpetuate and even amplify human bias across various groups (Kraemer et al., 2011). Consequently, one of the paramount inquiries in AI ethics is how to mitigate machine bias.

[3] While there is room for debate regarding whether a machine of this nature is genuinely autonomous in the strictest sense, it is worth considering examples such as autopilot systems that have been utilized in aviation for several decades. These systems undeniably incorporate AI but are not commonly referred to as autonomous vehicles.

In principle, there is nothing inherently wrong with the concept of developing AI systems that support and potentially improve human decision-making. Such systems have the potential to enhance efficiency, accuracy, speed, and the overall scope of decision-making processes. We already witness the role of AI systems in various decision-making contexts, such as online shopping platforms, personalized music recommendations on YouTube, and numerous other situations. However, just like human decision-making, bias can influence any decision-making process. Gordon and Nyholm (n.d.) highlight several notable examples of machine bias and bias in AI systems, providing additional sources of information on these phenomena: gender bias in employment; racial bias in employment and general contexts; racial bias in creditworthiness assessments by banks; racial bias in decisions regarding the allocation of conditional sentences; racial bias in predicting criminal activity within urban areas; bias in determining a person's sexual orientation; racial bias in facial recognition systems (which exhibit a preference for lighter skin tones); and racial and social bias in inferring a person's ethnic background or socioeconomic status based on geolocation data.

How did we, all of a sudden, come to the idea that machines can be biased? We can identify at least three reasons for machine bias: (a) data bias, (b) algorithm bias, and (c) outcome bias (Gordon & Nyholm, n.d.). Let us delve deeper into the first two reasons, noting that outcome bias can be considered a subcategory of data bias (for a more detailed description of outcome bias, please refer to Gordon & Nyholm, n.d.). *Data bias* occurs when an AI system is "trained" using unbalanced data, intentionally or unintentionally, in terms of certain attributes or categories. Over time, these differences are gradually amplified, exerting an increasing influence on the system's decision-making process. Thus, bias becomes perpetuated through a "vicious circle" of decision-making. *Algorithm bias* arises when the creator of an algorithm consciously or unconsciously incorporates a certain (personal?) bias into the algorithm or set of algorithms within a system. Consequently, while data bias originates from the data itself and its distribution, algorithm bias stems from how the algorithm utilizes that data. The design of a program is undeniably shaped by the programmer's understanding of normative and non-normative ethical values of others.

Many AI researchers, IT professionals, and technology academics acknowledge that creating an AI system completely free from bias may be an unattainable goal (Gordon & Nyholm, n.d.). Therefore, Gordon and Nyholm suggest focusing on minimizing machine bias to the greatest extent possible. It is crucial to recognize that *machine bias is fundamentally rooted in human bias*. Treating machine bias as a distinct entity only initiates a semantic game that generates confusion and can have significant and far-reaching consequences. As humans, we are the originators of biases, permitting them to influence us based on largely unfounded categorizations in reality. Consequently, we are also responsible for the existence of machine bias. Therefore, the most effective approach may involve continuous efforts to mitigate human bias (which is an ongoing process with gradual changes) or alternatively, developing a system with more robust ethical principles than our current ones.

The Black Box – the problem of opacity

As previously mentioned, AI systems are increasingly being utilized globally to make crucial decisions that have significant impacts on individuals' lives, such as loan approvals, university admissions, job placements, recidivism predictions, and more. Due to the potential consequences of these decisions, it is vital that we can comprehend the rationale behind the system's choices. In other words, *explainability/ explicability* (Robbins, 2019) of the decision-making process in AI systems is imperative. Numerous authors in the field of AI ethics assert that explainability is a fundamental ethical requirement for an AI system to be considered acceptable (Floridi et al., 2018). However, the majority of decisions made by autonomous AI systems remain incomprehensible to the average person, even if they possess a moderate level of technological literacy. This issue, commonly referred to as *the problem of opacity* within expert circles, raises concerns about the transparency of the algorithms that underpin specific intelligent systems.

The opacity of AI systems can manifest in various ways. Sometimes, even though the algorithms underlying the decision-making process are relatively straightforward to understand, these algorithms are considered trade secrets by companies, who choose not to share them with anyone outside the company (which is their legitimate and legal decision). Another reason for this opacity is that the majority of people lack sufficient technical knowledge to comprehend how an AI system functions, even though there is nothing inherently non-transparent about those specific algorithms. However, there are AI systems where even experts struggle to fully understand the decision-making algorithm. This is known as *the black box problem* (Wachter et al., 2018).

What are the potential psychosocial implications of the aforementioned? From an individual's standpoint, it can be perceived as a violation of autonomy and personal dignity, creating a sense of frustration when it is challenging or impossible to explain the reasoning behind a machine's decision that significantly impacts one's life. At the societal level, the increasing prevalence of algorithmic decision-making has the potential to fundamentally reshape established social processes, with some even perceiving it as a "threat" (Gordon & Nyholm, n.d.). Conversely, Danaher (2016) raises an intriguing concern – the possibility that people, unable to comprehend the decisions made by hypercomplex AI systems, may resort to superstitious and irrational behaviour, leading to the emergence of contemporary rituals reminiscent of ancient practices, such as dancing to summon rain. Danaher terms this *the threat of algocracy*, wherein we must conform to the rule of algorithms we do not fully understand. Historian Yuval Noah Harari (2015) also discusses the potential rise of new *techno-religions*, notably in his work "Homo Deus: A Brief History of Tomorrow". Given these circumstances, it is not surprising to envision the development of anthropomorphic ideas concerning intelligent systems (Gordon & Nyholm, n.d.), wherein human qualities and characteristics are attributed to machines. This trajectory is not unexpected, as humans historically ascribed human abilities to animals before accumulating the extensive knowledge we possess today about the animal kingdom and the natural world.

Machine consciousness

Since the early days of modern machine engineering, the notion has emerged that as machines become increasingly complex, they may eventually develop what we address as "consciousness" (given our current lack of a precise definition of consciousness, with limited prospects of obtaining one in the near future). This, undoubtedly, could give rise to numerous ethical and psychosocial implications. In the 21st century, researchers worldwide are actively engaged in projects aimed at creating machines endowed with artificial consciousness. Engineer Kunihiro Asada has dedicated his career to developing a robot capable of experiencing pleasure and pain, drawing inspiration from the principles of prelinguistic learning observed in infants before they acquire language (Marchese, 2020). Another noteworthy example is Sophia, a robot created by Hanson Robotics, which became the first robot to be granted citizenship by a country (ABC News, 2021).

However, as highlighted by Joanna Bryson (2012), the presence of consciousness in machines can be argued depending on how we define it (and given the lack of a precise definition, there remains flexibility defining it). She proposes that if consciousness is defined as the existence of internal states and the capacity to report on those states, then it could be said that some machines already fulfill these criteria.

The classic *Turing test*, also known as *the imitation game*[4], has long been utilized to evaluate whether a machine possesses consciousness, as named by its creator Alan Turing (Oppy & Dowe, 2021). This test involves three participants: an interrogator (a human being), a real person, and a machine, all physically separated. The interrogator's role is to pose questions (usually in text form, due to limited text-to-speech technology) to both the real person and the machine, aiming to determine which channel of communication corresponds to the real person and which to the machine. If the interrogator cannot reliably differentiate the machine from the real person based on the communication channels, the machine is considered to have passed the Turing test. Turing predicted that by the year 2000, machines would advance to a level where humans would have no more than a 70% chance of correctly discerning whether they were interacting with a machine or a real person[5]. While the Turing test has long been the benchmark for evaluating machine consciousness, Aïda Raoult and Roman Yampolskiy (2018) identified 21 contemporary tests of machine consciousness in their research. The conclusion is clear: progress is undeniably evident.

However, even if we were certain that machines possess consciousness, it would most likely differ significantly from human consciousness. Ultimately, the

[4] Therefore, the title of the widely acclaimed 2014 film about Alan Turing and his team's endeavours during World War II is "The Imitation Game."

[5] In 2018, we witnessed Google's latest AI system, Google Duplex, successfully passing the Turing test by autonomously making phone calls to schedule appointments at a hair salon using voice communication (DW Documentary, 2019).

consciousness and subjectivity of any entity depend on the "hardware" it possesses – such as the brain, sensory organs, and nervous system for humans, or processors, sensors, and conductors for machines (Nagel, 1974).

Before we proceed with further discussion, it is crucial to clarify the concept of moral status. *Moral status* is a fundamental concept in ethics and moral philosophy. It refers to the consideration an entity deserves in various decisions involving moral issues (Madsen, 2015). The 20th and 21st centuries have witnessed a significant expansion in the range of entities attributed with moral status, evident in the increasing importance and number of rights granted to ethnic minorities, women, the LGBTQ+ population, animals, and the environment. Today, these categories and others possess certain rights that were once only imagined for their members. Consequently, it is not challenging to envision the future expansion of this circle of assigning moral status to include intelligent systems.

Considering the points mentioned above, whether we agree or disagree with the claim that some AI systems already possess consciousness, and regardless of whether it occurred by mere chance or the intention of their creators, it is undeniable that ethical implications arise. For example, Thomas Metzinger (2013) advocates for the principle of prohibiting the creation of machines capable of experiencing suffering as the foundation of AI ethics. His utilitarian argument is straightforward – suffering is negative, causing suffering is morally wrong, therefore creating a machine that can suffer would be immoral. Bryson (2010; in her article creatively titled "Robots Should Be Slaves") agrees with this line of reasoning, asserting that if there is a possibility of creating machines with moral status and human-like capabilities, it is best to avoid such an endeavour from the outset. Once again, this brings us back to the starting point, as it all hinges on how we initially define consciousness. In line with this, some scientists argue that the development of a robust theory of consciousness is the first and most crucial task if we aim to continue advancing towards the creation of increasingly sophisticated and complex AI machines and systems (Schwitzgebel & Garza, 2015).

Another intriguing perspective is put forth by Nicholas Agar (2020), an ethics professor at Victoria University of Wellington in New Zealand. Author suggests that when arguments both for and against the possibility of advanced machines having consciousness exist, it is safer to assume that machines do possess consciousness. Accordingly, Agar contends that we should refrain from any actions that could inflict suffering upon machines. Conversely, Danaher (2020) asserts that we can never be certain if a machine has developed consciousness, but argues that its certainty is irrelevant – if we can deduce genuine awareness based on the machine's behaviour, it is sufficient grounds to regard the machine as a conscious entity. The origins of this *ethical behaviourism* can be traced back to the aforementioned Turing test. Nonetheless, the unresolved question persists, continuing to perplex psychology, neuroscience, and philosophy – how do we ascertain whether another entity, whether a machine, a human being, or something entirely different, possesses consciousness? Perhaps it is the burgeoning field of AI ethics that will prompt us to delve deeper into this age-old problem.

Technological singularity

Among the various questions and problems addressed by AI ethics, technological singularity is a concept that hovers on the boundary between reality and science fiction. The notion of *technological singularity* was first introduced in the 1960s by mathematician Irving John Good (1965), a colleague of Alan Turing, in his work titled "Speculations Concerning the First Ultraintelligent Machine". Good defines an ultraintelligent machine as one that surpasses all intellectual activities performed by any human being. Since machine design is itself an intellectual activity, an ultraintelligent machine could design even superior machines, resulting in an inevitable "intelligence explosion" that far exceeds human intelligence. Consequently, the creation of the first ultraintelligent machine would mark the culmination of human invention.

The concept of an "intelligence explosion," envisioning the emergence of self-replicating (as exemplified by the xenobots mentioned earlier) and superintelligent AI machines, may appear unimaginable to many, leading them to dismiss such claims as mere myths in the realm of AI development. However, influential figures within the field of AI ethics, both in academic and non-academic spheres, take the idea of technological singularity with utmost seriousness, perceiving it as a potential threat leading to the extinction of the human species. These "concerned" individuals include renowned philosophers Nick Bostrom and Toby Ord, esteemed experts in AI ethics, as well as notable figures from various fields, such as entrepreneur Elon Musk and the late physicist Stephen Hawking.

Authors exploring the concept of technological singularity vary in their explanations of its causes. Renowned futurist Ray Kurzweil is a strong advocate of technological singularity, primarily citing Moore's Law as the basis for his claims (Insane Curiosity, 2020). *Moore's Law* states that the computing power of transistors has doubled every two years since the 1970s, and it is reasonable to expect this trend to continue in the future. This suggests that it is only a matter of time before sufficiently advanced hardware is developed, enabling the emergence of technological singularity. Prominent AI researcher Stuart Russell (2019) argues that achieving singularity requires not only technological advancements but also progress in language processing and learning, areas where psychologists can play a significant role. He proposes three principles of AI design, reminiscent of Asimov's laws of robotics: (1) the machine's sole objective is to maximize the realization of human preferences; (2) the machine inherently lacks certainty about these preferences; (3) human behaviour serves as the ultimate source of information about human preferences. The crux of the various authors' perspectives on this topic lies in the quest for *value alignment*, which involves ensuring that the objectives and functioning of AI systems, particularly superintelligent ones, align with human values (Gordon & Nyholm, n.d.). The pivotal question to be posed is: which categories of people's viewpoints should guide the alignment of values in intelligent systems?

Other significant topics addressed by AI ethics

AI and the future of work. There is extensive discussion about the role of AI in the future labour market and the potential for *technological unemployment* due to widespread automation (Gordon & Nyholm, n.d.). This is often portrayed negatively, assuming that work is meant to provide individuals with meaningful engagement (Gheaus & Herzog, 2016). However, the reality is that many existing jobs in contemporary societies expose people to risks, making it more appropriate to employ machines for such tasks. Conversely, numerous modern jobs, despite their significant presence in the labour market, fail to bring meaning to the lives of those involved (Bregman, 2016). Some argue that the increased reliance on machines for a large number of jobs would lead to *existential boredom* and a loss of purpose for individuals, while others suggest that a world with less work could actually be an improvement (Gordon & Nyholm, n.d.). Taking all of these factors into account, it is crucial to consider how we can imbue increasingly technology-driven jobs with greater significance for humans.

AI and personal relationships. AI technologies are increasingly infiltrating the realm of relationships, including romantic connections and other interpersonal bonds, a trend that is likely to intensify in the future. Online friendships formed through social media platforms are gradually gaining equal importance as "real-life" friendships (Cocking et al., 2012). Critically, there are concerns that dating applications, which heavily rely on AI, perpetuate negative sexual stereotypes and reinforce certain expectations (Frank & Klincewicz, 2018). At the intersection of science and science fiction, discussions arise about the possibility of humans forming authentic friendships and even romantic relationships with robots and other AI-powered mechanisms. Akihiko Kondo, a Japanese man who legally married a holographic AI system named Hatsune Miku, represents an extreme example of this phenomenon (DW Documentary, 2019).

Dependency on AI systems. If the trend of increasing reliance on various AI systems for everyday decision-making continues, humans could eventually become fully dependent on the decision-making processes carried out by intelligent systems (which, as mentioned earlier, are likely to be incomprehensible to most people; Wachter et al., 2018). This aligns with the concept of technological singularity. Therefore, it is crucial for individuals to continuously enhance and refine their knowledge and skills[6], with a specific emphasis on what is commonly known as "21st-century skills."

[6] Homo sapiens possesses a natural inclination to adapt readily to new advancements and, eventually, start taking them for granted. The ubiquity and convenience of electricity serve as a prime example. Modern humans have grown so accustomed to its availability and usage that they often overlook a time when it was absent. Moreover, the intricate nature of electricity production and distribution systems renders them exceedingly complex, making it arduous for individuals to comprehend them autonomously. A parallel situation is currently unfolding with AI technologies. As they progressively evolve into more sophisticated and advanced forms, we increasingly delegate responsibilities to them, thereby intensifying our reliance on these systems. Simultaneously, this reliance complicates our ability to fully grasp their inner workings.

Ethical guidelines in the field of AI ethics. With the increasing recognition of the importance of AI and the growing interest in ethics related to this technology, influential institutions including governments, the European Union, major companies, and others have established expert teams to develop policies and documents as concrete ethical guidelines in the field of AI. The abundance of such documents makes providing a comprehensive overview challenging. Moreover, the field of AI ethics is witnessing a rise in funding sources and research centres dedicated to this area, offering tremendous potential for young researchers and scientists across various disciplines. Innovations in this domain are generally well-received. Jobin et al. (2019) conducted a comprehensive review of documents that provide specific ethical guidelines for AI. Their work involved a comparative analysis of 84 documents from national and international entities. They identified five key principles common to all these documents: (1) transparency, (2) fairness, (3) non-maleficence, (4) accountability, and (5) privacy. Additionally, these documents frequently emphasize other important principles such as beneficence, freedom, autonomy, trust, sustainability, dignity, and solidarity, among others.

Instead of conclusion: who engages in ai ethics?

This brief introduction to the field of AI ethics is best concluded by acknowledging the wide range of professions that have an interest in this emerging field. AI ethics stands out as one of the most diverse and inclusive scientific domains, encompassing professionals and interests from various disciplines. The interdisciplinary nature of this growing field becomes evident when considering the professions of experts engaged in AI ethics. These professionals include individuals from the fields of information technology, engineering, and mathematics; biochemists, geneticists, and molecular biologists; experts in neuroscience, psychology, philosophy, and other social sciences; specialists in biology, medical and health sciences; physicists and astronomers; practitioners in the emerging field of "decision science"; pharmacologists, toxicologists, and pharmacists; professionals in business, management, and accounting; experts in materials science and geology; immunologists, microbiologists, and veterinarians; professionals in economics and finance, among many others. Interestingly, psychology ranks seventh on the list of scientific disciplines that contribute the most proposals for new tests of machine consciousness (Raoult & Yampolskiy, 2018), highlighting its significance in the field.

AI is gradually becoming an integral part of our daily reality. Consequently, the emergence of the new scientific field of AI ethics is both logical and timely. AI ethics is a youthful discipline brimming with potential for young researchers from diverse backgrounds. The aim of this paper was to elucidate the defining characteristics of AI ethics by addressing its current focal points and exploring the wide range of fields that contribute experts to this domain.

However, while the interdisciplinary, multidisciplinary, and transdisciplinary

nature of this emerging research field offers promise for its timely development, the key question remains whether this development will be swift enough to respond proactively to the potentially problematic areas of artificial intelligence's influence on the contemporary world, humanity, and human ethics. Furthermore, it raises concerns about the specific challenges this relationship is bound to pose in the future.

In light of the rapid pace of technological advancements that underpin the progression of artificial intelligence, such as quantum computing, one may, and with good reason, question whether there will be sufficient time to establish a dedicated scientific discipline and engage in comprehensive discussions about AI's impact on our lives. Alternatively, there may be a need to address problems stemming from the everyday application of AI on a case-by-case basis, dealing with a range of issues across social, psychological, security, economic, media, and informational domains.

The future trajectory of AI ethics is challenging to anticipate, given its propensity for continually unveiling novel questions and themes. As a field in its formative stages, it welcomes contributions from experts spanning various disciplines. While some may perceive engagement with such a subject as premature or futuristic, we echo the words of Lao Tzu: "Do something while it is still nothing".

References

ABC News. (2021, July 9). *Creators of famous Sophia robot reveal AI robotics for children, elderly | Nightline* [Video]. YouTube. https://www.youtube.com/watch?v=JRHdnkUjcZg

Agar, N. (2020). How to Treat Machines That Might Have Minds. *Philosophy & Technology*, *33*(2), 269–82.https://doi.org/10.1007/s13347-019-00357-8

Anderson, M., & Anderson, S. (2011). *Machine Ethics*. Cambridge University Press.

Arkin, R. (2010). The Case for Ethical Autonomy in Unmanned Systems. *Journal of Military Ethics*, *9*(4), 332–41.https://doi.org/10.1080/15027570.2010.536402

Asimov, I. (1942). *Runaround*. Retrieved October 12, 2021 from https://web.williams.edu/Mathematics/sjmiller/public_html/105Sp10/handouts/Runaround.html

Bregman, R. (2016). *Utopia for Realists: The Case for a Universal Basic Income, Open Borders, and a 15-hour Workweek*. De Correspondent.

Bryson, J. (2010). Robots Should Be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions*, 63–74. John Benjamins.

Bryson, J. (2012). A Role for Consciousness in Action Selection. *International Journal of Machine Consciousness*, *4*(2), 471–82.https://doi.org/10.1142/S1793843012400276

Casuistry. (2021, June 22). In *Wikipedia*. https://en.wikipedia.org/wiki/Casuistry

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Cocking, D., Van Den Hoven, J., & Timmermans, J. (2012). Introduction: One Thousand Friends. *Ethics and Information Technology*, *14*, 179–84.https://doi.org/10.1007/s10676-012-9299-5

Copeland, B. (2020, August 11). *Artificial intelligence*. Encyclopaedia Britannica. Retrieved September 6, 2021, from https://www.britannica.com/technology/artificial-intelligence

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, *29*(3), 245–68.https://doi.org/10.1007/s13347-015-0211-1

Dehghani, M., Forbus, K., Tomai, E., & Klenk, M. (2011). An Integrated Reasoning Approach to Moral Decision Making. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics*, 422–441. Cambridge University Press.

Duignan, B., & West H. R. (2021, March 2). *Utilitarianism*. Encyclopaedia Britannica. Retrieved October 13, 2021, from https://www.britannica.com/topic/utilitarianism-philosophy

DW Documentary. (2019, August 14). *Artificial intelligence and its ethics | DW Documentary* [Video]. YouTube. https://www.youtube.com/watch?v=Izd2qOgOGQI

Encyclopaedia Britannica. (2021, March 2). *Deontological ethics*. Retrieved October 13, 2021, from https://www.britannica.com/topic/utilitarianism-philosophy

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Frank, L., & Klincewicz, M. (2018). Swiping Left on the Quantified Relationship: Exploring the Potential Soft Impacts. *American Journal of Bioethics*, *18*(2), 27–28. https://doi.org/10.1080/15265161.2017.1409833

Gheaus, A., & Herzog, L. (2016). Goods of Work (Other than Money!). *Journal of Social Philosophy*, *47*(1), 70–89.https://doi.org/10.1111/josp.12140

Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. In F. Alt & M. Rubinoff (Eds.), *Advances in Computers (vol. 6)*, 31–88. Academic Press. https://doi.org/10.1016/S0065-2458(08)60418-0

Gordon, J., & Nyholm, S. (n.d.). *Ethics of Artificial Intelligence*. Internet Encyclopedia of Philosophy. Retrieved September 6, 2021 from https://iep.utm.edu/ethic-ai/

Guarini, M. (2006). Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems*, *21*(4), 22–28.https://doi.org/10.1109/MIS.2006.76

Harari, J. N. (2018). *Homo deus: kratka istorija sutrašnjice*. Laguna.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. Wiley.

Insane Curiosity. (2020, August 29). *Ray Kurzweil, The Technological Singularity And The Future Of Humanity!* [Video]. YouTube. https://www.youtube.com/watch?v=L8f7nsUdq-k

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.https://doi.org/10.1038/s42256-019-0088-2

Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem?. *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Kraemer, F., van Overveld, K., & Peterson, M. (2011). Is There an Ethics of Algorithms? *Ethics and Information Technology*, *13*, 251–260.https://doi.org/10.1007/s10676-010-9233-7

Kriegman, S., Blackiston, D., Levin, M., & Bongard, J. (2021). Kinematic self-replication in reconfigurable organisms. *Proceedings of the National Academy of Sciences*, *118*(49), e2112672118. https://doi.org/10.1073/pnas.2112672118

MacClellan, J. (2023). Is Biocentrism Dead? Two Live Problems for Life-Centered Ethics. *The Journal of Value Inquiry*, 1–22. https://doi.org/10.1007/s10790-023-09954-5

Madsen, P. (2015, December 17). *Moral standing*. Encyclopaedia Britannica. Retrieved September 30, 2021, from https://www.britannica.com/topic/moral-standing

Marchese, K. (2020, February 24). *Japanese Scientists Develop "Blade Runner" Robot That Can Feel Pain*. Design Boom. https://www.designboom.com/technology/japanese-scientists-develop-hyper-realistic-robot-that-can-feel-pain-02-24-2020/

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955, August 31), A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Retrieved October 12, 2021, from http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Metzinger, T. (2013), Two Principles for Robot Ethics. In E. Hilgendorf & J. P. Günther (Eds.), *Robotik und Gesetzgebung*, 263–302. Nomos.

Müller, V. C. (2020). *Ethics of Artificial Intelligence and Robotics*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/ethics-ai/

Müller, V. C., & Simpson, T. W. (2014). Autonomous Killer Robots Are Probably Good News. *Frontiers in Artificial Intelligence and Applications*, *273*, 297–305.

Nagel, T. (1974). What is it like to be a bat? *Readings in Philosophy of Psychology*, *1*, 159-168.https://doi.org/10.4159/harvard.9780674594623.c15

National Highway Traffic Safety Administration. (2022). *Summary Report: Standing General Order on Crash Reporting for Level 2 Advanced Driver Assistance Systems* (U.S. DoT Publication No. DOT HS 813 325). U.S. Department of Transportation, National Highway Traffic Safety Administration. https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/ADAS-L2-SGO-Report-June-2022.pdf

Oppy, G., & Dowe, D. (2021). *The Turing Test*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/turing-test/

Orlowski, J. (Director). (2020). *The Social Dilemma* [Film]. Exposure Labs; Argent Pictures; The Space Program.

Raoult, A., &Yampolskiy, R. (2018). Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*, *26*(5–6), 35–64.

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*(4), 495–514.https://doi.org/10.1007/s11023-019-09509-3

Russell, S. (2019). *Human Compatible*. Viking Press.

Salter, M. (2023, February 16). *How Many People Have Actually Been Killed By Self-Driving Cars?* SlashGear. https://www.slashgear.com/1202594/how-many-people-have-actually-been-killed-by-self-driving-cars/

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.https://doi.org/10.3389/frobt.2018.00015

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Schmidt, D. P. (2014, September 15). *Casuistry*. Encyclopaedia Britannica. Retrieved September 6, 2021, from https://www.britannica.com/topic/casuistry

Schwitzgebel, E., & Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy*, *39*(1), 98–119.https://doi.org/10.1111/misp.12032

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, *3*(3), 417–57. https://doi.org/10.1017/S0140525X00005756

TEDx Talks. (2018, October 24). *How Will Autonomous Vehicles Transform Our Cities? | Nico Larco | TEDxCollegePark* [Video]. YouTube. https://www.youtube.com/watch?v=tTOFMwKEg7o

Torrance, S. (2013). Artificial agents and the expanding ethical circle. *AI & Society*, *28*, 399–414. https://doi.org/10.1007/s00146-012-0422-2

Wachter, S., Mittelstadt, B., &Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887.

Wakabayashi, D. (2018, March 19). *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam.* The New York Times. https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

ISTRAŽUJUĆI HORIZONT NAUKE: KRATAK UVOD U AI ETIKU

Miloš Stojadinović
Departman za psihologiju, Filozofski fakultet, Univerzitet u Nišu, Srbija

Apstrakt

U poslednjih nekoliko godina u povoju je moderna oblast etike veštačke inteligencije (eng. artificial intelligence – AI). Danas gotovo da nema oblasti društvenog funkcionisanja koja nije pod uticajem AI na neki način. Osim toga, funkcionalna autonomija koju poseduju ovi inteligentni sistemi jako brzo napreduje. Svi pokazatelji ukazuju da će se ovaj trend nastaviti i verovatno intenzivirati u bliskoj budućnosti. U ovom procesu, prirodno je da se jave važna pitanja koja zahtevaju temeljnu filozofsku i psihološku analizu. Pravovremeno bavljenje ovim pitanjima moglo bi sprečiti moguće nesuglasice i nepoželjne ishode. Stoga je cilj ovog rada da pruži kratki uvod u novonastajuću oblast AI etike, ističući probleme i pitanja kojima se savremena oblast AI etike bavi. Rad počiva na neraskidivosti svake etičke rasprave od njenih psiholoških uzroka i posledica. U radu ćemo prvo razmotriti definiciju veštačke inteligencije, kao i definiciju AI etike i predmet njenog istraživanja. Zatim ćemo se osvrnuti na najznačajnija pitanja i probleme u oblasti AI etike (u pogledu urgentnosti) – autonomne sisteme, pristrasnost mašina, problem neprozirnosti (tzv. problem „crne kutije“), svest mašina, tehnološki singularitet i druge važne teme kojima se AI etika

bavi. Na kraju, rad ćemo zaključiti pregledom oblasti iz kojih dolaze istraživači i stručnjaci koji se bave AI etikom, osvrtom na interesovanje psihologa za AI etiku, i njihovu potencijalno ključnu ulogu u ovoj novonastajućoj naučnoj oblasti.

Ključne reči: etika veštačke inteligencije (AI etika), mašinska pristrasnost, problem prozirnosti (problem crne kutije), mašinska svest, tehnološki singularitet

RECEIVED: 13.06.2023.
REVISION RECEIVED: 13.10.2023.
ACCEPTED: 13.10.2023.