Humanology, *Vol. 1*, n. 2, 2025
ISSN 3104-2570 (online)
DOI: 10.46793/HumanologyI-2.071V
Received: 03/01/2025
Accepted: 13/06/2025
Published online: 07/09/2026

Divna Vuksanović[1]
Full Professor
Faculty of Dramatic Arts
University of Arts in Belgrade

ARTIFICIAL INTELLIGENCE AND "BLACK BOXES" OF DIGITAL TOTALITARISM

Abstract: *Current processes of digitization across all areas of life have not been sufficiently examined. In this context, it seems that in the age of late capitalism, which heavily relies on the use of new technologies and artificial intelligence, the manner in which intelligent systems are utilized and who controls them has become highly questionable. Knowledge, in this regard, appears to play a crucial role, concerning both the manipulation of these systems and the delegation of decision-making power to them. However, the use of knowledge as power paves the way towards digital totalitarianism, which will be discussed in this article. Additionally, the text explores the potential humanization of these processes, taking into account the qualities of transparency and human participation in explaining the activities of artificial intelligence. All these processes are viewed from the perspective of the relationship towards so-called other consciousnesses, illustrated through the thought experiment of the "Chinese Room" and the metaphor of "black boxes."*

Keywords: *digital totalitarianism, artificial intelligence, "black boxes," knowledge, capitalism*

[1] vuksanovic.divna@gmail.com

The phrase “digital totalitarianism” emerged in theoretical discussions at the end of the 20th century and the beginning of the 21st century, alongside related concepts that describe constructs such as “digital fascism,” “digital feudalism” and “digital concentration camps.” Attempts at digital interventions in all spheres of social life do not only manifest in the educational system and in the entire social reality, as a question of the necessity for social development, progress, and culture, as is often presented in the media; they also reflect efforts to establish digitization as a totalitarian form of governance that is technically supported by the internet, digital communications, and social networks, as well as algorithms and artificial intelligence (AI), viewed through the lens of ubiquitous surveillance, control, and manipulation. In this sense, this phrase refers both to older concepts that were elaborated in a philosophical or literary context by Hannah Arendt, Michel Foucault, Aldous Huxley, George Orwell, and others, and to newer theories—such as Shoshana Zuboff’s approach to surveillance capitalism.

In practice, these theories have been partially confirmed primarily thanks to Edward Snowden, a former collaborator of the American National Security Agency (NSA),who “decoded” or revealed a series of secret mass surveillance programs in 2013, conducted by the NSA in collaboration with other government agencies and corporations, using platforms and companies such as Google, Facebook, Apple, Microsoft, and others. These programs had a huge reach and enabled surveillance of communications in the global space, including not only states and business competitors but also ordinary

citizens, which sparked a massive public debate about privacy, freedom, and the abuse of state power. Snowden's revelations were direct evidence of "mass surveillance" and marked the beginning of the era of so-called "smart totalitarianism."[2] On that occasion, a decade later, at a cryptocurrency conference (2022), Snowden, participating via video link from Moscow in the proceedings organized by the website "CoinDesk" stated the following: "Before 2013, there were experts and scientists who understood that mass surveillance was possible and that it was likely happening. However, this was treated as a conspiracy theory within institutions and the general public. These were more doubts, speculations. And in 2013, it became clear that this was a reality, that the world had changed." (DW, 2023).

Around the same time, initially as a pilot project related to finance (2011), and later extending to other areas of social life (2014), a social credit system was

[2] Yahoo, for example, as mentioned in the article "Mass Surveillance and 'Smart Totalitarianism'" intends to patent „smart billboards" that would be placed along highways, at airports and ferry terminals, and in general, in public transportation systems, as well as in bars and hotels, at intersections, and in other public and private spaces. These digital advertising panels would rely on a range of invasive surveillance technologies, such as mobile towers, applications, images, video cameras, vehicle navigation systems, satellites, drones, microphones, motion detectors, and biometric sensors, including fingerprint, retina, and facial recognition devices. Yahoo's smart jumbo billboards would be used to identify specific individuals in order to ascertain their demographic data, as well as their socio-economic status, thereby exploiting their needs and habits (in a consumer sense) based on the collected data. Yahoo has termed this process "grouplization" while some critics have labeled this exploitation of personal data for corporate gain as "Stasi capitalism" (Spannos 2016).

developed in China. The wider public became aware of the phenomenon of the Chinese social credit system only in 2019 when media outlets and analysts initiated discussions on the subject. The media began reporting on the mechanisms of implementing the system, as well as the potential implications for other countries in the future, particularly in the context of digital surveillance and control. The social credit system in China is, in fact, a state project that utilizes a combination of technologies such as surveillance, artificial intelligence, and data analysis to evaluate or score the behavior of citizens and companies. It functions as an economic and social mechanism for scoring individuals based on their activities, including financial transactions, public behavior, compliance with laws and contracts, and even personal and social interactions. Currently, this system is only partially implemented, meaning it has not yet been fully integrated into Chinese society. In this regard, Chinese sources claim that citizens mostly support this method of media surveillance and control, which allegedly serves as a stimulus for moral improvement for the people of China. In contrast, sources outside of China, from both ethical and political perspectives, generally condemn this treatment of citizens, as it is based on centralized power and the loss of individual privacy and freedom. Examples of extreme pro and contra reactions include the uprising of the Uyghurs, an ethnic minority from Xinjiang, whose struggle against the Chinese regime is also a rebellion against the social credit system on one side, while, on the other, certain circles in the West express disbelief that such a scoring system even exists in China.

In relation to the previous statements, the question arises regarding the role of artificial intelligence in enforcing a regime of authoritarian control, moderating

behavior, and manipulating data. It seems that artificial intelligence (AI) plays a very significant role in the context of realizing so-called digital totalitarianism, as it enables sophisticated and automated forms of surveillance, analysis, and manipulation of vast amounts of data (big data) collected about citizens. In this sense, AI technologies contribute to the strengthening of totalitarianism in the digital space through advanced algorithms for pattern recognition, behavior prediction, and decision-making without human intervention. Thus, artificial intelligence is, in fact, a key technology that makes digital totalitarianism possible and potentially sustainable. Through surveillance and control, algorithmic profiling (behavior analysis and modeling to predict future actions or potential threats), information manipulation, and automated decision-making, AI enables authoritarian regimes and corporations to collect, analyze, and utilize data about individuals in ways that were previously unimaginable. The relevance of AI's use for digital totalitarianism primarily lies in its ability to support mass surveillance, information manipulation, and shape citizens' behavior through automated processes (without their knowledge and consent), raising a series of serious questions about privacy, freedom, and human rights in the digital age.

These broader social processes that are in progress and could eventually lead to the more widespread implementation of artificial intelligence for the purpose of establishing societies of total digital surveillance and control are not possible without understanding the mechanisms by which algorithms operate, whether they are used as tools that assist humans or as autonomous systems that can function independently of human beings. Therefore, our examination will focus first on epistemological questions and only then on the issues of

social use of AI, which can contribute to the loss of human freedom on a large scale.

To gain a deeper understanding of these processes, we will attempt to connect the operation of AI in the pursuit of digital totalitarianism with Austin's concept of "other minds." In this regard, we will focus on the crucial ideas from both domains of theoretical research: the problem of knowledge about other minds (and the position of so-called "epistemological uncertainty"), as well as how digital totalitarianism can utilize technology to establish control and influence over the behavior and minds of other individuals.

First, if we recall the oracle at Delphi and the task posed not only to philosophers but to anyone on the path of knowledge—namely, knowledge primarily understood as self-awareness—we can rightly transform this task into an infinite striving for knowledge as such, or a love of wisdom. Thus, this striving is primarily recognizable as a process of self-discovery, which can then be assumed to pave the way for further knowledge. However, how can one seek further when the first condition (self-discovery) has not been fulfilled? The path that philosophy has traversed to this day regarding this same task can be seemingly expanded to other subjects, but it always remains the same initial unknown and uncertainty regarding it. Therefore, it is certain that in order to explore artificial intelligence, we must first understand our own, based on which we can make any claims about it (as intelligence). In other words, we presuppose that something is intelligence by analogy with our own, even though we know little about it, meaning we navigate among assumptions.

A similar dilemma, but directed towards other "intelligences" has been posed by those interested in exploring not only our own minds but also the so-called "other minds." For instance, philosopher, scientist, and diver Peter Godfrey-Smith, in his work *Other Minds: The Octopus and the Evolution of Intelligent Life*, argues that the entire nature, throughout the evolution of species, becomes self-aware, pointing to octopuses, which have a multitude of neurons in their bodies (tentacles) that seem to think for themselves (Godfrey-Smith: 2017). The question, of course, is whether this neuronal awareness is truly consciousness and whether—and, primarily, how octopuses "think." When considering whether other species think and how they do so, we can only speculate starting from ourselves, but do we truly know how we think? Of course, we pose this question to ourselves as a distinct "thinking" species (homo sapiens). On the other hand, if we compare an octopus to intelligent algorithms, although the octopus is a product of (intelligent?) nature, and algorithms constitute the essence of AI, we would find that what we humans call "consciousness" which has the brain as its relevant physiological basis, is not linked to intelligent behavior in algorithms and octopuses, but rather to the appropriate arrangement of neurons or neuronal "networks."

In the rich tradition of philosophy, thought (or consciousness) has been a subject of inquiry for many philosophers, from Parmenides and his understanding of being, metaphorically represented as a solitary thinking sphere, through Descartes, who thematized thought itself, to German classical philosophy and Hegel, who demonstrated through the dialectical movement of concepts that the entire reality is rational. It is well known, of course, that "other minds" were also addressed

by Austin, who posed questions about how we know that we know anything about so-called other minds, as explored in his essay "Other Minds." To connect Austin's analysis of other minds, drawn from the article "Other Minds" with the concept of digital totalitarianism, we will focus on the key ideas from both domains of inquiry: one concerning the problem of knowledge about other minds (epistemological uncertainty), and the other on the ways in which so-called digital totalitarianism uses technology to gain control over and influence the behavior and minds of other people.

To explore this connection, we will treat artificial intelligence (AI) as an "other mind" approaching it similarly to how Austin approached any other mind. The distinction, however, lies in the fact that—when it comes to AI—this other mind is not "natural." In other words, it is not a matter of an octopus or another human being. Thus, by analogy, we might ask whether we know, and more importantly—how we know, for instance, that artificial intelligence is angry (assuming this question is meaningful at all). Or, redefined: such a question could translate into what we know about AI's thought processes and on what basis we draw conclusions about them. On the other hand, as our creation, this "other mind" such as artificial intelligence, is likely to regard us as "other minds" as well.

Austin's work, as is well known, highlights the following epistemological problem: how can we be certain that we understand or even have access to the inner thoughts, intentions, and feelings of other people? Digital totalitarianism represents an attempt by governments and corporations to eliminate this epistemological uncertainty by using artificial intelligence and surveillance technologies to gain comprehensive insight into the thoughts, feelings, and

behaviors of other minds—i.e., people. When addressing the question of how we know that we know how someone feels, what they think, etc., Austin, using anger as an example, demonstrates the difference between knowledge and probability, that is, between certainty and assurance on the one hand, and their opposites on the other. "First of all," the author argues, "it is certain that we sometimes say we know that another person is angry, and we also distinguish these situations from those where we merely say we believe they are angry. For, of course, we do not for a moment assume that we always know, nor about all people, whether they are angry or not, nor that we are capable of discovering it" (Austin 1980: 171-172).

If we accept this as true, we could conclude that, regarding the anger of an "other mind" instead of making assertions, we can only offer more or less well-founded assumptions. What is, in our view, even more intriguing is that recognizing the anger of a particular person applies exclusively to that individual and within a recognizable context of observation. In contrast, how do we behave when it comes to artificial intelligence, conceived as an "other mind"? Here, the concept of "mind" is undoubtedly understood as a crude reduction—specifically, to intelligence, or, as will later become evident, to intelligent behavior. Fundamentally, we can, it seems, rightfully assume that what pertains to mind or consciousness is not necessarily of anthropomorphic origin. For, as we have already indicated, intelligence can also be possessed by other natural organisms and, in our understanding, by entities that are artificially generated. Thus, humans are not exclusively "other minds" to one another, and here we are examining additional possibilities.

To summarize: John Austin, in his essay *Other Minds*, asserts that we can draw conclusions about the mental states of others only based on external manifestations; we cannot directly know their inner content. In other words, it is legitimate to say that someone appears angry, but we cannot know with certainty how that person truly feels or what they are thinking. This highlights the issue of epistemological distance that exists between our experience and the experience of others. Applying this analogy to artificial intelligence, our understanding of the "content" of AI is quite similar—we cannot precisely know what AI "thinks" or "feels," especially since AI does not possess subjective consciousness like humans do. Instead of discussing content, we are in a position to observe the behavior of algorithms or intelligent systems, their responses, and decision-making processes, fully aware that these manifestations are the products of programmed models, not conscious experiences. Essentially, we conditionally grant ourselves the right to claim that we understand how AI functions in terms of *input-output* systems and how it uses data to make decisions. However, we cannot speak about the "content" of AI's consciousness or its intentions in the same way we do for humans. Admittedly, familiarity with a specific system provides us with insights about AI akin to those we might have about an individual – for example, recognizing situations where we might describe someone as angry. This approach emphasizes the parallels and limits of understanding "other minds" whether human or artificial.

If the consciousness of an octopus, human consciousness, and artificial intelligence—all together, but also taken individually—are considered "other minds" what, then, is the actual difference? The differentiation, of course, is primarily established by the

one who observes and compares these "other minds." At first glance, a paradox emerges. Unlike the octopus—and especially unlike the human mind, where we speculate with varying degrees of probability about whether it is angry—we have a pre-established trust in our knowledge when it comes to artificial intelligence. Thus, in a certain sense, AI is indeed an *"other mind"* but distinct from the human mind, given that it is artificially created. With this in mind, we usually assume that we know more about AI systems than about human consciousness, as we ourselves have created these systems and are familiar with their internal mechanisms, algorithms, and code. Moreover, we understand how data is processed, which models are used for decision-making, and we are capable of tracking every step of their operation.

However, if we position ourselves to compare with human consciousness, assuming that we know something about the "other mind" only based on external signs, we can similarly say that artificial intelligence algorithms externally signal "responses" without us knowing how it (AI) "experiences" those processes, if we can even speak of any "experience" at all. Although we relatively understand the mechanisms and technical workings, this also represents the limit of our knowledge of artificial intelligence. Just as with humans, we know details about the structure of the brain, neurons, synapses, and the ways impulses are transmitted, yet we do not fully understand how subjective experience or consciousness arises in humans. In this sense, our understanding of how the brain functions might be comparable to our understanding of how algorithms function in artificial intelligence—we have insight into the technical aspects, but consciousness remains a mystery.

In the same way that we understand the physical structures of the human brain but do not comprehend how subjective consciousness emerges from them, we face a similar problem with AI. In the case of artificial intelligence, we know its basic algorithms, mathematical models, and methods of data processing, but this does not reveal what, if anything, constitutes an inner consciousness or experience within those processes. Our understanding of AI, much like our understanding of human consciousness, remains at the level of external manifestations—behavior, responses, actions—but that "inner world," if it exists, remains epistemologically elusive. Thus, both entities, human consciousness and AI, remain domains of the unknown for us, even though we know a great deal about their physical or technical foundations.

However, some researchers would disagree with the claim that we cannot "read" other minds, especially when it comes to so-called superintelligent AI, or autonomous, strong versions of artificial intelligence. Interpreting Harris's views from the article titled: "Reading The Minds of Those Who Never Lived. Enhanced Beings: The Social and Ethical Challenges Posed by Super Intelligent AI and Reasonably Intelligent Humans" (Harris 2019), Sven Nyholm argues that in the mentioned text, the philosophical "problem of other minds" is reduced to the interaction between human beings (reasonable beings) and artificial intelligence, which could become superintelligent (Nyholm, 2019). This, according to our understanding, excludes the consideration of the "nature" of any consciousness, whether ours or others'. Instead of knowledge about consciousness, it is reduced to behavior, that is, to reciprocity and interaction with enhanced artificial intelligence.

To explain this, we will use the following analogy. Let's say that we extract a substance from nature about which we know nothing, except how it reacts with certain natural elements, and then we mix it with a substance that is artificially (laboratorially) produced, for which we know in principle how it reacts in a laboratory environment. Then, under controlled conditions, we mix the first and second substances multiple times, observing their interaction, from which we draw both specific and general conclusions. The first open question regarding this experiment is: what does "under controlled conditions" mean? The second question is: what can we learn from the interaction of these two substances, and how many times should we repeat the experiment (this or a similar one) before we can conclude that we understand their interaction and that it is under control?

Haris, as explained by Nyholm, considers what could be called *"bilateral mind-reading."* This reciprocity concerns those human beings who "read the thoughts" of artificial intelligence, on the one hand, and intelligent systems that "read the thoughts" of human beings, on the other. From this two-way "mind-reading," primarily focusing on the possibilities of superintelligent AI, further potential ethical and social implications are drawn (Nyholm, 2019:1). Following this, in our opinion, is a significant interpretive difference between Harris's and Nyholm's understanding, because before presenting the possible consequences (social and ethical) of mutual "reading" between humans and smart machines, Nyholm argues that we should first ask, regarding robots and other intelligent machines, whether their algorithms are even intelligent, and especially whether they are "intelligent in a human way" and whether we have any evidence of that (Nyholm 2019:3). In our view within a paradigm that is humanitarian, understood in the broad

sense of the term, consciousness is primarily defined as human, and thus the consequences of its possible action—ethical and social—follow from this.

And finally, following Austin, what could we say about Harris's "mind-reading" as a certain knowledge that we gain through interaction with machines, as well as those with us? In the subsection of the article "Other Minds" titled: "If I know, I can't be wrong," Austin actually wants to suggest that it is not the same when we know something and when we think we know something. In the first case, if we know something, it necessarily implies that we are confident in our knowledge and that we are right. However, if our knowledge is uncertain, can we still be right? Here's what the author says about this: "The final problem regarding the challenge 'How do you know?' directed at the one who uses the expression 'I know,' should be addressed by considering the statement 'If you know, you cannot be wrong'. Certainly, if what has been said so far is correct, then we often have the right to say that we *know*, even in cases where it later turns out we were wrong—and indeed, it seems that there is always, or almost always, a possibility that we may be wrong" (Austin 1980: 165). In other words, even when we think we know something, and there is a basis for it, although we have the right to claim that we know something (such as the content of others' thoughts or consciousness), we can be wrong. And should we, in fact, claim to know something (that we do not know) while also potentially making mistakes, especially when the consequences are serious in social or ethical terms? Such questions are very important, and sometimes they can be decisive when it comes to AI autonomous decision-making systems.

In the following part of the article, and in order to reassess the potential of AI in the context of actions that favor totalitarian systems of surveillance, control, and decision-making, we will analyze Searle's thought experiment, which deals with "strong" AI systems, i.e., the very artificial superintelligence for which it can be assumed that it possesses certain cognitive properties. Searle, in fact, conceived the Chinese Room experiment to reflexively test the idea of whether a computer, simply because it processes data and produces correct answers (based on the operation of algorithms), can possess consciousness or understanding of what it does. In this thought experiment, the author is supposedly in an isolated (locked) room, receiving external information in the form of Chinese characters (which he does not understand); he then uses rules (algorithms) to combine them according to instructions in English, and finally gives correct answers to those outside the room, without any understanding of the meaning of what is written (Searle, 1980). This experiment leads to the conclusion that correct answers from the supercomputer/"strong AI" are not the same as human understanding of a problem, thus challenging the assumption of machines as conscious entities (Searle 1980).

The connection between Austin's theses and Searle's Chinese Room problem actually lies in the impossibility of proving the inner state of another entity—whether that entity is a human (Austin) or a machine (Searle). Hence, the key relation between Austin's discussion of other minds and Searle's argumentation regarding the so-called Chinese Room is based on doubt about conclusions drawn about some internal state (of the mind) that are derived from external manifestations. While Austin is skeptical about the human capacity to discern another's consciousness through behavior and

speech, Searle, on the other hand, criticizes the idea that algorithmic processing of information can be conscious. Both emphasize the gap between what is externally visible (behavior, speech, or algorithmic processing) and what is truly happening inside (consciousness, understanding).

The connection between Austin's and Searle's research with the idea of digital totalitarianism can be observed in the context of examining how symbolic manipulation can be used in modern digital systems for control and surveillance, without any deeper understanding or empathy toward human beings and their needs. Digital totalitarianism, as a system in which ruling structures use advanced technologies, such as artificial intelligence, to monitor, predict, and control the behavior of citizens, can be built on a similar principle as the "Chinese Room." This implies that AI systems track, analyze, and manipulate data without truly understanding its meaning or the (social, ethical, and other) consequences of their actions. In digital totalitarianism, machines can make decisions that directly affect people's lives; however, such decisions are not and cannot be the result of genuine reflection, but merely data processing. Intelligent machines cannot truly make moral judgments, nor are they aware of human authentic needs, rights, and freedoms. Just like in Searle's experiment, AI systems "act" as if they understand what they process, but their manipulation of symbols (in this case, information about humans) can be deeply dehumanizing.

What is dangerous in such a system is the potential for the abuse of data manipulation that machines possess or could possess. If elites in positions of power use artificial intelligence in this way, they can justify decisions and

actions based on "objective" data, while, in reality, there is a huge space for errors, manipulation, and ethical, political, and other abuses of such systems. Citizens, or victims, could be labeled, sanctioned, or controlled based on algorithms that do not understand the true context of their actions or intentions, thus creating a system that is mechanical and impersonal, but presented as rational and fair. Furthermore, if the context is capitalist (primarily referring to the economy, not ideological bias), it becomes parasitic in relation to those it monitors, controls, and exploits.

Shoshana Zuboff, in her book *The Age of Surveillance Capitalism*, referring to Marx and earlier times when capitalism, resembling vampiric existence, "fed on the labor" of the proletariat, thereby generating surplus value, i.e., profit, claims that in the current so-called surveillance capitalism, which, as the author states in the subsection of the first part titled "What Is Surveillance Capitalism?," is "parasitic" and "self-referential" (Zuboff 2019), exploitation continues, but this time in the digital space. It is, of course, realized in a much subtler manner – through the use of non-thinking machines that observe, direct, and ultimately exploit our experiences. Surveillance capitalism defines capitalism in such a way that control is no longer exercised over the means of production, but over tools for modifying human behavior. The bearer of these modifications, Zuboff suggests, is the company Google, followed by other related internet corporations. In other words, capitalism, as the dominant socio-economic formation of our time, through the omnipresent digitalization primarily dictated by the economy, represents the main threat to the rights and freedoms humanity has gained so far; and we add: the technological foundation for introducing digital forms of dictatorship in the modern era. According to

Shoshana Zuboff, the goal of the surveillance capitalist order is the prediction and modification of human behavior (*behavior modification*) in order to achieve profit through control.

But the power of surveillance capitalism is not only in knowing our behavior but also in monopolizing the information that results from user behavior (personal human experience); hence, the collection, condensation, analysis, and design of these data are transformed into the so-called "behavioral surplus" that is traded on the market (Digital Literacy and AI, 2022).

What, however, do we, as researchers, know about technologies programmed to manage our behavior, and how can we even come to such knowledge? If we only have knowledge based on theoretical assumptions, which we prove or refute through thought experiments, while on the other hand, reality occurs about which we know little or nothing in terms of managing new AI technologies—how can we break through or remove the cognitive barrier about surveillance, control, and management of our behavior if this, among other things, is a business secret of leading *high-tech* companies? In pursuit of these uncertainties and unknowns, we will, in the following text, attempt to connect Searle's experiment with the concept of the *"black box"* borrowed from Wiener's cybernetics.

Namely, within Wiener's cybernetics, the "black box" is conceptualized as a system for which it is not necessary to know its internal structure, yet it can still be studied based on input and output information. For illustration, we will provide the explanation given on the website of the Faculty of Computer Science in Belgrade: "When mentioning the term 'black box' many think of devices that record what happens in airplanes and are

extremely important for analyzing events that occurred before an accident. In English, the term is also used for small and minimally equipped theaters. However, the black box is also an important concept in the world of artificial intelligence. In the field of artificial intelligence, black boxes refer to internal processes that occur in artificial intelligence systems, which are completely inaccessible and unknown to users. As we know, you can give such systems a query and receive an answer, but you have no insight into the system code or logic that created the answer, i.e., the output" (Faculty of Computer Science, 2024).

If we compare Searle's experiment with the briefly outlined concept of the black box, we will notice the following similarities. In his "Chinese Room" experiment, Searle tries to demonstrate that understanding language or "consciousness" is not merely a matter of correctly processing symbols, but it requires understanding the meaning (semantics) of what is said or written. The experiment is a demonstration of the idea that a certain program can simulate knowledge without actually possessing it. The "black box" theory, now widely known in engineering, refers to systems that operate in such a way that the user does not fully understand them. In the black box, the input and output information are known, but the internal processes remain hidden or unknown. This theory is often used to describe artificial intelligence, where the output (decision or result) is known, but the way in which AI arrived at the appropriate outcome often remains opaque and difficult to understand, even by programmers and AI trainers. Similar to how in the "Chinese Room" a person inside the room simply follows rules without understanding what they are doing, the "black box" can perform complex processes without the ability to uncover how

and why those decisions were made. In both cases, the problem of "transparency" arises, and the question is—how can we claim that a system (human or artificial) truly understands what it is doing, if our access to internal processes is limited or completely unavailable? Thus, this issue can trigger a discussion about how complex systems, like AI, can simulate intelligence or understanding, with epistemological boundaries similar to those Searle emphasizes in his experiment.

One attempt to overcome this problem and provide users with insight into how the artificial intelligence they use functions is defined as the tendency to move from the use of "black box" systems to algorithms that are transparent, i.e., to "white box" (*White box AI*) or "glass box" (*Glass Box AI*) systems. This shift aims to achieve what is called explainable artificial intelligence (*XAI*), with the goal of ensuring transparency and accountability in AI's actions through the development of models that can explain how AI systems make their own decisions (Johal, 2023). With XAI, users can gain insight into the reasons behind the recommendations and decisions made by artificial intelligence, making them more trustworthy and understandable. It should be emphasized that XAI should not only satisfy our curiosity but also ensure that its use is, colloquially speaking, harmless. Especially important is that the use of XAI ensures that artificial intelligence aligns with human values and ethics. This would be, from a values perspective, a humanistically centered artificial intelligence (Johal 2023).

In recent articles on the potential humanistic use of explainable artificial intelligence (XAI), a shift in approach is noted, moving from a technocentric to a humanistic and sociotechnical focus. Traditional XAI

was actually focused on algorithmic transparency, that is, explaining how AI makes decisions by opening the "black box." However, the development of more complex models over time has shown that it is difficult to "open" these systems and fully understand their internal processes. Hence, the fundamental question when it comes to the explainability of AI is – for whom should it be understandable? "In fact, the challenges of designing and evaluating AI in the 'black box' depend on who the human in the loop is" (Ehsan & Riedl 2020: 450). In other words, the question arises about the social recognizability of AI's work, as well as potential abuses.

The posed challenge has led to the emergence of so-called "Human-Centered Explainable AI" *(HCXAI)*, which focuses on human-centered aspects of explainability, arguing that AI systems do not function in a vacuum, but within a human environment (Ehsan & Riedl, 2020). Thus, intelligent systems are believed to be part of "sociotechnical" frameworks in which humans are key to the interpretation and use of these systems (Ehsan & Riedl, 2020). These frameworks are shared by both humans and machines, and are supposedly developed for mutual benefit. Therefore, the moment of explainability depends not only on the technological transparency of the system's operation but also on how artificial intelligence is used and interpreted in a social context. In other words, the use of HCXAI advocates an approach that incorporates the values of both users and designers, suggesting that explanations do not have to come only from algorithms but can also be found outside them; because here lie humans and the human environment, who are crucial for interpretation and decision-making in the real world (Ehsan & Riedl 2020).

From the above, it is clear that HXAI generally aims for artificial intelligence to be more transparent and human-centered, thanks to human explanations of the decisions made by AI. However, if this transparency is backed by malicious, manipulative processes and interests, the integrity of HXAI and the individuals, companies, or regimes behind it can be called into question. In such systems, there is a risk of abuse, meaning someone might attempt to use AI system explanations for manipulation, which is problematic, as information can be (mis)used to subtly influence people. In this context, the question arises: does someone truly want to help the user understand decision-making processes, or do they want to shape their thinking and, consequently, behavior in a direction that benefits them? This assumption brings the discussion back to the idea of possible abuse of ignorance and information, as even technology that appears to serve transparency may have a hidden agenda. Based on the above, the relevant question is: what is the true purpose of these explanations and who actually controls them?

Assume that manipulation of AI systems is profitable. In this case, it is hard to imagine that those benefiting from it would be honest about the methods used. Precisely because they are part of the system, the manipulator (whether personal, corporate, or institutional) is in a position to use explanations as a facade, reassuring users that everything is transparent and in their best interest, while actually aiming for control and manipulation. In such a scenario, the explanations are not real explanations but strategies to create a false sense of security or trust in AI. These actions create the illusion of understanding, reducing user suspicion in the absence of real disclosure of the underlying mechanisms of manipulation. This

mechanism can become self-sustaining—the system provides users with information that limits their ability to recognize manipulation, while at the same time pretending to help them better understand decision-making processes. In other words, such a system can operate by creating an illusion of control and choice, while keeping users within pre-set "decision-making" options.

Who creates such a system, based on the asymmetry of knowledge (and power)? At this point, we will introduce an argument based on the division into so-called algorithmic elites, on one side, and users of opaque AI systems, on the other. Such terminology is often used in critical analyses that deal with the opacity of algorithmic systems and the power based on them, as part of the contemporary discourse on technological and political power, viewed in the context of the actions of large technology companies, as well as those regimes that control and create algorithms shaping current societal trends. The concept of technological (algorithmic) elites refers, as the name suggests, to a limited number of people or organizations (usually large tech companies) that create and control algorithmic systems, or possess the knowledge and capacity to design algorithms that govern key aspects of modern life: financial markets, social networks, internet searches, employment, crediting, judicial processes, education, media, etc. These elites, as opposed to so-called ordinary users, possess disproportionate power and achieve significant societal influence and profit by using algorithmic tools to make important decisions in a manner that is not transparent to the broader public. Their power stems from access to vast amounts of data and the ability to use it to manipulate markets, social interactions, and political processes.

It is not uncommon for the mentioned elites to avoid oversight by the state or the public, since their algorithms, as we have said, are opaque or closed ("black boxes"). Very few people understand how they actually work, which allows the techno-elite to control processes that affect the daily lives of billions of people. In short, these elites control social infrastructures in a way that prevents ordinary users from accessing information about how algorithms reach certain conclusions. This creates a huge advantage for the elites over users and allows them to maintain a monopoly on algorithmic infrastructure. Furthermore, algorithmic elites have concentrated power in the hands of tech giants like Google, Facebook, Amazon, and others. These companies control the infrastructure through which information, products, and services are disseminated worldwide. It is clear that algorithmic elites can abuse their power for their own benefit, without being held accountable to laws or rules that protect societal interests.[3] Additionally, these algorithms can act discriminatively, spread misinformation, or target specific groups of people without their knowledge or consent.

In the context of the analysis of digital totalitarianism, it is important to critically reflect on algorithmic elites, as the power of algorithms to influence human behavior and political processes can be considered a form of technological authoritarianism, where they take on the

[3] A well-known example is the case of Facebook (Meta) concerning the abuse and mass sale of user data, influencing elections, spreading fake news, censorship, violence, and the exploitation of children (discussions in the U.S. Congress), as well as evading taxes, for which neither the CEO of Facebook, Mark Zuckerberg, nor his company have been held accountable.

role of decision-making that was previously reserved for democratic institutions. The irresponsibility of these elites and the inability of ordinary citizens to directly monitor these processes lead to the emergence of digital totalitarianism. This situation paves the way for the realization of what Frank Pasquale critically discusses in his book *The Black Box Society* (Pasquale 2015), namely the potential creation of a *"society of black boxes."*

Pasquale in this book thoroughly analyzes how opaque algorithms control key social processes, primarily financial flows, information, and the judiciary. In his study on black boxes, the author delves into the strategies that allow them to remain closed to the public. In this sense, keeping in mind that knowledge is power, Pasquale argues that the deconstruction of black boxes in *big data* is not a simple task (Pasquale 2015:6). Even if the algorithmic elite were willing to expose their methods to the public, the modern internet and banking sector would still pose barriers to understanding the methods of manipulation in relation to end users (Pasquale 2015: 6). For, the conclusions drawn by internet corporations and bankers, for instance, about employee productivity, the relevance of websites, or favorable investments, are synthesized through complex formulas devised by "legions of engineers and guarded by a phalanx of lawyers" (Pasquale 2015: 6).

Understanding and monitoring the development of AI technologies in relation to the phenomenon of black boxes is important to ensure they are used in a way that respects human rights and democratic values. However, if digital technologies, particularly artificial intelligence, are abused to achieve complete control over information, communication, and human behavior, in this context, the society of black boxes could anticipate or represent an

early stage of the implementation of digital totalitarianism. This would also mean that technologies like AI, which function as black boxes, enable corporations or governments to make non-transparent decisions without taking responsibility, thereby limiting human freedoms and abusing power (knowledge) to the detriment of the majority of people.

More precisely, the perception of a society of black boxes in the context of interconnected digital technologies, especially AI, can be seen as a new form of networked ("rhizomatic") digital order in which people, institutions, and systems are connected through complex and often invisible networks of decision-making. In such a society, the networking of black boxes would represent multiple layers of interactions and communications that are opaque and difficult to understand by external observers. In this case, advanced artificial intelligence would become the core of managing the networked society, but in a way that the manner in which it makes decisions is often not transparent.

Furthermore, in this interpretative context, black boxes would not only be algorithms but also the systemic decisions made by artificial intelligence, as well as the way these decisions are applied. This means that individuals' daily activities could be controlled and moderated through AI systems, without clear insight into the criteria or processes that shape these decisions. Individuals within this system would not be able to understand or manage these processes, leading society into a phase where algorithms, rather than people, become the authorities of knowledge and social practice. Such a dynamic, if not "illuminated" and controlled, could lead to digital totalitarianism, in which people are subject to surveillance and controlled through a system

they do not understand, while simultaneously being unaware of how to protect themselves from its effects.

However, the question of transparency in the functioning of black boxes is merely superficial, while the real problem lies in techno-capitalism, which uses technology as a tool for control, exploitation, and the maintenance of systemic power within the capitalist socio-economic order. In other words, the digital network of black boxes does not need to be transparent; it certainly operates in a way that enables centralized power and profit in the hands of technocratic and capitalist elites. This new, more lucid way of surveilling and controlling individuals and entire communities includes sophisticated forms of manipulation, where black boxes (AI systems, algorithms) serve corporate and capitalist interests, keeping users in a subordinated position of ignorance and servitude. Thus, technology becomes a means for further exploitation of resources. Instead of traditional labor, capitalism now uses digital systems to extract value from data, algorithms, and "smart" digital interactions. The networking of black boxes within the capitalist system serves to more efficiently manage and control the workforce and consumers. This surveillance is not carried out only through traditional control mechanisms but through technology that possesses autonomous decision-making powers. In this way, the illusion of freedom is created for individuals, while they are, in reality, increasingly integrated into networks of surveillance, control, and exploitation.

Therefore, we believe that demystifying these processes, as well as critically reflecting on them, is the first step in addressing the problems brought by techno-capitalism. In this context, understanding how artificial

intelligence technology is used to establish and maintain power, moving towards enslavement and digital totalitarianism, as well as exploring possibilities for alternative and more humane approaches to this issue, can help in building a fairer and more democratic system that aims to achieve a greater degree of freedom, both for the individual and for specific social communities of our time.

References

1. Austin J. L. (1980). Tuđe svesti, in: Svest i saznanje, Beograd: Nolit, p.p.141-184.
2. "Defining Surveillance Capitalism" (2022), in: Digital Literacy and AI, Pepperdine Libraries, https://infoguides.pepperdine.edu/digitalliteracy/surveillancecapitalism,15.10.2024.
3. Deset godina kasnije: Edvard Snouden i njegova otkrića, (2023). DW, https://www.dw.com/sr/deset-godina-kasnije- edvard-snouden-i-njegova-otkri%C4%87a/a-65837392,
4. Ehsan, U. Riedl, M. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach, https:// www.researchgate.net/publication/340438092_Human-centered_Explainable_AI_Towards_a_Reflective_Sociotechnical_Approach, pdf.
5. Godfrey-Smith, P. (2017). Other Minds: The Octopus and the Evolution of Intelligent Life, London: William Collins.
6. Johal, A. (2023). From Black Box to Glass Box: The Evolution of Explainable AI: The Human Side of AI: Why Explainable AI is Essential for Human-Centered AI, 16.

7. Harris, J. (2019). Reading the minds of those who never lived. Enhanced beings: The social and ethical challenges posed by super intelligent AI and reasonably intelligent humans. Cambridge: Cambridge Quarterly of Healthcare Ethics 28(4), p.p. 585–591.
8. Nyholm, S. (2019). Other minds, other intelligences: the problem of attributing agency to machines. Cambridge University Press: Cambridge Quarterly of Healthcare Ethics , 28 (4), p.p. 592-598; pdf, p.p. 1-8.
9. Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information, Cambridge, Massachusetts, London, England: Harvard University Press.
10. Spannos, Ch. (2016). Mass Surveillance and "Smart Totalitarianism," ROAR: Issue #4: 'State of Control,' 4. 10. 2024.
11. „Šta je veštačka inteligencija u crnoj kutiji?" (2024). Računarski fakultet Univerziteta Union u Beogradu, https://raf.edu.rs/citaliste/sta-je-vestacka-inteligencija-u- crnoj-kutiji/, 16. 10. 2024.
12. Vuksanović, D. (2022). Philosophy in the Time of Media and Technological-information Madness, in: In medias res: časopis filozofije medija, Vol. 11 No. 20, Zagreb: Centar za filozofiju medija i mediološka istraživanja, p.p. 3292, https://www.centar-fm.org/inmediasres_eng/index. php/divna-vuksanovic-philosophy-in-the-time-of-media- and-technological-information-madness, 20.10. 2024.
13. 13. Zuboff, Shoshana (2019).The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, we.riseup.net Crabgrass, na stranici: https://we.riseup.net, pdf.