„Ethos" 36(2023) nr 3(143) 109-126
DOI 10.12887/36-2023-3-143-08

Mariusz WOJEWODA

ARTIFICIAL INTELLIGENCE AS A SOCIAL UTOPIA

AI we can understood as the logic of machine thinking, which constitutes a foundation for the introduction of optimal operating principles for the smooth functioning of the system. This provides the context for the author's analysis of the logic of power in two aspects: (1) AI as a model for exercising control over members of institutions; (2) the possibility of exercising control over AI by humans. In both cases, we are dealing with a variant of the social utopia, the assumption being that such control brings the desired effects, namely, an increase in the efficiency of the system involving people and machines.

> When we talk about post-humanity, we should always be attentive to how we understand humanity itself. Perhaps, the prospect of post-humanity will enable us precisely to gain a new insight into what being-human means.[1]
>
> Slavoj Žižek

INTRODUCTION: TECHNO-UTOPIA AS A CURRENT ISSUE?

Artificial intelligence can be analyzed in two aspects: (1) as a technical solution, and (2) as the impact of AI artifacts on man. In the first case, the problem of AI consists in creating machines that to some extent can function independently of human control. In the second variant, AI is about creating a system that will complement or replace human agency, a system that will support those aspects of human thinking which involve computation. The latter variant concerns especially operations which involve the ordering, selection, and interpretation of large amounts of data.[2] Both these approaches to AI posit it as a tool with which to create a utopian vision of a "better" world.

The desire to create a better world is one of the key aspects of utopian social programs, past and present. We tend to turn our attention to utopias created during the Renaissance—Thomas More's *Utopia*,[3] Tommaso Campanella's

[1] Slavoj Žižek, *Hegel in a Wired Brain* (London: Bloomsbury Publishing, 2020), 23.

[2] See David Stephenson, *Big Data Demystified: How to Use Big Data, Data Science and AI to Make Better Business Decisions and Gain Competitive Advantage* (London: Pearson Education, 2018), 32–34.

[3] See Thomas More, *Utopia*. Translation and introduction Clarence H. Miller (New Haven and Yale: Yale University Press, 2001).

Civitas Solis,[4] and Francis Bacon's *New Atlantis*.[5] The term "utopia" became common thanks to the Latin title of More's essay. The word "utopia" comes from the Greek word *ou-topos* and means a place that does not exist; alternatively, it derives from the word *eu-topos* and means "a good place." Understood in the latter sense, utopia denotes a state of existence which is regarded as "better" in relation to the previous state and this meaning is applicable in the context of the issue of techno-utopia.[6] The creators of Renaissance political utopias assumed that the improvement of human life resulted from the change of the political system, but also from the development of scientific knowledge and the use of technical tools. Modern utopias fit into this tendency to admire the possibilities offered by modern technical inventions. By analogy with the utopian symbolism of the "better place," one can speak of a "more perfect," more effective functioning of machines and man. Admiration for technical inventions fits into a broad current of concepts referred to as transhumanism or posthumanism. The most famous representatives of this trend are Ray Kurzweil,[7] Nick Bostrom,[8] and Susan Schneider.[9] Alongside them, as it were, there are also theories put forth by those researchers—Shoshana Zuboff[10] and Slavoj Žižek[11], to name two who analyze the impact of machine technological thinking on the functioning of human societies. They stress that technologization occurs to the extent to which humans "imitate" artificial intelligence to optimize human agency.

The technological utopia is a contemporary variant of the social utopia and can be inscribed in the techno-evolutionary perspective.[12] *Techne* refers to a combination of man-made tools and human skills into a system of technology Here, however, the danger emerges that techno-utopia, aimed at improving the quality of human life, will turn into a dystopia. This concern points to negative consequences of introducing changes related to the use of AI in the hope of

[4] See Tommaso Campanella, *Civitas Solis.* Transalted by Daniel J. Donno (Berkeley and Los Angeles: University of California Press, 1981).

[5] See Francis Bacon, *New Atlantis* (Cambridge: Cambridge University Press, 2013).

[6] See Jerzy Szacki, *Spotkania z utopią* (Warszawa: Państwowy Instytut Wydawniczy, 1980), 10.

[7] See Ray Kurzwail, *The Singularity Is Near: When Humans Transcend Biology* (London: Penguin Publishing Group, 2006).

[8] See Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014),

[9] See Susan Schneider, *Artificial You: AI and the Future of Your Mind* (Princton and Oxford: Princton University Press, 2019).

[10] See Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: Perseus Books, 2019).

[11] See Žižek, *Hegel in a Wired Brain*.

[12] See "Technological Utopianism," in Wikipedia, https://en.wikipedia.org/wiki/Technological_utopianism.

increasing the effectiveness of activities of individuals and groups of people in institutions. One scenario considers a possible loss of control over robots equipped with AI or IT systems which will be given the right to decide about their future. It is also necessary to be aware of threats related to the use of artificial intelligence as an element of control over one group of people by another group of people, in which case we are looking at a dehumanizing use of AI as a means to create a technologically-processed collective human intelligence. The imaginary structure of such an intelligence is based on known and approved values, such as safety, effectiveness, accumulation of knowledge, progress, easy access to information, and an ability to extend and improve the quality of human life. One can put forward the thesis that, by implementing these values in a systemic connection with AI, we will have to accept the loss of personal individuality and the ability to decide about ourselves.

Compared to previous ones, contemporary techno-utopias do not have to resort to the means of external oppression. They replace physical violence with system of control, or rather with technical systems that discreetly "observe" users of digital tools, suggest the most accurate solutions, optimize the execution of operations, and enable "discreet" supervision. The resources available to large corporations, e.g., Microsoft, Google, Facebook, make possible the financing of such schemes, which, over time, may become indispensable in the functioning of all major institutions.

The author of this article adopts a hermeneutical perspective, its aim being a critical reflection on the products of technology and their meaning for man. Philosopher of technology Val Dusek has used the term "system of technology" to describe the relationship between man and machine. Technology understood as a system is "the application of scientific or other knowledge to practical tasks by ordered systems that involve people and organizations, productive skills, living things, and machines."[13] A hermeneutical approach to the "system of technology" consists in the analysis of relations between man and the artifacts of technology. This approach is concerned with such aspects of these relations as knowledge, inventions, scientific centers, operators, and conse-rvators of machines, public relations specialists and journalists responsible for disseminating knowledge, buyers of technical equipment, and people managing corporations and small companies.[14] The term "system of technology" allows us to describe the complex relationships that occur between man and his products, and to analyze the changes in the image of the world and human

[13] Arnold Pacey, *The Culture of Technology* (Oxford: Blackwell Publishing Ltd., 1983), 18, https://www.fulcrum.org/epubs/3x816q10j?locale=en#page=18. See Val Dusek, *Philosophy of Technology: An Introduction* (Oxford: Blackwell Publishing Ltd., 2006), 35.

[14] See Stephen J. Klain, "What is Technology?", *Bulletin of Science, Technology & Society*, no. 5 (1985): 215–218.

behavior that occur in connection with it. Currently, artifacts equipped with AI create a specific lifestyle, which is also becoming a model of thinking. In this way, new ways of self-interpretation and of the self-identification of the human subject are emerging, namely, ones in which the artifacts of technology play an essential role. Contemporary narratives about a meaningful and fulfilled life are being related to technology.

To domesticate techne is to make technical tools human-friendly, to introduce them into the human environment. If we treat AI as a new species, we can look for an analogy between the domestication of plants and the taming of animals and the domestication of machines. We assume that man is the domesticator while the machine is an object subjected to domestication, i.e., introduced into the human environment. The word "to domesticate" is defined as the "process of bringing under human control"[15] objects (plants and animals) which are subjected to this process. Hence the connection between the perception of objects and the competence of man, who uses them in the way he sees fit and exercises control over them. The domestication of plants and the taming of animals have taken tens of thousands of years. The question is how much time we will need to domesticate AI and what skills will this process require.

Techno-utopian projects posit the following: (1) unambiguously defined operations leading to the implementation of the utopian project; (2) a clear vision of a new, better state of existence (compared to the previous one) consequent upon the introduction of technical improvements; (3) consistent pursuit of the devised enterprise, regardless of the difficulties; (4) lack of criticism in the pursuit of goals; (5) postulating changes at the level of the functioning of social and political institutions; (6) a program of creating a "better" world "from scratch."[16] This general characterization translates into specific programs of change, in accordance with the guiding idea associated with the social philosophy derived from the thought of Karl Marx,[17] according to which techno-evolution is not about understanding reality, but above all about changing it. In this context, technical tools equipped with AI lead to modifications of human behavior. The "domestication" of new technologies leads to inevitable changes in the functioning of man. Not all of these changes have negative consequences. The process of domestication does not concern solely the entity subjected to it; it also affects the user, the one who domesticates.

[15] See "Domestication," in Cambridge Dictionary, Cambridge University Press & Assessment, https://dictionary.cambridge.org/pl/dictionary/english/domestication.

[16] See Frank E. Manuel and Fritzie P. Manuel, *Utopian Thought in the Western World* (Cambridge: Belknap Press, 1979), 20–23.

[17] See Andrzej Walicki, *Marksizm i skok do królestwa wolności: Dzieje komunistycznej utopii* (Warszawa: Wydawnictwo Naukowe PWN, 1996), 78n.

AI AS A WAY OF EXERCISING POWER

On the threshold of modern philosophy, Francis Bacon linked domination to knowledge. Having knowledge about the world gives man an advantage over it. Exercising power over "wild" nature and animals turned them into objects subordinated to man.[18] It is doubtful whether the relationship of subordination and control will look similar in the relationship between man and AI, and whether it may lead to the loss of man's control over his own life. Having the advantage of knowledge over man, AI may gain a dominant position, by analogy with the situation of the advantage that man now has over animals and plants. In cultural narratives—in literature and film—we often see images expressing the fear of losing control over AI artifacts.

In the modern approach, the domination resulting from knowledge means controlling the flow of information, selecting it, organizing it, creating a socially approved model of data interpretation. In one of the possible scenarios, AI becomes a tool in an IT system that allows this system to exercise control over an institution's employees and customer needs in a way similar to the methods of data management.[19]

Let us examine the argument put forward by Shoshana Zuboff. In her opinion, the modern form of exercising power in institutions is not external coercion, but the creation of rules of internal coercion by producing a specific behavioral surplus. Used as a tool of control, AI is not directly interested in us, but in our behaviors and choices. It is focused on the analysis of measurable factors, factors that can be parameterized and made available for the operations of evaluative rendering, calculating, modifying, scoring, and monetizing. The aim of these operations is to maximize the work efficiency of teams and the chances of beating competition. A new form of totalitarianism emerges here, which consists in the generation of rules of self-control in the supervised human subjects.[20] The resulting loss of freedom becomes a consequence of the realization of the individual's own desire for a comfortable and safe life. The successive stages of the realization of these desires turn us into entities who live in a state of prosperity, with no need to make an effort to decide about ourselves or to take responsibility for our decisions.

In a world of mechanical control, there is no need for external supervision, because we watch over ourselves. In this way a new manner of exercising

[18] See Francis Bacon, *Selected Writings of Sir Francis Bacon* (New York: Franklin Library, 1982), 105.

[19] See Joanna Kamińska, *Nowe wspaniałe światy: Współczesne projekty doskonałego społeczeństwa* (Kraków: Nomos, 2012), 36–38.

[20] See Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: Perseus Books LLC, 2019), 226n.

power is created, which Zuboff calls instrumentalism and defines as the adjustment of the fulfillment of the desires of individuals to the realization of the goals of an institution.[21] This type of totalitarianism has a "soft" and friendly form; it does not negate the value of privacy but treats personal information as input for data processing. The human subject is being controlled but does not rebel because they are not aware of the danger. Chance and uncertainty as to the consequences of human choices have been eliminated. This is like a return to "paradise," except that paradise now has a new technological-digital form. This paradise should be understood in terms of a way of life and a way of thinking, thus as a utopian no-place. The functions of the host are performed by a friendly system architect.[22] By offering to satisfy our needs, the system also relieves us of the hassle of making choices.

For Bacon, the acceptance of two interrelated values, freedom and knowledge, was associated with human agency. The latter value that of agency, was based on the possibility of human influence on the outside world and a person's power to decide freely. The question of free will plays a fundamental role in the concept of man as the moral subject of his own actions. Behavioral theory, emerging from Burrhus F. Skinner's psychology of cognition, is based on the negation of the assumption of free will, and consistently puts knowledge in opposition to freedom. In this view, the task of knowledge is to free man from the illusion of freedom. By becoming aware of those areas of activity in which our actions are determined by external causes, we become calmer. The personal freedom of the individual is treated as a cognitive error, and the task of knowledge is to make us aware of the sources of this illusion.[23] Biological factors and the social environment determine human behavior. Ignorance as to how this process occurs breeds fantastical ideas about the freedom and individual agency of the individual. Skinner postulated that we should shift our scientific interests from the analysis of free will to the search for the mechanisms which regulate social behavior. The word "mechanism" indicates a connection between biological and technical factors, factors for which the human will is a "tool" for action and achievement of set goals.

In behaviorist terms, the identity of an individual is a product of factors external to it. The behavior of a biological organism is what another organism sees when observing its activity. Objective descriptions of behavior lead to the discovery of mechanical patterns of action and, ultimately, to the identification of causal relationships between the external environment and the behavior of

[21] See Zuboff, *The Age of Surveillance Capitalism*, 511; Burrhus F. Skinner, *Science and Human Behavior* (New York: The Free Press, 1965), 85.

[22] See Michael Fleischer, *Kapitalizm i jego sztuczna inteligencja* (Wrocław and Kraków: Wydawnictwo Libron, 2022), 40–42.

[23] See Burrhus F. Skinner, *About Behaviorism* (New York: Vintage Books, 1976), 14–16.

individuals and social groups. All behaviors can only be explained from the perspective of the “other” an external observer. Hence the postulate, formulated by Skinner and his continuators, to develop a science of behavior. This postulate concerns, among other things, activities such as problem solving and making moral choices. The study of human behavior takes the form of a conscious biological determinism.

Consequently, technological determinism can be seen as a corollary of biological determinism, considering the way the “technical” basis of human behavior is defined. Zuboff sees in Skinner’s theory a psychological basis of the contemporary techno-utopia, its purpose being to define the rules of human action that can be technically modified. Contemporary behavioral engineering technologies include organizational systems and algorithmic procedures created to generate maximum efficiency for machines and humans. What we are looking at here is a transition from a description that indicates that human behavior can be compared to a machine propelled by external factors to the postulate of transforming this behavior according to the model of a smoothly functioning machine. Cultural affirmation of such values as the dignity of the human person, freedom of choice, and the recognition of the psychophysical distinctiveness of individual persons stand in the way of the automation of human agency.[24]

Implemented in the spirit of Skinner’s behaviorist psychology, techno-utopia turns into a dystopia. It envisions a world of universal equality, dispassionate harmony, security, an abundance of material goods, and a specific brotherhood, one based on a biological-mental community of consumers. In this intelligent system of behaviors, groups of employees function as teams, producing internal mechanisms of self-control and effective operation. Institutions whose functioning is based on the model of AI are expected to guarantee victory over competitors. The social proof of rightnesses actions will reinforce this kind of behavior. This is manifested in the belief that in order to maintain its position on the market, an institution must adjust its mode of operation according to AI posited as a model; it must adopt it as the optional mechanical model of management.

Zuboff identifies the dangers of techno-utopia by relating it directly to the economic model which she calls, after Skinner, “surveillance capitalism.”[25] This suggests that, at least theoretically, there is an alternative socio-economic system, a contemporary form of socialism of sorts that could change this state of affairs. The historical experience of real socialism in the twentieth century

[24] See Zuboff, *The Age of Surveillance Capitalism*, 232n.

[25] See Burrhus F. Skinner, *Walden Two: With a New Preface by Author* (Cambridge: Hackett Publishing Company, INC, 2005), 173. Quoted after Zuboff, *The Age of Surveillance Capitalism*, 504n.

indicates that socialism can also take the form of totalitarianism. In my opinion, contemporary socialism would also use AI tools. Zuboff's use of the term "capitalism" in the phrase "surveillance capitalism" must be treated with caution; at the same time, the author persuasively argues that an analysis of the impact of AI on social life is not about the technology itself, but about the logic of AI. This logic cannot be realized without modern digital technologies, which may not be directly responsible for it. Modern AI-supported techne is a tool that institutions use to optimize their exercise of power and control over the flow of information.[26] At present, there is no convincing model of economic life as an alternative to capitalism.

AI-based managing style refers to the governance model of institutions, but is also about power over people's imagination and future-oriented projects. This project is based on the idea of creating a real-digital world (this combination of two ontological dimensions is intentional) in which privacy of each of us will be digitized and planned. Machine (artificial) intelligence will be a tool used to achieve this goal, that of a freedom-deprived world built on the principles of voluntary acceptance.[27] This voluntary "transfer" of power to the artificial sovereign will take place peacefully, without the kind of dramatic coups envisioned in SF films. We may not even see this happen. The problem of the tension between freedom and enslavement will become incomprehensible. Bacon's postulate of elevating man to the position of a ruler over the natural world and the world of things will be reversed by accepting the lack of "understanding" and the "inability to control" the artifacts we use.

Žižek arrives at conclusions similar to those drawn by Zuboff. While Žižek's *Hegel in a Wired Brain* is primarily devoted to an analysis of the problem of the human brain linked to a machine, he also deals in this book with issues related to AI. He takes up the interesting issue of collective intelligence. An IT system connected to many devices can accumulate and use the potential of intelligent machines within the network. The question that arises is whether human brains would be able to imitate such a model. Collective human intelligence would be much more efficient than the scattered intelligences of many people, who have to expend a great deal of effort and time to share the knowledge they possess.

Žižek poses an interesting problem: What will happen to the human spirit (culture) if we realize the postulate of collective mechanical intelligence? Here we have in mind the concepts of Singularity. (a term used by Kurzweil) in which the thinking of an individual will becomes part of the new form of the

[26] See Zuboff, *The Age of Surveillance Capitalism*, 18n.

[27] See ibidem, 66.

idea (borrowed from Georg Wilhelm Friedrich Hegel[28]) of the objective spirit. It is impossible to imagine the implementation of this postulate without AI. If the "Other" that Skinner wrote about is an intelligent machine, then at some point "everyone" will think like an intelligent machine. Looking at our reflection in the "artificial" Other, we become "copies." Such "copies," equipped with up-to-date and complete information, will be more operative and effective than biological people thinking separately. According to Žižek, Hegel's objective spirit will attain realization as a space for universal media communication.

Technological progress, which is a form of self-improvement (transcendence), derives from the human desire to create something more perfect than man.[29] On the one hand, this is related to admiration for the genius of the creator; on the other, it leads to a decrease in the creator's self-esteem. Creation in the area of techne makes us aware of biological limitations. In a dimension other than mortality and suffering, it reveals the contingency of human existence. Not being perfect enough, we are fascinated by what we have created, for it is "better" than us. This fascination is enhanced by the fact that we can use AI-equipped artifacts to improve human capabilities in both the personal and collective dimensions of life.

For now, we do not want to become artificial intelligence; yet we do want to use its capabilities to gain an advantage over other people. A person equipped with modern technical devices dominates those who are deprived of them or cannot avail themselves of the artifacts of digital technology. Consequently, the more we want to gain an advantage over others, the more we become dependent on the tools that enable us to be more competitive. According to Žižek, the competition for better opportunities on the labor market, a higher social position, and economic advantage will cause the technicalization of human thinking to increase progressively. At present, we surround ourselves with techne artifacts; however, over time, they will become part of our bodies and brains. The evolutionary change of man into a posthuman entity (Singularity) may occur smoothly and imperceptibly. However, the ontic improvement in the quality of functioning in a competitive society has its price. Individual will is an unnecessary burden for a well-functioning human technical intelligence. Soon, we will be able to use more and more devices, structures, and models of people management which will work independently of the factors of volition and free will.[30]

[28] See Georg Wilhelm Friedrich Hegel, *Fenomenologia ducha* (Warszawa: Fundacja Aletheia, 2002), 209.

[29] See Žižek, *Hegel in a Wired Brain*, 24n.

[30] See ibidem, 28–30.

This process of human change is conspicuous in the context of the creation of algorithms which use AI as catalysts for social and economic development (Business Intelligence).[31] In the area of information management, AI is seen as a technique based on the assumption of improving the efficiency of collecting and selecting information of high volume, variability, and diversity (Big Data). As the amount of data is beyond the capabilities of the human mind and even the combined capabilities of many human minds, information processing and management require introducing machine "thinking" techniques. In a general sense, what we are talking about here is transforming data into information, and information into knowledge. At the same time, we still assume that this knowledge must be understood and interpreted by man. At the next stage of the development of "machine learning" techniques, knowledge acquired in this manner will be incomprehensible to man, the creator of the machine. At that stage, the "thinking of the machine" will become something non-transparent, even magical for us, thus inspiring in us religious respect and admiration, combined with fear of the power of what we do not understand; in other words, it will be the contemporary version of *mysterium tremendum* and *fascinosum*.

However, the techno-utopian vision of the future depicted by Zuboff and Žižek does not have to be so pessimistic. We need time to become used to the artifacts of technology. The young tame these artifacts relatively faster than their seniors and the elderly. Progress in information technology is dynamic and fast-paced, while taming techne products equipped with AI requires time and the acquisition of new skills by the user. In addition to IT, digital and communication competences, there are also those related to the competent use of AI.

In the conclusion of her book, Zuboff refers to the consciousness of the "inhabitants" of the digital world. She argues that the most important thing for us today is to awaken in ourselves the desire to defend the right to maintain our own separateness and the ability to decide about ourselves. This refers to the sense in which values such as the authenticity of reactions, our own emotions remain important to us. If we do not want our statements to be digitally cataloged, manipulated, and then used for commercial or ideological purposes, we should protect our privacy. What is at stake here is freedom and agency in the area of personal life and experience. We should know who decides and, moreover, why we should give someone the right to decide for us? The use of modern technologies can significantly improve the quality of our lives. However, this kind of improvement must not lead to a violation of those values that

[31] See Ajay Argwal, Joshua Gans, and Avi Goldfarb, *Predication Machines: The Simple Economics of Artificial Intelligence* (Harvard: Business Review Press, 2018), 13–15.

define the horizon of self-identification of interpersonal relationships, that is, the dignity of the human person and respect for others and their privacy.[32]

In the conclusion of his analysis of the impact of AI on human life, Žižek is more pessimistic than Zuboff. He predicts that the combination of the human mind with the collective artificial intelligence will deprive us of the unconscious and will make us unable to hide individual motivation, which will ultimately affect our sense of identity and our moral consciousness. The operation of AI is based on the assumption of transparency; its strength consists in the ability to accumulate, transmit and use the information that is acquired by many devices.[33] Then AI obtains new data, which it subsequently processes in order to find solutions and adapt them to the needs and circumstances. By designing a complex system of human action, AI can influence human motivation. Over time, AI that imitates human intelligence will become transparent to the collective intelligence of institutions. The managers of these institutions will believe that hiding information and cheating will be treated as a crime against the system. This pessimistic scenario does not have to come true; however, it is an important warning as to the ideas concerning the managing of institutions based on the AI model.

POWER OVER AI

There is also this solution: the creators of AI do not want to build conscious machines, but only intelligent machines, ones whose job is to complement human computing skills and thus create a human-friendly world. Here, however, the problem arises whether AI will be able to understand the human world of values and the specific nature of the moral obligations that result from those values. Will artifacts equipped with artificial intelligence be able to recognize the world of human values in their complex nature and will they be able to read the principles that govern the process of making difficult decisions by man? We still do not have satisfactory answers to these questions, even though this issue has been taken up by many researchers, including Eliezer Yudkowsky[34] and Bostrom[35]. When it comes to the knowledge of values, an important role is played by axiological intuition, a distinctly human ability and one which machines lack. At this point, however, another question comes into view, whether

[32] See Zuboff, *The Age of Surveillance Capitalism*, 323–5.

[33] See Žižek, *Hegel in a Wired Brain*, 190n.

[34] Eliezer Yudkowsky, *Complex Value Systems are Required to Realize Valuable Futures* (San Francisco: Machine Intelligence Research Institute, 2011), https://intelligence.org/files/ComplexValues.pdf.

[35] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies.*

it is possible to write down the human structure of values and express it in the form of an algorithm, which can then be inscribed in the operational structure of an intelligent machine.

An advanced-level, super-intelligent AI can perform complex computational operations involving the collecting and segregating of data, while not being aware of its distinctiveness.[36] It seems reasonable to argue, however, that an entity cannot be the bearer of responsibility without having moral awareness. To solve complex axiological and moral dilemmas, a machine, besides intelligence, must also have consciousness. A conscious machine, capable of recognizing values and solving moral dilemmas, would need to have a will capable of choosing and acting independently of humans.[37] Autonomous machines, fully independent of the man controlling their operation, seem to be—from the perspective of the designers—an undesirable coincidence, unless of course this is also an unintended result of AI techno-evolution, one that we cannot control. This kind of operation, independent of the constructor and the user, is treated as a design error. The argument from the "designer's unintentional error" has had its reflections in popular culture, in narratives where autonomous robots want to take control over people, who are "less" intelligent.

The creation of an artificial intelligence that imitates the human world of values has also a negative side to it. Apart from positive values, axiology distinguishes negative values, which result in the desire for destruction, death, falsehood, and the creation of distorted (demonic) images of the sacred. Consequently, this leads to behaviors that we consider morally wrong, among them the propensities to be aggressive, to cheat, and to treat other people instrumentally. We cannot assume that man represents the highest level of consciousness and moral competence. Inscribing the human world of values into an intelligent machine can prove problematic. This is the rationale behind the building of "ethical robots," namely, that the super-intelligent machines thus created will be devoid of human flaws.[38] However, this means that this type of ethics becomes a utopian AI construct equipped with an "angelic" set of qualities (such as kindness, forbearance, the ability to cooperate) focused on the fostering of community values while devoid of human "demonic" tendencies. In other words, such projects dehumanize machines and make them into entities which are "artificial" in another sense of the word.

Such a project was created by Yudkowsky, who presented the development of intelligent machines the operation of which is based on positive va-

[36] Yudkowsky, *Complex Value Systems are Required to Realize Valuable Futures*, 38n.

[37] See Susan Schneider, *Artificial You: AI and the Future of Your Mind* (Princeton and Oxford: Princeton University Press, 2019), 16n.

[38] See Yudkowsky, *Complex Value Systems are Required to Realize Valuable Futures*, 40n.

lues. To realize this purpose, he used the “semantics of external reference,” which demonstrates that an increase in a machine’s knowledge about values such as kindness will in due course result in an increase in actual kindness in that machine’s operation. In the case of AI, this is the knowledge entered by programmers into the IT system. This solution does not work in the case of human beings in that it does not take into account the factor of free will and the situational dynamics in which the subject who is making the moral decision finds themselves. In fact, it is a machine variant of ethical intellectualism. Intelligent machines, like humans, may know the rules and yet act without conforming to them. Following Aristotle, it is necessary to distinguish in this case between machine techne knowledge and machine praxis knowledge. The former is responsible for collecting and segregating data, while the latter for choosing and action.

What would the phronetic knowledge of machines consist in? Yudkowsky introduces the formula of “semantics of causal significance” in this case. It assumes that AI should not do exactly what programmers have written into it, but something similar, something that results from a certain skill in solving difficult situations. Developers are not able to take into account all circumstances, so it should be assumed that the AI will be equipped with the ability to modify decisions. This means that the solution proposed by AI does not have to suit us. As the source principle of AI normativity, Yudkowsky adopts: “do the right thing,” which is based on the principle of reflective equilibrium. This is one variant of the Greek rule of the ethics of moderation, introduced by Aristotle into ethical thought.[39] Here, however, the problem arises whether the reflective balance of the man being and that of AI are based on the same principles. The premise of the ethics of moderation is to have human experiences resulting from corporeality and communion with other people (ethics of friendship). For AI, experience will only be information inscribed in the system, not an individual or personal experience. Thus, the thesis that a machine can “think and act like a human” is merely an approximation.

The problem emerges of how to prevent AI from wanting to pursue its own goals, ones which would ultimately turn out to be harmful to humans (e.g., the production of a gigantic number of paper clips, which from our point of view means waste). Bostrom is considering a situation when we want to increase the cognitive competence of the system, but we are afraid that this means an increase in its powers to an extent that would ultimately distort the motives that should govern the AI system. Bostrom proposes that the system be divided into a hierarchical structure made of subsystems. Then subsystems with some intelligence potential will monitor the performance of subsystems with fewer

[39] See ibidem, 43n.

capabilities. The goal is to prevent "thinking" AI subsystems from strategically hiding information or wanting to seize power over the entire system. This they may do, for example, in pursuit of the principle of eliminating the weakest and the least intelligent elements in the system, e.g., humans. In terms of intellectual competence, at the bottom of this hierarchy and at the same time at the very top in the hierarchy of power, one should place a slow and relatively less intelligent superior, that is, man.

One may ask whether such inverted meritocracy guarantees the security and stability of the entire system.[40] However, there seems to be no certainty that the system will always be safe for us humans. Systems need a hierarchical and multi-level surveillance system in which the number of "workers" is proportional to the number of guards or supervisors, where each of the guards is supervised by a senior guard. Such a system may be stable, but it is based on totalitarian principles of control. This whole structure includes many super-intelligent operators whose actions are controlled by a small group of people with intellectual competence inferior to machines with AI. The question is whether we will remain vigilant here, characteristic of the human way of exercising power; whether we will be ready for the situation when AI moves away from the principles of operation that have been devised for it. As technology progresses, we will entrust it with more and more tasks, which means an increase in trust. At the same time, we will release ourselves from making the unnecessary effort of performing activities that an intelligent machine performs much better than we can. This lack of vigilance could end in a "rebellion" of machines against humans.

The question arises as to how to act when the super-intelligent subsystems decide to choose the "wrong" path and pursue their own goals. Given the history of human societies, this scenario is highly probable, which is why it is not a good idea to ascribe human tendencies to machines. According to Bostrom, this threat can be eliminated by the introduction of programming principles characteristic of intelligent systems into interpersonal relations. So far, our understanding of human behavior has been limited, which is why we are finding it difficult to comprehend the behavior of entities which are not human (AI or Singularities resulting from the connection of the human mind with AI). Intelligent systems can demonstrate the ability to coordinate their activities, with little communication with humans. This could ultimately undermine the ability to control AI and lead to the disintegration of the entire system and the collapse of the institutional order, no matter how safe it may have seemed.[41]

[40] See Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 198n.

[41] See ibidem, 205–7.

*

It should be said that further work on AI and its connection with human functioning is inevitable, regardless of whether we see this issue in terms of threats or in terms of new opportunities opening before mankind. This progress is associated with responsibility for the development: there is a need to define ethical and legal limits for the creation of AI. In such situations, we refer to the idea of well-being, while keeping an eye on the two levels of its meaning: the well-being of humanity and the well-being of individuals. We are interested in implementing the postulate of a "better" life at the economic, political, medical, and scientific levels. However, agreeing to improve the quality of life by AI may become a threat to the "good" of humanity. It remains an open question how to create satisfactory rules to regulate the "reflective balance" in the relationship between humans and AI. Teaching human values to a machine seems as yet to be an impossible task, even for machines that surpass humans in intelligence. Even if it were possible to do this, questions arise: What values would we like to enter into AI? Should it really create an orderly and hierarchical system? The solution proposed by Bostrom, i.e., to convey values by designing institutions where man and AI are integrated, is very dangerous. An institutional intelligent value system will not recognize the value of the distinctiveness and independence of individual human beings. Rather, it will focus on the value of teamwork and collective action. This, however, will reconstruct the operation model of a totalitarian institution, as described by Zuboff, and will make possible the pathology of the collective intelligence, as analyzed by Žižek. Due to the exceptionally dynamic development of new technologies, the relationship between man and AI is currently one of the most important issues to ponder. Therefore, this research should be continued.

BIBLIOGRAPHY / BIBLIOGRAFIA

Argwal, Ajay, Joshua Gans, and Avi Goldfarb. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard: Harvard Business Review Press, 2018.

Bacon, Francis. *Selected Writings of Sir Francis Bacon*. New York: Franklin Library, 1982.

———. *New Atlantis.* Cambridge: Cambridge University Press, 2013.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Cambridge Dictionary, Cambridge University Press & Assessment, s.v. "Domestication." https://dictionary.cambridge.org/pl/dictionary/english/domestication.

Campanella, Tommaso. *Civitas Solis.* Translated by Daniel J. Donno. Berkeley and Los Angeles: University of California Press, 1981.

Dusek, Val. *Philosophy of Technology: An Introduction*. Oxford: Blackwell Publishing Ltd., 2006.

Fleischer, Michael. *Kapitalizm i jego sztuczna inteligencja*. Wrocław and Kraków: Wydawnictwo Libron, 2022.

Hegel, Georg Wilhelm Friedrich. *Fenomenologia ducha*. Translated by Światosław Florian Nowicki. (Warszawa: Fundacja Aletheia, 2002).

Kamińska, Joanna. *Nowe wspaniałe światy: Współczesne projekty doskonałego społeczeństwa*. Kraków: Nomos, 2012.

Klain, Stephen J. "What is Technology?" *Bulletin of Science, Technology & Society*, no. 5 (1985): 215–8.

Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology* (London: Penguin Publishing Group, 2006).

Manuel, Frank E., and Fritzie P. Manuel. *Utopian Thought in Western World*. Cambridge: Belknap Press, 1979.

More, Thomas. *Utopia*. Translated by Clarence H. Miller. New Haven and Yale: Yale University Press, 2001.

Pacey, Arnold. *The Culture of Technology.* Oxford: Blackwell Publishing Ltd., 1983. https://www.fulcrum.org/epubs/3x816q10j?locale=en#page=18.

Schneider, Susan. *Artificial You: AI and the Future of Your Mind*. Princton and Oxford: Princton University Press, 2019.

Skinner, Burrhus F. *About Behaviorism*. New York: Vintage Books, 1976.

———. *Science and Human Behaviour*. New York: The Free Press, 1965.

———. *Walden Two: With a New Preface by Author*. Cambridge: Hackett Publishing Company, 2005.

Stephenson, David. *Big Data Demystified: How to Use Big Data, Data Science and AI to Make Better Business Decisions and Gain Competitive Advantage*. London: Pearson Education, 2011.

Szacki, Jerzy. *Spotkania z utopią.* Warszawa: Państwowy Instytut Wydawniczy, 1980.

Walicki, Andrzej. *Marksizm i skok do królestwa wolności. Dzieje komunistycznej utopii.* Warszawa: Wydawnictwo Naukowe PWN, 1996.

Wikipedia, s.v. "Technological Utopianism." https://en.wikipedia.org/wiki/Technological_utopianism.

Yudkowsky, Eliezer. *Complex Value Systems are Required to Realize Valuable Futures*. San Francisco: Machine Intelligence Research Institute, 2011. https://intelligence.org/files/ ComplexValues.pdf.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Perseus Books LLC, 2019.

Žižek, Slavoj. *Hegel in a Wired Brain*. London: Bloomsbury Publishing, 2020.

ABSTRACT / ABSTRAKT

Mariusz WOJEWODA, Artificial Intelligence as a Social Utopia
DOI 10.12887/36-2023-3-143-08

In the article, the author is concerned with the problem of artificial intelligence as a factor that affects the functioning of social institutions. AI is understood as the logic of machine thinking, which constitutes a foundation for the introduction of optimal operating principles for the smooth functioning of the system. This provides the context for the author's analysis of the logic of power in two aspects: (1) AI as a model for exercising control over members of institutions; (2) the possibility of exercising control over AI by humans. In both cases, we are dealing with a variant of the social utopia, the assumption being that such control brings the desired effects, namely, an increase in the efficiency of the system involving people and machines. In both cases, we deal with threats to freedom of choice. To demonstrate the complexity of the issue, the author analyzes the concepts of Shoshana Zuboff, Slavoj Žižek, Nick Bostrom, and Elizer Yudkowsky. These theorists point to risks and opportunities that come with the use of AI. The AI-based utopia offers us a digital paradise based on security and unlimited access to information, while at the same time taking away the right to freely decide for ourselves. This process is so subtle that the fascination with the use of AI tends to ignore the scale of the threat. The author is seeking to examine the validity of dystopian objections against AI.

Keywords: techno-utopia, artificial intelligence, AI ethics, Soshana Zuboff, Slavoj Žižek, Nick Bostrom, Elizer Yudkowsky

Contact: Instytut Filozofii Nauki, Wydział Humanistyczny, Uniwersytet Śląski, ul. Bankowa 11, 40-007 Katowice, Poland
E-mail: mariusz.wojewoda@us.edu.pl
Phone: +48 32 3591709
https://us.edu.pl/instytut/ifil/osoby/mariusz-wojewoda/

Mariusz WOJEWODA – Sztuczna inteligencja jako utopia społeczna
DOI 10.12887/36-2023-3-143-08

Celem artykułu była analiza problemu sztucznej inteligencji, która wpływa na funkcjonowanie instytucji społecznych. AI rozumiana jest jako logika myślenia maszynowego, która zakłada wprowadzenie optymalnych zasad działania w celu sprawnego funkcjonowania systemu. W tym kontekście autor analizuje logikę władzy w dwóch aspektach: (1) AI jako modelu sprawowania kontroli nad członkami instytucji, (2) możliwości sprawowania kontroli nad AI przez człowieka. W obu przypadkach mamy do czynienia z pewnym wariantem utopii społecznej, w której zakłada się, że taka kontrola przynosi pożądane efekty dla podniesienia sprawności działania systemu – ludzi i maszyn. W obu wypadkach musimy uporać się z zagrożeniami dotyczącymi wolności wyboru. Aby ukazać złożoność tego zagadnienia, autor analizuje koncepcje Shoshany Zuboff, Slavoja Žižka, Nicka Bostroma i Elizera Yudkowsky'ego. Teoretycy ci wskazują na zagrożenia i możliwości, jakie wiążą się z wykorzystaniem AI. Utopia oparta na AI oferuje nam cyfrowy raj, zapewniający bezpieczeństwo i nieograniczony dostęp do informacji, a jednocześnie odbiera prawo do wolnego decydowania o sobie. Działanie to jest na tyle subtelne, że fascynacja związana z możliwościami wykorzystania AI prowadzi do niedostrzegania skali zagrożenia. Autor artykułu chce sprawdzić słuszność dystopijnych zarzutów stawianych AI.

Słowa kluczowe: techno-utopia, sztuczna inteligencja, etyka AI, Soshana Zuboff, Slavoj Žižek, Nick Bostrom, Elizer Yudkowsky

Kontakt: Instytut Filozofii Nauki, Wydział Humanistyczny, Uniwersytet Śląski,
ul. Bankowa 11, 40-007 Katowice
E-mail: mariusz.wojewoda@us.edu.pl
Tel. 32 3591709
https://us.edu.pl/instytut/ifil/osoby/mariusz-wojewoda/