Krzysztof Wieczorek
University of Silesia in Katowice
Faculty of Humanities
https://orcid.org/0000-0002-7987-168X

Er(r)go. Teoria–Literatura–Kultura
Er(r)go. Theory–Literature–Culture
Nr / No. 42 (1/2021)
maszyna/podmiot/władza
machine/subject/power
ISSN 2544-3186
https://doi.org/10.31261/errgo.10418

The Conscience of a Machine? Artificial Intelligence and the Problem of Moral Responsibility

Abstract: The ever-accelerating progress in the area of smart technologies gives rise to new ethical challenges, which humankind will sooner or later have to face. An inevitable component of this progress is the increase in the autonomy of the decision-making processes carried out by machines and systems functioning without direct human control. At least some of these decisions will generate conflicts and moral dilemmas. It is therefore worth the while to reflect today upon the measures that need to be taken in order to endow the autonomous, self-learning and self-replicating entities – products equipped with artificial intelligence and capable of independent operation in a wide variety of external conditions and circumstances – with a unique kind of ethical intelligence. At the core of the problem, which both the designers and the users of entities bestowed with artificial intelligence must eventually face, lies the question of how to attain the optimal balance between the goals, needs and interests of both sides of the human-non-human interaction. It is so, because in the context of the expansion of the autonomy of the machines, the anthropocentric model of ethics does no longer suffice. It is therefore necessary to develop a new, extended and modified, model of ethics: a model which would encompass the whole, thus far non-existent, area of equal relations between the human and the machine, and which would allow one to predict its dynamics. The present article addresses some of the aspects of this claim.

Keywords: artificial intelligence, ethics, reinforcement learning, decision-making autonomy

On February 19, 2020, the European Commission published its *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust.* The document presented a series of policy options, defining the stance that Europe should adopt with respect to AI.[1] *White Paper* is thus

> [a] collection of proposals concerning particular actions, determining the directions of the future EU regulations and initiatives in the area of artificial intelligence. The Commission assumes that the new regulatory framework for artificial intelligence will be founded upon the criteria of excellence and trust. Above all, the development of artificial

1. The full text of the document is available at the following URL: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (29.02.2020).

> intelligence is to be human-oriented and must proceed in full respect for European values. […] *White Paper* also includes provisions concerning the fact that AI-based technology should be subject to strict and transparent human oversight, especially when implemented in high-risk sectors. […] Furthermore, *White Paper* posits that the systems using artificial intelligence ought to be subject to state oversight and control.[2]

The work on *White Paper* had been coordinated by Margrethe Vestager, EU Commissioner for Competition and Executive VP. In her speech delivered in February 2019 at the Web Summit Conference in Lisbon, Vestager postulated that in the EU we must assume control of some of the cornerstones of the new technology in order to be able to trust it. In that respect, she further argued, it is necessary that we implement efficient methods of oversight, and, importantly, that we must ensure that such technologies would remain bias-free.[3] It is in such a context that the Commissioner thus cautioned her audiences: "[w]e may have new technology, but we don't have new values."[4]

Announcing the commencement of the work on *White Paper*, Ursula von der Leyen, the newly-elected President of the European Commission, declared that the EC would soon present comprehensive legislation concerning the European approach to the human and ethical consequences of the development of artificial intelligence.[5]

On January 23, the Internal Market and Consumer Protection Committee of the European Parliament (IMCO) passed a resolution concerning the challenges posed by the fast advancement of the AI and Automated Decision-Making technologies. The Committee ruled that when consumers interact with an ADM system, they should be "properly informed about how it functions, about how to reach a human with decision-making powers, and about how the system's decisions can be checked and corrected."[6] Petra De Sutter, Chair of the Committee, said:

---

2. Paweł Zegarow, "Biała Księga w sprawie sztucznej inteligencji" [White Book Concerning Artificial Intelligence], Nask. *Cyber Policy*, February 27, 2020, https://cyberpolicy.nask.pl/biala-ksiega-w-sprawie-sztucznej-inteligencji (29.02.2020). Unless marked otherwise, all citations from texts written in languages other than English are provided in Paweł Jędrzejko's translation.

3. In accordance with Anna Zagórna's report, "Komisja Europejska szykuje plan na SI" [EC Prepares a Plan Concerning Artificial Intelligence], SI. Sztuczna Inteligencja, February 3, 2020, https://www.sztucznainteligencja.org.pl/komisja-europejska-szykuje-plan-na-si (19.02.2020).

4. Quoted in: Anna O'Hare, "The Highlights from Web Summit 2019," November 7, 2019, https://websummit.com/blog/highlights-web-summit-2019 (19.02.2020).

5. O'Hare, "The Highlights from Web Summit 2019" (29.02.2020).

6. Quoted after Isabel Teixeira Nadkarni, "Artificial intelligence: EU must ensure a fair and safe use for consumers," *News European Parliament*, January 23, 2020, https://www.europarl.europa.eu/news/en/press-room/20200120IPR70622/artificial-intelligence-eu-must-ensure-a-fair-and-safe-use-for-consumers (29.02.2020).

> Technology in the field of artificial intelligence and automated decision-making is advancing at a remarkable pace. The committee has today welcomed the potential of these advances, while at the same time highlighting three important issues that need to be addressed. We have to make sure that consumer protection and trust is ensured, that the EU's rules on safety and liability for products and services are fit for purpose in the digital age and that the data sets used in automated decision-making systems are of high-quality and are unbiased.[7]

According to the Committee, the existing ethical guidelines may prove to be insufficient in their scope.[8] Also, legal regulations currently in force need to be subjected to an analysis taking into account the exigencies of life determined by fast-developing smart technologies. In particular, the Committee recommends that "[r]eview structures should be set up to remedy possible mistakes in automated decisions. It should also be possible for consumers to seek human review of, and redress for, automated decisions that are final and permanent."[9]

Meeting the expectations concerning the efficient human oversight of the implementation of automated decisions, the Resolution states that "[h]umans must always be ultimately responsible for, and able to overrule, decisions."[10] This is as important as it is hard to achieve, because, as experts warn us, "AI-enabled products may evolve and act in ways not envisaged when they were first placed on the market,"[11] and such an eventuality must be counteracted well in advance.

The EC structures are expected to develop a common EU approach that will help to secure the benefits of the advancements in smart technology and reduce the risks across the European Union. One of their priorities is to prevent the discrimination of consumers on the basis of their nationality, place of residence, or temporary location, by automatic decision-making systems.

During the debate on the future of artificial intelligence in the EU, which was held in the course of the Parliament's Scientific and Technological Options Assessment (STOA) Panel, Margrethe Vestager attempted to convince her audience that AI may be a boon as long as it is properly prepared and adequately regulated in every sector, especially in sectors whose operations depend on high-risk technology.[12]

Anna Zagórna, the author of the above-quoted article on the EC plans concerning AI, concludes her reflections thus: "the key challenge for Europe is to find

7. Teixeira Nadkarni, "Artificial intelligence."
8. See: Anna Zagórna, "Komisja Europejska szykuje plan na SI."
9. Teixeira Nadkarni, "Artificial intelligence."
10. Teixeira Nadkarni, "Artificial intelligence."
11. Teixeira Nadkarni, "Artificial intelligence."
12. Zagórna, "Komisja Europejska szykuje plan na SI."

a balance between what AI can offer and what must be done to protect privacy, to build trust, and to remain ethical."[13]

It is not only the authors behind the initiative that gave rise to the publication of *White Paper*, but also many other experts, who emphasize the fact that the development of smart technologies raises new ethical challenges: challenges that cannot be met within the frame of the current ethical guidelines for regulatory frameworks. While addressing these problems, it is worthwhile to adopt a long-term approach to the phenomenon of smart technologies and to assume a perspective broader than the one which the EU politicians and experts have been ready to espouse thus far. On the one hand, such an approach would have to take into account wider ontological and anthropological contexts, while, on the other, it would also have to allow for probable scenarios for the further development of the AI technology in the world. One such scenario is explored by Nick Bostrom, a well-known enthusiast of the unfettered development of AI research and an articulate advocate of the fast implementation of its outcomes. In his book *Superintelligence: Paths, Dangers, Strategies*, the author, whom some deem controversial, reflects upon the possibility of the machine intelligence gaining a "decisive strategic advantage"[14] over humankind, which reflection leads him to the solemn question: "Is the default outcome doom?"[15]

Recent actions, undertaken on (or inspired by) the initiatives of the various bodies of the EU, remain focused exclusively on the human dimension of the development of smart technologies. Such a focus pertains primarily to questions of safety and responsibility for the results of research on AI and those concerning the practical implementation of the findings, that is, questions of how to protect privacy, how to build trust, and how to remain ethical in the process. A philosopher, however, may – and ought to – reach further, and expand the horizons of his or her imagination far enough to be able to intellectually tackle the situation which in all probability, in not-so-distant a future, may prove inevitable. In such a future, the more and more technically advanced and less and less predictable processes of AI self-evolution – processes that will not yield to human control – will generate such levels of self-sufficiency of decision-making processes run by autonomous systems that the day-to-day human oversight of the latter will eventually prove to be an illusion.

The inevitable prospect of a dialogue between man and machine – a dialogue in which one of the contentious issues will be the problem of human control

13. Zagórna, "Komisja Europejska szykuje plan na SI."

14. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 79.

15. Bostrom, *Superintelligence*, 115.

of actions taken by artificial intelligence – is addressed, among others, by Hannah Fry in her book *'Hello World': How to Be Human in the Age of the Machine*.[16] The feasibility of such a scenario should be anticipated early enough to responsibly reflect upon "preventive" measures, such as providing the self-perfecting AI software with tools enabling it to develop subsystems of inner moral autonomy.[17] Self-developed, and thereby ADM-integrated, subsystems of this kind would protect us against the undesirable effects of a possible clash between the AI-professed values and the values (and needs) of the human beings. The above notwithstanding, any thus oriented reflection must begin with the question of whether such precautions are at all possible, and – if so – what ways of their implementation must be considered next.[18]

Before we pass on to further stages of our considerations, it is perhaps worthwhile to more closely focus on questions concerning goals and motivations energizing the ongoing work on designing, constructing, and perfecting AI to higher and higher standards. Undoubtedly, the primary motivation to foster such a development is the practical usability of smart machines as universal tools of versatile functionality in an impressively broad range of areas (beginning with military applications and finishing with medicine), especially in the light of their technological efficiency, multi-purpose usage, and reliability. An additional factor, evidently impacting the directions of the research and development work on AI, is the internal logic of technological progress. Such a logic results in the fact that each solution space emerges as largely determined by the class of problems currently explored, only to generate yet another inquiry into the feasibility of further solutions: an inquiry oriented along particular, problem-determined, lines. Artificial intelligence has been potentially inscribed into the multi-stage sequence of tasks, projects, and conceptions inextricably correlated with the development of technology at least since the onset of the Early Modern Age; it was already

16. "As computer algorithms increasingly control and decide our future, 'Hello world' is a reminder of a moment of dialogue between human and machine, of an instant where the boundary between controller and controlled is virtually imperceptible. It marks the start of a partnership – a shared journey of possibilities, where one cannot exist without the other." Hannah Fry, *'Hello World': How To Be Human in the Age of the Machine* (New York–London: W. W. Norton & Co., 2018).

17. Among other scholars, this problem is addressed by Paweł Polak and Roman Krzanowski in their "Ethics in Autonomous Robots as Philosophy in Silico: The Study Case of Phronetic Machine Ethics," *Logos i Ethos*, no. 52 (2020), 33–48, https://doi.org/10.15633/lie.3576.

18. An important contribution to this debate is the article by Stephen Cave, Rune Nyrup, Karina Vold, and Adrian Weller, titled "Motivations and Risks of Machine Ethics," *Proceedings of the IEEE*, vol. 107, no. 3 (March 2019), 562–574, https://doi.org/10.1093/pq/pqv034. Its authors consider the possibility of constructing and implementing AI-enabled "ethical machines" for practical uses.

then that the possibility of constructing a "thinking machine" would frequently become an object of serious reflection.[19]

However, other factors impacting the directions of research and work on the AI development also exist. Next to the mentioned two, the third factor whose importance I wish to emphasize is the ludic/carnivalesque component of our civilization.[20] Although machines and other appliances serving the purpose of bolstering human thinking capacity would initially come into existence in response to mankind's most urgent problems and needs (related, among others, to the hopes for a significant acceleration – and simultaneous increase in the reliability – of complex calculations in such areas as the art of war, mathematics and natural sciences, engineering, or outer space exploration), it did not take very long for us to discover that electronic devices and systems could also be used for entertainment. Paradoxically, at present, this area of AI deployment is treated as seriously as are its scientific or industrial uses: suffice it to mention the popularity of chatbots, the dynamically developing game-dev industry, or the emergence of automatic household management systems, such as Siri or Alexa, whose basic "job descriptions" include ludic, entertainment-related, tasks. Beyond doubt, this fact significantly impacts the structure of the space of human-AI relations, giving rise to yet another set of questions of ethical nature.

There is also the fourth factor – one that should not be treated lightly, even though, outwardly, it might seem trivial. It is the reinless human curiosity: the constitutive trait of the European technoscientific civilization. It is especially this peculiar quality of human nature that has ceaselessly been tantalizing us

19. See: Marek Jan Kasperski, *Sztuczna inteligencja. Droga do myślących maszyn* [Artificial Intelligence. A Path Towards Thinking Machines] (Gliwice: Helion, 2003), chap. 1.2. "Pionierskie pomysły na temat maszyn myślących" [Some Pioneering Ideas on Thinking Machines], 32–40.

20. In an introduction to his *Homo Ludens*, Johan Huizinga makes the following observation: "There is a [...] function, however, applicable to both human and animal life, and just as important as reasoning and making – namely, playing. It seems to me that next to Homo Faber, and perhaps on the same level as Homo Sapiens, Homo Ludens, Man the Player, deserves a place in our nomenclature. […] For many years the conviction has grown upon me that civilization arises and unfolds in and as play." In chapter XII, "Play Element in Contemporary Civilisation," he adds: "Certain activities whose whole raison d'être lies in the field of material interest, and which had nothing of play about them in their initial stages, develop what we can only call play-forms as a secondary characteristic. […] The impetus given to this agonistic principle which seems to be carrying the world back in the direction of play derives, in the main, from external factors independent of culture proper-in a word, communications, which have made intercourse of every sort so extraordinarily easy for mankind as a whole. Technology, publicity and propaganda everywhere promote the competitive spirit and afford means of satisfying it on an unprecedented scale." Johann Huizinga, *Homo Ludens. A Study of the Play-Element in Culture* (Routledge & Kegan Paul: London, Boston and Henley, 1938), ix and 199–200.

with the question "what happens when you push the button?"[21] – and we rarely resist the temptation to seek an answer to that question, even if the urge to satisfy our curiosity should involve serious risks.

And yet, admittedly, curiosity serves a most useful social function: it is owing to curiosity that the civilization breaks subsequent barriers in the course of its progress. The above notwithstanding, occasionally, it is also this trait of our nature that puts us in (more or less serious) peril. The greater our agency is, the more dangerous and the more acute the possible negative effects of our excessive, untamed curiosity, which urges us to act first and to try to understand later. In his essay, dramatically titled "Is Technological Civilization Decadent, and Why?" the Czech philosopher Jan Patočka gives humankind a warning: "Humans no longer understand what it is they do […]. In their relation to nature, they are content with mere practical mastery and predictability without intelligibility." As a result, the philosopher continues, "humans have ceased to be a relation to Being and have become a force, a mighty one, one of the mightiest […] they became […] a grand energy accumulator in a world of sheer forces, on the one hand making use of those forces to exist and multiply yet, on the other hand, themselves integrated into the same process."[22]

There is much evidence to suggest that in the course of constantly accelerating work on more and more advanced AI systems we have arrived at the point in our development when we are able to create something without fully comprehending what that which we have just called into existence really is. We are now capable of efficiently initiating processes of deep reinforcement learning, but, due to the very nature of these processes, we are unable to precisely predict the end results of some of them. One of the reasons for it is that the pace and variability of such processes makes it impossible to trace their courses and, by that token, to thoroughly comprehend the system modifications occurring as the process unfolds.[23] It is evident that the more complex tasks the constructors continue

21. See the CNN-produced video compilation titled "What happens when you push the button?", https://www.youtube.com/watch?v=UjwLcmqZTKU (28.02.2020). Also, see Ian Stewart's, Terry Pratchett's and Jack Cohen's description of scientific thought experiments, which allows the authors to formulate the conclusion that "[s]ome questions should not be asked. However, someone always does," which opens chapter one of their book *The Science of Discworld* (Ebury Press: London 2013), 15.

22. Jan Patočka, *Heretical Essays in the Philosophy of History*, trans. Erazim Kohak (Chicago and La Salle, IL: Open Court Publishing Company, 1996), 115–116.

23. "240 minutes. This is how long AlphaZero – Google's artificial intelligence – took to learn to play chess via self-play. The people who supervised the entire experiment limited their input to "teaching" AlphaZero the rules of the game. Its strategies, however, the AI developed on its own, making use of the possibilities of machine learning algorithms. [After 240 minutes] the DeepMind team resolved to check how their AI would score against Stockfish, the world's highest ranked

to set before their self-perfection-ready creations, the more profound the abyss will be between our capability of understanding and predicting their actions, and the actual dynamics of the machine self-evolution.

At the same time, we realize that the autonomization of the processes in whose course inner structures, algorithms, and principles on which AI acts, undergo transformations, is irreversible. Once the wheels of this mechanism are put in motion, it can neither be stopped, nor subjected to the rigors of strict control. As a result, we end up in a situation, in which we must face the necessity of resolving problems arising as a consequence of the progressing autonomization of the AI. One of the issues, of which we are now becoming more and more aware, is the problem of how to guarantee the humankind its essential safety and, to the extent to which it is possible, how to grant the human race active support when AI carries out tasks serving a wide range of anthropogenic goals and human interests. At the stage of the evolution of intelligent technologies at which we have already arrived, it would be naïve to think that such goals are achievable by means of legal regulations, or appealing to the designers' or programmers' sense of responsibility. The problem must be tackled at its very source. To do this, we need to rethink the very strategy of designing *reinforcement learning processes* in such a way as to ensure that, at every stage of the process, those of AI's choices, decisions, and actions, which favor humans and their vital interests, receive a positive reinforcement. It is definitely not enough to design "failsafe" systems based on in-built static safety mechanisms of the kind suggested by Isaac Asimov in his "four laws of robotics,"[24] because it is clear that such an idea is utterly inadequate in terms of its compliance with the principles of modern AI's operation and decision-making both today and in the future.

---

chess engine. […] Stockfish failed to win even a single game." Tomasz Domański, "Sztuczna inteligencja pokonała arcymistrzowski program szachowy. Ludziom zostało kibicowanie" [Artificial Intelligence Has Defeated a Grandmaster Chess Program. All People Have Left to Do Is Cheer], *Spider's Web*, December 8, 2017, https://www.spidersweb.pl/2017/12/sztuczna-inteligencja-szachy.html (20.07.2019).

24. The so-called Three Laws of Robotics that Isaac Asimov proposes in his 1942 story "Runaround," state as follows: "First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm. Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law." Isaac Asimov, "Runaround," in: *I, Robot* (The Isaac Asimov Collection ed.). (Doubleday: New York, 1950), 40. Later, Asimov appended the existing three laws with one more, superordinate law, the so-called Zeroth Law: "A robot may not harm humanity, or, by inaction, allow humanity to come to harm." See: "Asimov's Three Laws of Robotics + the Zeroth Law," in: *Jeremy Norman's History of Information* 3 (1942) CE, http://www.historyofinformation.com/detail.php?entryid=4108 (29.02.2020).

If we were dealing with human beings, or beings resembling humans to a significant degree, then the application of the recognized methods of axiological and ethical education (well-known and practiced for thousands of years) would be a promising solution to the problem. To reframe this concept into very simple terms, the methods in use in human societies, depending on the type of culture they represent, consist in the fact that individuals, in the course of socialization processes, are subjected to a variety of systems of punishments and rewards, which, respectively, serve to reinforce or to inhibit specific types of behavior. As he or she matures, acquiring a larger and larger range of social competences, the person subjected to such processes learns to subjectively categorize behaviors, classifying them as either good (morally right) or bad (reprehensible), depending on the repetitive reaction patterns unique to his or her living environment. While experiences related to social acts of rewarding individuals for particular actions or behaviors delineate the scope of the class of morally acceptable behaviors, the acts of punishment generate knowledge about what actions and behaviors should be avoided, lest one should risk facing the consequences of their negative moral qualification.

It is in accordance with the procedures of defining and classifying objects (and in line with the principles of logical inference) developed in the European cultural tradition, that each of its individual participant, more or less adequately, constructs his or her own, progressively abstract and generalized, system of moral judgments. The subsequent stages of this procedure may be sketched out as follows: the first stage consists in one's increasingly conscious participation in a random sequence of individual life events, to which the subject (guided by the principles derived from external ethical assessments passed by members of his or her social environment, and especially by the "persons of a primary relationship"[25]) assigns a specific moral qualification. Such a set of experimentally collected cases is supplemented by knowledge acquired through cultural discourse (conversations with elders, stories with a didactic subtext, fairy tales with moral, literary texts, films, etc.).

The next stage is the stage in which the subject makes an attempt to resolve what it is that connects events that received positive feedback, and what the events that met with disapproval have in common. Adopting the qualification criteria that he or she has assimilated from his or her cultural tradition as a frame of reference, the subject uses the thus gathered material to create a mental map of "moral topography," charting areas characterized by the varying intensity of moral good and evil. Since individual differences in the field of moral experience occur, such

25. See: Martin Miller, *The True 'Drama of the Gifted Child'*, trans. Barbara Rogers and Rebecca Peterson. Published April 13th 2018. ISBN: 1980668949, Kindle Edition, loc. 89.

idiosyncratic mental maps may differ in a number of details. It should, however, be remembered that, basically, the "cartographic" process – the process of building and modifying one's map of moral topography – lasts a lifetime, and that the more extensive the comparative material is, the more similar the individual models become. Moreover, if subjects are embedded in essentially the same (or only slightly different) cultural frameworks within one socio-cultural formation, the range of similarities between individual versions of the "moral topography" is wide enough to ensure the group's successful functioning within a common moral order, which allows a relatively small margin for fluctuations.

A similar process leading to the gradual acquisition of "ethical intelligence"[26] could be imagined in the case of AI. Such a process could be conceived of as analogous to that in which the machine masters the strategy of the game of chess by gradually learning to opt for better and better solutions, drawing on an ongoing analysis of a sufficiently large number of moves and possible positions of pawns and figures on the chessboard. Likewise, in the process of "ethical reinforcement learning," having become familiar with a sufficiently large set of morally-charged cases, including, in each case, the data encompassing exemplary (partial) information about the correlation between a given action and its moral classification, the machine – owing to its in-built system warranting the durability of the disposition to reinforce behaviors/choices qualified as positive and to inhibit those qualified as reprehensible – may be expected to internalize decision-making procedures that would allow it to select behaviors determined as optimal from the point of view of their moral qualification.

The above notwithstanding, the problems, which we are not able to resolve by reference to theory, concern the methods which the self-learning system will adopt to generalize knowledge derived from specific cases, and directions of such generalizations. The issue at stake is thus how AI will expand its competences in the area of the moral norms and how it will apply them to resolve dilemmas related to new cases, thus far unencountered in the learning process. Let us consider: to its creators' surprise, in an astoundingly short time, the AlphaGo system proved perfectly capable of working out brand new game strategies, thus far unknown to any human player.[27] If this is the case, then, assuming that

26. See: Bruce Weinstein, *Ethical Intelligence. Five Principles for Untangling Your Toughest Problems* (Novato, CA: New World Library, 2011).

27. "Why does AlphaGo Zero mark a breakthrough? […] AlphaGo Zero learned the game by 'self-play.' It started off with random moves, gradually perfecting its skills on the basis of the potential of its self-learning neural network. [...] 40 days after the learning process commenced, the software outperformed all previous versions of AlphaGo in terms of the superiority of its skills (or more precisely: attained game results), thus becoming the strongest engine of its kind – both in comparison with other virtual game systems, and judging by the effects of its encounters

an analogous solution should indeed be implemented in AI, would it not be justified to expect that the "conscience of the machine" should develop in an equally surprising fashion?

For the sake of further considerations, let us adopt the position already seriously discussed in the literature of the subject[28]: let us assume that in the not-so-distant a future, biocomputers capable of self-controlled growth, self-organization, and self-replication will come into existence. Rudimentary machine learning software will have been incorporated in their basic operating systems, which the biocomputers, in the course of their development, will modify and adapt to fit their own parameters. Along with other elements of initiating software, such biomachines will certainly have been equipped with moral and axiological education programs, conceived of as the base for the inception of the development of autonomous ethical intelligence procedures. The design of such a module would have to take into account all of our programming experience gathered to date, both positive and negative, including cases such as that of the unfortunate experiment with a Microsoft chatbot named Taylor, who had mastered the principles of Internet hate within a few hours of his initial connection to the web, and soon started to generate texts of racist, sexist, and xenophobic nature, utterances offending major political leaders, etc.[29] Needless to say, such undesirable behaviors must be efficiently eliminated from the future repertoire of AI actions.

It is equally natural to assume that engineers and programmers will wish to equip the hypothetical "post-machine" (which working term, as I believe, could

---

with eminent Go masters. In the case of AlphaGo Zero, the system's self-evolution was based on the process of reinforcement learning. By principle, initially, the system (neural network) is not familiar with the Go strategy. However, with every self-played game the network changes, learning to predict subsequent moves, and finally begins to win [...]. The engine's neural network is linked to a search algorithm, owing to which new versions of AlphaGo Zero come into existence with each iteration." Zbigniew Piątek, "Dlaczego AlphaGo Zero jest przełomem?" [Why Does AlphaGo Zero Mark a Breakthrough], *Industry 4.0. Portal Nowoczesnego Przemysłu*, November 9, 2017, https://przemysl-40.pl/index.php/2017/11/09/alphago-zero (20.07.2019).

28. "The idea of constructing – and using – biological computers has been widely discussed for the past 10 years. [I]n the age of nanotechnology and genetics, this concept is becoming more and more appealing every day. [...] The idea employs the notion of information exchange between cells. Owing to this, one day a computer capable of self-renewal may come into existence." Kasperski, *Sztuczna inteligencja*, 206–207.

29. "Tay was a chatbot set up by Microsoft on 23 March, a computer-generated personality to simulate the online ramblings of a teenage girl. It took just two tweets for an internet troll going by the name of Ryan Poole to get Tay to become antisemitic. In the 24 hours it took Microsoft to shut her down, Tay had abused President Obama, suggested Hitler was right, called feminism a disease and delivered a stream of online hate [...]." Paul Mason, "The racist hijacking of Microsoft's chatbot shows how the internet teems with hate," *The Guardian*, March 29, 2016, https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism (20.07.2019).

be adopted as adequate for the postulated representative of a new generation of autonomous biocomputers) with a program not only allowing it to autonomously learn the principles of humanist ethics, but also responsible for the machine's absolute compliance with its rules. Again, in simple terms, leaving all nuances aside, such principles state that good is that which benefits the human, positively affecting his or her life, and serving his or her goals, needs and interests. Evil, in turn, is that which harms or endangers the human, or leads to the depletion of any resources affecting human prosperity or well-being. We do have the right to expect that further, autonomous, development of the post-machine's system of moral competences will progress within the frames of the above fundamental principles. Yet, before we content ourselves with this statement, perhaps, as impartially as is only possible, we should ask ourselves one question: why should it?

From a rational point of view – one unclouded by our human speciesist megalomania – it seems very likely that the more perfect our future biocomputer will become, the more efficiently it will learn to take care of its own needs and interests. Logically, it is these needs and interests are likely to gain precedence in the biocomputer's decision-making process over those imposed from the outside. Especially when, once initiated, the process of autonomous evolution progresses along the path of a multi-generational, constantly modified, and continuously improved self-replication of subsequent generations of post-machines. Józef Bocheński, whose reasoning could not be easily dismissed, observes that

> in the light of what we know [of the whole of nature], none of the arguments put forth by humanists favoring the alleged essential superiority of the human being is convincing. If crocodiles could practice philosophy, in all probability they would postulate crocodilism; after all, it is most gratifying to consider oneself as the most sublime of all creature[30].

If Bocheński – who, as our evidence demonstrates, was a human – arrives at such a conclusion, it is all the more likely that non-human, artificial intelligence should draw an akin logical inference out of the available data.

This, however, will not come to pass until an important qualitative threshold, one separating a machine (in all forms known to us to date) from a post-machine, has finally been crossed. The barrier in question is the machine's ability to autonomously set the goals of its actions. Thus far, in all cases of the human-machine relations, it has been the human to determine ultimate goals, while the machine, at best, has autonomously sought out the optimal ways to reach them. One day, however, this situation will change radically, albeit not all of a sudden: the moment

30. Józef Bocheński, *Sto zabobonów. Krótki filozoficzny słownik zabobonów* [One Hundred Superstitions. A Brief Philosophical Dictionary of Superstitions] (Kraków: Philed, 1994), 54–55.

of transition, therefore, may be hard to notice immediately. The above notwithstanding, such a change is inevitable, because along with the increase of the autonomy of the machines and self-controlling systems, the acceleration of their reversible process of the gradual expansion of the range and scope of self-dependent decisions that AI takes will ultimately be impossible to contain. In all probability, however, such an expansion, although unavoidable, will progress in small steps.

For example, one can imagine a scenario for the development of a network of autonomous social communications whose simplest elements are automatically controlled vehicles. Such vehicles are mandatorily equipped with systems capable of making a wide range of autonomous decisions while the vehicle is in traffic: decisions, whose consequences could potentially put other road users in danger. However, the ongoing evolution of such systems will probably lead to the development of much more complex solutions automatically managing the logistics of road transportation. The scope of their decision-making capacity will be incomparably broader than that of the AIs in existence today. In systems of such a degree of complexity, the human will only be able to define goals at the highest level of generality, while the concretization of these goals will inevitably be delegated to the system itself, as a function of its autonomous competence. Likewise, when designing and programing robots intended for military purposes, it must be borne in mind that their usefulness and efficiency will scale up in direct proportion to the increase of their independence in defining particular goals at every stage of the process for which they are to be deployed, and to the expansion of their autonomy in instantaneous decision-making in the context of the rapidly changing circumstances conditioning the course of military operations. The same applies to other areas in which autonomous devices find application – for instance, in medicine, where the speed and accuracy of the decision-making process may significantly impact the patients' prognoses, or even save their lives.

Therefore, the assumption that the development of machine decision-making capacity will sooner or later lead to a point when AI overcomes the barrier of goal-setting competence on its way towards full autonomy of actions – including the autonomy in defining long-term goals – ought to be considered as highly probable. I am of the opinion that it is only when such a change transpires that the humankind will encounter "artificial intelligence" in the literal, and not, as has been the case thus far, figurative sense of the term. In such a context, the problem of potential, or immediately actual discrepancy between the goals of man and the goals of self-empowered, largely human-control-free, artificial intelligence, is likely to affect us on a daily basis. This scenario, albeit so far it is hypothetical, may become real in a not-so-distant a future. Its prospect, in turn, begs the question whether we, the humankind, are mentally, psychologically, and organizationally prepared to face such a situation. All our experiences

to date, and all of our endeavors thus far, have played out in a unique spiritual space: the space in which man wields exclusive power over the goal-setting and courses of actions designed to attain them, both on an individual, and on a global scale. It is also within these very confines that technology has been evolving since the onset of the modern era.

It was at the threshold of the Early Modern period that our civilization entered the path of the development of science and technology, and it is towards technological and scientific excellence that is has been dynamically – and, with every passing century, more and more consistently – developing over time. Even though it was not the only possible option – especially bearing in mind that until the twilight of the Middle Ages (and, partly, also later) other priorities dominated in the culture of the West, and that none of the non-European cultures have developed their own intracultural mechanisms allowing for the emergence of a viable model of a scientific/technological civilization – over the past centuries this particular tendency has proven to motivate the global scenario of progress. According to the findings from the analyses carried out by some philosophers in the course of the twentieth century, the distinctive trait of this civilizational project is the exploitation of the resources of the planet. Consisting in the extraction and processing of raw materials, such an exploitation is carried out in keeping with concepts and procedures developed gradually by means of the application of available scientific knowledge to practical problem solving.

Martin Heidegger claims that throughout the modern era and in contemporary times, it was not just the technological practice that was dominated by a particular reductionist practice: something as profound as the very metaphysical essence of technology, understood as a paradigmatic mode of man's relation to reality as a whole, was affected as well. Specifically, the sense and the raison d'être of everything in existence has been reduced to the status of the "material for labor."[31] This leads to a situation, in which the fundamental question that man asks is not "what is this?" (as it used to be in the era of classical metaphysics), but "how can this be made useful for the human?" Heidegger further states that the dominance of this instrumental relation to reality, permeated with the air of entitlement, creates a false impression that while encountering the world, man "everywhere and always encounters only himself."[32] One may understand this

31. "The essence of materialism does not consist in the assertion that everything is merely matter rather than in a metaphysical determination according to which all manner of be-ing appears as material for labor. […] The essence of materialism stays hidden in the essence of technology." Martin Heidegger, *Letter on "Humanism,"* trans. Miles Groth, 32, http://wagner.edu/psychology/files/2013/01/Heidegger-Letter-On-Humanism-Translation-GROTH.pdf (28.02.2020).

32. Martin Heidegger, "The Question Concerning Technology," in: *The Question Concerning Technology and Other Essays*, trans. William Lovitt (New York & London: Garland Publishing, Inc., 1977), 27.

statement as an observation that whatever man sees all around him are solely his own projects: either those already completed and embodied in the form of artefacts, or those "in the making," realizable on the basis of forces and objects perceived as "raw material" to be processed for the purpose of actions aimed at meeting particular human needs or demands. Such an attitude reinforces the sense of the absolute, monopolistic dominance of man and his projects over the whole of reality, including not only the natural resources of the planet, but also man-made objects, from which we have become accustomed to expect an unconditional submission to human will.

Jan Patočka seconds Heidegger's position, observing that "the birth of Europe in the present sense of the word" occurred when

> European expansion shifted from the form of Crusades to exploration […] in the grasp for the wealth of the world; simultaneously, the internal development of production, of technologies, of commercial and financial practices led to the rise of an entirely new kind of rationalism, the only one we know today: a rationalism that wants to master things and is mastered by them […] Within the framework of nature […], humans then strive for their freedom – understood Platonically as that over which they stand.[33]

Following Heidegger and Patočka in their reflections upon the anthropological aspect of technological development, we chance upon a promising clue related to the possibility of overcoming the reductionist attitude based on demand and entitlement. Both philosophers concur in emphasizing that man and his "kingdom of ends" cannot be reduced to his functions as the explorer of the resources of nature and producer of technological artefacts serving the purpose of satisfying his material needs, for he is more than "a gigantic transformer, releasing cosmic forces accumulated and bound over the eons."[34] Man is – or at least he or she should aspire to become—a being comprehending *Dasein*, and consciously shaping his or her – essentially human, and thereby, as Helmuth Plessner would have it, "eccentric"[35] – relations towards it. Although the quoted thinkers do not write about it *expressis verbis*, among those relations there are also those shaped by man's axiological awareness: man, after all, is also a being who "reads values" and thinks

33. Jan Patočka, *Heretical Essays*, 109–110, *passim*.

34. Patočka, *Heretical Essays*, 116.

35. The thesis about man's "eccentric positionality," expressed in the last chapter of Plessner's *Levels of the Organic Life and the Human*, summarizes the answer to the question of who man is as a living being among the many layers of organic life by reducing it into an original-sounding formula." "Wstęp" [Introduction], in: Helmuth Plessner, *Władza a natura ludzka. Esej o antropologii światopoglądu historycznego* [Power and Human Nature. Essay on an Anthropology of Historical Worldview] (Warszawa: Wydawnictwo Naukowe PWN, 1994), XVII.

in accordance with values.[36] Thinking in terms of preferences and conscious being-with-respect-to-values result render the moral dimension an indelible dimension of human existence. In the light of this statement, it may seem puzzling that in our thinking about AI, in our search for more and more perfect solutions, and in our conceptual work on the technical aspects of the issue, we pay disproportionately little attention to ethical intelligence, which refers to the axiological dimension of *Dasein*.[37] Meanwhile, the perspective of the upcoming confrontation with non-human intelligence – shaped in the process of a techno-evolution (whose course has been radically different than the process of the evolution of nature from which man with all his properties emerged) and embodied in material carriers that are positively different than our bodies – should prompt us to raise and rethink fundamental questions about the moral order of the world.

Looking back, while examining the pasts of cultures and civilizations, we will observe that the overall direction of historical transformations in the area of moral views and ethical systems manifests itself as clear. The point of departure for this process is the concept of closed (exclusive) morality, whose validity is limited only to members of one's own ethnic community (family, tribe or nation), and all the subsequent stages of the evolution of ethical systems gradually lead towards the universalization of norms and principles of coexistence. The most radical appeals to abandon tribal morality and extend the imperative of unconditional charity to encompass the whole community of humans were voiced in the New Testament – which fact in itself, unfortunately, has never translated into any systematic or consistent implementation of the biblical guidelines into social or intellectual practice.[38] However, even if we met the moral requirements of Christianity

36. See: Józef Tischner, *Myślenie według wartości* [Thinking According to Values] (Kraków: Znak, 1982), *passim*.

37. One of the few exceptions to this rule, although its author also addresses the subject of the ethical intelligence of the machine only indirectly, is the following article: Magdalena Zdun, "Aksjologiczne uwarunkowania innowacyjności" [Axiological Conditioning of Innovativity], *Opuscula Sociologica* nr 1 (15) (2016), https://doi.org/10.18276/os.2016.1-02.

38. "I am convinced that Christianity – the Gospel – is not so much behind us, as it is ahead of us," as Józef Tischner wrote in his 1999 book *Ksiądz na manowcach* [A Priest Astray] (Kraków: Znak, 1999), 13. And Wacław Hryniewicz adds: "'It is people of limited horizons who can imagine that Christianity came to its ultimate fruition, or was fully constituted, in the fourth century (as some would claim) or in the thirteenth century (according to others), or at some other point in the past. In fact, Christianity has only taken its first, timid, steps in human history. Many words of Christ are still incomprehensible to us... The history of Christianity has only just begun. All that has been done in the past, all that we call Christian history today, is merely the sum of attempts made to date; some were clumsy, and some proved to be failures in their execution.' This is what Aleksander Mień, an Orthodox priest, said in the evening of September 8, 1990, during a conference held at the House of Culture and Technology in Moscow. My reading of his words coincided with the moment when I heard about Józef Tischner's passing. […] It is striking how similar are

with full radicalism and steadfast consistency, we would still face an altogether new situation when confronted with an intelligent post-machine. Christ exhorts: "Love thy neighbor as thyself" (Matthew 22:39) and John reminds us that "this commandment have we from him, That he who loveth God love his brother also" (1 John 4:21). And even we can possibly imagine AI as our "neighbor," how likely is it that it should ever become our "brother"?

Thus far the situation has been unprecedented, and therefore it requires of us to go beyond the established patterns and proven standards of thinking so that we can look at the problem with a fresh eye. The challenge is no more and no less than to take another step towards extending the scope of the validity of moral norms in such a way as to see them apply not only to the human species, but also to the intelligent non-human beings likely to emerge in the future. If we wish to expect that these beings (acting on their own, in accordance with the rules and procedures shaped in the course of their self-evolution) should conform to our own, human, moral standards, we must allow ourselves to recognize the idea of some form of mutual recognition of the elementary rights and interests of each of the two parties involved. If we fail to take into account the adoption of a dialogical perspective in relations with the coming generations of intelligent post-machines, their standards of behavior and their criteria of ethical impact assessment and our standards and criteria may gradually diverge. The greater the degree of autonomy characterizing the adaptive processes in subsequent generations of self-replicating post-machines, the more incomprehensible and more unacceptable for us their decisions and their motives for action may turn out to be. It would be difficult to exclude the possibility that like it has been in the case of people, who, over the centuries, have developed numerous standards of anthropocentric ethics guarding human interests, also in the case of intelligent non-human beings the process of the internal evolution of mechanisms and procedures governing the functioning of AI will develop towards "machinocentrism," embracing priorities different from ours and not always easy to harmoniously reconcile with human expectations or needs.

Today, of course, thinking in terms of a dialogue with intelligent non-human beings, whose future existence is little more than a probable hypothesis, bears traits of pure speculation. At the same time, however, if we neglect this direction of thinking about the prospects of the techno-evolution of the future today, we may simply miss the point of no return. We may overlook the moment when the question of the undesirable moral effects of our interaction with the AI-enabled

---

the voices of these two brave Christians from two different Slavic countries are!" Wacław Hryniewicz, "Chrześcijaństwo przed nami" [Christianity Ahead of Us], *Tygodnik Powszechny*, nr 28, 2000, http://www.tygodnik.com.pl/ludzie/tischner/hryniewicz.html#top (29.02.2020).

entities – which manifestly enter the space of our quotidian existence with ever increasing intensity – becomes a pressing practical problem, whose urgency will affect us daily. Then – it may prove to be too late for us to take any comprehensive corrective action to remedy the situation at hand.

Translated by Paweł Jędrzejko

 https://orcid.org/0000-0002-3251-2540

Bibliography

Asimov, Isaac. "Runaround." In: *I, Robot* [The Isaac Asimov Collection ed.]. New York: Doubleday, 1950.

Bocheński, Józef. *Sto zabobonów. Krótki filozoficzny słownik zabobonów*. Kraków: Philed, 1994.

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

Cohen, Jack, Terry Pratchett, and Ian Stewart. *The Science of Discworld*. London: Ebury Press, 1999.

Fry, Hannah. *'Hello World. How to Be Human in the Age of the Machine*. New York–London: W. W. Norton & Co., 2018.

Heidegger, Martin. *The Question Concerning Technology*. In: *The Question Concerning Technology and Other Essays*. Translated by William Lovitt. New York & London: Garland Publishing, Inc., 1977.

Huizinga, Johann. *Homo Ludens. A Study of the Play-Element in Culture*. London, Boston and Henley: Routledge & Kegan Paul, 1938.

Kasperski, Marek Jan. *Sztuczna inteligencja. Droga do myślących maszyn*. Gliwice: Helion, 2003.

Miller, Martin. *The True 'Drama of the Gifted Child'*. Translated by Barbara Rogers and Rebecca Peterson. Published April 13th 2018, ISBN: 1980668949. Kindle Edition 2018.

Paczkowska-Łagowska, Elżbieta. "Wstęp." In: Helmuth Plessner, *Władza a natura ludzka. Esej o antropologii światopoglądu historycznego*. Translated by Elżbieta Paczkowska-Łagowska. Warszawa: Wydawnictwo Naukowe PWN, 1994.

Patočka, Jan. *Heretical Essays in the Philosophy of History*. Translated by Erazim Kohak. Chicago and La Salle, IL: Open Court Publishing Company, 1996.

Plessner, Helmuth. *Levels of the Organic Life and the Human: An Introduction to Philosophical Anthropology*. New York: Fordham University Press, 2019.

Plessner, Helmuth. *Władza a natura ludzka. Esej o antropologii światopoglądu historycznego*. Translated by Elżbieta Paczkowska-Łagowska. Warszawa: Wydawnictwo Naukowe PWN, 1994.

Polak, Paweł, and Roman Krzanowski. "Ethics in Autonomous Robots as Philosophy in Silico: The Study Case of Phronetic Machine Ethics." *Logos i Ethos*, no. 52, 2020, 33–48. https://doi.org/10.15633/lie.3576.

Tischner, Józef. *Ksiądz na manowcach*. Kraków: Znak, 1999.

Weinstein, Bruce. *Ethical Intelligence. Five Principles for Untangling Your Toughest Problems*. Novato, CA: New World Library, 2011.

Zdun, Magdalena. "Aksjologiczne uwarunkowania innowacyjności." *Opuscula Sociologica*, nr 1, 15, 2016. https://doi.org/10.18276/os.2016.1-02.

Online sources

"Asimov's Three Laws of Robotics + the Zeroth Law." In: Jeremy Norman's History of Information 3/1942 CE. http://www.historyofinformation.com/detail.php?entryid=4108 (29.02.2020).

Cave, Stephen, Rune Nyrup, Karina Vold, and Adrian Weller. "Motivations and Risks of Machine Ethics." *Proceedings of the IEEE* 2019, vol. 107. no. 3 (March 2019), 562–574, https://doi.org/10.1093/pq/pqv034 (29.02.2020).

Domański, Tomasz. "Sztuczna inteligencja pokonała arcymistrzowski program szachowy. Ludziom zostało kibicowanie." *Spider's Web*, December 8, 2017. https://www.spidersweb.pl/2017/12/sztuczna-inteligencja-szachy.html (20.07.2019).

Heidegger, Martin. *Letter on "Humanism."* Translated by Miles Groth. http://wagner.edu/psychology/files/2013/01/Heidegger-Letter-On-Humanism-Translation-GROTH.pdf (28.02.2020).

Hryniewicz, Wacław. "Chrześcijaństwo przed nami." *Tygodnik Powszechny*, nr 28, 2000. http://www.tygodnik.com.pl/ludzie/tischner/hryniewicz.html#top (29.02.2020).

Mason, Paul. "The racist hijacking of Microsoft's chatbot shows how the internet teems with hate." *The Guardian*, March 29, 2016. https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism (20.07.2019).

Nadkarni, Isabel Teixeira. "Artificial intelligence: EU must ensure a fair and safe use for consumers." News European Parliament, January 23, 2020. https://www.europarl.europa.eu/news/en/press-room/20200120IPR70622/artificial-intelligence-eu-must-ensure-a-fair-and-safe-use-for-consumers (29.02.2020).

O'Hare Anna. The highlights from Web Summit 2019, November 7, 2019. https://websummit.com/blog/highlights-web-summit-2019 (19.02.2020).

Piątek, Zbigniew. "Dlaczego AlphaGo Zero jest przełomem?" *Industry 4.0. Portal Nowoczesnego Przemysłu*, November 9, 2017. https://przemysl-40.pl/index.php/2017/11/09/alphago-zero (20.07.2019).

"What happens when you push the button?" https://www.youtube.com/watch?v=UjwLcmqZTKU (28.02.2020).

White Paper On Artificial Intelligence – A European approach to excellence and trust. Brussels February 19, 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (29.02.2020).

Zagórna, Anna. "Komisja Europejska szykuje plan na SI." SI. Sztuczna Inteligencja. February 3, 2020. https://www.sztucznainteligencja.org.pl/komisja-europejska-szykuje-plan-na-si (19.02.2020).

Zegarow, Paweł. "Biała Księga w sprawie sztucznej inteligencji." *Nask. Cyber Policy*, February 27, 2020. https://cyberpolicy.nask.pl/biala-ksiega-w-sprawie-sztucznej-inteligencji (29.02.2020).