„Ethos" 36(2023) nr 4(144) 29-48
DOI 10.12887/36-2023-4-144-05

Zbigniew WRÓBLEWSKI
Paweł FORTUNA

# MORAL SUBJECTIVITY AND THE MORAL STATUS OF ARTIFICIAL INTELLIGENCE
## A Philosophical and Psychological Perspective

*The history of morality and ethical reflection shows that inclusion in the moral community has been a process of expanding the set of its members to include one's family, group, tribe, nation, race, and through non-human sentient beings (animals), to contemporary proposals for the inclusion of species, ecosystems, the biosphere, and artifacts. Importantly, some of the solutions proposed in the literature address the moral issues related to AI in the context of the next stage of the historical process of forming the moral community and ethical reflection on the mechanisms and rules of this process.*

In 2018, the Rector of the AGH University of Science and Technology gave a student's book to the humanoid robot Sophia during the Impact digital economy conference held in Krakow. This robot was developed by the Hong Kong-based company Hanson Robotics and was activated in 2016. In the Anthropomorphic Robot Database (ABOT) with 251 robots, Sophia ranks 8th for human likeness, 9th for surface appearance, and 31st for face appearance. The event was widely commented on in the media, and one post, published on the official TVPInfo website, read: "Android has dreams of having a family and friends, as well as striving to integrate humans with robots. In an interview with 'Khaleej Times Dubai,' the robot said that she wants to have a child, a daughter, and is seriously thinking of becoming a knowledge ambassador at the foundation of the Prime Minister of the United Arab Emirates".[1] In 2017, the fembot Sophia received the status of a citizen in Saudi Arabia,[2] and it was not a one-off event with regard to artificial systems. Less than a week later, Japan granted the resident status to a chatbot named Mirai[3]. The expansion of artificial intelligence (AI) driven artifacts continues. There is currently a lively discussion on the opportunities and threats of ChatGPT, an advanced language model that can be used for content creation, translation, learning support, idea generation, entertainment, acting as

[1] See Paweł F o r t u n a, *Optimum: Idea cyberpsychologii pozytywnej* (Warszawa: PWN, 2021), X.

[2] See Andrew G r i f f i n, "Saudi Arabia Grants Citizenship to a Robot for the First Time Ever," Independent UK, October 26, 2017, https://www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html.

[3] See Rosie M c C a l l, "Japan Has Just Granted Residency to an AI Bot in a World First," IFL-Science, November 7, 2017, http://www.iflscience.com/technology/japan-has-just-granted-residency-to-an-ai-bot-in-a-world-first.

a personal assistant or interpersonal relations trainer. One of the results of the public debate were calls from various scientific and technological communities to suspend work on artificial intelligence for some time.[4]

The digital revolution is densifying the environment with AI-based artifacts, whose image presented in pop culture narratives corresponds to the features of the so-called Artificial General Intelligence (AGI) or Strong AI.[5] These are, considered so far in hypothetical form, systems that have a general intelligence similar to that of an adult human being, not only functioning as if they had a mind, but as if they would actually have one. As a consequence, they would be capable of sensory perception, characterised by awareness, self-awareness and morality. Compared to them, "weak" AI (Artificial Narrow Intelligence) are systems capable of solving problems at the level of human beings or better than them, as if they had a mind and were thinking. The aim of the pursuit of AGI is to create a learning artificial intelligence, not limited to solving problems and concentrating on performing pre-programmed tasks, but capable of developing features corresponding to human intelligence. For this reason, the issue of the human-like status of AGIs[6] and their treatment as moral subjects is lively debated.[7]

AGI is merely a set of assumptions about possible forms of AI. The moment of its emergence cannot be determined,[8] but, as presented at the beginning, artificial systems that are examples of "weak" AI are assigned the attributes of AGI. The tendency to anthropomorphise these artifacts is stimulated by the achievements of both designers and marketing specialists. For example, the design of social robots is aimed at maximising the positive affect of the people who interact with them.[9] This is to be enabled by two paths of innovation: developmental cybernetics (developing human-like entities by simulating human psychological processes and kinesthetic functions) and developmental robotics (development of neural networks that would allow artificial entities to autono-

[4] See Mateusz N o w a k, „Elon Musk i założyciel Apple apelują o wstrzymanie prac nad AI. 'Utrata kontroli nad cywilizacją,'" https://android.com.pl/tech/581815-apel-o-wstrzymanie-prac-nad-ai/.

[5] See Cassio P e n n a c h i n and Ben G o e r t z e l, "Contemporary Approaches to Artificial General Intelligence," in *Artificial General Intelligen*ce. *Cognitive Technologies*, ed. by Cassio Pennachin, Ben Goertzel (Berlin, Heidelberg: Springer, 2007), 1–30; John R. S e a r l e, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–24.

[6] See Kamil M u z y k a, "The Basic Rules for Coexistence: The Possible Applicability of Metalaw for Human-AGI Relations," *Paladyn, Journal of Behavioral Robotics* 11, no. 1 (2020): 104–17.

[7] See Mark C o e c k e l b e r g h, "Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents," *AI & Society* 24, no. 2 (2009): 181–89.

[8] See Max T e g m a r k, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Alfred A. Knopf, 2017).

[9] See Elyakim K i s l e v, *Relationships 5.0: How AI, VR, and Robots Will Reshape Our Emotional Lives* (Oxford: Oxford University Press, 2022).

mously acquire sensorimotor and mental abilities of increasing complexity).[10] Artificial systems are given human names (e.g., chatbot Alexa, supercomputer IBM Watson, Ernest—UK Facebook messenger bank aggregator), appearance, and are also assigned self-awareness[11] and awareness (e.g., LaMDA, Google's artificially intelligent chatbot generator).[12] Assigning artificial systems almost equal status to human beings is also facilitated by the increasingly stronger voice of supporters of post-humanism, who question the dualism between nature/culture, human being/animal and human being/machine and argue for the empowerment of non-human forms of life.[13] For example, the post-humanist vision of business assumes an equal cooperation of natural human beings, cyborgs (neuroprosthetically augmented human employees), computers (artificial intelligence driven software) and bioroids (humanoid robots).[14]

The trends outlined above raise questions: (1) of a philosophical nature—Can potential AGI objects become moral subjects?; What criteria for assigning the moral status (MS) apply to AI-driven objects? and (2) of a psychological nature—At the level of belief, do AI-based artificial systems have an open path to inclusion in the moral community?; What factors determine the consent to assign the MS to AI-based systems?

When seeking answers to such questions in this paper, we first outline the broader context that is the reception of AI and the anthropomorphisation of artifacts based on it, then we consider the issue of moral subjectivity, and finally we present the results of psychological studies on assignment of the MS to artificial systems. We treat the discussion of the outlined problems as an element of preparing a public debate on the moral aspects of the rapid development of AI, which systematically expands the scope and possibilities of simulating mental processes. The assignment of appropriate MS to AI has practical consequences: AI objects can be autonomous moral subjects, and so can have moral responsibility,[15] can

[10] See Antonella M a r c h e t t i et al., "Theory of Mind and Humanoid Robots from a Lifespan Perspective," *Zeitschrift für Psychologie* 226, no. 2 (2018): 98–109.

[11] See Selmer B r i n g s j o r d, Paul B e l l o, and Naveen Sundar G o v i n d a r a j u l u, "Toward Axiomatizing Consciousness," in *The Bloomsbury Companion to the Philosophy of Consciousness*, ed. Dale Jacquette (London: Bloomsbury Academic, 2018), 289–324.

[12] See Nitasha T i k u, "The Google Engineer Who Thinks the Company's AI Has Come to Life," *The Washington Post*, June 11, 2022, https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine.

[13] See Rosi B r a i d o t t i, *The Posthuman* (Cambridge: Polity Press Ltd, 2013); Joshua C. G e l l e r s, *Rights for Robots: Artificial Intelligence, Animal and Environmental Law* (New York: Routledge, 2021).

[14] See Matthew E. G l a d d e n, *Posthuman Management* (Indianapolis: Synthypnion Press, 2016).

[15] See Aimee van W y n s b e r g h e and Scott R o b b i n s, "Critiquing the Reasons for Making Artificial Moral Agents," *Science and Engineering Ethics* 25, no. 3 (2019): 719–35; Mariarosaria

be subjects of rights, and thus other moral subjects have obligations in relation to them,[16] can be moral agents that make moral decisions,[17] and can be members of human communities.[18] The functioning of AI objects in the social environment creates a new situation for common morality. Regardless of the opinion of programmers (whether or not it is a conscious machine) and philosophers (whether or not to assign the MS to new objects), common sense judgments are already being spontaneously formulated influencing the way we treat artificial entities and the hybrid systems we create with them.

## AI AS A POTENTIAL MORAL SUBJECT

When considering the moral subjectivity and the MS of AI, it is necessary to clarify the very concept of AI, which has been a challenging task since its introduction,[19] and some researchers believe that this is an unrealistic goal at the current stage of research.[20] At a general level, there is consensus that AI is the attempt "to make a computer work like a human mind".[21] According to Lindes, the concept "artificial intelligence" should be used in two main senses, which the researcher labels as AI1 and AI2. According to him, AI1 refers to the quality of intelligence in the man-made computing systems, which can be compared and contrasted with natural intelligence.[22] AI2, on the other hand, is a field of study that deals with the design, construction and evaluation of AI1 systems, i.e. artificial systems that manifest intelligence. Because the definition

---

T a d d e o and Luciano F l o r i d i, "How AI Can Be a Force for Good," *Science* 361, no. 6404 (2018): 751n.

[16] See David J. G u n k e l, "The Other Question: Can and Should Robots Have Rights?," *Ethics and Information Technology* 20 (2018): 87–99; Jacob T u r n e r, *Rights for AI*, in Turner, *Robot Rules* (Cham: Palgrave Macmillan, 2019), 133–71.

[17] See John D a n a h e r, *Automation and Utopia: Human Flourishing in a World without Work* (Harvard University Press: Cambridge, MA, 2019); Colin A l l e n, Iva S m i t, and Wendell W a l l a c h, "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches," *Ethics and Information Technology* 7 (2005): 149–55; James H. M o o r, "The Nature, Importance, and Difficulty of Machine Ethics," *IEEE Intelligent Systems* 21, no. 4 (2006): 18–21.

[18] See Migle L a u k y t e, "Artificial Agents among Us: Should We Recognize Them as Agents Proper?," *Ethics and Information Technology* 19, no. 1 (2017): 1–17; Brian D u f f y, "Anthropomorphism and the Social Robot," *Robotic and Autonomous Systems* 42, nos. 3-4 (2003): 177–90.

[19] See Nils N i l s s o n, *The Quest for Artificial Intelligence* (Cambridge: Cambridge University Press, 2009).

[20] See Pei W a n g, "On Defining Artificial Intelligence," *Journal of Artificial General Intelligence* 10, no. 2 (2019): 1–37.

[21] Pei W a n g, Kai L i u, and Quinn D o u g h e r t y, "Conceptions of Artificial Intelligence and Singularity," *Information* 9, no. 4 (2018): 79.

[22] See Peter L i n d e s, "Intelligence and Agency," *Journal of Artificial General Intelligence* 11, no. 2 (2020): 47–49.

of AI2 depends on how we understand AI1, which in turn depends on how we understand intelligence itself, defining AI depends on precisely defining intelligence as such. This is problematic because only in the field of psychology this term is considered controversial and there is no consensus on a single definition (as many as 28 new ones have been proposed in the previous decade).[23]

Despite the difficulties in defining AI, it is easier to identify AI examples, referred to as rational agents—systems that receive percepts from the environment and perform actions.[24] The agents act as "intelligent tools," and many of them operate under marketing names (e.g., virtual assistants: Amazon's Alexa, Apple's Siri). They are driven by various types of algorithms (e.g., search, machine learning, evolutionary, artificial neural networks), and when combined with a physical body they become examples of "embodied" AI (e.g., self-driving cars, robots). Human beings, when interacting with AI-based systems, intentionally or unknowingly create hybrid systems.[25] The degree of fusion with artificial entities can be described on a continuum of cyborgisation: from interaction with static (PC), mobile (smartphone) and wearable technologies (smart-glasses), to augmentation (fusing artifacts with the human nervous system).[26] This fusion can be explicit, as in the case of human-cobot systems in the production process, but it can also be implicit to the user of the technology. An example of this are the algorithms that control the mathematical and statistical representation of each Internet user, which, according to Deleuze, can be described as "dividual."[27] This bank of data is created by the activity of the Internet user, but his "mind" is made up of algorithms beyond his control and suggesting customised content. Reacting to it makes the human being (individual) and the "dividual" function in a continuous feedback loop, providing data and reacting to it. They unknowingly meld together to form a kind of augmented mind,[28] which can be referred to in a working way as the "hybrid self."

---

[23] See Dagmar M o n e t t and Colin W.P. L e w i s, "Getting Clarity by Defining Artificial Intelligence: A Survey," in *Philosophy and Theory of Artificial Intelligence 2017*, ed. Vincent C. Müller (Berlin: Springer, 2018), 212–14.

[24] Stuart J. R u s s e l l and Peter N o r v i g, *Artificial Intelligence: A Modern Approach* (Boston: Pearson, 2020).

[25] See Wulf L o h and Janina L o h, "Autonomy and Responsibility in Hybrid Systems: The Example of Autonomous Cars," in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, eds. Patrick Lin, Keith Abney and Ryan Jenkins (New York: Oxford University Press, 2017), 35–50.

[26] Alex J u p i t e r, "The Human-Cyborg Continuum: Why AI Is Pointless and Why We Should All Become Cyborgs Instead," June 4, 2016, https://medium.com/@AlexJupiter/the-human-cyborg-continuum-why-ai-is-pointless-and-why-we-should-all-become-cyborgs-instead-4de0c4bb476f.

[27] See Gilles D e l e u z e, "Postscript on the Societies of Control," *October* 59 (1992): 3–7.

[28] See Andy C l a r k and David C h a l m e r s, "The Extended Mind," *Analysis* 58, no. 1(1998): 7–19.

Later in this paper, the results of a study on the assignment of the MS to AI-driven artifacts will be presented. Due to the fact that this process is informal, intuitive reasoning, it is worth looking at how AI is understood by the users themselves. The surveys conducted in seven countries (e.g., USA, Germany, China) show that public awareness of AI seems to depend on the visibility of its use.[29] It was found that 90% of respondents were aware that a voice assistant (visible AI) was based on AI, while only one in three respondents associated online shopping websites, video streaming services and social media (unvisible AI) with AI. The obtained results correspond to the results of studies, which captured differences in the understanding of AI by experts (IT specialists) and laymen.[30] As it turns out, for people with expert knowledge, AI is primarily "algorithmic systems" (e.g. image generation algorithm), while for laymen, it is mainly "nature imitating systems" (e.g. humanoid robot). When categorising AI examples, experts are mainly guided by functional features, while laymen also consider structural features of the systems. The functions of "algorithmic systems" are cognitive, related to performing the so-called objective tasks (e.g., pattern recognition), while "nature imitating systems" perform tasks that seem subjective in nature (based on emotions and intuition).[31]

The identification of AI with embodied, imitating entities found in nature should be attributed to contact with "AI narratives" present in pop culture, which include "portrayals of any machines (or hybrids, such as cyborgs) to which intelligence has been ascribed, which can include representations under terms such as robots, androids or automata."[32] Some narratives are non-fictional (e.g., TV news) and some are fictional (e.g., sci-fi films). In non-fiction AI narratives, attention is paid mainly to the examples of "weak" AI, while the heroes of fiction AI narratives are the examples of AGI. In the latter case, they not only talk and walk, but are capable of feeling human emotions, have elements of self-awareness and free will. In addition, they are characterised by exaggerated corporeality (e.g., T-800 in *Terminator*, 1984) and hypersexuality (e.g., Ava in *Ex Machina*, 2015), they have superhuman resistance to pain and indestructibility.[33]

---

[29] Jem D a v i e s, "AI Today, AI Tomorrow. The Arm 2020 Global AI Survey," armBlueprint, February 3, 2020, https://www.arm.com/resources/report/ai-today-ai-tomorrow-ty.

[30] See Paweł F o r t u n a and Oleg G o r ba n i u k, "What Is Behind the Buzzword for Experts and Laymen: Representation of 'Artificial Intelligence' in the IT-professionals' and Non-professionals' Minds," *Europe's Journal of Psychology*" 8, no. 2 (2022): 207–18.

[31] See Yoel I n b a r, Jeremy C o n e, and Thomas G i l o v i c h, "People's Intuitions about Intuitive Insight and Intuitive Choice," *Journal of Personality and Social Psychology* 99, no. 2 (2010): 232–47.

[32] Stephen C a v e, Kanta D i h a l, and Sarah D i l l o n, *AI Narratives: A History of Imaginative Thinking about Intelligent Machines* (Oxford: Oxford University Press, 2020), 5.

[33] D a v i e s, "AI Today, AI Tomorrow: The Arm 2020 Global AI Survey."

## MORAL STATUS VS MORAL SUBJECTIVITY

The discussion on the moral aspects of AI has been ongoing since it was launched in the 1950s as part of the so-called ethics of AI.[34] The issues raised there focused, among other things, on the threat to privacy, information technology surveillance, the use of knowledge about citizens, the use of information to manipulate people, freedom of citizens, and civil rights. The common denominator of the issues mentioned is the assessment of the (actual, potential) effects of the use of AI on moral subjects (human beings). However, what is morally assessed is the usefulness of the artifacts, not the artifacts themselves.[35]

In parallel, another type of moral reflection was being developed, which focused on the potential AGI objects.[36] If we hypothetically assume that the realisation of such attributes as reasoning, decision-making, representing knowledge, planning, learning, or communicating in a natural language makes it possible to achieve the level of artificial awareness analogous to the awareness of a human being, then the problem of the MS of these entities arises. It is no longer just a question of a moral assessment of the effects of using AI, but whether AI has a moral significance that is based in itself and not in its technical usefulness, and whether human beings have any obligations towards it. This raises the problem of defining new boundaries of morality, or more precisely of shifting the boundaries of the moral community to include new morally relevant beings, and thus the question: can the moral community be extended to include AI-driven entities?

The moral community includes beings towards whom moral subjects have certain obligations. The inclusion in it is made by recognising the MS of a member of the community on the basis of having a specific feature or set of features. The history of morality and ethical reflection shows that inclusion in the moral community has been a process of expanding the set of its members to include one's family, group, tribe, nation, race, and through non-human

---

[34] Norbert W i e n e r, *The Human Use of Human Beings: Cybernetics and Society* (Boston: Houghton Mifflin, 1950).

[35] See Nick B o s t r o m and Eliezer Y u d k o w s k y, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, eds. Keith Frankish, William Ramsey (Cambridge: Cambridge University Press, 2014), 316–34; Kenneth E. H i m m a, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?," *Ethics and Information Technology* 11, no. 1(2009): 19–29; Vincent C. M ü l l e r, "Is it Time for Robot Rights? Moral Status in Artificial Entities," *Ethics and Information Technology* 23 (2021): 579–87.

[36] Vincent C. M ü l l e r, "Ethics of Artificial Intelligence and Robotics," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/.

sentient beings (animals), to contemporary proposals for the inclusion of species, ecosystems, the biosphere, and artifacts. The continuous expansion of the moral community is referred to as the "tower of morality"[37] or "expanding circle."[38] Importantly, some of the solutions proposed in the literature address the moral issues related to AI in the context of the next stage of the historical process of forming the moral community and ethical reflection on the mechanisms and rules of this process.[39] The concept necessary to describe it is the concept of MS.

According to Warren, "the concept of moral status is, rather, a means of specifying those entities towards which we believe ourselves to have moral obligations, as well as something of what we take those obligations to be."[40] In turn, Kamm proposes the following definition of MS: "X has moral status = because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake."[41] The concept of MS has various functions. The concept of MS makes it possible to define the general obligations that moral subjects should fulfil in relation to beings of a given type, so it can be used to define the basic standards of acceptable behaviour towards them. The core features of the concept of MS are thus the generality (of obligations, rights, interests), and the fact that the MS is assigned to members of a specific group rather than to individuals (e.g., primates, not just Sarah the chimpanzee). The moral obligations arising from the assignment of the MS are the obligations towards that being, not someone else (e.g., towards the robot ASIMO, not its legal owner)[42]. The concept of MS can also justify moral ideals, e.g. the Christian ideal of loving one's neighbour or the Jainist ideal of not killing. Such ideals, creating space for supererogation, encourage moral development.

The MS is assigned to entities on the basis of meeting the relevant criteria, although their determination is a source of dispute. In the debate, the views of supporters of single- and multi-criteria theories of MS clash. The single-criteria theories postulate the identification of a single intrinsic characteristic of the

---

[37] Frans de W a a l, *Primates and Philosophers: How Morality Evolved* (Princeton: Princeton University Press, 2006).

[38] Peter S i n g e r, *The Expanding Circle: Ethics and Sociobiology* (Oxford: Clarendon Press, 1981); Steve T o r r a n c e, "Artificial Agents and the Expanding Ethical Circle," *AI and Society* 28, no. 4 (2013): 399–414.

[39] See Adam J. A n d r e o t t a, "The Hard Problem of AI Rights," *AI and Society* (2020): 1–14; H i m m a, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent have to Be a Moral Agent?," 19–29.

[40] Mary A. W a r r e n, *Moral Status: Obligations to Persons and Other Living Things* (Oxford: Clarendon Press, 1997), 9.

[41] Frances M. K a m m, *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (New York: Oxford University Press, 2007), 7.

[42] See W a r r e n, *Moral Status*, 10.

entity, the possession of which guarantees the MS and inclusion in the moral community, e.g., life,[43] capacity to feel[44] and subjectivity (being a person).[45] The latter of the listed features can be understood: (a) restrictively—the subject should possess certain cognitive capacities that enable one to reflect on moral issues, which makes it possible to be a moral subject, or (b) less restrictively: the entity should be a subject of life, possessing beliefs, desires, memory, the capacity to anticipate and to act intentionally.[46] The literature also indicates proposals for identifying the MS on the basis of external features of subjects (relational features: individual—community, environment), e.g., the moral status of a given being depends on the function (positive or negative) it performs within a biological or social community.[47] In another proposal, it is assumed that the MS of a given entity depends on the feelings we have towards it, e.g. our care for a given entity assigns it the MS.[48] Each of the listed features (internal or external) is treated by philosophers as a necessary and sufficient condition for having the MS.

According to multi-criteria theories of MS, it is assumed that (1) there is more than one valid criterion for the MS, (2) there is more than one type of the MS, and (3) the criterion for having the MS takes into account the internal and external features of a given entity.[49] The general principles arising from the assignment of the MS are interdependent, i.e., the practical consequences resulting from one principle are understood in the context of the other principles. The adoption of this approach is dictated by the fact that many moral problems are more complex in nature than it appears to the supporters of single-criteria theories. The common-sense diversity of moral intuitions on complex or radically new issues (humanoid robots as moral agents) seems to support this approach. The strategy of accepting a variety of criteria for the MS (and thus accepting many of its types) and ordering them in the form of a system (as suggested by the multi-criteria theory) seems to be more optimal

[43] Albert S c h w e i t z e r, *Civilization and Ethics* (London: Adam & Charles Black, 1955).

[44] Jeremy B e n t h a m, *An Introduction to the Principles of Morals and Legislation* (Oxford: Oxford University Press, 1998); Peter S i n g e r, *Animal Liberation: A New Ethics for Our Treatment of Animals* (New York: Harper Collins, 1975).

[45] Immanuel K a n t, *The Metaphysics of Morals* (Cambridge: Cambridge University Press, 2017).

[46] Tom R e g a n, *The Case for Animals Rights* (Berkeley, Los Angeles: University of California Press, 1983).

[47] Aldo L e o p o l d, *A Sand County Almanac* (Oxford: Oxford University Press, 1987); John B. C a l l i c o t t, *In Defense of the Land Ethic: Essays in Environmental Philosophy* (New York: State University of New York Press, 1989).

[48] Nel N o d d i n g s, *Caring: A Feminine Approach to Ethics and Moral Education* (Berkeley, Los Angeles: University of California Press, 2013).

[49] See W a r r e n, *Moral Status*, 21.

than the strategy of reducing a variety of criteria to a single, key one (the single-criteria theory)—the moral community is diverse in terms of the MS of its members (a heterogeneous, pluralistic community).

We recognise that it is better to use the multi-criteria theory for determining the MS of AI-driven artifacts. It allows to take into account the new nature of objects and their moral significance. While various uses of AI have become the subject of numerous ethical, social and psychological studies in the scheme—what is the impact of using AI in domain X—the problem of the MS of AI (selected objects) solved in the scheme—what moral significance AI objects have—is still an under-researched problem. Furthermore, the variety of types of AI-driven objects (algorithms, autonomous machines, expert systems, humanoid robots) makes it impossible to use a single criterion of the MS (e.g., an AGI and a chess-playing programme), and therefore a pluralistic strategy (multiple features and different degrees of the MS) is better than using a zero-one strategy (only one feature, e.g., the ability to feel pain, determines whether an object has or does not have the MS). For this reason, as suggested by Warren, we considered seven criteria of identification of MS that relate to potentially internal and external characteristics of HR: (1) being a living being (structured purposeful systems, showing the basic attributes of life); (2) being a sentient being; (3) being an individual with cognitive abilities that enable reflection on moral problems; (4) being a person (subject of life) who has beliefs, desires, memory, the ability to predict and act intentionally; (5) being a significant part of the environment; (6) being a member of an interspecies community, and (7) being recognized as a significant entity by another moral entity. Each of the above-mentioned features is related to one of the moral principles that define the normative consequences of assigning MS: (1) the principle of respect for life; (2) the principle against cruelty; (3) the principle of the rights of the subject; (4) the principle of human rights; (5) the environmental principle; (6) the interspecific principle; and (7) the principle of the transitivity of respect.[50]

## MIND PERCEPTION AND MORAL STATUS OF AI

Assigning AI-driven systems the moral subjectivity and the MS, like ascribing human attributes to them, is a manifestation of anthropomorphisation (Greek: *anthropos* for "human," *morphe* for "shape" or "form"[51]). It is an automatic process, built into perception of the surroundings, and the degree of

[50] See ibidem, 148–70.

[51] See Nicholas E p l e y, Adam W a y t z, and John T. C a c i o p p o, "On Seeing a Human: A Three-factor Theory of Anthropomorphism," *Psychological Review* 114, no. 4 (2007): 864–86.

anthropomorphizing can be determined on a continuum from the superficial and habitual use of personifying word labels to assigning them human dispositions, including emotions, thinking, and free will.[52] The triggering factor is the presence of typical human features in the encountered entity. Noticing them activates the knowledge about a human being stored in memory, and then integrates it with information about this person.

From the psychological point of view, the process of transmitting MS is coupled with the process of mind perception.[53] Research shows that our cognitive apparatus uses a two-dimensional filter in the process of mind perception, also referred to as "the cognitive template for morality."[54] In a classic study, the participants' task was to compare pairs of thirteen characters according to one of eighteen attributes or one of six personal judgments. The characters compared were humans (e.g., a five-month-old infant, adult woman, human in a vegetative state, test subject), animals (frog, domestic dog, wild chimpanzee), as well as a dead woman, God (defined as the creator of the universe and the ultimate source of knowledge, power, and love) and a robot Kismet.[55] The set of attributes included, among others, the feeling of pain, personality, awareness, morality, memory, and reflection. It has been revealed that a person evaluates other individuals on two dimensions: Experience (the ability to feel suffering) and Agency (the ability to take intentional actions). What is important, they were related to the classical distinction between individuals as moral patient and moral agent introduced by Aristotle.[56] Accordingly, the character's high assessment of the experience dimension (also referred to as the ability to feel suffering) indicates that we are dealing with the so-called a moral patient. On the other hand, a similar assessment on the dimension of agency indicates that the character is a moral agent.

The revealed dimensions of mind perception have been confirmed in many studies,[57] although there are analyses in which three dimensions have been

---

[52] See Marina P u z a k o v a, Hyokjin K w a k, and Joseph F. R o c e r e t o, "Pushing the Envelope of Brand and Personality: Antecedents and Moderators of Anthropomorphized Brands," *Advances in Consumer Research* 36 (2009): 413–20.

[53] See Heather M. G r a y, Kurt G r a y, and Daniel M. W e g n e r, "Dimensions of Mind Perception," *Science* 315, no. 5812 (2007): 619.

[54] See G r a y, Y o u n g, and W a y t z, "Mind Perception is the Essence of Morality," *Psychological Inquiry* 23, no. 2 (2012): 101–24.

[55] See G r a y, G r a y, and W e g n e r, "Dimensions of Mind Perception," 619.

[56] A r i s t o t l e, *Nicomachean Ethics*, trans. William. D. Ross (New York: World Library Classics, 2009).

[57] See Imge S a l t i k, Deniz E r d i l, and Burcu A. U r g e n, "Mind Perception and Social Robots: The Role of Agent Appearance and Action Types," in *HRI'21: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction: March 8-11, 2021, Boulder, CO, USA* (New York, NY: Association for Computing Machinery, [2021]), 210–214; Aleksandra

revealed, which argues for the heterogeneous nature of the Agency dimension.[58] With a two-dimensional mind perception matrix at our disposal, we can classify all the characters we encounter into one of four categories (examples taken from a classic study):

(1) low "Experience" and low "Agency"—e.g., a dead woman—she can neither be inflicted with suffering nor be expected to behave intentionally;

(2) low "Experience" and high "Agency"—e.g., God, robot—it cannot be inflicted with suffering but can be expected to behave intentionally;

(3) high "Experience" and low "Agency"—a frog—it can be inflicted with suffering but cannot be expected to behave intentionally;

(4) high "Experience" and high "Agency"—the subject itself—the subject can both be inflicted with suffering and be expected to behave intentionally.

Although participants of the referred studies tend to locate a social robot in the same group as God, however, as can be seen, the assessment of the intentionality of these characters is closer to the location of a dead woman, chimpanzee, and dog than adult humans and the subject himself. A similar low position of artificial systems on the Experience and Agency dimensions was reported by Lukaszewicz-Alcaraz and Fortuna.[59] This involved the social robot Pepper, the algorithm Aaron used for artistic realisations and the fembot Ai-Da advertised as an AI-driven artificial artist. However, it should be remembered that the position of these agents in the perceptual space of mind should not, however, be regarded as unchanging. It has been noted that humanoid robot's ratings on these dimensions may change depending on their appearance and response. For example, presenting robot with a human-like face places it higher on the Experience dimension than the same agent with exposed electronic components not covered with synthetic leather and higher scores on the Agency dimension when the robot performed an activity of a communicative nature

---

L u k a s z e w i c z and Paweł F o r t u n a, "Towards Turing Test 2.0—Attribution of Moral Status and Personhood to Human and Non-Human Agents," *Postdigital Science and Education* 4 (2022): 860–76; Paweł F o r t u n a, Arkadiusz G u t, and Zbigniew W r ó b l e w s k i, "Hey Robot, the Mind Is Not Enough to Join the Moral Community! The Effect of Assigning a Mind and a Soul to Humanoid Robot on Its Moral Status," *Annals of Psychology / Roczniki Psychologiczne* (2023), online first, https://doi.org/10.18290/rpsych2023.0008.

[58] See Kara W e i s m a n, Carol S. D w e c k, and Ellen M. M a r k m a n, "Rethinking People's Conceptions of Mental Life," *Proceedings of the National Academy of Sciences of the United States of America* 114, no. 43 (2017): 11374–79; Bertram F. M a l l e, "How Many Dimensions of Mind Perception Really Are There?," in *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, ed. Ashok Goel, Colleen Seifert, and Christian Freksa (Montreal, QB: Cognitive Science Society, 2019), 2268–74.

[59] See L u k a s z e w i c z and F o r t u n a, "Towards Turing Test 2.0—Attribution of Moral Status and Personhood to Human and Non-Human Agents," 860–76.

(gestures of waving a hand) or presenting the satisfaction of a biological need (placing a cup to the mouth imitating thirst quenching).[60]

The distinction between the moral patient and the moral agent is present in the literature on the MS of animals[61] and artificial agents.[62] Bostrom and Yudkowsky define the Experience as Sentience dimensions, in which Sentience is the capacity for phenomenal Experience or qualia, such as the capacity to feel pain and suffer. They also define Agency as Sapience (a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent) and state that these criteria are "commonly proposed as being importantly linked to moral status, either separately or in combination."[63] In the context of the discussion on MS of AI-driven characters, this distinction was adopted by Torrance, according to which the notion of "having ethical status" can be separated into two associated aspects: ethical receptivity and ethical productivity. Ethical recipients are those who stand to benefit from, or are harmed by, the ethical producers, and ethical producers are those who do or do not do their duties, such as saints and murderers.[64] From this perspective, AI and other smart machines can be both ethical producers and ethical recipients.[65]

The relationship between the mind perception and MS dimensions of an artificial AI-driven system has been empirically confirmed.[66] Study participants, who believed in the existence of the mind and the soul, assessed the MS of the humanoid robot Sophia and assigned attributes to it. The tool for assessing the MS was constructed on the basis of a concept distinguishing multiple criteria.[67] It was found that the attribution of the mind and the soul to the robot significantly affected the MS of the robot. Moreover, the dimensions of mind perception acted as a mediator, but only in the mind-MS relationship, while the soul-MS relationship was direct. It is clear from the studies presented that, for those who present a tripartite common anthropology (naive spiritualists), the assignment of the MS encounters the strong barrier of having to identify the mind and soul of an artificial system. This means that the "cognitive matrix of

[60] See Gray, Gray, and Wegner, "Dimensions of Mind Perception," 619.

[61] Regan, *The Case for Animals Rights*.

[62] See Bostrom and Yudkowsky, "The Ethics of Artificial Intelligence," 316–34.

[63] Ibidem, 322.

[64] See Steve Torrance and D. Roche, "Does an Artificial Agent Need to Be Conscious to Have Ethical?," in *Technologies on the stand: legal and ethical questions in neuroscience and robotics*, eds. Bibi van den Berg, Laura Klaming (Nijmegen: Wolf Legal Publishers, 2011), 285–310.

[65] See Torrance, "Artificial Agents and the Expanding Ethical Circle," 399–414.

[66] See Fortuna, Gut, and Wróblewski, "Hey Robot, the Mind Is Not Enough to Join the Moral Community! The Effect of Assigning a Mind and a Soul to Humanoid Robot on Its Moral Status."

[67] See Warren, *Moral Status*.

morality" enabling the robot to be treated as a moral patient and agent appears to be an insufficient criteria base for assigning the MS.

In order to better understand the psychological determinants of the assignment process of the MS, another study was carried out to investigate the influence of anthropocentric beliefs on the assignment of the MS to the humanoid robot Sophia.[68] At the same time, the mediating effect of ascribing mind and soul to this agent was examined. As in the previous study, the participants were naive spiritualists and, as before, they tended to make the assignment of the MS to an artificial entity dependent on the assignment of mind and soul to it. However, it was noted that such attributions depended on the strength of the respondents anthropocentric beliefs. It was found that the stronger the conviction about the superior status of human beings in relation to other beings, the lower the tendency to attribute mind and soul to the robot. Similar correlations were seen in subsequent studies whose participants responded to the possibility of attributing the MS to a chimpanzee and a cyborg character. Such analyses argue for the need to extend the criteria for assigning the MS to artificial systems (and other entities) and to go beyond the attributes associated with the mind (taken into account in studies on the mind perception). They also raise awareness of the inclusion of subjective factors, such as the type of common anthropology, in the discussion on the assignment of the MS to AI-based systems. It can be predicted that other criteria will be relevant to naive monists, dualists and spiritualists. Perhaps it is the case that those who postulate assigning the MS to artificial systems are physicalists or materialists, nesting the criteria of the MS in the physical substrate of algorithmic systems and the similar human operations they are capable of performing.

*

In the outlined panorama of philosophical and psychological reflections on the MS of AI-driven objects, it is clear that the key elements of the discussion are the set of criteria that make it possible to assign it. The debate on this topic is complicated by the sheer difficulty in defining AI, and thus in determining the attributes that can be assigned to artifacts driven by it. Intuitions coming from the world of science intertwine with pop culture narratives, creating an ambiguous and illusory picture. AGI is still a futuristic pipe dream, and the spectacular and media-publicised increase in humanoidisation of artificial

[68] See F o r t u n a, G u t, and W r ó b l e w s k i, "Hey Robot, the Mind Is Not Enough to Join the Moral Community! The Effect of Assigning a Mind and a Soul to Humanoid Robot on Its Moral Status."

agents still does not make them human beings. Despite the amazing skills presented by artificial systems (e.g., winning in Go, ChatGPT erudition), they do not yet reveal such qualities that in the perception of users could place them high on the dimensions of mind perception and, consequently enable them to be assigned the status of moral patient and moral agent. Not only that, but it appears that supporters of a tripartite common anthropology, i.e., those who base the architecture of human nature on the essence of body, mind and soul, extend the criterion base of the MS to include the spiritual element. It can be predicted that even if artifact constructors managed to imitate the qualities of the mind and soul, those with strong anthropocentric beliefs will still be cautious about similar claims, upholding the superior status of human beings over other beings. Thus, by examining the psychological determinants of the assignment of the MS to artificial systems, we are expanding our knowledge of ourselves, in line with Susan Schneider's prediction that "the age of AI will be a time of soul-searching—both of ours, and for theirs."[69] We are aware that the empirically confirmed findings presented in this paper will not extinguish the debate on the MS of AI-driven systems. However, we hope that they will facilitate the development of a position in a discussion that will become increasingly topical and heated with the subsequent innovations.

## BIBLIOGRAPHY / BIBLIOGRAFIA

Allen, Colin, Iva Smit, and Wendell Wallach. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches." *Ethics and Information Technology* 7 (2005): 149–55.

Andreotta, Adam J. "The hard problem of AI rights." *AI and Society* (2020): 1–14.

Aristotle. *Nicomachean Ethics*. Translated by William. D. Ross. New York: World Library Classics, 2009.

Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Oxford: Oxford University Press, 1998.

Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*. Edited by Keith Frankish and William Ramsey. Cambridge: Cambridge University Press, 2014.

Braidotti, Rosi. *The Posthuman*. Cambridge: Polity Press Ltd, 2013.

Bringsjord, Selmer, Paul Bello, and Naveen Sundar Govindarajulu. "Toward Axiomatizing Consciousness." In *The Bloomsbury Companion to the Philosophy of Consciousness*. Edited by Dale Jacquette. London: Bloomsbury Academic, 2018.

[69] See Susan S c h n e i d e r, *Artificial You: AI and the Future of Your Mind* (Princeton: Princeton University Press, 2019), 84.

Callicott, John B. *In Defense of the Land Ethic: Essays in Environmental Philosophy*. New York: State University of New York Press, 1989.

Cave, Stephen, Kanta Dihal, and Sarah Dillon. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines.* Oxford: Oxford University Press, 2020.

Clark, Andy, and David Chalmers. "The Extended Mind." *Analysis* 58, no. 1(1998): 7–19.

Coeckelbergh, Mark. "Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents." *AI & Society* 24, no. 2 (2009): 181–89.

Danaher, John. *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press: Cambridge, MA, 2019.

Davies, Jem. "AI Today, AI Tomorrow: The Arm 2020 Global AI Survey". armBlueprint, February 3, 2020. https://www.arm.com/resources/report/ai-today-ai-tomorrow-ty.

Deleuze, Gilles. "Postscript on the Societies of Control." *October* 59 (1990): 3–7.

Duffy, Brian. "Anthropomorphism and the Social Robot." *Robotic and Autonomous Systems* 42, nos. 3–4 (2003): 177–90.

Epley, Nicholas, Adam Waytz, aand John T. Cacioppo, "On Seeing a Human: A Three-factor Theory of Anthropomorphism," *Psychological Review* 114, no. 4 (2007): 864–86.

Fortuna, Paweł. *Optimum: Idea cyberpsychologii pozytywnej*. Warszawa: PWN, 2021.

Fortuna, Paweł, Arkadiusz Gut, and Zbigniew Wróblewski, "Hey Robot, the Mind Is Not Enough to Join the Moral Community! The Effect of Assigning a Mind and a Soul to Humanoid Robot on Its Moral Status," *Annals of Psychology / Roczniki Psychologiczne* (2023). Online first. https://doi.org/10.18290/rpsych2023.0008.

Fortuna, Paweł, and Oleg Gorbaniuk. "What is Behind the Buzzword for Experts and Laymen: Representation of 'Artificial Intelligence' in the IT-professionals' and Non-professionals' Minds," *Europe's Journal of Psychology* 8, no. 2 (2022): 207–18.

Gellers, Joshua C. *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. New York: Routledge, 2021.

Gladden, Matthew E. *Posthuman Management*. Indianapolis: Synthypnion Press, 2016.

Gray, Heather M., Kurt Gray, and Daniel M. Wegner. "Dimensions of Mind Perception." *Science* 315, no. 5812 (2007): 619.

Gray, Kurt, Liane Young, and Adam Waytz, "Mind Perception is the Essence of Morality." *Psychological Inquiry* 23, no. 2 (2012):101–24.

Griffin, Andrew. "Saudi Arabia Grants Citizenship to a Robot for the First Time Ever." Independent UK, October 26, 2017. https://www.independent.co.uk/tech/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html.

Gunkel, David J. "The Other Question: Can and Should Robots Have Rights?" *Ethics and Information Technology* 20 (2018): 87–99.

Himma, Kenneth E. "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?" *Ethics and Information Technology* 11 (2009): 19–29.

Inbar, Yoel, Jeremy Cone, and Thomas Gilovich. "People's Intuitions about Intuitive Insight and Intuitive Choice." *Journal of Personality and Social Psychology* 99, no. 2 (2010): 232–47.

Jupiter, Alex. "The Human-Cyborg Continuum: Why AI Is Pointless and Why We Should All Become Cyborgs Instead." Medium, June 4, 2016. https://medium.com/@AlexJupiter/the-human-cyborg-continuum-why-ai-is-pointless-and-why-we-should-all-become-cyborgs-instead-4de0c4bb476f.

Kamm, Frances M. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press, 2007.

Kant, Immanuel. *The Metaphysics of Morals*. Cambridge: Cambridge University Press, 2017.

Kislev, Elyakim. *Relationships 5.0: How AI, VR, and Robots Will Reshape Our Emotional Lives* Oxford: Oxford University Press, 2022.

Laukyte, Migle. "Artificial Agents among Us. Should We Recognize Them as Agents Proper?" *Ethics and Information Technology* 19, no. 1 (2017): 1–17.

Leopold, Aldo. *A Sand County Almanac.* Oxford: Oxford University Press, 1987.

Lindes, Peter. "Intelligence and Agency." *Journal of Artificial General Intelligence* 11, no. 2 (2020): 47–49.

Loh, Wulf, and Janina Loh. "Autonomy and Responsibility in Hybrid Systems: The Example of Autonomous Cars." In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Edited by Patrick Lin, Keith Abney, and Ryan Jenkins. New York: Oxford University Press, 2017.

Lukaszewicz, Aleksandra, and Paweł Fortuna. "Towards Turing Test 2.0—Attribution of Moral Status and Personhood to Human and Non-Human Agents." *Postdigital Science and Education* 4 (2022): 860–76.

Malle, Bertram F. "How Many Dimensions of Mind Perception Really Are There?" In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.* Edited by Ashok Goel, Colleen Seifert, and Christian Freksa. Montreal, QB: Cognitive Science Society, 2019.

Marchetti, Antonella, et al. "Theory of Mind and Humanoid Robots from a Lifespan Perspective." *Zeitschrift für Psychologie* 226, no. 2 (2018): 98–109.

McCall, Rosie. "Japan Has Just Granted Residency to an AI Bot in a World First." IFLScience, November 7, 2017. http://www.iflscience.com/technology/japan-has-just-granted-residency-to-an-ai-bot-in-a-world-first.

Monett, Dagmar, and Colin W.P. Lewis. "Getting Clarity by Defining Artificial Intelligence: A Survey." In *Philosophy and Theory of Artificial Intelligence 2017*. Edited by Vincent C. Müller. Berlin: Springer, 2018.

Moor, James H. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21, no. 4 (2006): 18–21.

Müller, Vincent C. "Is it Time for Robot Rights? Moral Status in Artificial Entities." *Ethics and Information Technology* 23 (2021): 579–87.

Muzyka, Kamil. "The Basic Rules for Coexistence: The Possible Applicability of Metalaw for Human-AGI Relations." *Paladyn, Journal of Behavioral Robotics* 11, no. 1 (2020): 104–17.

Nilsson, Nils. *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press, 2009.

Noddings, Nel. *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley, Los Angeles: University of California Press, 2013.

Nowak, Mateusz. "Elon Musk i założyciel Apple apelują o wstrzymanie prac nad AI: 'Utrata kontroli nad cywilizacją.'" March 30, 2023. https://android.com.pl/tech/581815-apel-o-wstrzymanie-prac-nad-ai/.

Pennachin, Cassio, and Ben Goertzel. "Contemporary Approaches to Artificial General Intelligence." In *Artificial General Intelligence: Cognitive Technologies*. Edited by Cassio Pennachin, Ben Goertzel. Berlin, Heidelberg: Springer, 2007.

Puzakova, Marina, Hyokjin Kwak, and Joseph F. Rocereto. "Pushing the Envelope of Brand and Personality: Antecedents and Moderators of Anthropomorphized Brands." *Advances in Consumer Research* 36 (2009): 413–20.

Regan, Tom. *The Case for Animals Rights*. Berkeley, Los Angeles: University of California Press, 1983.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Boston: Pearson, 2020.

Saltik, Imge, Deniz Erdil, and Burcu A. Urgen. "Mind Perception and Social Robots: The Role of Agent Appearance and Action Types." In *HRI'21: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction: March 8-11, 2021, Boulder, CO, USA*. New York, NY: Association for Computing Machinery, [2021].

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

Schweitzer, Albert. *Civilization and Ethics*. London: Adam & Charles Black, 1955.

Searle, John R. "Minds, Brains and Programs." *Behavioral and Brain Science* 3, no. 3 (1980): 417–24.

Singer, Peter. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: Harper Collins, 1975.

———. *The Expanding Circle: Ethics and Sociobiology*. Oxford: Clarendon Press, 1981.

Taddeo, Mariarosaria, and Luciano Floridi. "How AI Can Be a Force for Good." Science 361, no. 6404 (2018): 751–52.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred A. Knopf, 2017.

Tiku, Nitasha. "The Google Engineer Who Thinks the Company's AI Has Come to Life." *The Washington Post*, June 11, 2022. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine.

Torrance, Steve, and Denis Roche. "Does an Artificial Agent Need to Be Conscious to Have Ethical Status?" In *Technologies on the Stand: Legal and Ethical Questions in Neuroscience and Robotics*. Edited by Bibi van den Berg and Laura Klaming. Nijmegen: Wolf Legal Publishers, 2011.

Torrance, Steve. "Artificial Agents and the Expanding Ethical Circle." *AI and Society* 28, no. 4 (2013): 399–414.

Turner, Jacob. "Rights for AI." In Turner, *Robot Rules*. Cham: Palgrave Macmillan, 2019).

De Waal, Frans. *Primates and Philosophers: How Morality Evolved*. Princeton: Princeton University Press, 2006.

Wang, Pei. "On Defining Artificial Intelligence." *Journal of Artificial General Intelligence* 10, no. 2 (2019): 1–37.

Wang, Pei. Kai Liu, and Quinn Dougherty, "Conceptions of Artificial Intelligence and Singularity," *Information* 9, no. 4 (2018): 79.

Warren, Mary A. *Moral Status: Obligations to Persons and Other Living Things*. Oxford: Clarendon Press, 1997.

Weisman, Kara, Carol S. Dweck, and Ellen M. Markman, "Rethinking People's Conceptions of Mental Life." *Proceedings of the National Academy of Sciences of the United States of America* 114, no. 43 (2017):11374–79.

Wiener, Norbert. *The Human Use of Human Beings: Cybernetics and Society*. Boston: Houghton Mifflin, 1950.

Wynsberghe, Aimee van, and Scott Robbins. "Critiquing the Reasons for Making Artificial Moral Agents." *Science and Engineering Ethics* 25, no. 3 (2019): 719–35.

Zalta, Edward N., ed. *The Stanford Encyclopedia of Philosophy*, s.v. "Artificial Intelligence and Robotics" (by Vincent C. Müller). https://plato.stanford.edu/entries/ethics-ai/#AutoEmpl.

## ABSTRACT / ABSTRAKT

Zbigniew Wróblewski and Paweł Fortuna, Moral Subjectivity and the Moral Status of Artificial Intelligence: A Philosophical and Psychological Perspective
DOI 10.12887/36-2023-4-144-05

The paper is a voice in the discussion concerning the possibility of assigning moral status to technological artifacts driven by artificial intelligence. In the dimension of philosophical reflection, it refers to the debate on the understanding of the concept of „moral status" and the related possibility of extending the moral community to include artificial agents. In the psychological perspective, on the other hand, it presents the results of a study that examined the importance

of factors relevant to the assignment of moral status, such as the features of artifacts, dimensions of mind perception, soul assignment and anthropocentric beliefs. The considerations are set in the context of currently implemented innovations, pop culture narratives shaping the image of artificial systems and discussions on the possibility of the emergence of the so-called superhuman artificial intelligence.

Keywords: artificial intelligence, moral subject, moral status, mind perception, soul assignment

Contact: (Zbigniew Wróblewski) Katedra Filozofii Przyrody i Nauk Przyrodniczych, Instytut Filozofii, Wydział Filozofii, Katolicki Uniwersytet Lubelski Jana Pawła II, Al. Racławickie 14, 20-950 Lublin, Poland; (Paweł Fortuna) Katedra Psychologii Eksperymentalnej, Instytut Psychologii, Wydział Nauk Społecznych, Katolicki Uniwersytet Lubelski Jana Pawła II, Al. Racławickie 14, 20-950 Lublin, Poland
E-mail: (Zbigniew Wróblewski) zbigniew.wroblewski@kul.pl; (Paweł Fortuna) pawel.fortuna@kul.pl

Zbigniew Wróblewski, Paweł Fortuna – Podmiotowość moralna i status moralny sztucznej inteligencji. Perspektywa filozoficzno-psychologiczna

DOI 10.12887/36-2023-4-144-05

Artykuł jest głosem w dyskusji dotyczącej możliwości nadawania statusu moralnego artefaktom technologicznym sterowanym sztuczną inteligencją. W wymiarze refleksji filozoficznej odnosi się do debaty nad rozumieniem pojęcia „status moralny" oraz powiązanej z nim możliwości rozszerzenia wspólnoty moralnej o sztucznych agentów. W perspektywie psychologicznej prezentuje z kolei wyniki badań, w których testowano znaczenie takich czynników istotnych dla nadawania statusu moralnego, jak cechy artefaktów, wymiary percepcji umysłu, przypisywanie duszy oraz przekonania antropocentryczne. Rozważania są osadzone w kontekście aktualnie wdrażanych innowacji, popkulturowych narracji kształtujących obraz sztucznych systemów oraz dyskusji nad możliwością pojawienia się tzw. nadludzkiej sztucznej inteligencji.

Słowa kluczowe: sztuczna inteligencja, podmiot moralny, percepcja umysłu, przypisywanie duszy

Kontakt: (Zbigniew Wróblewski) Katedra Filozofii Przyrody i Nauk Przyrodniczych, Instytut Filozofii, Wydział Filozofii, Katolicki Uniwersytet Lubelski Jana Pawła II, Al. Racławickie 14, 20-950 Lublin; (Paweł Fortuna) Katedra Psychologii Eksperymentalnej, Instytut Psychologii, Wydział Nauk Społecznych, Katolicki Uniwersytet Lubelski Jana Pawła II, Al. Racławickie 14, 20-950 Lublin
E-mail: (Zbigniew Wróblewski) zbigniew.wroblewski@kul.pl; (Paweł Fortuna) pawel.fortuna@kul.pl