„Ethos" 36(2023) nr 3(143) 92-106
DOI 10.12887/36-2023-3-143-07

Paweł POLAK
Roman KRZANOWSKI

HOW TO TAME ARTIFICIAL INTELLIGENCE? A Symbiotic AI Model for Beneficial AI

In this paper, we posited that in light of the possible risks presented by the uncontrolled development of AI, the original goal of AI that was proposed at a workshop at Dartmouth University in 1956 and then reaffirmed in many subsequent publications should be revised, and a new goal based on the concept of Beneficial AI should be adopted. Furthermore, we proposed that the framework in which we could conceptualize Beneficial AI and its development can be broadly based on the domestication or taming of animals, something that humans have been doing for thousands of years.

The development of artificial intelligence (AI) has been, and continues to be, mostly dominated by engineers, computer scientists, and technological visionaries.[1] The notion of AI and the foundational goals that have chartered its course for decades were conceived at a workshop at Dartmouth University in 1956, with this being attended by John McCarthy, Marvin L. Minsky, Claude E. Shannon, and John Holland.[2] McCarthy and his colleagues proposed establishing a new discipline called artificial intelligence with the objective of "creating computer systems that could learn, react, and make decisions in a complex changing environment."[3] These original goals were restated and reconfirmed half a century later in 2006 at an event called *The Dartmouth College Artificial Intelligence Conference: The Next 50 Years*.[4] Current perspectives for the objectives of AI still echo these initial dreams (e.g., "AI is computer science, which aims to develop intelligent machines that can mimic human behavior."[5]). The

[1] See, e.g., Eugénio O l i v e i r a, "Beneficial AI: The Next Battlefield," *Journal of Innovation Management* 5, no. 4 (2018): 6–17; Stuart J. R u s s e l l, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Penguin Books, 2020).

[2] See John M c C a r t y, Marvin L. M i n s k y, Nathaniel R o c h e s t e r, and Claude E. S h a n n o n, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 31, 1955, http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

[3] See R u s s e l l, *Human Compatible*.

[4] See James M o o r, "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years," *AI Magazine* 27, no. 4 (2006): 87–9.

[5] See "Future of Artificial Intelligence—Javatpoint," JavaPoint, 2021, https://www.javatpoint.com/future-of-artificial-intelligence.

predictions, however, have become more specific in promising great advances in many areas of life, although not without danger.[6]

Some AIexperts go even further in their claims: "The research suggests that our shared AI future is, in part, about our becoming cyborgs. It's a slow, symbiotic coevolution of our own choosing, and it's already begun."[7] These and other similar predictions have been made by programmers, computer scientists, engineers, businessmen, and journalists, while philosophers, humanists, and other non-technical folk have been conspicuously absent from these discussions. Only sci-fi writers occupying the fringes of science have voiced concerns about the benefits of AI that were spelled out (implicitly) in McCarthy's manifesto, at least if AI is left untamed in the hands of technocrats.[8]

Over the past few years, however, more and more people beyond the core AI research teams have begun to perceive potential problems with AI,[9] forcing the AI community to engage in an open discussion about the very nature of this technology. The risks of AI are usually reduced to several clear, well-defined distinct issues, such as job security, safety, privacy, democracy, health, poverty, and so. Yet the real problem with AI is more fundamental, more philosophical rather than related to a specific application or implementation. Indeed, the problem concerns the foundational assumptions behind AI technology itself: Why, and for what, are we developing AI?

Some sixte odd years after that workshop in Dartmouth, the goal of AI remains a nuanced version of McCarthy's 1956 vision of "creating computer systems that could learn, react, and make decisions in a complex changing environment," although it is now couched in the jargon of current technology.[10]

---

[6] See, e.g., Katharine G a m m o n, "5 Ways Artificial Intelligence Will Change the World by 2050," USC News, December 04, 2017, https://news.usc.edu/trojan-family/five-ways-ai-will-change-the-world-by-2050/; Janna A n d e r s o n and Lee R a i n i e, "Artificial Intelligence and the Future of Humans," Pew Research Center: Internet, Science & Tech (blog), December 10, 2018, https://www.pewresearch.org/internet/2018/12/10/artificial-intelligence-and-the-future-of-humans/; Ashley S t a h l, "How AI Will Impact The Future Of Work And Life," Forbes, March 10, 2021, https://www.forbes.com/sites/ashleystahl/2021/03/10/how-ai-will-impact-the-future-of-work-and-life/.

[7] Mike B e c h t e l, "The Future of AI: Seeing the Forest for the Trees, and the Forests Beyond," Deloitte AI Institute, 2021, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/process-and-operations/us-ai-institute-future-of-ai.pdf.

[8] See Roman K r z a n o w s k i and Paweł P o l a k, "The Future of AI: Stanisław Lem's Philosophical Visions for AI and Cyber-Societies in Cyberiad," *Pro-Fil* 22, no. 3 (2021): 39–53.

[9] "Stephen Hawking, Elon Musk, Steve Wozniak, Bill Gates, and many other big names in science and technology have recently expressed concern in the media and via open letters about the risks posed by AI, joined by many leading AI researchers." Max T e g m a r k, "Benefits & Risks of Artificial Intelligence," Future of Life Institute, 2016, https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/.

[10] The terms currently encapsulating AI objectives are Artificial General Intelligence or Human Intelligence (see, e.g., Melanie M i t c h e l l, *Artificial Intelligence: A Guide for Thinking Humans*

Thus, we make the claim that the problem with AI lies in its original manifesto, the fundamental 1956 vision for why we pursue AI; the vision that in fact did not evolve. We also posit that in order to avoid the current and future problems that AI technology is likely to bring, in addition to its numerous benefits, our fundamental vision of AI should shift from those original 1950s objectives to something new.[11] The question, though, is to what should it shift?

To avoid or at least limit the risks of AI, we should drop the ambition of building super-intelligent, transhuman systems that mimic or exceed human capacities.[12] In trying to blindly replicate and improve upon human intelligence, according to the 1956 vision, we should heed one of Kant's warnings: "Out of the crooked timber of humanity, no straight thing was ever made."[13] We should instead focus on building AI systems that will be beneficial to humanity in the most general sense. In other words, AI systems should benefit us not just in a particular field but rather in principle, in essence. Even if it sounds like a cliché, AI systems should be good to us as a technical requirement.

The objective of AI research should be to create AI that benefits us in all walks of life rather than creating a perfect reckoning system, which was McCarthy's original idea for AI. We can label such AI systems as "Beneficial AI," and this is what this paper is all about. The way to do this, we propose, is based on the analogy of taming animals, so we will "tame" AI. The metaphor of taming (or domestication) will be taken in its basic sense, where it influences the relationship between humans and other biological organisms. We therefore work under the assumption that the process of taming an animal can in principle be helpful in bringing about the desired model of Beneficial AI, especially as we are apparently very good at taming wild things.[14]

(London: Penguin, 2019); Michael Timothy B e n n e t t, "Computable Artificial General Intelligence," arXiv, May 30, 2022, http://arxiv.org/abs/2205.10513; Gary M a r c u s, "Artificial General Intelligence Is Not as Imminent as You Might Think," Scientific American, June 6, 2022, https://www.scientificamerican.com/article/artificial-general-intelligence-is-not-as-imminent-as-you-might-think1/; Ragnar F j e l l a n d, "Why General Artificial Intelligence Will Not Be Realized," *Humanities and Social Sciences Communications* 7, no. 1 (2020): 10; Samu K u m p u l a i n e n and Vagan T e r z i y a n, "Artificial General Intelligence vs. Industry 4.0: Do They Need Each Other?," *Procedia Computer Science* 200 (2022): 140–50).

[11] See also O l i v e i r a, "Beneficial AI"; Russell, *Human Compatible.*

[12] For the incorrect focus of our AI-development strategies, see also R u s s e l l, *Human Compatible*.

[13] John B a n v i l l e, "Foreword," in Isaiah Berlin,*The Crooked Timber of Humanity: Chapters in the History of Ideas*, ed. Henry Hardy (Princeton and Oxford: Princeton University Press, 2013), XI–XVIII, Jstor, http://www.jstor.org/stable/j.ctt2tt8nd.3. Kantian concerns about replicating and improving upon the human mind were somewhat reflected by Russell when he stated that "the very definition of success in AI is wrong," meaning that we are developing AI under flawed objectives (see R u s s e l l, *Human Compatible*, 13).

[14] See, e.g., Jared M. D i a m o n d, *Guns, Germs and Steel: The Fates of Human Societies* (London: Cape, 1997); Jessica L e a r, "Our Furry Friends: The History of Animal Domestication,"

The concept of Beneficial AI is not new, because it has been elaborated upon in several publications.[15] However, the idea of "taming AI" as a means to progress toward Beneficial AI has not been discussed thus far.

WHY DO WE NEED TO REVISE THE GOALS OF AI?

AI has provided many very useful solutions, so it is therefore the technology that is currently being pinned as the greatest hope for improving human life. At the same time, however, the proliferation of AI systems is causing increasing risks.[16] The main philosophical determinants of these threats can be boiled down to four fundamental philosophical problems.

First, people are increasingly losing their decision-making power (i.e., a loss of primary decision agency). Second, they are also gradually losing control over their cognitive processes (i.e., a loss of primary epistemic agency). Third, they are gradually losing control over technical systems and the development of new technology (i.e., a loss of control). Fourth, AI systems have an overwhelming advantage when it comes to processing large amounts of data, allowing them to solve problems more efficiently and gain a significant advantage (i.e., the supremacy of computing power).

The lack of clarity surrounding the risks posed by AI has partly resulted from the peculiar ideology driving the development of AI. The project to create a "synthetic human," which is an explicit or implicit premise of the AI discussion, appeals to myths and metaphysical longings, and it may even be a challenge to the Creator. However, it should be noted that what is actually being undertaken is the construction of an ideological envelope for the AI program, because it effectively obscures the actual goals and problems by diverting people's attention to intriguing but irrelevant aspects. After all, the very use of the term "artificial intelligence" is very ideologically loaded, and this is particularly evident when comparing it to the program's original label, namely cybernetics.

Journal of Young Investigators, February 17, 2012, https://www.jyi.org/2012-february/2017/9/17/our-furry-friends-the-history-of-animal-domestication.

[15] See, e.g., "Beneficial AI 2017," Future of Life Institute, January 12, 2017, 2022, https://futureoflife.org/bai-2017/; O l i v e i r a, "Beneficial AI"; R u s s e l l, *Human Compatible*; Pedro F e r n a n d e s, Francisco C. S a n t o s, and Manuel L o p e s, "Norms for Beneficial AI: A Computational Analysis of the Societal Value Alignment Problem," *AI Communications* 33, nos. 3–6 (2020): 155–71; "Is 'Provably Beneficial' AI Possible?" ITU Hub, September 29, 2020, https://www.itu.int/hub/2020/09/is-provably-beneficial-ai-possible/.

[16] See, e.g., T e g m a r k, "Benefits & Risks of Artificial Intelligence."

It should also be noted here that according to the rationality of technical activities, AI is designed with the aim of achieving specific, often hidden, business goals. This makes it possible to develop the program, although such a choice leads to a reduced axiology for the activities undertaken. More and more voices claim that it is possible to account for other hierarchies of values and goals in the design of technical systems and thus humanize technology, but this has not led to a wider discussion, probably because of the limited awareness of the risks. Instead, a significant problem is a lack of ideas for an AI model that will benefit societies in the long term rather than just profit a narrow group in the short term. In other words, the current business model for AI development has no competitors, and this is a serious problem that needs to be confronted when addressing the threats of AI.

TOWARD BENEFICIAL AI

PRELIMINARIES

To organize the philosophical considerations surrounding AI, we should first pose some basic questions. In particular, we should question what we expect from AI:

(1) Do we need superhumans?
(2) Do we desire perfect slaves?
(3) Do we want synthetic humans?
(4) Is a symbiotic coexistence appealing?

The answers to such questions reveal important differences between philosophers, tech visionaries, and AI researchers and engineers, so we need to clarify some fundamental questions. We should therefore pose some philosophical questions rather than the previous set of questions:

(1) What kind of AI do we need as humanity?
(2) What kind of relationships do we need?
(3) What values should be preferred?
(4) Is the anthropocentric viewpoint on AI justified?

The first question sets the perspective for the whole deliberation, while the subsequent questions are refinements of this perspective, which we will call *Beneficial AI*. As we understand it, this concept represents the AI that we need as humans for beneficial development in the long term.

BENEFICIAL AI

What is Beneficial AI? Stuart J. Russell's definition states that a beneficent machine, one driven by Beneficial AI, realizes our objectives rather than its

own.[17] Of course, it would be simpler if we knew what we really wanted,[18] but this is not the case. Thus, Russell proposes some tentative guidelines under which beneficially inclined AI systems should operate. He qualifies his proposal by admitting that these are just guidelines rather than rules of any sort, because he fears that these may be taken like Isaac Asimov's notorious laws of robotics, which were originally proposed in Asimov's work *I, Robot*[19] and amended several times. Such an approach risks pushing the whole idea of Beneficial AI down the rabbit's hole.[20]

Russell's rules for Beneficial AI, which are not indented as laws,[21] state, firstly, that the machine objective is to maximize the realization of human preferences. Secondly, they assert that the machine does not know initially what these preferences should be. Thirdly, they posit that the machine learns these preferences from human behavior. Russell is fully aware that we do not actually know how to do this, technically, conceptually, or otherwise, but he is sure that if we want to avoid the potential calamities of unbridled AI development, we must pursue this endeavor.

The concept of Beneficial AI has also been elaborated in the Asilomar AI Principles.[22] This list of recommendations from the Beneficial AI Conference is a lengthy one,[23] but a few of the more important ones include:

(1) Ethics: AI systems should be designed and operated such that they are compatible with the ideals of human dignity, rights, freedoms, and cultural diversity.

(2) Value alignment: Highly autonomous AI systems should be designed such that their goals and behaviors are guaranteed to align with human values throughout their operation.

(3) Shared benefits: AI technologies should benefit and empower as many people as possible.

---

[17] See R u s s e l l, *Human Compatible*.

[18] See ibidem.

[19] See Isaac A s i m o v, *I, Robot* (Garden City, New York: Doubleday & Company, Inc., 1950).

[20] This expression is used especially in the phrase "going down the rabbit hole" or "falling down the rabbit hole." It is a metaphor for something that transports someone into a wonderful (or troublingly) surreal state or situation (see "Rabbit Hole,"Dictionary.com, https://www.dictionary.com/e/slang/rabbit-hole/). The expression dates back to the famous 1865 classic *Alice's Adventures in Wonderland* by Lewis Carroll (see Lewis C a r r o l l, *Alice's Adventures in Wonderland* Cambridge: Cambridge University Press, 1865), who was less famously a mathematician.

[21] See R u s s e l l, *Human Compatible*, 172.

[22] "Asilomar AI Principles," Future of Life Institute, August 11, 2017, https://futureoflife.org/2017/08/11/ai-principles/.

[23] See "Beneficial AI 2017."

As we said, the list is long, with it comprising twenty three areas grouped into research issues, ethics and values, and long-term issues.[24] These ideas are certainly on the mark, and one could say they benefit the discussion about AI. Anyway, we more or less know what Beneficial AI should be, but the problem is that we are not sure how to realize it. This is where the concept of domestication or taming comes in.

A SYMBIOTIC PERSPECTIVE ON BENEFICIAL AI

An important task for contemporary philosophers should be to search for ways to understand which AI models will meet the generally defined requirements of Beneficial AI. In principle, the task at hand seems practically impossible, since it would require first solving problems that have plagued humanity for centuries. The absence of any reasonable hope for generally solving this problem does not automatically lead to skepticism, though, because this incredibly complicated issue can be simplified in a non-trivial way by imitating a successful strategy from the history of *Homo sapiens*' development. This is admittedly an inductive inference, but the mechanisms of biological cooperation and domestication are still widely used and play an important role in people's lives. So, can such a strategy be used for AI? In other words, is it valid to analogize embodied, biosemiotic AI systems to biological species? The arguments raised below 1indicate that the proposed approach could be justified, but every approach to AI should be evaluated separately. Whether it is adequate, and to what extent, can only be established through experience. Nevertheless, the problem to be solved is so weighty, and the prospect so promising, that it is worth taking a risk and testing out this theory.

Now, let's try to look at autonomous AI systems as a specific species that coexists in the human environment. The biological perspective then gives, through analogy, a wide range of concepts for describing non-competitive (symbiotic) relationships. It is noteworthy that from such a perspective, the beneficence (i.e., usefulness) of an AI system can be understood as being analogous to that of animals, namely not as the utility of a tool but rather as a beneficial coexistence. Note also that such a biologically inspired perspective offers more possibilities than the model of an artificial slave, which is assumed in Asimov's famous laws of robotics.[25]

---

[24] See "Asilomar AI Principles."

[25] For a critique of Asimov laws' application to robotics, see, e.g., Susan Leigh A n d e r s o n, "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics," in *Machine Ethics*, ed. Michael Anderson and Susan Leigh Anderson (Cambridge: Cambridge University

Symbiotic AI therefore appears to be a very promising model for creating useful AI. Symbiotic relationships are well-tested as a strategy for enhancing the developmental capabilities of organisms, and this approach draws attention to the fact that AI systems are, whether we want them to be or not, part of the operating environment of modern humans. Biological metaphors then make it possible to effectively shape inter-species relations, including those with artificial species, through taming/domestication by introducing an important conceptual and theoretical framework for Beneficial AI concepts.

TAMING AI

AI SYSTEMS: ARTIFACTS OR SPECIES?

Let us start our investigation about how to interpret the AI phenomenon with a brief review of the arguments and counterarguments. The arguments in favor of regarding AI systems as mere tools first assert that they are sophisticated, extremely complex technical artifacts. They also point out that AI systems fulfill roles as tools, as they were intended to. It is also clear that AI systems are a result of typical technical processes, such as design, manufacture, testing, and use. These arguments are generally valid, although they are certainly one-sided. It should also be noted that they are entirely accurate for some AI techniques. The conclusion to be drawn is that the concept of AI is overly broad, and to avoid confusion in further considerations, only a select class of technical solutions that are labelled as AI should be considered. In this paper, we choose to focus on autonomous robotic AI systems (i.e., embodied AI) that have been designed from a biosemiotic perspective. For such systems, we will therefore address the arguments for regarding them as an artificial species.

First, their autonomy excludes them from being treated as typical tools, because tools lack autonomy and independent initiative, so the actions of a human determine whether a tool functions properly. Secondly, the adaptability of these AI models, along with their autonomy, makes the relationship with such artifacts different from that with typical tools, which have a strictly defined purpose, a scope for correct behavior, and no independent decision-making abilities. Thirdly, machines can enter into more complex interpersonal relationships. Even simple and unsophisticated systems like ELIZA have strongly engaged people in peculiar social relationships. Fourth, in such systems, it is possible for them to take over some organizational and social functions (e.g.,

---

Press, 2011), 285–96; Peter W. S i n g e r, "Isaac Asimov's Laws of Robotics Are Wrong," Brookings(blog), 2010, https://www.brookings.edu/opinions/isaac-asimovs-laws-of-robotics-are-wrong/.

care for the elderly). Finally, we are observingevident bioinspiration in the latest paradigms for AI technology[26], such as embodied AI and the biosemioticparadigm in AI.

It is worth paying some attention to the biosemiotic perspective adopted here.[27] This is based on the acceptance of diverse, species-determined forms of intelligence, and it also assumes the possibility of reconstructing biosemiotic relations in artificial systems. Such an AI system will be an embodied autonomous system with its own internally generated field of meaning. The failures of the AGI (artificial general intelligence) concept flow from a misunderstanding of the distinctiveness of the intelligence of an artificial "machine species."

Incorrect assumptions about AI[28] exclude the possibility of understanding the goals of the AI program. Thus, if the goal of AI is not to create an "artificial human," what are the actual goals?

TAMING AND DOMESTICATION OF ANIMALS

Taming and domestication of animals have been important cultural activities throughout history and facilitated the fascinating development of *Homo sapiens*. Such processes enabled the development of human societies by providing stable sources of food, materials, and energy. Dogs were the first species to be domesticated about fifteen thousand years ago. Another momentous event in human history was the domestication of sheep and goats between 9,000 and 7,000 BC.

Coexistence and cooperation between different, often very different, species are made possible through fundamental biological relationships that can be labelled under the umbrella term s y m b i o s i s. Symbiosis is defined as a non-antagonistic, long-term relationship, and two basic forms manifest:[29]

(1) m u t u a l i s m, where both species gain benefits;

(2) c o m m e n s a l i s m, where an asymmetric benefit occurs (i.e., where one species benefits while the other species is neither benefitted nor

[26] More about AI paradigms, see: Roman K r z a n o w s k i and Paweł P o l a k, "Ontology and AI Paradigms." *Proceedings* 81, no. 1 (2022): 119. https://www.mdpi.com/2504-3900/81/1/119.

[27] See Anna S a r o s i e k, "Biologiczne i semiotyczne nurty w dziedzinie sztucznej inteligencji," *Transformacje* 1–2, nos. 88–89 (2016): 294–306; Anna S a r o s i e k, "The Role of Biosemiosis and Semiotic Scaffolding in the Processes of Developing Intelligent Behaviour," *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, no. 70 (2021): 9–44.

[28] See, e.g., Brian Cantwell S m i t h, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: The MIT Press, 2019); Michael W o o l d r i d g e, *The Road to Conscious Machines: The Story of AI* (London: Penguin, 2021).

[29] See Pierre Joseph van B e n e d e n, *Animal Parasites and Messmates* (London: Kegan Paul, Trench, 1883).

harmed—a typical example is how dogs follow a commensal pathway into domestication).

The possibility of analogously applying biological concepts to artificial systems depends upon whether AI can be treated as a synthetic species. This will be the subject of the next part of the article.

TAMING AND DOMESTICATING AI

The prospect of taming synthetic species and integrating them into symbiotic relationships brings many new challenges. First, it is relativized to just a certain group of AI systems, although it should also be noted that there is still no consensual understanding of the concept of AI autonomy. Second, the ability to design the traits of artificial systems means the relationship can be developed more quickly than with biological organisms, where generations of artificial selection may be required. This is where one of the critical problems arises: The wide scope of feature design and the ability to design quickly makes the time required for reception (i.e., adaptation, achieving cooperation) a critical factor. In this respect, taming AI and creating a symbiotic relationship with it will be incomparably more difficult than it is for biological species.

It is also worth raising questions about the specific subjective conditions for taming such an "artificial species." Recall that we refer here to an embodied AI equipped with its own internally induced semantics.

The most important propositions for such conditions can be summarized as follows:

(1) social acceptance of AI as a separate species (IT worldview)[30];

(2) a lack of psychological barriers (e.g., uncanny valley);

(3) no self-reduction of humans, where relationships are shaped like relationships with a tool rather than a living organism (Bolter's thesis)[31];

(4) a need for critical thinking to accelerate cultural evolution (a critical time factor!);

(5) openness to change in the image of the world (i.e., the image of humanity, machines, society).

---

[30] See Witold M a r c i s z e w s k i and Paweł S t a c e w i c z, *Umysł—Komputer—Świat. O zagadce umysłu zinformatycznego punktu widzenia* (Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2011), http://libra.ibuk.pl/book/101353.

[31] See Jay David B o l t e r, *Turing's Man: Western Culture in the Computer Age* (Chapel Hill: University of North Carolina Press, 1984).

*

In this paper, we posited that in light of the possible risks presented by the uncontrolled development of AI, the original goal of AI that was proposed at a workshop at Dartmouth University in 1956 and then reaffirmed in many subsequent publications should be revised, and a new goal based on the concept of Beneficial AI should be adopted. Furthermore, we proposed that the framework in which we could conceptualize Beneficial AI and its development can be broadly based on the domestication or taming of animals, something that humans have been doing for thousands of years.

We have to recognize that taming of animals is a process different in kind, not in a degree, from taming of AI systems. In fact, apart from broad similarities, the details how to tame AI systems are not immediately obvious. The details are not obvious as in taming of AI we do not have the experience of thousands of years of taming animals and other human beings to inform us. There is a lot, if not everything, to learn how to do it right.

The issue of taming AI also opens up interesting fields of ethical research, for example: can the complex human–AI symbiotic relationships even be meaningfully described by utilitarian ethics without falling into unacceptable simplifications? What kind of ethics would be appropriate for such an analysis?[32] Clarifying the concepts involved in ethical issues of taming of AI is certainly necessary to responsibly address such questions. This should become the subject of future studies.

AI systems are an important part of the widely understood human environment, so it is necessary to rethink the goals of AI development. Indeed, this is one of the most important tasks facing philosophy today (i.e., philosophy in technology). As we indicated, symbiotic relationships could be used to shape useful AI. The analysis shows that the critical factor in the face of differentiation in AI is the reception time factor, because differentiated AI systems makes it impossible to undertake a general analysis to find general solutions. The requirements for AI systems in symbiotic relationships must therefore be formulated for a specific context.

On the other hand, it is possible to formulate some general subject requirements for humans in these symbiotic relationships. It is imperative to develop the emerging concept of Beneficial AI, despite its numerous difficulties, due to its critical importance for humans.

Such an approach admittedly does not solve the philosophical problems associated with the development of AI, but it quite effectively diverts attention

[32] The authors wanted to thank the anonymous reviewer for bringing these issues to their attention.

from them. On the other hand, the numerous dangers and threats associated with the rapid development of AI technology, both real and imaginary, bring some important philosophical themes back into the discussion.

The progressive technologization of individual and social life makes humanizing the technological sphere a particularly important issue. The task of philosophy is therefore to set goals for our rapidly developing technology, so it will serve the long-term interests of modern human life, in contrast to the hitherto increasingly pronounced dehumanization process. If this task is not undertaken, the future will be shaped along the lines of a technocratic project, or it will be susceptible to strong irrational elements hidden within uncritical visions of technological development and fueled by wishful thinking.

BIBLIOGRAPHY / BIBLIOGRAFIA

Anderson, Janna, and Lee Rainie. "Artificial Intelligence and the Future of Humans." Pew Research Center: Internet, Science & Tech (blog), December 10, 2018. https://www.pewresearch.org/internet/2018/12/10/artificial-intelligence-and-the-future-of-humans/.

Anderson, Susan Leigh. "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics." In *Machine Ethics*.Edited by Michael Anderson and Susan Leigh Anderson. Cambridge: Cambridge University Press, 2011.

"Asilomar AI Principles." Future of Life Institute, August 11, 2017. https://futureoflife.org/2017/08/11/ai-principles/.

Asimov, Isaac. *I, Robot*. Garden City, New York: Doubleday & Company, Inc., 1950.

Banville, John. "Foreword." In Isaiah Berlin, *The Crooked Timber of Humanity*: *Chapters in the History of Ideas*. Edited by Henry Hardy. Princeton and Oxford: Princeton University Press, 2013. http://www.jstor.org/stable/j.ctt2tt8nd.3.

Bechtel, Mike. "The Future of AI: Seeing the Forest for the Trees, and the Forests Beyond." Deloitte AI Institute, 2021. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/process-and-operations/us-ai-institute-future-of-ai.pdf.

van Beneden, Pierre Joseph. *Animal Parasites and Messmates*. 3rd ed. London: Kegan Paul, Trench, 1883.

"Beneficial AI 2017." Future of Life Institute, January 12, 2017. https://futureoflife.org/bai-2017/.

Bennett, Michael Timothy. "Computable Artificial General Intelligence." arXiv, May 30, 2022. http://arxiv.org/abs/2205.10513.

Bolter, Jay David. *Turing's Man: Western Culture in the Computer Age*. Chapel Hill: University of North Carolina Press, 1984.

Diamond, Jared M. *Guns, Germs and Steel: The Fates of Human Societies*. London: Cape, 1997.

Dictionary.com (s.v. "Rabbit Hole"). https://www.dictionary.com/e/slang/rabbit-hole/.

Fernandes, Pedro, Francisco C. Santos, and Manuel Lopes. "Norms for Beneficial AI: A Computational Analysis of the Societal Value Alignment Problem." *AI Communications* 33, nos. 3–6 (2020): 155–71.

Fjelland, Ragnar. "Why General Artificial Intelligence Will Not Be Realized." *Humanities and Social Sciences Communications* 7, no. 1 (2020): 10.

Gammon, Katharine. "5 Ways Artificial Intelligence Will Change the World by 2050." USC News, December 04, 2017. https://news.usc.edu/trojan-family/five-ways-ai-will-change-the-world-by-2050/.

"Future of Artificial Intelligence—Javatpoint." JavaPoint, 2021. https://www.javatpoint.com/future-of-artificial-intelligence.

"Is 'Provably Beneficial' AI Possible?" ITU Hub, September 29, 2020. https://www.itu.int/hub/2020/09/is-provably-beneficial-ai-possible/.

Krzanowski, Roman, and Paweł Polak. "The Future of AI: Stanisław Lem's Philosophical Visions for AI and Cyber-Societies in Cyberiad." *Pro-Fil* 22, no. 3 (2021): 39–53.

———. "Ontology and AI Paradigms." *Proceedings* 81, no. 1 (2022): 119. https://doi.org/10.3390/proceedings2022081119.

Kumpulainen, Samu, and Vagan Terziyan. "Artificial General Intelligence vs. Industry 4.0: Do They Need Each Other?" *Procedia Computer Science* 200 (2022): 140–50.

Lear, Jessica. "Our Furry Friends: The History of Animal Domestication." Journal of Young Investigators, February 17, 2012. https://www.jyi.org/2012-february/2017/9/17/our-furry-friends-the-history-of-animal-domestication.

Marciszewski, Witold, and Paweł Stacewicz. *Umysł—Komputer—Świat: O zagadce umysłu z informatycznego punktu widzenia*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2011. http://libra.ibuk.pl/book/101353.

Marcus, Gary. "Artificial General Intelligence Is Not as Imminent as You Might Think." Scientific American, June 6, 2022. https://www.scientificamerican.com/article/artificial-general-intelligence-is-not-as-imminent-as-you-might-think1/.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 31, 1955. http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. London: Penguin, 2019.

Moor, James. "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years." *AI Magazine* 27, no. 4 (2006): 87–91.

Oliveira, Eugénio. "Beneficial AI: The Next Battlefield." *Journal of Innovation Management* 5, no. 4 (2018): 6–17.

"Rabbit Hole." In Dictionary.com. https://www.dictionary.com/e/slang/rabbit-hole/.

Russell, Stuart J. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin Books, 2020.

Sarosiek, Anna. "Biologiczne i semiotyczne nurty w dziedzinie sztucznej inteligencji." *Transformacje* 1–2, nos. 88–89 (2016): 294–306.

———. "The Role of Biosemiosis and Semiotic Scaffolding in the Processes of Developing Intelligent Behaviour." *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, no. 70 (2021): 9–44.

Singer, Peter W. "Isaac Asimov's Laws of Robotics Are Wrong." Brookings (blog), 2010. https://www.brookings.edu/opinions/isaac-asimovs-laws-of-robotics-are-wrong/.

Smith, Brian Cantwell. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press, 2019.

Stahl, Ashley. "How AI Will Impact The Future Of Work And Life." Forbes, March 10, 2021. https://www.forbes.com/sites/ashleystahl/2021/03/10/how-ai-will-impact-the-future-of-work-and-life/.

Tegmark, Max. "Benefits & Risks of Artificial Intelligence." Future of Life Institute, 2016. https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/.

Wooldridge, Michael. *The Road to Conscious Machines: The Story of AI*. London, UK: Penguin, 2021.

ABSTRACT / ABSTRAKT

Paweł POLAK, Roman KRZANOWSKI, How to Tame Artificial Intelligence? A Symbiotic Model for Beneficial AI

DOI 10.12887/36-2023-3-143-07

The paper presents a vision of tamed AI, an AI technology that would maximize its benefits to society and individuals. The need for such a technology has been discussed at length under the umbrella term of Beneficial AI, yet there have been no proposals for how to actually implement such a system. In this paper, we discuss the concept of domesticating (or taming) AI, how this would apply to AI technology, and how it could lead to a symbiotic human—machine system. We refer to this proposed AI concept as a symbiotic AI model for Beneficial AI.

Keywords: tamed AI, Beneficial AI, symbiotic systems, symbiotic AI, future of AI, AI objectives, mitigating AI risks

Contact: Department of History and Philosophy of Science, Faculty of Philosophy, Pontifical University of John Paul II, ul. Kanonicza 9, 31-002 Cracow, Poland

E-mail: (Paweł Polak) pawel.polak@upjp2.edu.pl; (Roman Krzanowski) wf@upjp2.edu.pl
(Paweł Polak) http://polak.wikidot.com/; (Roman Krzanowski) https://www.researchgate.net/profile/Roman_Krzanowski
ORCID: (Paweł Polak) 0000-0003-1078-469X; (Roman Krzanowski) 0000-0002-8753-0957

Paweł POLAK, Roman KRZANOWSKI – Jak oswoić sztuczną inteligencję? Model symbiotycznej SI jako realizacja dobroczynnej SI

DOI 10.12887/36-2023-3-143-07

W niniejszym artykule zaprezentowano wizję oswojenia sztucznej inteligencji (SI jako techniki, która maksymalizuje korzyści płynące dla używających jej zarówno społeczności, jak i jednostek). Dotychczas o potrzebie takiej techniki dyskutowano, używając pojęcia dobroczynnej SI (ang. beneficial AI), choć wciąż nie pojawiły się konkretne propozycje, w jaki sposób zaimplementować taki system. W opracowaniu zaproponowano wykorzystanie analogii wybranych klas systemów SI do organizmów żywych, co pozwala na rozważenie możliwości udomawiania lub wdrażania takich systemów SI. Strategia domestykacji lub oswajania jest wypróbowaną metodą przynoszącą długofalowe korzyści gatunkowi ludzkiemu, dlatego w artykule podjęto kwestię, w jaki sposób można zastosować tę koncepcję w technice SI oraz jak należy projektować symbiotyczne systemy człowiek–maszyna. W rezultacie sformułowano symbiotyczny model SI dla dobroczynnej SI.

Słowa kluczowe: oswojona SI, dobroczynna SI, systemy symbiotyczne, symbiotyczna SI, przyszłość SI, cele AI, ograniczanie ryzyka związanego z AI

Kontakt: Katedra Historii i Filozofii Nauki, Wydział Filozofii, Uniwersytet Papieski Jana Pawła II, ul. Kanonicza 9, 31-002 Kraków
E-mail: (Paweł Polak) pawel.polak@upjp2.edu.pl; (Roman Krzanowski) wf@upjp2.edu.pl
(Paweł Polak) http://polak.wikidot.com/; (Roman Krzanowski) https://www.researchgate.net/profile/Roman_Krzanowski
ORCID: (Paweł Polak) 0000-0003-1078-469X; (Roman Krzanowski) 0000-0002-8753-0957