FILOSOFIJA. SOCIOLOGIJA
2024. T. 35. Nr. 1, p. 5–13

ISSN 0235-7186 eISSN 2424-4546
DOI: https://doi.org/10.6001/fil-soc.2024.35.1.2

Black Boxes that Curtail Human Flourishing are no Longer Available for Use in Artificial Intelligence (AI) Design

JOHN W. MURPHY
University of Miami, 1320 S Dixie Hwy, Coral Gables, 33146 Florida, United States
Email: j.murphy@miami.edu

CARLOS LARGACHA-MARTÍNEZ
Fundación Universitaria del Área Andina, Carrera 14A No.70 A-34, Bogotá, Colombia
Email: clargacha@areandina.edu.co

AI is considered to be very abstract to a range of critics. In this regard, algorithms are referred to regularly as black boxes and divorced from human intervention. A particular philosophical maneuver supports this outcome. The aim of this article is to (1) bring the philosophy to the surface that has contributed to this distance between AI and people and (2) offer an alternative philosophical position that can bring this technology closer to individuals and communities. The overall goal of the analysis in this paper is the humanising of AI by addressing the shortcomings of conceptualising algorithms as black boxes.

Keywords: dualism, artificial intelligence, explainability, linguistic turn, anti-dualism

INTRODUCTION

The aim of this paper is to clear up an important facet of artificial intelligence (AI). To the public and many users, AI is already shrouded in mystery. Many exaggerated claims have been made about AI and how this technology will affect the future. The result is that many everyday persons and experts worry about the social impact of AI. The casual talk that often occurs regularly about 'black boxes' only adds to the current intrigues and uncertainty. Nonetheless, referring to algorithms as black boxes is quite commonplace (Savage 2022; Power 2023).

The problem is that black boxes cannot be entered and their contents known. When algorithms are described in this manner, these devices are removed from scrutiny. As a result, AI is thought by many persons to operate *in camera*. Any attempt to explain how they function is beyond human capability and futile. Their logic is simply unavailable for serious examination and cannot be reconstructed on demand.

In this way, algorithms are placed in a domain similar to Kant's noumenal realm (Kant 2007). Two outcomes occur because of this placement. The first is that algorithms exist beyond

human experience. Their operation thus eludes human comprehension. And second, also like in Kant's typology, anything that exists in the noumenal domain has a unique status. Specifically, the content of this space is worth more than elements in the phenomenal realm that are tied to experience. The key distinction is that the noumenal sphere is not contaminated or limited by quotidian concerns.

The trust of this paper is to examine the philosophical maneuver that justifies black boxes and the separation of computer technology, specifically algorithms, from human experience. A dualism is accepted, but is seldom examined in the context of AI, that provides this technology a patina of autonomy. This sort of investigation is necessary to reduce the threat of AI and enable this technology to serve human needs. Alternatively, a non-dualistic position is required to achieve this aim, which will also be discussed.

This critique of black boxes follows a trend in philosophy inaugurated by Dreyfus (1972), Dreyfus and his brother (1986), and Winograd and Flores (1986). These authors adopted phenomenology to undercut the autonomy of AI and provide computers with a human grounding. The non-dualistic strategy that guides this paper is somewhat broader than phenomenology, and could be classified as poststructural (Belsey 2022). In both cases, however, the point is to understand AI as an extension of human agency.

These distinctions between regions, based on their respective ties to experience, are dubious in view of contemporary philosophy. A dualistic maneuver is made that is considered to be illegitimate. In this regard, L. Wittgenstein (1990) declared that persons must remain silent about whatever cannot be spoken about or known. Considering this advice, black boxes can be explored or they do not exist. Referring to them as inaccessible, therefore, is a maneuver that no longer has any integrity. There are simply too many contradictions involved.

There are demands on the horizon for an explainable AI (XAI) (Varma 2021; McNamara 2022). The thrust of this movement is that buying into black box imagery is no longer acceptable. Some ethicists, in fact, declare that algorithms that cannot be reviewed and explained should not be used (Durán and Jongsma 2021). The basic objection is that persons who are affected have the right to know the operational logic of these devices and how decisions are reached.

The point is not that people necessarily become experts. What they need is to trust algorithms, as the majority of people trusted many vaccines, although they did not understand the bio-chemistry. Thus, the '*black-box*' analogy is not a good descriptive. Transparency and humbleness must be present. Stating that machine-learning techniques are '*opaque*', and uncritically accepting this denomination is not a good strategy. In this case, opaque means that 'even experts with relevant equipment cannot determine why and how inputs are transformed into outputs' (Stahl et al. 2021: 383).

The question that guides this discussion is the following: how are black boxes established? Without addressing this issue, algorithms will remain black boxes and considered to be beyond explanation. To be successful, this investigation must be both philosophical and historical. By adopting this two-pronged approach, the philosophical gambit that makes these boxes possible, along with their utility, can be clearly understood. Subsequent to these revelations, steps to make algorithms more accessible and explainable will remain a challenge but this task may seem feasible. To critics of AI, explainability is essential to trusting this technology (Hamon et al. 2020). Opening algorithms to intense scrutiny has thus become a high priority.

AI AND BLACK BOX UTILITY

The immediate task is to clarify the rationale behind the invocation of black boxes. Specifically, what is their *raison d'etre*? The thinking that supports their existence and use is important but is rarely mentioned in debates about the merits of AI. When using the phrase black box, a particular intention is put into motion. Usually, the focus is on inputs and outputs, with little concern for the internal functioning of the mechanism or organism in question (Estevez 2022). This emphasis is thought to lead to exactness.

Black boxes have been adopted in various disciplines, such as economics, engineering, chemistry and psychology. In psychology, for example, behaviourists were prominent in this regard. They wanted to transform psychology into a science and move away from the influence of Freudianism and other speculative theories of the *psychē.* They wanted to avoid imputing content to the mind that could not be verified (Buckley 1989; McLeod 2007). Persons, accordingly, were accorded the status of black boxes.

As black boxes, nothing would be attributed to persons that could not be publicly reviewed. No longer should reference be made to dubious notions such as consciousness, a self, and most notably the unconscious. These factors, according to behaviourists, only detract from sound analysis. Accordingly, the only valid explanations should be attributed to stimuli and responses, otherwise known as inputs and outputs.

In economics, black boxes have performed a similar function particularly with large amounts of data. Predictive power appears to increase with black box models. As a consequence of the increased connections that are possible, black boxes can make more of the data that are available. The problem, however, is that analysis may begin to drift away from the original information. Thus reality may be obscured by the data, although analysis is improved.

In aerospace engineering, the use of the black box is also applauded to determine what happened and to learn from accidents. However, this method does not provide a holistic view, for example, the past/present/future of an accident. Neither is the context revealed. Without this additional information, not much can be learned about actual events from these models. Although some simulations and projections can be made, these analyses tend to be very sterile.

Examples such as psychology and economics, and similar gambits in other fields, set the stage for speculating about explanations (Mullainathan and Spiess 2017). While trying to be scientific, legitimacy was ascribed to unknown elements to explain behaviour or events. What AI has done is to follow a well-worn path. Although black boxes can come to be treated as problematic, the use of AI is proceeding with almost universal applause.

References to black boxes arise most often because of the complexity accorded to algorithms (Cassawers 2020; Savage 2022). What is going on in these devices is thought to be too complex for humans to handle. Particularly at the level of so-called deep learning, the issue of dimensionality comes into play (Shashmi 2021). That is, as inputs are processed through the layers of nets, geometric relationships are established that are beyond the capability of humans to visualise. Also, transmissions can occur non-linearly and recurrently. The result is the claim that these processes cannot be recounted, and thus the rules that are followed are dismissed as mysterious. At this point, this activity becomes a black box.

The so-called weak black boxes can be reversed engineered, whereas strong or deep boxes allegedly cannot. The earlier categories that set this process in motion are accessible, while later or deeper directions are thought to involve internal modifications that are difficult, or nearly impossible, to retrace. These later actions do not adhere to predefined scripts, in

the strict sense, and therefore are thought to defy rules. The modifications, that take place and are regularly linked to learning, are attributed to almost spontaneous actions that cannot be easily reproduced.

Many critics of black boxes claim that the 'veiled logic' that is, in fact, operating is accessible (Mir and De Blanc 2023). Those who are interested in discovering these operations have to merely look under the hood, so to speak. But regularly, this advice is dismissed as useless. What is going on at the deeper functions of algorithms is pushed aside as beyond human comprehension. Although there is little doubt that increased complexity requires difficult work in retracing these activities, inaccessibility is another question. Clearly, time consuming and tedious interventions are needed to reveal the logic that is in play.

Developers know that 'flags' can be installed during coding, so that explainability can be managed. But this work is time consuming. In the end, ethics and decency are killed by profits. For example, the documentation debt is a cost that designers want to avoid. But as Bender et al. (2021: 615) declare, 'without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones'. Hence their solution is 'to budget for documentation as part of the planned costs of dataset creation, and only collect as much data as can be thoroughly documented within that budget'. (Bender et al. 2021: 615).

Explainability, in fact, can go in a variety of directions when attempting to explore these mysterious boxes (Miller 2019). For example, does the entire operation need to be revealed? The answer to this question leads to issues of time and effort. And then in the end, resources have an important role in determining how much can be explained. Explainability is thus not merely a matter of taking a close look at the design of an AI platform. There may be many logistical constraints that stifle a thorough examination (McDermid et al. 2021). Additionally, various stakeholders and interests make explain ability a multidimensional task.

Nonetheless, at the heart of this discussion is whether anything is ever beyond human comprehension. What comes to mind at this point is Nietzsche's (1997) declaration that nothing human is foreign to persons. Applied to algorithms, his idea is that humans invented these devices, so why are they readily treated as black boxes? In this regard, increased effort does not lead automatically to inaccessibility, because some sort of leap seems to be taking place that stifles the investigative spirit and abilities of persons.

Here is where philosophy comes on the scene. As suggested earlier, a philosophical position is available that allows black boxes to make sense, that is, to become a refuge for processes that defy easy explanation. Exposing this philosophy is necessary to facilitate bringing real transparency to algorithms. As long as black boxes are considered legitimate, limits to explainability will be entertained that extend beyond logistical issues. Simply put, algorithms may be viewed as inevitably beyond human knowledge and control.

BLACK BOXES AND PHILOSOPHY

Historically, the recognition of black boxes has accomplished a couple of aims. In general, a safe space to operate became available. This place allowed for the establishment, for example, of an objective space divorced from contingency, a location where reason can operate, and a reliable foundation for knowledge. Additionally, operations can be dumped there for a variety of reasons. In the case of AI, black boxes allow for complex processes to be ignored or hidden, while useful knowledge is put into practice without revealing anything.

Black boxes are accorded the status of the nominal realm described by Kant. What this designation means is at a specific domain placed beyond human experience and accessibility. The content of this sphere, accordingly, remains unknown but powerful. While in Kant's scheme the noumena cannot be known, this realm represents reality. To acquire valid knowledge, persons must strive to gain access to this domain. Failure to do so means that only limited knowledge is available to humans.

By making a distinction between everyday existence (phenomena) and the noumena, Kant is following a trend in Western philosophy. Throughout most of this history persons have had to transcend, or overcome, the everyday world to acquire reliable knowledge and correct ethical guidance. This advice is sustained by quite a fiction.

This tendency became evident *circa* 1600 with the writing of Descartes (Gombey 2007). What Descartes did is make obvious the dualism that was operating behind the scenes in earlier philosophies. While earlier writers speculated about a theory or ethereal foundations of knowledge and ethics, Descartes made a straightforward tact that avoided this ambiguity and hypothesising. He argued that the mind (*res cogito*) and matter (*res extensa*) could be separated. What he made explicit is the dualism that enables spaces to exist where absolute foundations can be positioned. With the mind separated categorically from what is known, a situation is available where these absolutes can exist. At least theoretically, there is some justification for the existence of places that are unaffected by the contingencies and limitations inherent to the mind.

Wheather dualism can ever be overcome, so that these obscured realities might be revealed, is another issue. What can be stated, however, is that these realms are possible due to this philosophy. Black boxes have been the beneficiaries of this dualism. Because of this philosophical principle, black boxes, for good or bad, can be assumed to operate beyond human awareness. Ineffable, non-contingent operations can thus be given credence without much consternation.

In black boxes, the focus is not competencies but execution. 'A simply leads to B', although this association can become quite convoluted. Nonetheless, a coherent system of operational logic is presumed to be present. Specific input leads to particular output that seems to make sense, in the absence of the intricacies of the connections being revealed. The dualism that supports these black boxes inspires this confidence. After all, in the purity of these spaces the exact logic that is attributed regularly to AI is possible (Poster 2001).

As long as this style of thinking continues, AI will remain a closed system; the mystery associated with this invention will continue. Indeed, any suggestion that humans can exert control of algorithms will be treated as wishful thinking, since a real investigation of these devices is imagined to be too daunting. Successfully navigating a black box, given human flaws, seems to be an impossible assignment. But both experts and laypeople have the right to know what AI is doing. This belief is expressed in multiple documents by governments and regional agencies, like the EU AI Act.[*] What is also happening is that AI ethics has been invented to analyse and avoid problems.

What contemporary philosophy adds to this discussion is important. Following the anti-dualistic stance that is taken, an honest portrayal of AI is possible. At least blaming the complexity of algorithms for a failure to investigate these devices will not appear to be a reasonable

[*] Visit https://artificialintelligenceact.eu. Art. 52 talks about 'New Transparency Obligations for Certain AI Systems'. For a summary, visit https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf?

decision. A black box rationale may not disappear right away, but the philosophy that supports their autonomy and opacity is challenged. Nietzsche's (1997) theme that nothing humans invent is beyond their comprehension may be taken seriously, thereby inaugurating a new relationship between individuals and communities, and AI.

NO MORE BLACK BOXES

Identifying the beginning of contemporary philosophy is difficult, although 1900 provides a relevant entry point. Around this time, important changes were taking place in many disciplines, including philosophy, the arts and physical sciences. What was emerging were challenges to the dualism that pervaded those areas of study. Writers in those areas were beginning to reveal what J. Gebser (1984) called a '*world without opposite*'. Along a similar line of thinking, G. Deleuze (1994) noticed that these developments were causing the world to disappear.

Of course, neither contemporary philosophy nor art movements were destroying reality. Nonetheless, the dominant realism became more difficult to maintain. That is, persons were no longer facing, or encountering, a world but were intimately involved in creating whatever is known or comes to be accepted as real. This change is witnessed, for example, in phenomenology, existentialism, surrealism, and quantum theory (Bakewell 2016). Stated simply, human agency and reality now are inextricably intertwined.

From the viewpoint of L. McTaggart (2011: xix), Descartes 'banished any kind of holistic intelligence … However, the latest scientific discoveries founded in quantum mechanics has shown that everything is relational.' Her point is that persons and the world are interconnected; there is no real separation. The result is an effect that is called superposition, whereby any physical event can have various meanings simultaneously until a human intervention occurs. A human bond appears to be necessary to hold together even the physical reality together.

E. Husserl (1964) contributes to this trend with his notion of intentionality, which he defined as '*consciousness is always conscious of something*'. With this somewhat arcane phrase, he placed the dualism proposed by Descartes in jeopardy. Similarly, the indeterminacy effect that W. Heisenberg (1958) revealed in his trenchant experiments on light call into question the separation of the knower from the known. What they both stress is that nothing is immune to the human presence and the concurrent influence, even physical reality.

Following these announcements, writers in philosophy and other humanities began to explore new approaches to language. Under the general heading of poststructuralism, these contributors no longer treated language as a tool that enabled humans to highlight or point to objects in their environments (Poster 1989). This traditional viewpoint, sometimes called the indexical thesis, was dualistic and treated these objects as things. The job of language, accordingly, is to simply pinpoint their location in a system of things. In Descartes' theory, these elements constitute the autonomous and objective *res extensa*, divorced from distractions provoked by humans.

Recognised as the linguistic turn, those newer philosophers, such as M. Merleau-Ponty and R. Rorty, were proposing that nothing escapes from the inventive character of language (Rorty 1967). More specifically, language does not point to anything but mediates all knowledge. The unavoidable conclusion is that reality is shaped by language. Values, beliefs and commitments, based on linguistic acts, organise the world that is considered to be real. For this reason, Merleau-Ponty (1973) characterised language as the prose of the world.

In the sciences, T. Kuhn (1970) began to recognise this influence of language when he declared that all findings are theory-laden. Given the pervasiveness of language, there is no longer

any hidden variables that can be introduced as explanatory factors. Nothing has the autonomy any longer to assume this role. Every setting is thus an interpretive sphere that projects possible realities and explanations consistent with the interpretations, or worldviews, that are in effect. In contemporary philosophical parlance, these worlds are interpersonally constructed.

S. Fish (1979) summarises this trend nicely when he suggests that now interpretation goes all the way down. As a result, the standard ultimate foundations have lost credibility – universal facts or rules are now embedded in language and deprived of any autonomy. Facts, for example, are no longer empirical references but accomplishments of individuals or groups (Pollner 1991). To echo Wittgenstein (1960; 1953) on this point, what is considered to be real depends on the '*language game*' that is being played. In view of contemporary philosophy, the reality of the world is both a linguistic invention and convention. Any reality that emerges and becomes paramount is both thin but substantial enough to attract attention and be taken seriously.

The problem is that language is disappearing in machine learning and AI because models like ChatGPT and BERT works on tons of 'words'. This approach is not about language, or words, but the correlational patterns among *things* called *words*. Large language models are used currently in an attempt to make AI literate, that is, to speak as persons do in everyday life. But persons do not speak by linking words to objects in the environment. Instead, in line with contemporary philosophy, persons are connected to the world through a seamless use of language. For this reason, critics such as Bender et al. (2021) contend that these models are operating outside of language and are not likely to ever perform like humans. What they are saying is that the language game is not currently a part of AI development

As might be suspected by now, black boxes do not survive this critique. In the absence of dualism, these places are mediated thoroughly by language and lose their unique, autonomous status. As a result, they should not be treated blithely as unknown or beyond comprehension. They are clearly in reach of examination, due to their fundamental connection to, and origin in, human agency. To borrow from Sartre (2001), anyone who claims that black boxes are beyond scrutiny is acting in '*bad faith*,' that is, denying the connection to human interests that is obvious. In a manner of speaking, consistent with anti-dualism, black boxes disappear as an excuse for not pursuing rigorous investigations of algorithms.

CONCLUSIONS

After the rejection of dualism, all algorithms are white boxes, or accessible. With nothing beyond human comprehension, complexity is no longer a reason to evade a difficult task. A complex or time consuming investigation is not synonymous with wandering in the unknown. Former black box actions can be traced to the interactions of variables and (re)tested; logic is operating that can be reported. For example, the filters can be identified that are supporting the decisions that group data, even at the deeper levels of neural nets. With no place to hide, the logic of decisions cannot remain concealed.

This shift prompts a new realisation about AI. Specifically, the recognition of black boxes only encourages the mystification of this technology. Explainable AI (XAI) is striving to avoid this inevitability (Bleicher 2017; Varma 2021). But this movement needs a guiding philosophy. The point must be made clear that no human creation is inherently unknowable if this attempt to reclaim AI is going to succeed. There is a story operative at the basis of algorithmic operations. These narratives may not follow an exact chain of rules, and may even skip steps. But nothing mysterious is happening.

On a practical note, one possibility is to apply '*adversarial design*' to the creation of algorithms (Morozov 2013). What Morozov is suggesting is to think against the grain in AI development. He believes that by exploring alternative design options, a critical awareness can be promoted about how to foster AI that is compatible with human desires and avoid the reductionism associated with black box strategies. His point is move beyond consensus and safe options – the usual format – to push the boundaries of AI development to include a wide range of individual or community experiences. In effect, Morozov (2013: 328) is referring to a philosophical shift that is necessary to increase personal and collective awareness and human flourishing. Different from Explainable AI, Morozov is attempting to offer an alternative, non-dualistic philosophy to support a humane AI.

Received 19 July 2023
Accepted 23 October 2023

References

1. Bakewell, S. 2016. *The Existentialist Café*. New York: Other Press.
2. Belsey, C. 2022. *Poststructuralism: A Very Short History*. Oxford, UK: Oxford University Press.
3. Bender, E. M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. 2021. 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAccT '21), 610–623. Available at: https://doi.org/10.1145/3442188.3445922 (accessed 18.07.2023).
4. Bleicher, A. 2017. 'Demystifying the Black Box that is AI', *Scientific American*. Available at: https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/ (accessed 12.05.2023).
5. Buckley, K. W. 1989. *Mechanical Man: John Broadus Watson and the Beginnings of Behaviorism*. New York: Guilford Press.
6. Cassauwers, T. 2020. 'Opening the "Black Box" of Artificial Intelligence', *Horizon*. Available at: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/opening-black-box-artificial-intelligence (accessed 12.05.2023).
7. Deleuze, G. 1994. *Difference and Repetition*. New York: Colombia University Press.
8. Dreyfus, H. L. 1972. *What Computers Can't Do*. New York: Harper and Row.
9. Dreyfus, H. L.; Dreyfus, S. E. 1986. *Mind over Machine*. New York: Free Press.
10. Durán, J. M.; Jongsma, K. R. 2021. 'Who is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI', *Journal of Medical Ethics* 47(5): 329–335. Available at: https://jme.bmj.com/content/medethics/47/5/329.full.pdf (accessed 12.05.2023).
11. Estevez, E. 2022. 'What is a Black Box Model? Definition, Uses, and Examples', *Investopedia*. Available at: https://www.investopedia.com/terms/b/blackbox.asp (accessed 12.05.2023).
12. Fish, S. 1979. 'A Reply to John Reichert: Or How to Stop Worrying and Learn to Love Interpretation', *Critical Inquiry* 6(1): 173–178.
13. Foucault, M. 1979. *Discipline and Punish: The Birth of the Prison*. New York: Pantheon.
14. Gebser, J. 1984. *The Ever-present Origin*. Athens, OH: Ohio University Press.
15. Gombay, A. 2007. *Descartes*. Malden, MA: Blackwell Publishing.
16. Hamon, R.; Junkelwitz, H.; Sanchez, I. 2020. *Robustness and Explainability of Artificial Intelligence*. European Union Technical Report. Luxembourg: Publications Office of the European Union. Available at: https://publications.jrc.ec.europa.eu/repository/handle/JRC119336 (accessed 12.05.2023).
17. Heisenberg, W. 1958. *Philosophy and Physics: Revolution in Modern Science*. New York: Harper.
18. Husserl, E. 1964. *The Paris Lectures*. The Hague: Nijhoff.
19. Kant, I. 2007. *Critique of Pure Reason*. New York: Penguin.
20. Kuhn, T. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
21. McDermid, J. A.; Jin, Y.; Porter, Z.; Habil, I. 2021. 'Artificial Intelligence Explainability: The Technical and Ethical Dimensions', *Philosophical Transactions of the Royal Society of London/A: Mathematical, Physical and Engineering Sciences*. Available at: https://pubmed.ncbi.nlm.nih.gov/34398656/ (accessed 12.05.2023).
22. McLeod, S. A. 2007. 'Behaviorist Approach to Psychology: Definition, History, Concepts, and Impact', *Simply Psychology*. Available at: https://www.simplypsychology.org/behaviorism.html (accessed 19.07.2023).
23. McNamara, M. 2022. 'Explainable AI: What is it? How does it Work? And what Roles does Data Play?', *NetApp*. Available at: https://www.netapp.com/blog/explainable-ai/ (accessed 12.05.2023).

24. McTaggart, L. 2011. *The Bond. How to Fix Your Falling-Down World*. New York: Free Press.
25. Merleau-Ponty, M. 1973. *The Prose of the World*. Evanston: Northwestern University Press.
26. Miller, T. 2019. 'Explanation in Artificial Intelligence: Insights from the Social Sciences', *Artificial Intelligence* 267: 1–38. Available at: https://doi.org/10.1016/j.artint.2018.07.007
27. Mir, R.; De Blanc, M. 2023. 'Open Data and the AI Black Box,' *Electronic Fronter Foundation*. Available at: https://www.eff.org/deeplinks/2023/01/open-data-and- ai-black-box (accessed 12.05.2023).
28. Morozov, E. 2013. *To Save Everything, Click Here*. New York: PublicAffairs.
29. Mullainathan, S.; Spiess, J. 2017. 'Machine Learning: An Applied Econometric Approach', *Journal of Economic Perspective* 31(2): 87–106.
30. Nietzsche, F. 1997. *Daybreak*. Cambridge, UK: Cambridge University Press.
31. Pollner, M. 1991. 'Left of Ethnomethodology: The Rise and Decline of Radical Reflexivity', *American Sociological Review* 56(3): 370–380.
32. Poster, M. 1989. *Critical Theory and Poststructuralism*. Ithaca, New York: Cornell University Press.
33. Poster, M. 2001. *What's the Matter with the Internet*. Minneapolis: University of Minnesota Press.
34. Power, R. 2023. 'No Black Boxes: Keep Humans Involved in Artificial Intelligence,' *Forbes*. Available at: https://www.forbes.com/sites/rhettpower/2023/01/15/no-black-boxes-keep-humans-involved-in-artificial-intelligence/?sh=3601be9674fa (accessed 12.05.2023).
35. Rorty, R. (ed.). 1967. *The Linguistic Turn: Recent Essays in Philosophical Method*. Chicago: University of Chicago Press.
36. Sartre, J.-P. 2001. *Being and Nothingness*. New York: Citadel Press.
37. Savage, N. 2022. 'Breaking into the Black Box of Artificial Intelligence', *Outlook*. Available at: https://www.nature.com/articles/d41586-022-00858-1 (accessed 12.05.2023).
38. Shashmi, K. 2021. 'Curse of Dimensionality – Curse to Machine Learning', *Toward Data Science*. Available at: https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33b-feb (accessed 12.05.2023).
39. Stahl, B. C.; Andreou, A. M.; Brey, P.; Hatzakis, T.; Kirichenko, A.; Macnish, K.; Laulhé Shaelou, S.; Patel, A.; Ryan, M.; Wright, D. J. 2021. 'Artificial Intelligence for Human Flourishing – Beyond Principles for Machine Learning', *Journal of Business Research* 124: 374–388. Available at: https://www.sciencedirect.com/science/article/pii/S0148296320307839?via%3Dihub (accessed 18.07.2023).
40. Varma, G. 2021. 'The Philosophy Behind AI Explainability', *Geek Culture*. Available at: https://medium.com/geekculture/the-philosophy-behind-ai-explainability-a774d084bbc3 (accessed 13.05.2023).
41. Winograd, T.; Flores, F. 1986. *Understanding Computers and Cognition*. Norwood, New Jersey: Ablex.
42. Wittgenstein, L. 1953. *Philosophical Investigations*. New York: MacMillan.
43. Wittgenstein, L. 1960. *The Blue and Brown Books*. New York: Harper and Row.
44. Wittgenstein, L. 1990. *Tractatus Logico-Philosophicus*. New York: Routledge.

JOHN W. MURPHY, CARLOS LARGACHA-MARTÍNEZ

Juodosios dėžės, ribojančios žmogaus klestėjimą, nebegalimos naudoti kuriant dirbtinį intelektą (DI)

Santrauka

Daugelis kritikų dirbtinį intelektą laiko labai abstrakčiu. Šiuo atžvilgiu algoritmai įprastai vadinami juodosiomis dėžėmis ir atskiriami nuo žmogaus įsikišimo. Tam tikras filosofinis manevras palaiko šį rezultatą. Šio straipsnio tikslai yra 1) atskleisti filosofinį požiūrį, prisidėjusį prie šio atstumo tarp DI ir žmonių, ir 2) pasiūlyti alternatyvią filosofinę poziciją, kuri gali priartinti šią technologiją prie individų ir bendruomenių. Bendras šio straipsnio analizės tikslas yra humanizuoti DI, pašalinant algoritmų, kaip juodųjų dėžių, konceptualizavimo trūkumus.

Raktažodžiai: dualizmas, dirbtinis intelektas, paaiškinamumas, kalbinis posūkis, antidualizmas