Scan to know paper details and
author's profile

On a New Ethics of AI and Moral Progress

I.E. Pris

ABSTRACT

The "new ethics" of artificial intelligence proposed by M. Gabriel is critically evaluated. It is argued that, unlike human intelligence, artificial intelligence (AI) is devoid of normative dimension, or, equivalently, of sensitivity to context. Gabriel's view conflicts with J. Benoist's contextual realist approach to ethics and T. Williamson's moral realism, according to which it is not principles that are primary but moral perception in context, paradigmatic examples of moral knowledge. The approaches of Gabriel, D. Andler, L. Floridi, S. Russell to AI are considered and compared. It is proposed to adopt Andler's principle of moderation. It is argued that AI systems imitate intelligence, agency, autonomy, ethics. A realistic conception of AI is contrasted with its idealistic conception.

Keywords: AI ethics, moral progress, autonomy, context, normativity, moral realism.

Classification: LCC Code: BJ1188

Language: English

Great Britain
Journals Press

LJP Copyright ID: 573314
Print ISSN: 2515-5784
Online ISSN: 2515-5792

London Journal of Research in Humanities & Social Science

Volume 25 | Issue 2 | Compilation 1.0

© 2025. I.E. Pris. This is a research/review paper, distributed under the terms of the Creative Commons Attribution- Noncom-mercial 4.0 Unported License http://creativecommons.org/licenses/by-nc/4.0/), permitting all noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

On a New Ethics of AI and Moral Progress

I.E. Pris

ABSTRACT

The "new ethics" of artificial intelligence proposed by M. Gabriel is critically evaluated. It is argued that, unlike human intelligence, artificial intelligence (AI) is devoid of normative dimension, or, equivalently, of sensitivity to context. Gabriel's view conflicts with J. Benoist's contextual realist approach to ethics and T. Williamson's moral realism, according to which it is not principles that are primary but moral perception in context, paradigmatic examples of moral knowledge. The approaches of Gabriel, D. Andler, L. Floridi, S. Russell to AI are considered and compared. It is proposed to adopt Andler's principle of moderation. It is argued that AI systems imitate intelligence, agency, autonomy, ethics. A realistic conception of AI is contrasted with its idealistic conception.

Keywords: AI ethics, moral progress, autonomy, context, normativity, moral realism.

Author: Institute of Philosophy of the National Academy of Sciences, Minsk, Belarus.

I. INTRODUCTION

1. I would distinguish three interrelated meanings of the term "AI ethics".[1] Firstly, the study of ethical issues related to the production and use of AI. Secondly, the study of the possibilities of creating intrinsically ethical AI, that is, ethical AI by its design. The most general principles of AI ethics are the same as in medical ethics (beneficence, non-maleficence, justice, (human) autonomy), plus the AI-specific principle of explainability. The principles may vary slightly [1; 2]. Thus, the authors of a recent article mention six principles: "freedom [they also talk about human agency, which encompasses freedom, autonomy, and dignity], privacy, fairness, transparency, accountability, and well-being (of individuals, society, and the environment)" [3, p. 1267–1268]. To these can be added harmlessness, responsibility and some other principles. These abstract principles are supplemented by more operational principles. Finally, thirdly, there is the question of an AI that would have the capacity to discover or produce new ethical values.

The essence of a new AI ethics, or a new Enlightenment ethics, proposed by the German philosopher M. Gabriel, as I understand it, is to create, in the process of global cooperation of different cultures with different values, a powerful ethical AI by its design, a kind of Alpha Buddha or Alpha Jesus, which would discover or at any rate help man to discover and socio-economically implement new moral facts and laws (including those concerning the AI itself), i.e. would actively contribute not just to radical changes in society, but to rationally controlled, scientifically guided moral progress. Such an AI is seen by Gabriel as a system for universalising morality, helping us to understand who we are as human beings, who we want to be and who we should become [4].

I have some reservations and concerns about this project, particularly regarding the possible loss of human autonomy, at least in part.

2. But first of all, what is the relationship between AI and human intelligence? I interpret the relationship between them in terms of a categorical distinction between the ideal (normative) and the real. This distinction can also be explained in terms of the Wittgensteinian rule-following problem. AI follows formal (machine) rules [5–6]. A similar view was

[1] The term "artificial Intelligence" (AI) is conveniently defined based on the way it is currently used primarily by specialists, but also by the wider public. This includes not only programs, algorithms, programmed computers and robots (AI systems), but also relevant laboratories, institutes, projects and so on. Usually, depending on the context in which the term is used, it is clear what we are talking about. In the future, perhaps the term will also denote some new common property shared by all AI systems: "artificial intelligence".

©2025 Great Britain Journals Press

defended by S. G. Shanker in his book "Wittgenstein's Remarks on the Foundations of AI" back in 1998 [7].

It is also consistent with the fact that for the French philosopher D. Andler, AI relates to humans like a shadow to a cowboy, and for Gabriel like a map to a territory [1; 4]. For Gabriel, AI is a model of thought. It has an artificial rather than neurobiological basis [8]. The discrepancy between me and Gabriel is that for him thought is real, something like a non-natural human sixth sense[2], not an informational process that has no reality of its own (from this point of view AI does not think), whereas for me it is ideal, but this implies its rootedness in reality, including neurobiological reality (according to the conceptual grammar of the concept of thought) [8].

Earlier I argued that AI is not intelligence and within the existing naturalistic paradigm it will never be, because it lacks a normative dimension, or equivalently, sensitivity to context. The idea of transhumanism is a myth. The so-called moment of singularity will never come [5; 6].[3] At the same time, the Promethean project of creating an autonomous AI in the image and likeness of a human is a threat and should be abandoned. D. Andler takes a similar position: context has a normative dimension, and intelligence is normativity[4], while AI is only capable of solving problems, which is a secondary task for human intelligence [1; 10].

M. Gabriel, on the contrary, defines intelligence as the ability to solve problems. In this sense, AI can be smarter than humans, although it does not possess the highest form of thinking – reflective thinking. Also, Gabriel sometimes says that no one knows what thinking/thought is. "If thinking is something more abstract, a process in reality not essentially tied to brains and their parts, AI systems could in principle become or already be real thinkers" [4]. (In this case the model (the AI system) would belong to the same reality as the target system (human thinking).)

According to the Italian philosopher L. Floridi, the question of whether AI thinks or not does not matter [11]. What matters is what AI does and is able to do. Floridi believes that AI does not think, but is an agent. AI is a new kind of agency. It is a non-human, mindless agency that transforms the environment and requires its transformation (semanticization). Otherwise, AI could not exist and be used. But if by agency we mean the ability to perform full-fledged actions, I wouldn't call AI systems agents. Actions, like judgments, are normative. Only humans are capable of them.

3. According to Gabriel's new moral realism, there are universal, a priori, absolute and unchanging moral principles, which are first discovered and then applied in a context external to them [12]. This neoclassical approach to morality contradicts the realist contextual approach of the French philosopher J. Benoist, which I share, and the moral realism of the British philosopher T. Williamson, who criticizes moral inferentialism [13; 14]. A more general position – moral principlism – is also problematic (different principles may contradict each other, be interpreted differently, and their applicability depends on the context). In fact, it is not principles that are primary, but moral perception in context, paradigmatic examples of moral knowledge [13].

The Williamsonian critique of internalism and coherentism in epistemology, as well as the Wittgensteinian critique of the notion of an absolute moral fact that would contain all its applications, should also be taken into account here. Ethics cannot do without ontology (moral facts), but neither can it be reduced to ontology. The factual, what is cannot tell us about the normative, about what ought to be. In other words, the introduction of a Platonizing (ideal),

[2] For this reason, for Gabriel, human intelligence is "artificial intelligence" (but certainly not in the sense in which we speak of AI) [8].

[3] Among the contemporary philosophers, the same point of view is held, for example, by M. Gabriel, D. Andler, L. Floridi, M. Bitbol. The opposite point of view is held, for example, by D. Chalmers [9].

[4] "Intelligence is not a thing, not a phenomenon, not a process and not a function, but a norm that applies to behavior: it qualifies the relationship between a human and her world, and in a way that is never objective and definitive (...)." [1, p. 12].

but non-metaphysical, dimension is necessary [13]. But this is precisely what AI is devoid of by definition.

Gabriel's AI new ethics seems to me to imply Gabriel's general approach to morality [12]. But if an AI is not sensitive to context (otherwise it would not be an AI, but a human being, or perhaps some autonomous non-human intelligence with non-human morality), much less a moral one, and the essence of morality is such sensitivity, the question arises about the possibility of implementing Gabriel's proposed program of moral progress with the help of an AI and the potential consequences of attempts to implement it. Perhaps Gabriel's moral project of "Be progressive!" should be replaced by a more moderate project.

4. Classical symbolic AI is a program, an algorithm, an extended logic. Connectionist AI, which replaced it, is an artificial neural network. The philosophy of the former is rationalism ("everything is logic!"); the philosophy of the latter is empiricism ("everything is perception!"), although it includes essential elements of symbolic AI. So-called "deep learning" and "large language models" (Chat GPT, etc.) are a contemporary development of connectionism. Presumably AI of the near future will synthesize both approaches. The philosophy of such hybrid AI can be conventionally compared to Kant's critical synthesizing rationalism and empiricism.[5] Accordingly, ethics can be built into AI from the top down (it seems that this approach is closer to Gabriel's one), but it can also be built into it from the bottom up, by training the AI on large amounts of empirical data.

Thus, S. Russell suggests an alternative to principlism. The essence of his approach is to orient AI ethics to human preferences, which would be revealed from statistical data on human behavior [20, ch. 7]. This approach – inductivism – is, as Andler notes, based on illusions. In fact, it is not possible to identify human preferences purely statistically, behavior is not determined by preferences alone, and finally, the future does not always have to be determined by the past – as something that has a high probability of occurrence (this is not true in crisis and intractable situations, as well as in science and art) [1, p. 223].

5. AI is a new kind of reality. However, it does not exist by itself (absolutely), but is integrated into socio-economic and material relations, practices, that is, it has real conditions for its existence. If we stop caring about it, it will disappear. AI is a complex technology. As is known, when a complex technology is used by a large number of independent agents, there are situations when not the agents control the technology, but the technology controls the agents, which indicates its reality.

There is a general problem of control of AI and, in particular, the problem of alignment of AI ethics and human ethics. We are not able to fully control AI. So we want at least the values of AI to match or harmonize with those of humans. This problem may turn out to be unsolvable [1, § 10.5].[6] The dilemma here is as follows: either we design AI systems that cannot solve complex problems

[5] Already after writing this article I learned that a similar comparison is made by R. Evans. He writes: "The neural network is the intellectual ancestor of empiricism, just as logic-based learning is the intellectual ancestor of rationalism. Kant's unification of empiricism and rationalism is a cognitive architecture that attempts to combine the best of both worlds, and points the way to a hybrid architecture that combines the best of neural networks and logic-based approaches" [15, p. 41]. Some believe that the Kantian categorical imperative can be formalized, algorithmized, and implemented in AI (see, e.g., [15–17]). Others conclude that the AI cannot be a Kantian moral agent in the real sense of the term because it cannot possess autonomy or the power of reasoning in the Kantian sense [18]. Within my contextual/normative approach, the latter conclusion is obvious. At the same time, AI that imitates an ethical agent is possible and has practical use. For example, the author of one article argues that AI can be (moral)reasons-responsive, make (moral) judgments, and make (moral) decisions. At the same time, he argues that AI cannot be an authentic, or responsible, (moral) agent [19]. While agreeing only with the latter, I note that authentic reasons-responsiveness, judgments, and decisions are normative, whereas for AI they are purely causal.

[6] The literature also discusses the "responsibility gap problem" related to the alignment problem, which raises the question of who bears responsibility for unpredictable actions performed by self-learning (quasi-)autonomous AI. In my view, the attempt to shift the responsibility, at least partially, to the AI is untenable.

that we cannot solve without AI help, but would like them to be solved, or we design AI systems that can solve complex problems, but at the same time turn out to be at least partially (quasi-)autonomous. The problem is that it is impossible to impose values on an (quasi-) autonomous system from the outside by definition. It chooses its own values and chooses whether or not to accept the values offered to it.

An aspect of the alignment problem is the problem of determining which human values should be prioritized for alignment, whose values should be encoded in AI systems. This is the problem of "value pluralism, in which different individuals and cultures hold diverse, conflicting and irreducible values. Undemocratic value alignment excludes the users from acting as full epistemic agents, and as a result, full moral agents" [21, p. 4, § 3]. It is difficult, if not impossible, to make AI simultaneously take into account the interests of society as a whole, different groups of people, and different individuals.[7] And also there are various normative ethical theories. A thought experiment with a quasi-autonomous (self-driving) car as a version of the classic thought experiment of the trolley problem illustrates this problem. Depending on the system of normative ethics embedded in the AI program – deontological or utilitarian, as well as their interpretations, – the AI will "act" one way or the other in some well-defined (corresponding to the AI algorithm) situations. (See analysis of the problem, in particular, in the Kantian perspective, for example, in [21, 24, ch. 6–8].)

6. But even if AI systems were relatively safe, we might become dependent on them, because once we lived in a world transformed for them, we could no longer do without them. This raises the question: Do we want to live in a world made for machines and not be able to do without them?

Andler, for example, puts forward the principle of moderation: "Use artificial intelligence only when the risks are reduced and the benefits are significant; use AI systems that are as simple as possible and capable of providing the expected service" [1, p. 224]. This principle, in particular, implies the following: Use AI only when its net contribution will be positive. Do not assign it tasks that can be accomplished without AI. Do not give it a humanoid appearance. Do not use it where human intelligence is required, i.e. not just the ability to solve problems. In particular, do not assign it tasks whose solution requires wisdom.

Quantum logic, in a sense, takes into account the inherent non-(pre)determinacy and contextuality of human decisions and actions. One can therefore assume that the quantum or the quantum-like AI based on it will be human-like [26]. But , according to my argument, it will never become intelligent and ethical, nor will it come close to a human being, because context is not reducible to logical operations.

AI imitates intelligence, ethics, autonomy, agency/action.[8] Conceptual confusions of the artificial and the natural, the ideal and the real have undesirable consequences, both theoretical and practical. One of the tasks of AI philosophy is precisely to separate one from the other, to emphasize as much as possible the differences between AI and humans. Anything that AI can or will be able to do, no matter how advanced, is not part of human nature. In other words, we need a realistic, not idealistic, conception of AI.

[7] The later philosophy of Wittgenstein is applied to the alignment problem in [22]. It is proposed to take into account psychological, social, and cultural contexts, their variability. While this approach allows us to reduce the severity of the problem, it is, I claim, based on an imitation of sensitivity to context. There is no genuine rule-following here in the sense in which Wittgenstein understands it. As for imitating Wittgensteinian AI, it is possible, but more difficult than imitating Kantian AI (see the attempts to use the resources of Kant's philosophy to improve the "cognitive" and "ethical" abilities of AI in [15; 23–25]).

[8] One might say, "But it's obvious!" And, from my point of view, it really is. The philosophical study of AI does not so much prove the absence of AI's genuine intelligence, ethics, etc., as it tries to reveal what is not AI, i.e., the nature of natural intelligence, human beings. Kant, as we know, considered the question "What is man?" to be the key question of philosophy. At the same time, the unpredictability of AI does not allow us to consider that AI is only an imitation of natural intelligence. AI systems can also be seen as a new kind of reality, for which traditional concepts acquire a different meaning. For example, one can introduce a non-anthropomorphic notion of a trustworthy AI [27].

REFERENCES

1. Andler A. Intelligence artificielle, intelligence humaine: la double énigme.–Paris: Gallimard, 2023. – 432 p.
2. Coeckelbergh M. AI ethics. – The MIT Press, 2020. – 248 p.
3. Brey P., Dainow B. Ethics by design for artificial intelligence // AI and Ethics. – 2024. – 4. – P. 1265–1277.
4. Gabriel M. The new ethics of AI. – 2024. – URL: https://www.youtube.com/watch?v=ynBU7untOoc.
5. Pris I. E. M. Gabriel's neo-existentialism against artificial intelligence // II Forum IT-Akademgrad "Artificial intelligence in Belarus". 12-13 October 2023 / Eds.: S. V. Kruglikov, S. N. Kasanin. – Minsk: The United Institute of Informatics Problems of the National Academy of Sciences of Belarus (UIIP NAS of Belarus), 2023. – 276 p. – P. 208–216. (In Russ.)
6. Pris I E. Artificial intelligence is not and will never be intelligence // Science and innovations. – 2024. – 9 (259). – P. 26–29. (In Russ.)
7. Shanker S. G. Wittgenstein's Remarks on the Foundations of AI. Routledge, 1998. – 280 p.
8. Gabriel M. Der Sinn des Denkens. – Ullstein Taschenbuch, 2018. – 368 S.
9. Chalmers D. The Singularity: A Philosophical Analysis // Journal of Consciousness Studies.–2010. – Vol. 17. – P. 7–65.
10. Andler D. The normativity of context// Philosophical Studies. – 2000. – Vol. 100. – P. 273–303.
11. Floridi L. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities.– Oxford University Press, 2023. – 272 p.
12. Gabriel G. Moralischer Fortschritt in dunklen Zeiten. – Ullstein: Berlin, 2020. – 352 S.
13. Pris I. E. Contextual moral realism // Siberian journal of philosophy. – 2023. – Vol. 21 –№ 4. – P. 5–28. (In Russ.)
14. Williamson T. Unexceptional moral knowledge // Journal of Chinese Philosophy. – 2022. – Vol. 49, № 4. – P. 405–415.
15. Evans R. (2022). "The Apperception Engine". In: Dieter Schцnecker/Hyeongjoo Kim (Eds.) Kant and Artificial Intelligence. De Gruyter, pp. 39–103.
16. Powers T., M.: Prospects for a Kantian machine. IEEE Intell. Syst. 21, 46–51 (2006)
17. Lindner, Felix and Bentzen, Martin Mose. 2019. "A Formalization of Kant's Second Formulation of the Categorical Imperative." *Journal of Applied Logic*. https:// arxiv. org/ abs/ 1801. 03160.
18. Chakraborty, A. Can artificial intelligence be a Kantian moral agent? On moral authonomy of AI system A. Chakraborty, N. Bhuyan // AI ethics.–2023.–URL: https://doi.org/10.1007/s43681-023-00269-6.
19. Gudmunsen Z. The moral decision machine: a challenge for artificial moral agency // AI ethics: [Электронный ресурс].–2024. – URL: https://doi.org/10.1007/s43681-024-00444-3 (date of access: 25.12.2024).
20. Russell S. Human compatible. Artificial intelligence and the problem of control / S. Russell. – New York: Penguin Books, 2019.
21. Huang L. T.-L., Papyshev G., Wong J. K. Democratizing value alignment: from authoritarian to democratic // AI ethics. – 2024. – URL: https://doi.org/10.1007/s43681-024-00624-1 (date of access: 12.12.2024).
22. Perez-Escobar J. A., Deniz Sarikaya D. Philosophical Investigations into AI Alignment: A Wittgensteinian Framework // Philosophy & Technology. – 2024. – 37:80.
23. McDonald F. J. AI, alignment, and the categorical imperative // AI and Ethics: [Электронный ресурс]. – 2023. – 3:337–344. – URL: https://doi.org/10.1007/s43681-022-00160-w.
24. Schцnecker D., H Kim (Eds.) (2022). Kant and Artificial Intelligence. De Gruyter.
25. Schlicht T. (2022). "Minds, Brains, and Deep Learning: The Development of Cognitive Science Through the Lens of Kant's Approach to Cognition". In: Dieter Schцnecker/ Hyeongjoo Kim (Eds.) Kant and Artificial Intelligence. De Gruyter, pp. 3–38.
26. Pris I.E. Quantum-like modeling and its philosophical foundations. Philosophy of science [Filosofia nauki]. 2024. No. 3(102). P. 109–129 (In Russ.)

27. Simion M., Kelp C. Trustworthy artificial intelligence // Asian Journal of Philosophy. – 2023. – 2:8. – URL: https://doi.org/10.1007/s44204-023-00063-5 (дата обращения 25.12.2024).