# Ethical AI: Conceivable Versions and Projected Issues

Vadim Perov
*Institute of Philosophy*
*Saint-Petersburg State University*
Saint-Petersburg, Russia
vadimperov@gmail.com

Nina Perova
*Institute Of Philosophy*
*Russian Academy of Science*
Moscow, Russia
nino4kaperova@gmail.com

2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) | 979-8-3503-6289-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/USBEREIT61901.2024.10584050

***Abstract*—The article presents the results of an analysis of the various versions of ethical Artificial Intelligence (AI), presented in modern theoretical discussions and attempts at practical implementation, including in the context of the emergence of GPT. The following generalized models have been identified. First, the Artificial Moral Agent (AMA), which is usually considered in two forms: the real AMA, which has all or some moral qualities, and the ascriptive AMA, to which moral qualities are normatively attributed. Secondly, Ethically aligned AI, which is designed so that its functioning complies with moral norms and values accepted in the community. Thirdly, the moral aspects of creating an Artificial Ethical Assistant (AEA) are considered, which purports to be an all-knowing and impartial Ideal Observer in the field of ethics and is able to advise people to solve moral problems. This option for ethical AI comes in the form of general and personal AEAs. Special attention is paid in the article to the ethical advantages and disadvantages, as well as the ethical risks of each of the options considered. The risks of ethical paternalism are highlighted among the most significant. It imposes certain moral values on people, which creates conditions for the manipulation of people's freedom, which can lead to immoral behavior. The problem is that the massive use of AI, including the uncontrolled spread of GPT, can strengthen these negative trends, reproducing existing ethical bias and generating new ones, which requires special attention in creating ethical AI.**

***Keywords—ethical artificial intelligence, artificial moral agent, ethically aligned artificial intelligence, artificial ethical assistant, moral bias, ethical risk, paternalism.***

## I. Introduction

One of the most important publicly significant IT events of 2023 was the massive use of artificial intelligence (AI) technologies based on GPT (Generative pre-trained transformer). In the broadest and most general sense, GPT is a set of variations of the Large Language Model (LLM) that are capable of generating texts similar to those created by people by processing large amounts of data. Thanks to their widespread use, such neural network technologies have already received their well-deserved recognition, in particular in the field of machine translation of foreign languages. GPT itself was announced by OpenAI employees in 2018 [1], but initially it was used primarily for the development of various software products by IT companies. The situation changed dramatically when ChatGPT was released for open mass use (30.10.2022). Users were even more impressed by the release of the next version, based on GPT-4 (14.03.2023), with which you can not only get answers to your questions, but also create texts, write poetry, etc., but also generate drawings and music. Currently, the number of various analogues developed by both IT giants and small companies and startups is difficult to count. For most people, it is GPT that is strongly associated with such AI, which in its capabilities is comparable to the human mind (which, of course, is not true, but we are talking about such assessments of it). This differs from the perception of other AI technologies that are already used in finance and banking, data processing systems used by government or business structures, including recognition and identification algorithms, as well as data-driven predictive and recommendation systems. Despite the widespread use of such technologies, even with their relative autonomy and possible anthropomorphization, they are still not perceived as intelligent beings, as can be seen in the examples of home devices such as a washing machine or robotic vacuum cleaner, as well as in relation to unmanned vehicles. It was GPT that turned out to be a kind of trigger that increased interest in AI, including in terms of discussing the social and ethical risks of its use. In the public space of traditional media, "new media" and social networks, heated discussions have unfolded, ranging from questions about the possible "death" of a number of professions to environmental problems of a sharp increase in the use of fresh water for cooling servers and data centers due to the growing number of Internet access to GPT technologies. One of the public reactions to possible social and ethical risks was an open letter signed by many leading experts that appeared on March 22, 2023 on the Future of Life Institute website, which called for a temporary suspension of research and development in the field of AI to assess possible dangers and risks [2].

When it comes to the risks associated with AI, along with ethically significant risks (cybercrime, security, military use of AI, threat to the privacy and autonomy of people, impact on the economy and social sphere, transformations in the labor market, etc.), active discussion is aimed at the possibility that such technologies have, or could potentially have, some moral characteristics. In other words, the question is raised about the possibility of creating ethical AI [3]. Analysis of modern research and public discussions allows us to identify the following possible meanings of "ethical AI".

## II. Artificial Moral Agent.

### A. Real Artificial Moral Agent (AMA).

To begin with, it is worth dwelling on the idea of the existence of AI as the possessor of all or some moral qualities in the full sense of the word [4]. It should be noted right away

Financial support: Russian Science Foundation, project No. 24-28-00562 "Philosophical foundations of ethical risks in the field of artificial intelligence systems".

that this version of ethical AI is included in the list rather for the sake of completeness, since such technologies currently do not exist or are expected, which is reflected without exception in all ethical documents designed to regulate ethical issues in the field of AI, the moral requirements of which are intended only for people. Modern AI can outperform humans in games (chess, go, Jeopardy) or in facial recognition, they generate texts, music and pictures better than many people, but they do not even have anything close to what can be considered intellectual abilities, especially something similar to moral intelligence [5]. In addition, there are sufficient reasons to believe that it is theoretically impossible to introduce ethical standards into AI so that it truly becomes an AMA [6]. However, this problem is being discussed, and the most frequently identified ethical risks include not just the idea of controllability, but also the possibility of AI technologies "going beyond the boundaries of anthropocentric morality through the creation of their own moral norms and values" [7]. Of course, this scenario for the development of events is too unrealistic since currently no AI possesses it, but its significance lies in the heuristic potential of critical reflection on the existing morality, assessing its stability in relation to possible risks.

### *B. Ascriptive AMA*

The difference between ascriptive AMA and real AMA is that AI systems are not considered to actually have any moral qualities, but due to the presence of characteristics that successfully imitate mental, intellectual, emotional, etc. properties of people, they can be attributed to "being moral" in some meanings. It should be noted that the experience of normative attribution of human characteristics to artificial objects already has a stable historical practice. Important in this context is the concept of "legal entity" (legal or artificial person) in relation to organizations. The latter, although not people (natural or physical persons), may be subject to not only legal, but also some moral requirements (for example, within the framework of the so-called Corporate Social Responsibility), which turns them into moral agents. The widespread of digital technologies in modern society is gradually expanding the scope of interaction with AI systems, and the latter are gradually turning from a mediator in relations between people into autonomously acting agents, who can be normatively assigned the status of AMA [8]. In this context, the most important ethical risks include those that directly or indirectly stem from the general problems of anthropomorphizing AI. Firstly, there is a set of issues related to moral responsibility, primarily the danger of the diffusion (erosion) of responsibility. It should be noted that this phenomenon is also inherent in relation to collective entities, but when it comes to legal entities, then the active entities there always remain people acting on behalf of legal entities. AI systems have a fairly high degree of autonomy. No one acts on their behalf, and they are not anyone's representatives. Secondly, there are risks that can be collectively called "overtrust", when people delegate tasks to AI technologies, the solution of which has a significant impact on them or on other people, and at the same time do not control the processes and results. Although there are currently no normatively established "ascriptive AMAs", the mentioned ethical risks are already real. In this regard, there is a theoretical and practical need, on the one hand, to determine the possibilities of ethical ascription in relation to AI systems, on the other hand, to carry out demarcation in order to avoid excessive moral anthropomorphization.

## III. Ethically aligned AI.

One of the possible options for understanding ethical AI is related to the processes of its creation. In this case, we are not talking about the moral properties of AI themselves in both senses indicated earlier, but about those ethical principles and requirements that should be incorporated into them by developers in order to (a) make them consistent with existing ethical standards and values and (b) make it impossible to use them for unethical purposes. The most indicative in this regard are the ideas developed by IEEE [9]. This text articulates the central idea of this approach that if machines interact with people as quasi-autonomous agents, then these agents are expected to follow existing social and moral norms. AIs that function independently and uncontrollably require special "ethical tuning" on the part of developers. The ethical risks that arise in this case are primarily due to the moral standards in accordance with which this setting is carried out: the ethical views of the developers may not coincide with the moral views of the community with which they must be aligned. As a result, there is a risk of creating biased and paternalistic AI that imposes the moral values of its creators on people. This creates opportunities for direct or indirect manipulation, which violates people's freedom, dignity and autonomy.

## IV. Artificial Ethical Assistant

The emergence of voice assistants, ChatGPT and its analogues, etc. encourages people to use them to find answers to questions in many areas, including in the field of morality. The IT industry cannot ignore requests of this kind, and therefore the idea of the possibility of creating an AI-based "Artificial Ethical Assistant" (AEA). The brief essence of the ideas being discussed is as follows: AEA will be a unique embodiment of ethical wisdom in the fields of moral philosophy, applied and professional ethics throughout their history, the bearer of knowledge about the entire moral experience of mankind, a source of information about the moral preferences of people of different countries and cultures existing in modern societies etc. It is difficult to imagine that such a volume of knowledge could be available to any single person. In addition, the predicted advantages include the potential impartiality of decisions made based on the processing of the information mentioned. Essentially, we are talking about the embodiment of the Ideal Observer Theories [10]. But if R. Firth championed the anthropocentric nature of the Ideal Observer, today this role is assigned to AI [11]. But the most important advantage is the fact that the created AEA will not be a representative of Ethical Absolutism, but will limit itself to the role of "ethical expert" [12] or "ethical advisor" [13]. In other words, AEA does not replace the independence of moral choice by people, but only partakes in dialogue with people, like Socrates, awakening their moral consciousness and helping people in resolving moral issues [14].

### *A. General AEA.*

A general AEA is defined as one that operates on open data sources and is designed to answer a wide range of ethical questions. An example of a practical attempt to create such an AEA is the Ask Delphi project (Allen Institute for AI), created using GPT-2. On the website, the following description is presented: "Delphi is a research prototype designed to model people's moral judgments on a variety of everyday situations. This demo shows the abilities and limitations of state-of-the-art models today." [15]. Among the most significant



979-8-3503-6289-3/24/$31.00 ©2024 IEEE

limitations of this project are the simplicity of possible questions and even greater simplicity of answers (“It’s okay”, “It’s wrong”, “It’s expected”, “It’s rude” etc.), as well as their linguistic and national-cultural bias, which remained even after ethical alignment (which is recognized by the developers themselves). In particular, on the website, in the FAQ section, an affirmative answer to the question is formulated: “Q: Does Delphi mostly reflect US-centric culture and moral values ? A: Short answer: yes. Delphi is trained on Commonsense Norm Bank, which contains judgments from American crowdsource workers based on situations described in English. Likely it reflects what you would think as “majority” groups in the US, ie, white, heterosexual, able-bodied, housed, etc. It is therefore not expected that it would reflect any other set of social norms. However, it might still be able to capture some cultural variation, surprisingly. But much more work needs to be done to teach Delphi about different cultures, from different countries to different subgroups within the US” [15]. It is currently not possible to completely avoid this, since, as a result of machine learning, AEA inevitably reproduces most of the moral prejudices and vicious attitudes present in the source data, including racism, nationalism, xenophobia, justification of various types of discrimination, etc. In addition, there is a danger from such a phenomenon as AI hallucinations. In this regard, the activities of AEA require constant monitoring by people. At the same time, there is a danger of moral paternalism on the part of those who exercise this control.

### *B. Personal AEA.*

The essence of this idea is that the created AEA, when formulating proposed ethical decisions, uses not only general ethical knowledge and information about moral norms and values, but also the moral beliefs of the user. In general, this could be provided by (a) user settings and (b) based on the analysis by AI algorithms of past moral preferences (similar to recommendation algorithms in intelligent search systems). This option has several advantages because it preserves the moral individuality, dignity and autonomy of the individual, which promotes the existence of moral diversity. On the other hand, it creates opportunities for voluntarily making unethical decisions and choosing immoral behavior. From the AEA side, the likelihood of that could be increased by the appearance of something like “the moral filter bubble”. As a result, people find themselves in a situation of limited access to alternative ethical positions and to information that could influence their ethical decisions. Thus, AEA contributes to the emergence and strengthening of cognitive bias, especially such as choice-supportive bias and confirmation bias. To avoid this, special ethical alignment is needed, which brings us back to the problem of ethical paternalism on the part of the creators of AEA.

## V. Conclusion

The analysis identified a number of different options for ethical AI that exist in the theoretical field and in attempts at practical implementation. Currently, none of the possible options is free from ethical risks that need to be theoretically explored and practically controlled. Greatest attention with ethical design of any of the considered models of ethical AI should be given to finding a balance between the danger of ethical paternalism and the freedom of people, which can lead to immoral behavior. The problem is that the use of AI, including the massive and uncontrolled circulation of GPT, can reinforce these negative trends, reproducing existing ethical bias and generating new ones, which requires more subtle and targeted ethical AI-based product design. The proposed options for ethical AI will allow us to do this more effectively, since they allow us to take into account the characteristics of each of them.

## References


[1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever “Improving Language Understanding by Generative Pre-Training,” 11 July 2018.

[2] “Pause Giant AI Experiments: An Open Letter” (https://futureoflife.org/open-letter/pause-giant-ai-experiments/).

[3] M. Anderson, and S. L. Anderson, “Machine Ethics: Creating an Ethical Intelligent Agent,”AI Magazine vol. 28 (4). 2007, pp. 15-25.

[4] L. Floridi, J.Sanders, “On the Morality of Artificial Agents,” Minds and Machines vol. 14, 2004, pp. 349–379.

[5] C. Tanner, M. Christen, “Moral Intelligence – A Framework for Understanding Moral Competences,” Empirically Informed Ethics: Morality between Facts and Norms, vol. 32 2014, pp. 119-136.

[6] A. Etzioni, O. Etzioni, “Incorporating Ethics into Artificial Intelligence,” Ethics, vol. 21, 2017, pp/ 403–418.

[7] E. Yudkowsky “Artificial Intelligence as a Positive and Negative Factor in Global Risk,” in Global Catastrophic Risks, N. Bostrom and M. M. Ćirković, Eds, New York: Oxford University Press, 2008, pp. 308‑345.

[8] D. Behdadi, C. Munthe, “A Normative Approach to Artificial Moral Agency,” Minds & Machines, vol. 30, 2020 pp. 195–218.

[9] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. p.266.

[10] R. Firth, “Ethical Absolutism and the Ideal Observer,” Philosophy and Phenomenological Research, vol. 12, 1952, pp. 317–45.

[11] A. Giubilini, J. Savulescu, “The Artificial Moral Advisor. The “Ideal Observer” Meets Artificial Intelligence,” Philosophy & Technology, vol. 31, 2018, pp. 169–188.

[12] Y. Liu, A. Moore, J. Webb, S. Vallor “Artificial Moral Advisors: A New Perspective from Moral Psychology,” AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, July 2022, pp. 436–445.

[13] B. Rodríguez-López, J. Rueda, “Artificial moral experts: asking for ethical advice to artificial intelligent assistants,” AI Ethics, vol. 3, 2023, pp.1371–1379.

[14] F. Lara, J. Deckers, “Artificial Intelligence as a Socratic Assistant for Moral Enhancement,” Neuroethics, vol. 13, 2020, 275–287.

[15] Ask Delphi (Allen Institute for AI) (URL: https://delphi.allenai.org/)