Ethics Documents in the Field of AI. Concepts, Achievements and Problems

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

Vladislav Golovkov
Institute Of Philosophy
Saint-Petersburg State University
Saint-Petersburg, Russia
golovkov.spb@mail.ru

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/USBEREIT61901.2024.10584031

***Abstract*—The development of AI systems and their widespread use in various spheres of people's lives and society have given rise to moral problems. As a result, a new field of applied ethics emerges - AI Ethics, which consists not only of scientific research into the ethical aspects of the design and use of AI systems but has practical implementation in various ethical documents: declarations, recommendations, ethical codes and ethical standards. In this article, in the course of studying existing ethical documents, a typology has been developed based on differences in their statuses, sources of origin and moral agency. The analysis made it possible to identify the main ethical issues discussed in AI technologies and identify the most frequently used concepts and terms. Particular attention is paid to the transparency, privacy, trust. The article formulates significant points of discrepancy in the understanding of these terms in various AI ethics documents. As a main conclusion, we can formulate the idea that currently most ethical documents have completed the stage associated with identifying the main problems and concepts of AI ethics, although there is some disagreement about their understanding. It is noted that the vast majority of ethical documents analyzed do not sufficiently develop the practical component, which should be the basis for making decisions to prevent and resolve moral problems arising in connection with the design and use of AI.**

Keywords—AI ethics, ethical documents, code of ethics, moral agent, trustworthy AI, Ethically Aligned Design AI.

I. INTRODUCTION

The current rapid development of digital technologies and their widespread adoption and use in almost all spheres of public and personal life poses certain of challenges. Addressing them requires certain normative regulation and ethical regulation in particular. Since the 1980s there has been issued a number of ethical documents in order to solve emerging problems and prevent potential ethical risks. These documents include declarations, recommendations, reports, ethical codes, etc. They are aimed at resolving both general ethical issues and issues in certain areas (e.g., the Internet, artificial intelligence, robots, big data, etc.). To date, there is no consensus of how to overcome ethical problems regarding the field of AI technologies. Nevertheless, the analysis of existing instruments allows, on the one hand, to highlight the already achieved results, and on the other hand, to identify the unresolved issues and to assess future prospects of the sphere. In the generalized form, the development and application of ethical documents involves the following stages determined by the specifics of the task: 1) identificating the key and/or controversial ethical issues ('pain points'); 2) coordinating positions on these issues, 3) developing regulatory framework for moral decision-making in ethically problematic situations. This paper examines the existing ethical documents in order to determine which of the abovementioned stages of development are most of them at.

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

II. THE VARIETY OF AI ETHICAL DOCUMENTS

By analyzing the existing ethical instruments aimed at solving moral problems in the field of development and use of AI we divided them into separate groups. The proposed typology is based on distinctions in their status, sources (i.e., the developers of the document) and moral agency (i.e., the recipients of the document). Individual examples of ethical documents for each group are listed below.

- "Proactive" etchical codes proposed by various researchers and groups of specialists and addressed to all stakeholders. The most well-known example from this group is "The Asilomar AI Principles", coordinated by Future of Life Institute (FLI) and developed at the Beneficial AI 2017 conference (Asilomar, USA) [1]
- Ethical documents developed by intergovernmental organizations, for example, Recommendation on the Ethics of Artificial Intelligence UNESCO (2021) [2] "Ethics Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence) EU, (2019) [3], "WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust", EU, (2020) [4] etc.,
- Recommendations of state authorities containing certain ethical elements: "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence", The Department of Defense USA (2019) [5]; "AI in the UK: ready, willing and able? - government response to the select committee report", The House of Lords, London, (2018) [6]; "Rome Call for AI Ethics" Vatican (2020) [7] etc.
- Ethical documents and recommendations issued by international professional organizations and societies in the field of digital technologies. One of most well-known documents of this type is "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems", Version 2. IEEE, 2017 [8]
- Ethical documents of the companies in the field of AI, for example, "Microsoft Responsible AI Standard",

v2. General requirements (2022), Microsoft [9], "AI Principles Progress Update 2023", Google, 2023, [10] etc.

- Ethical codes of corporate communities, such as "A code of ethics for the use of data". The Big Data Association, Russia (2019) [11]; "AI Ethics Code", AI Alliance Russia (2022) [12] etc.
- Standards IEEE 7000 series, such as IEEE P7001 – Transparency of Autonomous Systems, IEEE P7003 – Algorithmic Bias Considerations, IEEE P7008 – Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems etc.
- The ethical component of user agreements, privacy policies and related regulations of the Internet resources that use AI, especially for the operation of cookies and GTP technologies.

III. AI Ethical documents: principles, concepts, problems

The brief overview of some AI ethical documents of various statuses, as well as the comparative analysis of them, are presented below as an example ("Ethics guidelines for trustworthy AI" [3], "Ethically Aligned Design" [8] and "AI Ethics Code" [12]).

A. "Ethics guidelines for trustworthy AI".

In April 2019 the European Commission published the document "Ethics guidelines for Trustworthy AI" as part of ethical research in the field of artificial intelligence technologies. The authors outline the fundamental rights and ethical principles that will form the basis for achieving trustworthy AI. The fundamental rights listed in this document include respect for human dignity, freedom of the individual, respect for democracy, justice and the rule of law, equality, non-discrimination and solidarity, citizens' rights. The fundamental ethical principles are the principles of: respect for human autonomy, prevention of harm, fairness, explicability. This legal and ethical foundation allows to specify seven key requirements for trustworthy AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; social and environmental wellbeing; accountability. [3] "Ethics guidelines for Trustworthy AI" could be described as a set of criteria for distinguishing between trustworthy and untrustworthy AI. In this perspective, the document is intended for people who evaluate and certify certain technological products. At the same time, it can also provide developers and users with some valuable information. The former will understand what kind of requirements their product must meet to be considered trustworthy by the EU; the latter, having familiarized themselves with the selection criteria, can form their own judgement about whether they consider certified AI technology products to be trustworthy. The document clearly emphasizes the importance of human agency, i.e., by moral agents it means people. However, these are not only the creators of technological products, but also those who test, approve or reject a particular product.

B. "Ethically Aligned Design".

Next up is the 2017 IEEE (Institute of Electrical and Electronics Engineers) document "Ethically Aligned Design". It sets out the ethical principles that should be taken into consideration when developing AI systems and introducing them into society, namely: human rights, well-being, accountability, transparency, awareness of misuse [8]. "Ethically Aligned Design" focuses entirely on the issue of ethical ways of creating AI systems. Thus, we immediately see that the emphasis is placed on not just the final product but on making ethical the very development of it, since the result is determined entirely by the role ethics plays in process of creating. This document was designed by a professional community for a professional community and is completely focused on AI designers and developers. The code implies that the process cannot be ethical without the ethics of the Actor, i.e., of the creator of technological products.

C. "AI Ethics Code".

This document was adopted in 2022 by the AI Alliance, whose members are leading technology companies in Russia. This association declares that its mission is to become the center for the AI development in Russia and to provide the technological leadership of the country and of the Alliance participants in the global technology market. It views the regulation of AI as one of the areas of its activity. Within this framework the Alliance promotes the development of legislation on personal data and of concepts for regulating AI, the ethics of the use of AI and its technical standardization. In connection with the above, the Alliance presents a code of ethics in the field of AI, proposing principles and norms that members of the association undertake to follow. These are the following: development and use of AI is human-centered; responsibility must be fully acknowledged when creating and using AI; only humans are responsible for all consequences; development of AI must be controlled. [12]. The AI Code of Ethics is positioned as an instrument with which the member companies of the AI Alliance comply. The purpose of this document is better regulating of this area and establising the basis for the exchange of best practices. Obviously, the text contains principles that members of the Alliance undertake to follow. Of course, this code of ethics is mainly aimed at the AI creators, but it is also intended for users. One of the key issues in the contemporary ethics of digital technologies and AI is where to trust them or not. From the authors' point of view, the proposed document will help to build this trust by declaring the principles and values that the Alliance participants pledge to adopt. The moral agents are represented both by the assosiation members and the individual employees of these companies. Companies act as ethical supervisors, called upon to undertake obligations to comply with the stated provisions as well as to monitor their implementation by employees and users. According to the code, the company's developer employees can act independently, which makes them individuate moral agents.

IV. Problems of Basic Concepts in AI Ethics Documents

The main challenge for researchers in the field of AI ethics relates to its definition. Artificial intelligence can be called a wide range of technologies, such as machine learning, deep learning, generative AI and discriminative AI, etc. However, ethical issues in the design and application of AI do not always directly depend on the diversity of types of AI. First, we can identify common characteristics that are common to all AI. In this general sense, artificial intelligence is a technology that performs certain functions similar to human actions, has some

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

degree of autonomy and is able to adapt to external changes. Secondly, ethics in the field of AI acts as a way to regulate the relationships and behavior of people that arise during the creation and use of AI. Just as the rules of the road are intended for drivers of any vehicle, regardless of the type of engine and transmission features, so ethical standards are common to any type of AI.

It should be stated that the documents under consideration are largely repetitive, stating similar principles and values and using the same ethical concepts. The principles articulated in most AI ethics documents can be applied to most areas of technology ethics. Fairness and non-discrimination, responsibility, respect for human rights and goodness are the ethical basis for the creation and use of any technology. It is worth noting that there is some understatement and uncertainty. For example, much attention is paid to the concept of good. However, the definition of a good is either implied as commonly known and accepted, or it is absent altogether. As a result, there is a shift towards the concept of harm (not causing harm), but the definition of harm also remains questionable. This circumstance can be considered as a significant drawback of all ethical codes in the field of technological ethics. At the same time, some problems of understanding and interpretation arise. In the context of ethics in the field of AI as a special technology, there are principles that need to be given particular attention.

A. Transparency.

Transparency is a crucial concept in the field of AI ethics. It refers to the ability of an AI system to provide clear and understandable explanations for its decisions and actions. In other words, transparency is the degree to which an AI system can be understood by a human. Transparency is vital since it helps establish trust between humans and AI systems. If we understand how an AI system works and why it makes certain decisions, we are more likely to trust it. This is especially important in such spheres as healthcare, where AI systems are increasingly being used to diagnose and recommend treatment. First, the doctor need to understand the entire chain of the process that led to the particular result that the AI produced. Otherwise, the physician cannot be sure what is the best treatment to prescribe for his or her patient. Second, the patients themselves have to trust this technology, including to provide informed consent, a requirement that can only be addressed to the doctor, not to the AI technology.

Unfortunately, there are already tragic examples of a missing AI system transparency that led to a crash. In 2018 and 2019, two Boeing airplanes of the same model (Boeing 737 Max 8) crashed due to incorrect information perception by the aircraft software. The AI that was installed in the airplane's flight control system was designed to correct possible misbehavior by pilots. However, the AI's incorrect actions in situations of correct operations by the pilots led to tragedies. Not only were the pilots unable to understand the AI's actions, but they were unable to control the airplane because the machine was focused more on controlling the AI system. It turned out that the pilots were eliminated from controlling the airplane, and the opaque actions of the AI in both incidents led to airplane crashes.

The emphasis is put on this concept in all the reviews documents. It is worth noting that Russian documents pay attention to the transparency of processes and use of AI systems, while foreign documents mostly consider the transparency of the systems themselves.

B. Privacy.

For technologies working with data privacy is probably the most important ethical issue. The functioning of smart devices involves processing a large amount of data, including personal data. It should be noted that privacy as an ethical issue has existed for a long time. However, in the era of big data, it is being re-emphasized with renewed vigor due to the widespread use of digital technologies, especially those using AI technologies. The emerging complexities have produced a variety of definitions of privacy, such as, privacy of the person, privacy of behavior and action, privacy of communication, privacy of personal data, privacy of thoughts and feelings, privacy of location and space, privacy of association [13].

The problem of information privacy derives from the ubiquity of digital technologies and AI. Hypothetically, the user has the right to agree or disagree with the privacy policy of a particular organization. In reality, this means the consumer has to choose between using certain information product or refusing it, which affects their consumer preferences. As a matter of fact, user agreements are formulated in favor of the company and allow it to manipulate the received data as it pleases, since these data facilitates building, promoting and developing a business to a great extent. Thus, a classic dilemma arises in regard to digital ethics: should one struggle for one’s personal data, abandoning the use of information products, or perceive the public availability of one’s personal data as a price for the opportunity to use a quality product for one’s own purposes?

However, even if this moral dilemma favors the comfortable use of digital services, it turns out that there is no control over the data shared by the user. Firstly, it is very rare for requests to share data to include an explanation of why it is necessary. This raises the suspicion of excessive data collection. Second, the fate of private data is beyond the control of users, which leads to the inability to dispose of their own data. Consequently, the absence of external ethical regulation of data use will inevitably lead to asymmetric power on the part of technology owners.

Certainly, all the reviewed documents cover the issue of privacy of personal data, since it is impossible to create ethical digital products and AI systems without resolving it.

C. Trust.

Finally, trust, in our opinion, is a central concept that appears in every instrument analyzed in this paper. Currently it has become the most discussed term in the discourse of ethics of digital technologies and AI, pushing the concept of responsibility from the first place. Trust is an important category, because without it, technology keeps causing concern and fear. The problem of trust is caused, among other things, by the above-mentioned problems - the lack of transparency and uncontrolled use of technology. However, the problem of trust is not only that many people do not trust AI, but also that many people overly trust this technology. Humanity has always been dependent on technology. It would be wrong to claim that only modern humans cannot get by (at a minimum, would have significant difficulties) without the technology that is current to them. Such a statement is true for people from any period of history, beginning at the dawn of civilization. However, the increasing complexity of technology increases people's dependence on it. Distrust or

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

skepticism of AI can slow its development. On the other hand, overreliance on AI can lead to negative consequences.

All ethical documents presented here are aimed precisely at increasing trust – both in the technology and in its creators. This is why companies create codes of ethics, form alliances and formulate their principles: the need to demonstrate that they deserve people's trust and do everything correctly and ethically.

V. Conclusion

If we highlight certain trends, we can see, first, that all of the examined texts prioritize safety and well-being of man. The relationships between AI agents and their actions are not considered an ethical issue if they do not harm people. In fact, if the interaction of two or more AI systems does not pose a danger to people, then such an interaction can include absolutely anything and not be regulated in any way. If an AI`s action does not concern a human being or humanity as a whole, then it stays outside the boundaries of ethics. However, some concepts (good, for example) remains quite vague, hence the difficulties in defining, comparing and developing strategies of its increase. The general abstractness of the wording also indicates that these texts were compiled for people. Second, the existing variety of AI systems should not affect ethics in this field; its norms are valid for all types of AI. Third, many of the principles discussed in the documents are common to all technologies. Specific in the context of AI are the issues of transparency, privacy and trust. Fourth, the main goal of compiling a list of certain principles and documents is in one way or another, to both improve the application of technologies in everyday life and to increase the level of trust among ordinary users. Fifth, all documents lack a section elaborating on practical application of their principles, which in our opinion is a significant drawback. As the main conclusion, we can argue that at the moment most ethical documents have accomplished the stage related to identifying the main issues and concepts in the field of AI ethics; some of the interpretations might be controversial. However, the most important task is to develop the consolidated practical recommendations for solving the AI ethics problems.

References

[1] Future of Life Institute, "The Asilomar AI Principles," *Future of Life Institute*, 2017. [Online]. Available: https://futureoflife.org/open-letter/ai-principles/. [Accessed: Feb. 29, 2024]

[2] United Nations Educational, Scientific and Cultural Organization (UNESCO), "Recommendation on the Ethics of Artificial Intelligence," United Nations Educational, Scientific and Cultural Organization (UNESCO), 2021. [Online]. Available: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence. [Accessed: Feb. 29, 2024]

[3] European Union (EU), "Ethics Guidelines for Trustworthy AI," *European Union (EU), European Comission, High-Level Expert Group on Artificial Intelligence*, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. [Accessed: Feb. 29, 2024]

[4] European Union (EU), "WHITE PAPER On Artificial Intelligence — A European approach to excellence and trust," *European Union (EU), European Comission*, 2020, [Online]. Available: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. [Accessed: Feb. 29, 2024]

[5] Department of Defense (USA), "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence," Department of Defense (USA), 2019. [Online]. Available: https://media.defense.gov/2019/oct/31/2002204458/-1/-1/0/dib_ai_principles_primary_document.pdf [Accessed: Feb. 29, 2024]

[6] "AI in the UK: ready, willing and able? — government response to the select committee report," Published by the Authority of the House of Lords, London, 2018, pp.184

[7] Pontifical Academy for Life, "Rome Call for AI Ethics," *Pontifical Academy for Life*, Rome, February 28th, 2020. [Online]. Available: https://www.vatican.va/roman_curia/pontifical_academies/acdlife/documents/rc_pont-acd_life_doc_20202228_rome-call-for-ai-ethics_en.pdf. [Accessed: Feb. 29, 2024]

[8] 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.

[9] Microsoft, "Microsoft Responsible AI Standard, v2. General requirements," *Microsoft*, 2022. [Online]. Available: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf. [Accessed: Feb. 29, 2024]

[10] Google, "AI Principles Progress Update 2023" *Google*, 2023. Available: https://ai.google/static/documents/ai-principles-2023-progress-update.pdf . [Accessed: Feb. 29, 2024]

[11] The Big Data Association, "A code of ethics for the use of data," *The Big Data Association*, 2019. [Online]. Available: https://rubda.ru/en/activites/code-of-ethics/. [Accessed: Feb. 29, 2024] [In Russian]

[12] AI Alliance Russia, "AI Ethics Code," *AI Alliance Russia*, 2022. [Online]. Available: https://a-ai.ru/wp-content/uploads/2021/10/Code-of-Ethics.pdf. [Accessed: Feb. 29, 2024]

[13] R.L Finn, D. Wright, D.M. Friedewald, "Seven Types of Privacy," in European Data Protection: Coming of Age. S. Gutwirth, R. Leenes, P. de Hert, Y.Poullet, Eds. Springer, Dordrecht, 2013, pp.3-32