MGIMO UNIVERSITY

AI CENTRE MGIMO
EMPOWER INNOVATIONS

2022

Andrey Ignatyev

AI ETHICS IN FOCUS ON INTERDISCIPLINARY RESEARCH AND DEVELOPMENT OF NATIONAL APPROACHES

приоритет2030
Лидерами становятся

MOSCOW STATE INSTITUTE OF INTERNATIONAL

RELATIONS

(UNIVERSITY) OF THE MINISTRY OF FOREIGN AFFAIRS OF RUSSIA

MGIMO Centre For AI

Andrey Ignatyev

“AI ETHICS IN FOCUS ON INTERDISCIPLINARY RESEARCH AND DEVELOPMENT OF NATIONAL APPROACHES”

Research paper

Moscow,2022

Author: Andrey Ignatyev, MGIMO Center for AI, researcher

Abstract

Research Paper "AI Ethics in Focus on Interdisciplinary Research and Development of National Approaches" provides a synthesis of the theoretical background of AI ethics, its relationship to other scientific fields and disciplines, in particular technoethics. An attempt is made to advance the systematization of the meaning and scope of AI ethics in relation to the practical issues of technology development at different stages of the life cycle, including the specificity of approaches to the problem by various actors. The author outlines priorities for further improvement of ethical instruments and soft law documents in the field of AI, including drawing attention to the importance of specifying and promoting national approaches at the international level.

Key words: artificial intelligence (AI), ethics, AI ethics, technoethics, normative ethics, applied ethics

Table of contents

Introduction 4

1. Linking AI ethics with philosophy of technology, technoscience and technoethics 5

2. Ethics as a branch of philosophy. The main ethical theories and systems 9

3. Ethical paradigms, interdisciplinarity and specificity of actors' approaches..... 12

4. Key risks and negative consequences that can be eliminated or minimized through the application of ethical practices in AI..... 16

5. Development of ethical and soft law instruments in the field of AI, with due regard to national interests..... 18

6. Publications and research used in the working papers 21

Introduction

The working papers present the role of ethics in AI, providing basic definitions of the subject matter, a brief analysis of the theoretical foundations of ethics as a branch of philosophy, reflecting the relationship of AI ethics with philosophy of technology, technoscience, technoethics, and other scientific disciplines.

To comprehend and systematize the discourse at the current stage of development of approaches to AI ethics, an attempt is made to specify what is meant by this concept and what is its scope in relation to the practical processes of creation, use and regulation of AI systems.

The paper also provides information on the main risks and challenges that can be minimized or eliminated through an ethical approach to the implementation of AI projects.

The author concludes with possible scenarios for ethical developments in soft law instruments, noting the relevance and importance of promoting national approaches at the international level.

1. Linking AI ethics with philosophy of technology, technoscience and technoethics

With the development of digital technologies and digital society, humanity is forced to address the philosophical foundations of civilizational development as such, to recognize its past, its scientific achievements and progress, its relationship with nature. On this basis, it is possible to define goals and prospects for the further development of the digital environment, both in the global sense and as applied to a wide range of practical tasks.

«In the digital era, digits are becoming an extensive global phenomenon and force. The ethical culture of digital globalization has provided not only a new space for cultural exchange and integration among nations, but also a new environment for the formation of new global ethical principles and concepts»[1].

In this regard, ethics as a branch of philosophy is the most convenient research field to discuss moral and ethical issues in the contour of the rapid exponential development of AI, which now concentrates the most breakthrough possibilities of end-to-end technologies and digitalization in general.

It is only natural that ethics should be addressed in connection with the development of computer science, automation and, later, "smart" machines - various automated and computer-based systems and digital technologies have begun to have a major impact on all aspects of human life in the last century. At the same time, the speed of introduction of digital technologies is so high that society does not always have time to give a comprehensive and justified moral assessment of all impacts and manifestations of technology in various political, social and other areas of life, as well as to make an unambiguous forecast of the prospects of safe and ethical use of the latest technologies in various spheres. “Information and Communication Technologies have not only dramatically changed personal behavior, lifestyles and interpersonal relationships, but also the perception and the notion of society itself and the information that may be collected about individuals with or without their consent. Technological changes also affect routine aspects of modern societies such e-administration, e-commerce, e-health, e-education, telework and e-voting, telephone communication, po- litical activities, consumers’ rights, private (intellectual) property, democracy etc..”[2].

Of particular importance today are the channels of perception of reality and the multiple sources of information that shape the world view and world outlook. Here

[1] Bao, Zonghao and Xiang, Kun, Digitalization and global ethics, 2006, Kluwer Academic Publishers, USA, issn 1388-1957, Volume 8, Issue 1, abstract, https://doi.org/10.1007/s10676-006-9101-7

[2] Ethics of information and communication technologies, European Commission, Publications Office of the European Union, 2012, ISBN 978-92-79-22734-9, doi:10.2796/135412012, p. 37

too, technological tools are playing an increasingly important role and in some cases are decisive for the education of a new generation - "knowledge acquired through intensive gadget learning is absorbed by young users 'with the milk of the mother'"[3].

Thus, the issues of ethics are closely connected and included in a more global problematic - the study of the socio- and techno-natural system, the study of man in the environment of technical reality, the processes of technological revolution and globalization. "For design activity and engineering creativity, technical reality is all material objects (and their information representation) which are created: a) by man directly or using technical products - tools and devices; b) by technical automatically deterministic devices with the ability to learn and assess the situation; any object, from atomic (molecular) to cosmic level, which has arisen as a result of any material change - human impact (directly or by each other)"[4]. As far as philosophy in the field of technoscience is concerned the relationship and mutual influence of human beings and technology is studied. The demonisation of technology is often seen as an aspect of this – "attributing to technology an inherently hostile attitude to man and the world around it. In many works, technical reality is endowed with the capacity for self-development and, in the ultimate case, with free will"[5].

A separate field of philosophical research is the philosophy of technology (the phrase "philosophy of technology" first appeared in 1877 in a book by the German geographer and philosopher Ernst Christian Kapp, "Grundlinien einer Philosophie der Technik" (Elements of a Philosophy of Technology). Kapp believed that it was in the words of the ancient Greek thinker Protagoras, «Man is the measure of all things», that the anthropological criterion was first formulated, and the core of human knowledge and activity formed[6].

The first Russian philosopher of technology, P. Engelmeyer, believed that "there is an interaction of two homogeneous forces: man influences the world, and the world influences man. The first side of this interaction (man's adaptation to nature) is clarified by the philosophy of natural science, the second side (man's adaptation of nature to his needs) is clarified by the philosophy of technology"[7].

In deepening the search for the relationship between ethics and technological development, it is inevitable to turn to technoethics, which by around 1980 had already emerged as a distinct academic and research area that studied the ethical aspects of technological development and its impact on human beings and society. Technoethics

[3] Andrey Mironov, Philosophy of science, technics and technologies, 2014. – 272 p. ISBN 978-5-317-04749-8, p. 125
[4] Kudrin B.I. Introduction to technology, 2-nd ed.,Tomsk: Tomsk State University Press, 1993 – 552 p. 507-508
[5] Andrey Mironov, Philosophy of science, technics and technologies, 2014. – 272 p. ISBN 978-5-317-04749-8, p. 10
[6] V. Gorokhov, Philosophy of Technology and Methodological Analysis of Technical Sciences, Humanitarian Portal, ISSN 2310-1792, Chapt. 2, p. 18 https://gtmarket.ru/library/basis/6067
[7] P. Engelmeyer. Technical Result of the 19th Century. - M., 1898, p. 101–103, p. 105–106

developed as a result of a merger of a number of disciplines and directions at the intersection of philosophy and applied ethics, which focus on science and various sectoral technologies (biotechnology, nanotechnology, informatics, information technology, computer science, etc.). The term "technoethics" was introduced by philosopher Mario Bunge to describe the responsibility of technologists and scientists to develop ethics as a branch of technology and to identify rational rules for the management of science and technological progress. Bunge argued that the current state of technological progress is determined by practices based on limited empirical evidence and learning by trial and error. In his view, the engineer must take not only technical, but also moral responsibility for everything he designs or executes: his artefacts must not only be optimally effective, but also not harmful, they must be useful, not only in the short term, but also in the long term[8].

Technoethics considers moral and ethical aspects in relation to human beings in the technosphere. The notion of technosphere can be explained as "technocentric approach - the result of interaction between technical objects and systems; ecocentric approach - the result of interaction between humanity and nature; anthropocentric approach - the organo-projection of humans or objectification of human relations in the course of their life activity"[9]. It should be noted that the outstanding Soviet academician V. Vernadsky developed the doctrine of the noosphere - a sphere of mind, a geological shell that emerged at a certain stage of the evolution of the biosphere - a sphere of life, which in its essence is much broader than the technosphere. The term 'noosphere' was introduced into scientific usage by scientist and philosopher Edouard Le Roy (1870-1954). E. Leroy wrote: "Starting with man, evolution is carried out by new, purely mental means: through industry, society, language, intellect, etc., and thus the biosphere passes into the noosphere"[10]. Noosphere can be considered as "a geographical shell of the globe in which transformations of matter, energy and information associated with the activities of an intelligent human being play a major role". Vernadsky's worldview is based on the idea that science, religion, and philosophy are three fundamental and independent forms of reason, each of which is designed to solve its own problems[11]. A number of modern scientists note the relevance and importance of further study of the noosphere, including in relation to AI problems and the paradigm of the "human-noosphere" relationship. In the case of a systematic development of already existing domestic developments, the contribution of Russian scientists to the global discussion could be very significant.

[8] Bunge, Mario. (1977). "Towards a Technoethics," Monist 60(1): p. 96–107
[9] N. Popkov, Technosphere as an object of philosophical research, disserCat, 2005 г., https://www.dissercat.com/content/tekhnosfera-kak-obekt-filosofskogo-issledovaniya
[10] Le Roy, E. L`exigence idealiste et le fait l`evolution. – Paris, 1927. – p. 195–196.
[11] Rezhabek B., Vernadsky's Doctrine on the Noosphere and the Search for a Way out of Global Crises, Journal of the Century of Globalisation. Issue No.1/2008, https://www.socionauki.ru/journal/articles/129838/

In the context of the next (modern) round of AI technology[12] development, the main discussions, as well as scientific and research materials in the field of ethical aspects of this technology are mainly concentrated in the thematic field, which can be conventionally designated within the broad English-language term "AI Ethics". This is due to the emergence of several existing and entirely new moral and ethical challenges, dilemmas and methodologies that concern all stages of the life cycle of AI systems[13], from purely philosophical, worldview problems of technology development to purely practical ones associated with the application of specific systems. At the same time, given the scale and potential impact of technology on society and the environment, this discourse touches upon virtually all areas of human significance - politics and social development, economics, science, education, issues of citizens' rights and freedoms, ecology, intellectual property protection, etc.

The Alan Turing Institute defines AI ethics as a set of values, principles and methods that use widely accepted standards of 'good' and 'bad' to guide moral behavior in the development and use of AI technologies[14]. By now, however, there is a wide variety of formulations and related approaches to this definition.

[12] Author's Note: In most cases, international and national experts agree that AI as a concept is appropriately understood as - i) an interdisciplinary scientific field; ii) a set of data processing technologies; iii) a property of a data processing system; a reflection of the scientific debate on the definition of the term AI is beyond the scope of this paper.
[13] «Ethical questions regarding AI systems pertain to all stages of the AI system life cycle, understood here to range from research, design and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination», Recommendation on the Ethics of Artificial Intelligence, UNESCO, https://unesdoc.unesco.org/ark:/48223/pf0000379920_rus.page=16
[14] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

2. Ethics as a branch of philosophy. The main ethical theories and systems

In considering ethical issues in the field of AI, one cannot ignore the general theories and concepts that have already been developed within ethics, an important branch of philosophy. Given the large number of philosophical schools and doctrines, let us list only the most significant concepts. First of all, it should be noted that ethics investigates the *sources of morality*, a question that is crucial in defining the *values* on which various ethical *principles are*, in turn, based.

The main theories concerning the sources or origins of morality can be presented in the following classification, compiled on the basis of the author's synthesis of multiple sources (books, research papers and articles):

1. Mythological and theological approaches, where the basis for morality is myths or divine precepts (commands); closely related approaches may also be based on other transcendental sources of morality.

2. Naturalistic theories, in which the source of morality is nature and biological patterns of development.

3. Sociological or philosophical-sociological theories (morality is derived from society and social organization).

4. Anthropological or philosophical-anthropological theories which are based on the idea that morality is a property and reflection of human nature, inherent in human beings.

If we refer to the typology of normative-ethical doctrines and theories (sometimes also called 'systems') we can list the following, the most common ones:

1. Consequentialist theories (including utilitarianism, egoism, pragmatism, common good theory, etc.)

2. Non-consequentialist theories (duty-based, rights-based approaches, etc). 3.

3. Agent-oriented theories (virtue approach or virtue theory).

A number of frameworks for making ethical decisions can be found in foreign studies, an example being a study by Brown University (USA): A Framework for Making Ethical Decisions[15].

[15] Brown University, Program in Science, Technology, and Society, 2013, A Framework for Making Ethical Decisions, https://www.brown.edu/academics/science-and-technology-studies/framework-making-ethical-decisions

Several basic studies, scientific and academic literature tends to identify three main directions in ethics:

- Philosophical (theoretical ethics or metaethics, description of the essence of morality and typology of morality);

- Normative (codification of moral values, substantiation of moral norms and principles, search for moral criteria and rules);

- Applied ethics (application of moral rules and beliefs to practical situations, professional ethics, most common in medicine, biomedicine, ecology, computer science, etc.).

The above classification and listed concepts, however, do not limit the further development of other diverse strands and subsections that can be scientifically substantiated. Moreover, there are other approaches to formalising a number of well-established definitions in scholarly works, including the structuring of ethical orientations.

In the philosophical literature, one can find many different attempts to explain the meaning of the term 'ethics' itself, not to mention the variety of theoretical concepts and postulates associated with this discipline. Ethics, as a complex, polysemous concept, encompasses many 'layers' and can be viewed from different perspectives, using different tools of knowledge and 'evidence'. The search for ethical norms or principles continues worldwide, both as a theoretical framework and for profile/application. Despite the large number of formalised ethical principles[16], designers and operators of AI systems are faced with the need to find practical mechanisms and solutions to meet societal expectations. Thus, situations arise where different principles conflict and contradict each other, sometimes causing complex ethical dilemmas. Ethical dilemmas refer to situations in which any available choice leads to a violation of an accepted ethical principle, but a decision must still be made[17]. In this regard, an approach that combines several of the above or other theories and takes into account the specific circumstances of the use of AI systems can be applied to the resolution of ethical dilemmas. The determining factor is the identification of

[16] Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2; Marcello Ienca, Effy Vayena, AI Ethics Guidelines: European and Global Perspectives (2020), https://rm.coe.int/cahai-2020-07-fin-en-report-ienca-vayena/16809eccac
Additional information on the content of AI principles and general analysis of soft law instruments in the field of AI is also presented in the Report of the MGIMO AI Centre, prepared as part of the XIII Congress of the Russian Association of International Studies in October 2021 - Ethics in Artificial Intelligence: from Discussion to Scientific Reasoning and Practical Application: Analytical Report / A. Abramova, A. Ignatyev, M. Panova, Moscow, MGIMO-University 2021, ISBN 978-5-9228-2488-0, https://aicentre.mgimo.ru/upload/ckeditor/files/Ethics-in-Artificial-Intelligence-From-Discussion-to-Scientific-Evidence-and-Practical-Application.pdf

[17] Keith Kirkpatrick, The moral challenges of driverless cars, 2015, *Commun. ACM*, 58(8):19–20, https://cacm.acm.org/magazines/2015/8/189836-the-moral-challenges-of-driverless-cars/fulltext

risks and a comprehensive assessment of the potential harm to all actors or assets involved in a particular practice case. Equity issues are of particular importance in this case. The two main legal approaches to "equity" and "fairness" are individual fairness and group fairness. Individual fairness is the equality of all under the law. It implies that everyone should be treated equally and not discriminated against on special grounds. Equality is recognized as an international human right. Group fairness relies on the fairness of the outcome. It ensures that the outcome does not differ in any systematic way for people who, based on a protected characteristic (such as race or gender), belong to different groups. It considers that differences and historical circumstances may result in different groups responding differently to situations. Approaches to group justice vary considerably from country to country. Some, for example, use so-called "positive discrimination"[18]. There are various studies that attempt to optimize the system of fairness approaches in AI. For example, IBM has developed a number of approaches to measure individual and group fairness (in this case, models for comparing, for example, individuals with similar characteristics or groups of people in roughly the same circumstances)[19]. Schemes are also proposed to identify and proactively address signs of injustice (also related to bias, discrimination, etc.) in relation to different phases and components of a project, such as "data fairness", "design fairness", "output fairness", "implementation fairness"[20].

Questions of fairness within the field of AI raise the whole spectrum of philosophical, legal, and scientific-practical approaches to the concept. In a broad sense, fairness can be seen first and foremost as a principle of law. From this perspective, fairness is "the most important category of classical jurisprudence, first and foremost of the theory of natural law. Attempts of its substantial universalization cannot be considered convincing due to the multiplicity of social and personal identities"[21].

Overall, in practice, it is clear that fairness issues at different stages of the life cycle of AI systems can only be successfully addressed in contextual terms, that is, in relation to the analysis of the specific circumstances in which certain AI systems are used to solve specific problems.

Ethical dilemmas and questions of moral choice continue to be the subject of research by philosophers and regulators both, they provoke acute scientific and professional controversy, especially in the paradigm of rapid scientific, technological, and social modernization, in the search for new meanings for society in the modern era. AI, as a

[18] Artificial Intelligence in Society, OECD library, 2019, Chapter «Philosophical, legal and computational notions of fairness and ethical AI».
[19] AI Fairness 360, https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/
[20] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute.
[21] I. Chestnov, The Concept of Fairness in Postclassical Legal Understanding, 2013, p. 1.

technology capable of replacing humans in many areas and of generating recommendatory decisions based on big data, will inevitably face the need for *ethical guidance and risk assessment,* above all from the perspective of human rights and freedoms and issues of discrimination.

3. Ethical paradigms, interdisciplinarity and specificity of actors' approaches

The previous sections have summarized some of the features of ethics in the field of AI, and then provided a basic and most general understanding of ethics as a branch of philosophy.

On the basis of this framework, it seems reasonable to try to identify and discuss the most important features that characterize ethics in the field of AI. Here we are confronted with rather difficult issues to address, primarily related to the nature of AI technology itself, the scale of its possible impact on society and the different levels (dimensions) of study and systematization. Thus, even debates about AI as a concept or a subject of research, as well as efforts to provide an agreed, compromise definition of AI face problems - many different definitions are discussed, the reasoning of which in each case has varying degrees of credibility and contains controversial points.

A very significant aspect in conceptualizing and constructing a theoretical and practical framework for AI ethics is probably the difficulty of combining and balancing philosophical and applied approaches.

We will briefly consider the involvement of philosophers, neuroscientists, legal experts, and design engineers in the construction of such a research base.

Thus, philosophers are interested in exploring essential, abstract constructs, searching for metaphysical foundations - the most diverse aspects of the development and use of AI technology are material for philosophical consideration. This seems crucial in view of the fact that the large-scale impact of AI-based automation cannot be divorced from the problems of social relations, it cannot be extra-social. Developing this thesis, we can postulate that AI, as a significant phenomenon in the development of social processes, should also be considered within the social and political sciences - anthropology, psychology, economics, politics, international relations, etc. Accordingly, questions of ethics in this field, in one way or another, require study in relation to political systems, economics, ideology and religion. Thus, it is very difficult to achieve consensus and a common philosophical understanding of the various ethical

dimensions in the field of AI. Moreover, the existing similarities and differences between personal ethics and social ethics impose additional difficulties and complicate systematization and universal theoretical constructions.

In this context, philosophers are not only concerned with the relationship between humans (individuals) and society, but also with the attempt to create artificial intelligence, which, in turn, requires an understanding of human nature as such and, in a narrower sense, a reflection on the comparison between natural and artificial intelligence. Our insufficient knowledge of the human being as a physical and spiritual entity, the lack of a clear understanding of brain processes and cognitive functions, and finally the difficulty of a rigorous scientific interpretation of the very concepts of mind, intellect, thinking, reasoning make it difficult for scientists to advance toward systematic scientific constructions in this field. Consequently, without a sufficient initial, coherent knowledge base, it is very difficult to build a coherent scientific conception in the ethical field as well. In this respect, AI has given a huge impulse to the development of neuroscience, first and foremost neurophysiology. Immersion in this field helps to understand the physical foundations of consciousness and brain activity, which, in turn, can affect the comprehension of ethical aspects of AI technology, both in a broad sense and in the subject-practical sense. Emerging neurotechnologies offer some opportunities to monitor brain processes and understand what causes certain behaviours - "many new neurotechnologies enable us to intervene in these processes, to change and perhaps to control behaviors, traits, or abilities"[22].

It can be predicted that in the future such possibilities of "reading the brain" and influencing it will only increase with the development of neuroscience. In this regard, it is impossible not to reflect the existence of such a discipline as neuroscience ethics or neuroethics, which, among other things, studies issues of morality and brain activity. Neuroethics, at an interdisciplinary level, examines the impact of neuroscience advances on human beings, their self-awareness, behavior, emotions, and social life. "The range of approaches adopted in neuroethics includes but is not limited to historical, anthropological, ethical, philosophical, theological, sociological and legal approaches"[23]. In neuroethics, it is possible to distinguish such areas as: - ethical problems of implementing neuroscience research; - study of the neural foundations of morality; - use of neurotechnologies as technologies for enhancing (improving) human beings[24].

[22] Roskies, Adina, "Neuroethics", The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/neuroethics/

[23] Handbook of Neuroethics, published by Springer Netherlands, January 2015, DOI 10.1007/978-94-007-4707-4, ISBNs 978-9-40-074706-7, 978-9-40-074707-4, 978-9-40-074708-1, Authors: Demertzi, Athina, Laureys, Steven, Editors Jens Clausen, Neil Levy, https://link.springer.com/book/10.1007/978-94-007-4707-4#about-book-reviews

[24] Olga V. Popova, Epistemology & Philosophy of Science, Volume 56, Issue 3, 2019, pages 153-168,

Thus, within the framework of the topic of AI, philosophers are faced with a whole layer of pressing questions of being, which, however, cannot be considered and resolved in isolation from other humanities and practical sciences and without reliance on experimental practice. The construction of contemporary philosophical constructs, among other things, imposes difficulties of terminological uncertainty for a number of concepts that also cannot be resolved without subject and interdisciplinary research.

In general, however, any ethical question in the field of AI in its philosophical dimension leads to a wider discourse, including such problems as the relation of mind to matter, thinking to being, spirit to nature and others.

Legal experts consider ethics in the field of AI to be predominantly an important starting point and an important source of shaping legislative initiatives. The complexity of legislation in the field of AI has led to the emergence and widespread use of soft law instruments that include ethical aspects[25]. Often, ethical considerations are included in recommendations and guidelines of various kinds for the development of technology in general. This trend has led to a wide range of issues being incorporated into the concept of 'AI ethics' in Western practice. AI Ethics has been invested in a wide range of issues. This factor is widening the scope of the issues under consideration, and the concept of AI Ethics is becoming increasingly vague and comprehensive. Among the tools that most specifically address ethical technology issues are the various AI codes, both general and industry specific. However, due to their nature and purpose, codes cannot serve as full practical guidance for developers at the system design stage.

From the developers' point of view, ethical aspects are important to ensure user confidence in the products produced, to increase competitiveness and to maintain the company's brand and image. It may be noted that in the current situation, when the introduction of technology outstrips its legal regulation, it is the developer who is responsible for the negative consequences of applying certain codes, algorithms, databases and, in general, for ethical decisions and actions in the course of product or program operation. Modern AI products are usually the result of a combination of a number of technical decisions involving not only programmers, but also hardware manufacturers, data suppliers, security specialists, etc. In this regard, ensuring ethics at the design stage can be realized through more applied tools (practical guidelines,

---

https://doi.org/10.5840/eps201956356, Human and Human Death as a Neuroscience Ethics Problem https://journal.iphras.ru/article/view/3719

[25] A. Abramova, A. Ignatyev, M. Panova, Ethics in Artificial Intelligence: from Discussion to Scientific Reasoning and Practical Application: Analytical Report, October 2021, Moscow, MGIMO-University, ISBN 978-5-9228-2488-0, https://aicentre.mgimo.ru/upload/ckeditor/files/Ethics-in-Artificial-Intelligence-From-Discussion-to-Scientific-Evidence-and-Practical-Application.pdf

instructions, recommendations, metrics) relevant to a particular class of systems. For example, developers are faced with the challenges of dealing with ethical issues when designing and setting up referral systems, systems for managing various Internet resources and predictive analytics systems.

Thus, for the developer, as an actor in the life cycle of systems, the most important tools are normative and technical documents, technical standards, methods or methodologies for assessing compliance with ethical norms and principles. The last three years have been characterized by the emergence of a large number of these kinds of tools, developed primarily by large technology companies. However, the preparation of such documents should in one way, or another be based on theoretical and experimental bases, which confirms the logic and desirability of moving from conceptual research to concrete practical application guidelines, starting from the design stage.

4. Key risks and negative consequences that can be eliminated or minimized through the application of ethical practices in AI

To date, there have been many studies, publications and other materials that detail the potential negative impacts of the introduction and use of AI systems in various fields around the world. For example, international instruments[26] developed by the European Union, Council of Europe, OECD, UNESCO and others provide a systematic and detailed analysis of the negative scenarios and manifestations that AI technology can potentially have - in these materials such analysis is usually placed in the sections on risks, threats, types of damage, harms or in the sections reflecting ethical concerns. In the documents of international organizations and in various studies, risk maps are presented with certain emphases, determined by the mandate of the site or its objectives. Summarizing at the level of meanings, it is possible to identify the following big blocks, which at the structural level demonstrate and provide an understanding of the problem.

1. limitation (infringement) of rights and freedoms, inequality and segregation, discrimination against individuals or particular social groups.

These negative consequences of using AI systems can arise from bias. Biases are broadly categorised and can be caused by poor quality data, biases of people involved in the life cycles of the systems, computational algorithms themselves, improper configuration, and use (targeting) of the systems, and other reasons. It is important to emphasize that various initial negative biases or faults in the AI system (in algorithms or data grids), human (developer, operator, etc.) biases, can be scaled up in the subsequent stages of using such systems, i.e. can cause a multiplicative effect.

Examples of inequalities, marginalization, violation of the privacy of citizens can also be included in this block of risks, including restrictions on personal autonomy, violation of human dignity, problems related to personal data protection, hyper-personalization in data collection, polarization, or other breaches in social communication.

[26] The OECD Recommendation of the Council on Artificial Intelligence – https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; The Recommendation on the Ethics of Artificial Intelligence - https://unesdoc.unesco.org/ark:/48223/pf0000379920; Feasibility study on a legal framework on AI design, development and application based on CoE standards - CAHAI, 17 December 2020 - https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da, Ethics Guidelines for Trustworthy Artificial Intelligence - https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

The Council of Europe study[27] also refers to the phenomenon of 'digital power concentration', where many applications are developed and deployed by the 'Big Five' (GAFAM - Google, Facebook, Microsoft, Apple and Amazon) - so that significant political power is concentrated in the hands of a few private companies that favour 'shareholder values over the common good, this can threaten the credibility of democratic states'.

2. Incorrect or poor-quality outputs from systems, negative impacts on the reliability and safety of various processes and reduced public confidence in the use of scientific innovation.

3. Negative impact on human cognitive abilities, risks of intellectual, cultural, and creative degradation, loss of autonomy. In this regard, we can talk about freedom of thought or cognitive freedom, i.e., the risk of " degradation of human agency" in the process of life.

4. Various kinds of manipulation of individuals and social groups, the imposition of certain patterns of behavior, the malicious or unintentional manipulation of public opinion and the negative impact on social processes, including through the powerful new tools of influence on the media and political processes in election campaigns.

5. Various disruptive socio-economic effects (transformation of the labour market, undesirable factors in fair competition, financial and commercial transactions of economic agents, etc.).

[27] Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law, Council of Europe, December 2020, p. 31, https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

5. Development of ethical and soft law instruments in the field of AI, with due regard to national interests

Based on the above analysis, it can be stated that the development of ethics in the field of AI is not only in the form of abstract principles and theoretical studies, but also has an impact on the further improvement of the legal system, including in terms of sensitive, socially relevant issues of AI regulation.

In addition, frameworks and approaches relating to ethical criteria are being actively incorporated into tools for assessing the safety and credibility of AI products - this is happening for all the diversity and ambiguity and flexibility of approaches to this concept. While system characteristics such as reliability, safety, functionality and efficiency can be supported by technical measurements and actual indicators (evidence base), analysing and defining the ethics of a system probably requires more sophisticated approaches, which, in many cases, can be quite controversial in terms of positions and perceptions of different social groups.

In assessing the ethics of an AI system, the issue essentially comes down to the following elements:

- an ethical assessment of the outputs of the system;

- the ethical responsibility of the individual involved in all life cycles of the system.

In this respect, developing some kind of universally acceptable 'ethical scale' seems to be a very difficult process which, in any case, will be based on the researcher's or regulator's perceptions of ethics as such. Such perceptions and positions need to be grounded in conceptual research and conclusions that take account of previous experience and, at the same time, draw on the experimental and theoretical basis of contemporary science and fit harmoniously into the picture of the modern world.

We are faced with thinking about contemporary processes as applied to philosophy in the field of AI - political, economic, socio-humanitarian, anthropological processes. And here we must come to an understanding of dialectical development - the development of morality, ethics, and fairness under the current conditions of the new technological paradigm.

In addition to the hasty universalisation of ethical norms, there is currently a trend towards the introduction of monitoring mechanisms and evaluation procedures (along the lines of AI Ethical Risk Assessment), which can also become certain instruments for competition and product promotion in regional or international markets. In such a broad and multidisciplinary research subject as AI, a conceptual, and ontologically sound approach will therefore deepen the basis for moving from general, correct but

impractical regulations and principles to applied specialization. Such specialization should take into account the characteristics and capabilities of specific types of AI systems, their purpose and conditions of use, and the specific societal groups potentially affected by such systems. Based on such an approach and adaptation to the specific domain, the various tools, and guidelines for assessing the AI ethics would generate more trust and interest.

Ethical issues in the field of AI can be considered not only at the level of the individual, social or community group, but also at the level of national entities and states. We thus come to the importance of thinking about ethical concepts at the national level. Only with such clear national concepts can we effectively engage in dialogue on harmonising basic ethical standards across national boundaries.

While it is difficult to "universalize" ethics in relation to AI, current trends show that certain concepts, constructs, and frameworks are likely to dominate the world, which in turn will be embedded in soft law documents as well as in standardization and certification documents. At the same time, at the actual moment, de facto the core of such documents is Western research, including developments by major technologically and industrial corporations in the US, Europe, and China. International organizations' instruments are mainly based on publications and research by scientists from Western reputable universities or authors affiliated with major technology companies.

Consequently, the world remains uneven in the participation of all national actors in the global debate on shaping the mainstream ethical norms and principles. The basic fairway for the development of ethical issues is still determined by the most economically advanced countries. A note of caution should be made that after about 2019, during the COVID-19 pandemic, Latin American states, the African continent, and other countries that are not among the economic leaders of the planet have begun to join this process.

Another factor is that, given the increasing position of large digital platforms, there is a tendency for certain points of confrontation in the approaches of state institutions and multinational corporations operating across borders. In this sense, it is becoming increasingly problematic to develop common rules for the implementation of a single monitoring mechanism for compliance with soft law instruments by all market players.

From such a situation arises the need to intensify national research work on AI ethics and to promote its results at the international level. This will, in particular, make it possible to more effectively defend the achievements of the national school of philosophy and increase its representation in international scientific communities and organizations. The inclusion of representatives of Russian technology companies in such research and the build-up of experience in developing targeted ethical guidelines

for various purposes would also create a favourable environment for the promotion of products both in Russia and abroad. The potential for the development of national science-based approaches to AI ethics is largely dependent on the creation of productive, interdisciplinary teams that are well organized and governed.

6. Publications and research used in the working papers

1. Bao, Zonghao and Xiang, Kun, Digitalization and global ethics, 2006, Kluwer Academic Publishers, USA, ISSN 1388-1957, https://doi.org/10.1007/s10676-006-9101-7

2. Ethics of information and communication technologies, European Commission, Publications Office of the European Union, 2012, ISBN 978-92-79-22734-9, doi:10.2796/135412012

3. Andrey Mironov, Philosophy of science, technics, and technologies, 2014., ISBN 978-5-317-04749-8

4. B. Kudrin, Introduction to technology, 2-nd ed., Tomsk: Tomsk State University Press, 1993

5. V. Gorokhov, Philosophy of Technology and Methodological Analysis of Technical Sciences, Humanitarian Portal, ISSN 2310-1792, https://gtmarket.ru/library/basis/6067

6. P. Engelmeyer. Technical Result of the 19th Century. - M., 1898

7. Bunge, Mario, Towards a Technoethics, 1977, Monist 60(1).

8. N. Popkov, Technosphere as an object of philosophical research, disserCat, 2005 г., https://www.dissercat.com/content/tekhnosfera-kak-obekt-filosofskogo-issledovaniya

9. Le Roy, E. L`exigence idealiste et le fait l`evolution. – Paris, 1927.

10. Vladimir Vernadsky, Scientific thought as a planetary phenomenon, Moscow: Nauka, 1991.

11. Vladimir Vernadsky, Problems of Biogeochemistry II. On the Fundamental Material-Energetic Distinction Between Living and Nonliving Natural Bodies of The Biosphere, 939, https://21sci-tech.com/articles/ProblemsBiogeochemistry.pdf

12. B. Rezhabek, V. Vernadsky's Doctrine on the Noosphere and the Search for a Way out of Global Crises, Journal of the Century of Globalisation. Issue No.1/2008, https://www.socionauki.ru/journal/articles/129838/

13. The Recommendation on the Ethics of Artificial Intelligence, UNESCO, https://unesdoc.unesco.org/ark:/48223/pf0000379920_rus.page=16

14. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529

15. Brown University, Program in Science, Technology and Society, 2013, A Framework for Making Ethical Decisions, https://www.brown.edu/academics/science-and-technology-studies/framework-making-ethical-decisions

16. Jobin, A., Ienca, M. & Vayena, E; The global landscape of AI ethics guidelines. Nat Mach Intell 1, 2019, https://doi.org/10.1038/s42256-019-0088-2

17. Marcello Ienca, Effy Vayena, AI Ethics Guidelines: European and Global Perspectives, 2020, https://rm.coe.int/cahai-2020-07-fin-en-report-ienca-vayena/16809eccac

18. A. Abramova, A. Ignatyev, M. Panova, Ethics in Artificial Intelligence: from Discussion to Scientific Reasoning and Practical Application: Analytical Report, October 2021, Moscow, MGIMO-University, ISBN 978-5-9228-2488-0, https://aicentre.mgimo.ru/upload/ckeditor/files/Ethics-in-Artificial-Intelligence-From-Discussion-to-Scientific-Evidence-and-Practical-Application.pdf

19. Keith Kirkpatrick, The moral challenges of driverless cars, 2015, Commun. ACM, 58(8), https://cacm.acm.org/magazines/2015/8/189836-the-moral-challenges-of-driverless-cars/fulltext

20. Artificial Intelligence in Society, OECD library, 2019, Chapter «Philosophical, legal and computational notions of fairness and ethical AI»

21. AI Fairness 360, IBM, https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/

22. Leslie, D., Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, 2019, The Alan Turing Institute

23. I. Chestnov, The Concept of Fairness in Postclassical Legal Understanding, 2013

24. Roskies, Adina, "Neuroethics", The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/neuroethics/

25. Handbook of Neuroethics, published by Springer Netherlands, January 2015, DOI 10.1007/978-94-007-4707-4, ISBNs 978-9-40-074706-7, 978-9-40-074707-4, 978-9-40-074708-1, Demertzi, Athina, Laureys, Steven, Editors Jens Clausen, Neil Levy, https://link.springer.com/book/10.1007/978-94-007-4707-4#about-book-reviews

26. Olga V. Popova, Epistemology & Philosophy of Science, Volume 56, Issue 3, 2019, pages 153-168, https://doi.org/10.5840/eps201956356, Human and Human Death as a Neuroscience Ethics Problem https://journal.iphras.ru/article/view/3719

27. The OECD Recommendation of the Council on Artificial Intelligence, OECD, 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

28. Feasibility study on a legal framework on AI design, development and application based on CoE standards - CAHAI, December 2020, https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

29. Ethics Guidelines for Trustworthy Artificial Intelligence, European Commission, the High-Level Expert Group on AI, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

30. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law, Council of Europe, December 2020, https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

31. Nemitz Paul, Constitutional democracy and technology in the age of artificial intelligence, 2018, Phil. Trans. R. Soc. A.3762018008920180089, http://doi.org/10.1098/rsta.2018.0089

32. Webb, A., The Big Nine: How the tech titans and their thinking machines could warp humanity, ISBN-13: 9781541773745, 2019, https://www.publicaffairsbooks.com/titles/amy-webb/the-big-nine/9781541773745/

33. Study on the impact of digital transformation on democracy and good governance, CoE CDDG (2021) 4 Final.

MGIMO Centre for AI was established to enhance international cooperation and support collaboration with all the actors of digital economy both at national and international levels. Our multidisciplinary research is focused on international cooperation agenda, national policies for AI and business opportunities. International trade and trade policy (prioritising digital trade), sustainable development, AI ethics are the key areas of our activities.

On the basis of MGIMO-University we promote an international AI expert platform with regular conferences and round tables, peer-revied articles and research papers. Our enlarging network of strategic partneships makes it possible to provide AI consulting and policy solutions both for business and goverment agencies.

The Centre was founded in October, 2021

MGIMO Centre for AI research paper collection

- Artificial intelligence in education
- Artificial intelligence for development

Monthly digest

Digital economy bulletin

Reports

Discussions on Artificial Intelligence Ethics: Development Tracks by Key Groups of Actors,2021

Annual conference AI Global dimension

MGIMO Digital Discussion club

Round tables&workshops

These documents& other Centre updates are available at https://aicentre.mgimo.ru/activities/research/papers

We hope to develop cooperation and we are open to any partnership offerings!

Our contacts

143007, Moscow Region, Odintsovo,
Novo-Sportivnaya street, 3
https://aicentre.mgimo.ru
E: aicentre@inno.mgimo.ru
P: +7 903 623-95-15
https://t.me/aicentremgimo

AI CENTRE MGIMO
EMPOWER INNOVATIONS

приоритет2030^
Лидерами становятся