Index of Artificial Intelligence Systems Ethics

Oleh Zaritskyi

National Aviation University, av. L.Guzara, 1, Kyiv, 03124, Ukraine

Abstract

The article deals with the current issues of classification of the challenges that arise and the principles that should been used, during the development and implementation of artificial intelligence systems. In the subject area of AI ethics, the notion of an AI system ethics index has been introduced. The author made a detailed analysis of ideas and methods in the subject area of ethics of artificial intelligence and proposed a general approach for quantifying the level of ethics of developed systems by classifying the main challenges, evaluating them and introducing compensatory measures. The approach reflects the general idea, which could been detailed by specialists from the respective subject areas. The research is purely theoretical in nature, summarizing existing ideas and principles, for the first time putting forward the idea of a quantitative assessment of the question of the ethics of artificial intelligence.

Keywords

AI system, AI ethics, ethics Index

1. Introduction

The intensive development of information technology in the last decade, especially in the field of artificial intelligence and hardware in the form of neurosynaptic and quantum computers, poses new challenges to society in its harmonious development in terms of moral and ethical issues, as well as information security. Numerous AI programs by the world's leading governments published in the past few years also highlight the urgency of the ethical issues that will arise as these technologies develop. “A Next Generation Artificial Intelligence Development Plan” strategizes China AI development by 2030. By 2025, it states that China will have major breakthroughs in AI theory and AI will become the driving force for industrial upgrading and economic restructuring. In addition, by 2030, China will become the world’s major AI innovation center [1]. In USA DARPA announces, “$2B+ investment plan to overcome limitations on AI technology” [2]. The “AI Next program” begins [3]. The Subcommittee on Information Technology of the U.S. House Committee on Oversight and Government Reform publishes a white paper on AI and its impact on policy [4].

The UK government publishes its “AI Sector Deal” which invests 950M pounds (1.2B USD) to support research / education, and enhance the UK’s data infrastructure [5].

Since 2014-15, public, private companies, educational and research institutions have begun to publish various regulatory documents, materials related to the ethical issues of the development, implementation and application of artificial intelligence systems. The importance of moral issues in information systems and AI are also been evidenced by the research on ethical issues highlighted in separate sections of the systematic AI index reports [6-8], which highlight several general principles that unite these documents, among them: confidentiality, accountability, transparency, and explainability.

The very fact that such documents appear shows that society is beginning to pay serious attention to such a difficult issue as ethics and human rights in the field of artificial intelligence. However, criticism should been noted, that has arisen from experts in ethics and human rights in connection with the possible ambiguous or inaccurate use of existing terms in this field.

Information Technology and Implementation (IT&I-2022), November 30 - December 02, 2022, Kyiv, Ukraine
EMAIL: oleh.zaritskyi@npp.nau.edu.ua
ORCID: 0000-0002-6116-4426

The abstract nature of the introduced principles does not allow us to speak about their adequacy from the point of view of a correct description of the subject area and, accordingly, about the possibility of their use to control compliance with ethical norms in the field of AI.

2. Related work

Research [6-8], covering more than a hundred papers produced by various organizations on AI ethics, identifies the 12 most frequently cited challenges to AI ethics (tabl.1).

Table 1
Challenges to AI Ethics

Ethical Challenges	Definition
Accountability	All stakeholders of AI systems are responsible in the moral implications of their use and misuse
Safety	Throughout their operational lifetime, AI systems should not compromise the physical safety or mental integrity of humans
Human Control	It assumes control by the developer and end-user in the development and use of AI systems, respectively
Reliability, Robustness, and Security	All systems designed and used must be reliable in use, resistant to external influences and meet information security standards
Fairness	The development of AI should refrain from using datasets that contain discriminatory biases
Diversity and Inclusion	Understand and respect the interests of all stakeholders impacted by your AI technology
Sustainability	The AI development must ensure the sustainability of our planet is preserved for future
Transparency	An AI system should be able to explain its decision making process in a clear and understandable manner
Interpretability and Explainability	Developed AI systems should be understandable in terms of their internal content (construction) and easily explainable in terms of their functionality
Multi Stakeholder engagement	Involves multiple independent stakeholders in the development and operation of AI systems
Lawfulness and Compliance	All the stakeholders in design of an AI system must always act in accordance with the law and all relevant regulatory regimes
Data Privacy	Users must have the right to manage their data which is used to train and run AI systems

This list is not exhaustive, but shows general trends in AI. Researches show that fairness, interpretability and explainability, transparency are most mentioned across all documents studied.

Research [9] presents core ethical principles *(i.e., respecting autonomy, avoiding harm & doing good, ensuring justice)* and the instrumental principles that primarily link to them. With about 100 sets of principles published as of today, it is easy to get lost in these separate but similar documents, so the "Dynamics of AI Principles" is tool for keeping track of, and systematize, the bewildering and growing

number of AI Principles out there. Private companies, governmental agencies, international organizations, research centers, and professional organizations had published AI principles.

In [10], much attention had been paid to the legal aspects of the development of AI systems from American legislation system. The report addresses issues such as: privacy, innovation policy, liability (civil), liability (criminal), agency, certification, labor and taxation.

In research report [11] authors mentioned that for millennia, waves of technological change have been perceived as a double-edged sword for the economy and labor market, increasing output and wealth but potentially reducing pay and job opportunities for typical workers. Thus, the study emphasizes the question (SQ11): How has AI affected socioeconomic relationships? We do not see an unambiguous answer. Perhaps the impact on the economy and the labor market is not as noticeable as expected, because AI has been localized in certain industries and countries and does not have the proper level of implementation, i.e. we did not get a critical mass. Thus, an analysis of recent research suggests several main areas of concern for ethics and human rights scholars: information security (human rights, adequate historical data for learning samples, etc.) and the impact on human and social well-being.

3. AI challenges and classification approaches

The research methodology involves the study of the key causes of disagreement between the fields of research in artificial intelligence and ethics, as well as the classification (formalization) of the basic concepts of the field of study. Obviously, all the disagreements between AI specialists and ethicists in its classical sense arise from different interpretations of AI terms in terms of ethics. It is necessary to turn to the definition of ethics and the tasks it addresses in a broad and narrow sense.

Ethics is a philosophical discipline whose subject is morality. Ethics has two main functions – moral-educational and cognitive-educational, so two areas could been distinguished in ethics - normative ethics, aimed at teaching about life, and theoretical ethics, studying morality. [12].

Theoretical ethics is a scientific discipline that examines morality as a special social phenomenon, finds out what it is, how morality differs from other social phenomena. Theoretical ethics studies the origin, historical development, regularities of functioning, social role and other aspects of morality. Its methodological basis is the knowledge, concepts and ideas concerning the scientific knowledge of morality. Normative ethics searches for a principle (or principles) that governs human behavior, guides one's actions, establishes criteria for evaluating the moral good, and a rule that can act as a general principle for all cases.

Applied (practical) ethics studies particular problems and the application of moral ideas and principles articulated in normative ethics to specific situations of moral choice. Applied ethics interacts closely with the social and political sciences and has a number of sections, e.g. business ethics, medical ethics, computer ethics, etc. Obviously, AI ethics could been classified as a section of applied (normative) ethics, which is very close to information (computer) ethics. Let us make a brief excursion into information ethics and consider its principles.

Broadly speaking, computer ("information" or "cyber ethics") ethics investigates the behavior of people who use information systems, based on which appropriate moral precepts and norms of behavior are developed [13]. Computer ethics covers almost all spheres of human activity and deals with technical, moral, legal, social, political and philosophical issues. The problems analyzed in it could been roughly divided into several classes:

1. Problems associated with the development of moral codes for users and developers.
2. Problems of protection of property rights, copyrights and basic human rights (rights to privacy and freedom of speech, obtaining and using information, the right to work, privacy and personal data, etc.) as applied to the field of information technology.
3. Cyber security, determination of the status of incidents and crimes, that is, predominantly legal problems, as a rule, formalized in the form of national legislative acts on information security.

Principles developed in computer ethics:

1. Privacy – a person's right **to autonomy and freedom in private life**, the right to be protected from intrusion by authorities and others.

2. Accuracy (accuracy) – compliance with the norms related to the accurate execution of the instructions for the operation of systems and information processing, honest and socially responsible attitude to their duties.

3. Property – inviolability of private property. Adherence to this principle means observance of the right of ownership of information and copyright norms.

4. Accessibility – the right of citizens to information, its accessibility at any time and in any place.

The principles of information (computer) ethics developed are very similar to the ethical principles of AI that we reviewed in the literature review, but they relate only to data processing and issues of security and ownership.

The only issue is that not all these principles have been systematized in the framework of a corresponding standard and each developer tries to take into account all possible cases of AI impact on society, which leads to their duplication, different interpretation of sometimes the same principles and challenges. For the same reason various so to speak analytics are also mixed, for example, direct impact on society and incorrect historical data, etc.

The main ethical challenges could been divided into several main groups (fig.1).

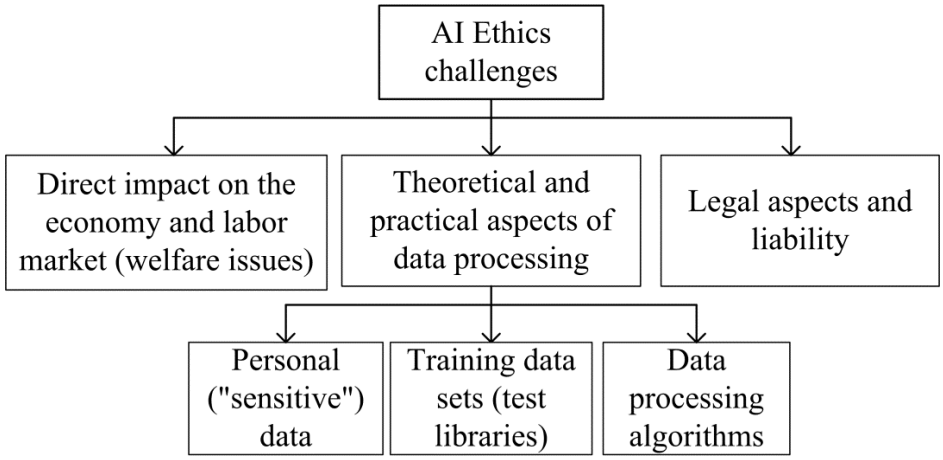

Figure 1: AI ethics challenges

4. Index of AI system ethics

All challenges and problematic issues that arise in one way or another during the development of AI systems can be attributed either to the issues of data and algorithms, or to those that affect the economy and society, or conflict with existing legal norms. Very often the challenges are complex and can simultaneously positively affect the economy as a whole, while also causing social contradictions, such as a couple of productivity and employment issues. In the aspect of data processing, special attention should been paid to the creation of test (training sets) of data, which require strict adherence to the principles: data neutrality, representative data, accuracy, reliability, openness and diversity.

Among the many approaches to classifying ethical principles in AI issues, the author would distinguish four large global groups (fig.2). Figure 2 shows the classification of ethical principles into four main groups, as well as the relationship of these groups to the principles described in papers [9-11]. Group “Safety and Security” includes all the principles that describe the security and protection of both data and its accuracy and reliability in order to create a secure information technology. A very important issue is the adequacy of historical data to create correct training samples, which affects both the manageability of AI and social responsibility. Group “Manageability (Controllability)” describes the principles that must been followed to create AI software that is manageable, efficient, understandable to the end user, and controllable by the end user. The principles of this group also imply a cautious attitude toward the prospective capabilities of the AI system are being creating, which have not yet been fully clarified by the developer. The development should been conducted at a high scientific and technical level. The system must be reproducible under different conditions. The developer must take into account all risks in operating conditions.

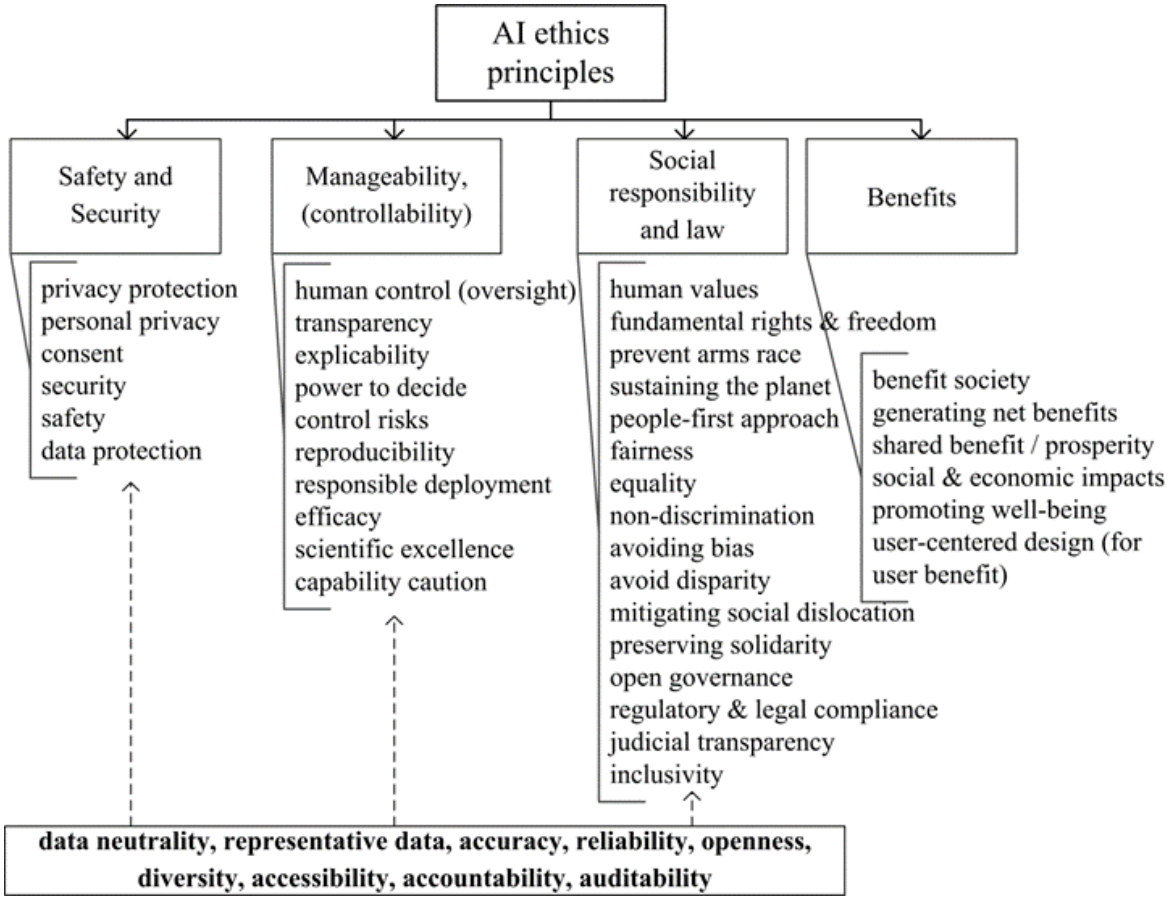

Figure 2: AI ethics principles

The group "Social Responsibility and law" includes principles that characterize the AI system to be developed in terms of compliance with social and legal norms of society. Group "Benefits" describes the principles for assessing the usefulness of the developed AI system from both a tangible and intangible point of view for the user and society as a whole.

There is a group of principles (bottom of Figure 2): data neutrality, representative data, accuracy, reliability, openness, diversity, accessibility, accountability, auditability that are common to the three groups. These principles relate to data, but their implementation lays the foundation for addressing safety, manageability, and social responsibility. Using the developed classifiers of challenges and principles, let us introduce an index of AI system ethics (1):

$$E_{AI} = \sum_{i=1}^{N} \left(\pm P_i + P_i^{IN} \right); \quad , \qquad (1)$$
$$\lim_{+P_i \to \max; -P_i \to 0} E_{AI} \to \max$$

$\pm P_i$ - is an evaluation of the principles that the AI system fulfills, (+) means positive impact, (-) – negative; P_i^{IN} - is a principle (initiative, IN) necessary to compensate for negative influences.

5. Summary and Conclusion

The main condition for using formula (1): the principles of development must correlate with the challenges posed by the development and minimize their negative impact on society from an ethical point of view. For example, automation can lead to increased productivity and, as a consequence, to job losses. This is a serious challenge with the highest ethical rating. We can rate the principle – fundamental rights and freedoms (right to work) on a maximum scale, for example – 5 points, $P_i = -5$. At the same time, it will reduce the length of the working day, improving social conditions (if there is

such a point in state programs, for example), so we can assess this initiative also on the maximum scale $\_P_i^{IN} = +5$. Thus, we are talking about the implementation of initiatives (recommendations for legislatures and businesses), which can minimize the impact of negative factors (challenges).

The author's top-level classification of AI ethics principles could been detailed by independent research and introduced as a standard after agreement with all stakeholders. The Index of AI Ethics is been considered as a general approach to the evaluation of developed AI systems. It requires further study in terms of a detailed classification of principles in the form of a final list, which will be included in the upper level groups proposed by the author. The development of evaluation scales for quantitative assessment of the ethics of AI systems is also been envisaged. The results of research in the field of ethics of artificial intelligence are closely intertwined with research and approaches to quantify the technological singularity proposed in work [14]. AI ethics in a broad sense could been seen as an applied part of general ethics, which examines the behavior of people who develop and use AI systems, as well as the impact of these systems on society. The result of this study is the principles and norms of morality designed both to solve private practical problems in the development process and to realize the harmonious development of society with maximum ethical benefit.

6. References

[1] A Next Generation Artificial Intelligence Development Plan, China copyright and media, Aug. 2017. [Online]. Available: https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan/.

[2] DARPA announces "$2B+ investment plan to overcome limitations on AI technology", Defense Advanced Research Projects Agency, 2018. [Online]. Available: https://www.darpa.mil/news-events/2018-09-07.

[3] AI Next program begins, Defense Advanced Research Projects Agency, 2018. [Online]. Available: https://www.darpa.mil/work-with-us/ai-next-campaign.

[4] Rise of the Machines: Artificial Intelligence and its Growing Impact on U.S. Policy, United states. Congress. House. Committee on oversight and government reform, 2007. [Online]. Available: https://www.hsdl.org/?abstract&did=816362

[5] AI Sector Deal. Policy pape. Gov.uk. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal.

[6] Y. Shoham, R. Perrault, E. Brynjolfsson, J. Clark, J. Manyika, J.C. Niebles, T. Lyons, J. Etchemendy, B. Grosz and Z. Bauer. Artificial intelligence Index. 2018 Annual Report. Steering Committee. Stanford University. 2018.

[7] Y. Shoham, E. Brynjolfsson, J. Clark, J. Manyika, J.C. Niebles, T. Lyons, J. Etchemendy, B. Grosz and S. Mishra. Artificial intelligence Index. 2019 Annual Report. Steering Committee. Stanford University. 2019.

[8] Y. Shoham, E. Brynjolfsson, J. Clark, J. Manyika, J.C. Niebles, T. Lyons, J. Etchemendy, B. Grosz and S. Mishra. Artificial intelligence Index. 2021 Annual Report. Steering Committee. Stanford University. 2021.

[9] TOOLBOX: Dynamics of AI Principles, AI ETHICS LAB. Aiethicslab.com.

[10] https://aiethicslab.com/big-picture/.

[11] Artificial intelligence and life in 2030. One hundred year study on artificial intelligence. Report of the 2015. Study panel, Stanford University, 2016. [Online]. Available: https://ai100.stanford.edu/2016-report.

[12] Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. Stanford University, 2021. [Online]. Available: https://ai100.stanford.edu/2021-report/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence.

[13] Julia Driver. Ethics: The Fundamentals. Wiley-Blackwell, 1st ed., 2006. 192 p.

[14] T. Bynum. Computer and Information Ethics. The Stanford Encyclopedia of Philosophy (Spring 2011 Edition), 2018.

[15] O. Zaritskyi, O. Ponomarenko. Quantitative assessment of technological singularity, The International Scientific and Technical Journal Problems of control and informatics, 2022. - №1. - P.93-111. DOI: http://doi.org/10.34229/1028-0979-2022-1-9.