Environment. Technology. Resources. Rezekne, Latvia
Proceedings of the 13th International Scientific and Practical Conference. Volume 2, 214-218

Ethical Development and Implementation of Artificial Intelligence

Aleksejs Zorins
Faculty of Engineering
Rezekne Academy of Technologies
Rezekne, Latvia
Aleksejs.Zorins@rta.lv

Peteris Grabusts
Faculty of Engineering
Rezekne Academy of Technologies
Rezekne, Latvia
Peteris.Grabusts@rta.lv

***Abstract* - The paper discovers an essence and importance of introduction of ethical dimension in all phases of artificial intelligence (AI): development of concept and source code, implementation in real-life applications and support and improvement of existing solutions. Modern society largely depends on cybertechnologies most of which are using elements of AI and ethical aspects of it is of paramount importance.**

Keywords* - *Ethical Artificial Intelligence, Artificial Intelligence, Ethics, Cyber Security, Safety of Artificial Intelligence.

I. INTRODUCTION

The paper gives some insights into importance of ethical development and implementation of artificial technology as a part of modern digital world. Several researchers stated that in the recent years (from near 20 till 100) a machine capable to perform on at least human level on all tasks will be developed [3, 4, 11, 14, 22]. Our society on practically all aspects are already depending on digital technologies and AI itself. Despite different predictions whether strong or clever AI will be developed, many researchers agree that this problem is very important and urgently needs serious attention from scientific and technological society [1,2,7,12,26]. Ethical side of the problem is being often forgotten or put to least important category.

This issue is becoming more and more important due to rapid development of AI both on software and hardware levels, wide implementation of AI in business, governmental structures, military, medicine, finance, and personal life. The research in this field is in a very beginning and at the same time is crucial to our development and safety. Ethical side of AI and digital technologies are crucially important due to several reasons, firstly, ethical principals are vital for normal life of whole humanity and for each its person, secondly, ethical problem in digital technologies has its own features.

Ethical principles are based on core values, are of a universal nature. For example, cross-cultural studies have shown the universality of values such as assistance to relatives, support of their group, mutually beneficial sharing costs and benefits, respect for elders, respect for private property [6]. However, core values get different readings depending on the context and refraction into specific ethical requirements. In this case context is the structure of everyday life, including technology, who shape it.

Knowledge of the ethical dilemmas of digitalization and AI, about the emerging rules of ethics of digital technologies, especially important for government, and those who provide their professional training and professional development. To receive positive effect decision-makers should be aware of technological development, understand what economic and social the consequences will be caused by their application. Public administration and interaction with citizens are also digitalized. The effectiveness of such interaction is not in the last place depends on whether the ethical risks.

The issue of ethical AI development and implementation can also be useful to those who are responsible for the development digital services, products, and systems (including in the public sector), targeting citizens as recipients of services and consumers or as workers. Potential conflicts and risks associated with ethical side of the use of technology, can be prevented, if you pay attention to this when designing a service or product.

Online ISSN 2256-070X
https://doi.org/10.17770/etr2021vol2.6549

II. Definition of Ethics and its Context in Artificial Intelligence

According to Merriam-Webster's dictionary ethics is "the discipline dealing with what is good and bad and with moral duty and obligation" [19].

Ethics shows what the consequences can be if a specific idea will be perceived by an individual or society, what kind of preconditions and prospects. Thus, ethics can be defined as a reflection of morality. From other means of social regulation - rights, traditions, customs - moral norms differ in that they involve freedom of choice and are regulated primarily by such feelings as shame, duty, remorse [28].

The main system of ethical principles is humanism. Precisely by the principles of humanism mankind is guided in formulating the most important international instruments such as the Universal Declaration of Human Rights [25]. Humanism involves caring for a specific person, striving society to create conditions for the satisfaction of individual needs and personal fulfillment. The main principles of humanism include:

- ✓ guarantees of fundamental human rights as a general condition for genuine private existence;
- ✓ support for the weak, going beyond conventional wisdom a given society about justice;
- ✓ formation of social and moral qualities that allow personality to self-actualize using social values [28].

Going back to IT side let us consider McCumber Information Security Model (widely known as Cybersecurity Cube), which gives us all aspects of information security (Fig. 1).

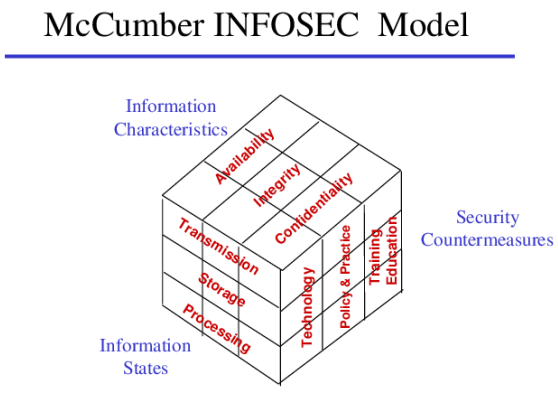

Fig. 1. Cybersecurity cube or McCumber INFOSEC model [17].

Looking from an ethical perspective we should treat all sides and aspects of this cube from an ethical dimension, taking into account ethical considerations developing each side of artificial intelligence system and implementing its security firstly from ethical side, starting from education of developers and ending with ethical support of such systems.

III. Responsibility of AI

The problem of responsibility for the actions of AI systems is the most important and most discussed among application of AI. Problem responsibility appears in areas of trust where a person has to rely on the actions of the AI system, especially in such areas like medicine, finance, politics, education, law enforcement etc.

Discussions suggest different approaches to the principles establishing responsibility for the actions of AI, including [27]:

- ✓ full exemption of anyone from responsibility for actions AI (by analogy with force majeure circumstances);
- ✓ partial exemption from liability (exemption specific person from any responsibility and the simultaneous payment of compensation to victims from various sources);
- ✓ responsibility through fault, arising only depending on fault a specific entity, for example a manufacturer, developer, person, responsible for training AI, owner, user, etc .;
- ✓ innocent responsibility (a certain person (most likely, manufacturer), as a general rule, is considered to be responsible for AI system actions);
- ✓ personal responsibility of robots subject to the endowment of robots legal personality (rights and obligations, status electronic personality).

There is still no consensus in AI responsibility problem, therefore it is valuable to give several interesting opinions on this topic, published in [25]:

Razin A.V., Doctor of Philosophy sciences, professor, head. department ethics of philosophical Faculty of Moscow State University M. V. Lomonosov: "There is a concept of shared responsibility in ethics: in one way or another degree responsibility is borne by all participants - and the developer of the system artificial intelligence, and its owner, and the user (if he has the ability to customize it), and the system itself. "

Karpov V.E., Cand. tech. Sci., Vice President of the Russian Association of Artificial Intelligence, head of laboratory of robotics NRC "Kurchatov Institute": "Most often, this responsibility is assigned to the "programmer", but he is only an operator who lays down the rules of behavior, determined by an expert, a specialist in some subject area. In the case of ethics, by some moral philosopher, for example. It is the expert who is the moral philosopher and is responsible for the essence of the system's behavior and what the logic should be based on decision making. What "moral code" will be provided; this will be implemented by a "programmer". "

Dushkin R.V., Director for Science and Technology of the Agency of artificial intelligence: "There are several options for who is responsible: the developer, the owner,

users and the artificial intelligence itself. The fourth option is radical, and that is what I preach. In most countries, the legislation is arranged in such a way that if something happened, the person who is to blame, compensates for the damage in money or falls to jail. Accordingly, an intelligent system can also be responsible for your mistakes with money. Let us take an autonomous drone or a car, it has certain needs: fuel, electricity, maintenance, parts and need in money for it all. He earns them with his functionality, continuous movement around the city, transporting passengers from point A to point B, for which passengers pay him, as they pay now Yandex or Uber. The machine spends less than it earns and saves money with which she will answer if something happens, and from her compensation for damage will be recovered. "

Milke V.E., PhD in Computer Science and Machine Learning, England Ruskin University (Cambridge, Great Britain): "Ethics in artificial intelligence is largely dependent on manufacturers of these solutions and very few from consumers or service companies - legal owners of systems artificial intelligence. We concentrate a lot on the question "Who is to blame if ...?". In fact, when the discussion begins on ethics of artificial intelligence, the broader question: how to develop artificial intelligence to avoid all kinds of risks. No need to invent something new, when there are ten commandments: do not steal, do not kill, honor your father and mother, etc. Many well-known ethical principles, for example from the Asilomar conference, written in modern technical language, but they say the same thing: do not steal data, do not harm your developments, system crashes in artificial intelligence must be discovered and explored, the superintelligence must be designed solely for the benefit of humanity, and self-learning systems should be under control of humanity, etc."

Neznamov A.V., Executive Director of Sberbank, founder of the Robopravo project, lead. researcher in Institute of State and Law of the Russian Academy of Sciences: "The term "artificial intelligence" has no single definition; therefore, it is rather difficult to talk about uniform rules of application of AI in all areas. Accordingly, the question responsibility also cannot be resolved unequivocally. Today only one rule could be called universal: the person is responsible. To a greater or lesser extent, but the responsibility for the actions of AI is a sole responsibility of a specific person. It seems necessary avoiding the two extremes: when there is no responsibility at all and when the responsibility is borne by the system of artificial intelligence. Both options seem completely irrelevant in the existing conditions".

IV. Explainability of Artificial Intelligence

In explainable or transparent AI a user should clearly and fully understand the output of an AI system and make corrections if necessary.

The concept of explainable AI (XAI) is shown in Fig. 2, which clearly shows that today the user usually does not know the answers to the following questions:

- ✓ why computer do it;
- ✓ why not something else;
- ✓ when it succeeded;
- ✓ when to trust a computer and how to correct errors.

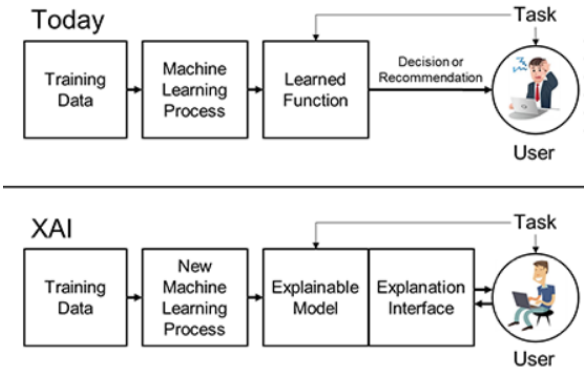

Fig. 2. Explainable AI [9].

In the case of XAI the user will have answers to all these questions.

Andres Holzinger presents an approach of a complete machine learning pipeline beyond algorithm development [13].

Wojciech Samek provides several reasons why explainability is so important for AI research and its safety aspects [23]. There reasons are: verification of the system; improvement of the system; learning from the system and compliance to legislation.

The sensitivity analysis (SA), Layer-Wise relevance propagation (LRP) and other methods make AI more transparent and explainable.

Unique algorithms created by developers are intellectual property and should not be fully disclosed (except for open source or individual cases stipulated by the contract for development). Accordingly, the explanatory component should work in such a way as to explain the results of a system, without revealing the entire process of its functioning. The presence of an explanatory component is an obligatory property of artificial system, otherwise trust in it and therefore its value is being questioned.

V. Bias in Artificial Intelligence

The problem of bias in AI systems is one of the most important in the application of AI. Biases and assumptions that are subtle at first glance can be hidden in data, and systems that are built on their basis can inherit both, which affects the objectivity of the system and makes its decisions unethical and biased. As a result, AI can make recommendations or take actions, which only reinforce and reproduce these biases. The fairness of algorithms is one of the most important areas in creating ethical AI.

However, there are also good examples of AI systems helping to tackle that bias, for example reducing racial inequality in the criminal justice system [15] or automated

financial underrating systems can be helpful for applicants with an undervalued credit history [8].

On the other hand, a violation of ethical standards in data collection, insufficient anonymization or insufficient validation of input data used to train AI systems may lead to discrimination of persons involved in that process.

For instance, trained on partially fictional case histories, IBM Watson sometimes makes deadly cancer treatment recommendations [21]. In the United States, the PredPol crime prediction program trained on an ethnically distorted sample, more often sends the police to the addresses where the representatives of ethnic minorities live [16]. Examining credit history when making hiring decisions can hurt disadvantaged citizens, although there is no proven link between quality of credit history and behavior at work [24].

Amazon discontinues its picking system staff after bias-related bias was found in the algorithm. The algorithm recognized patterns of words in the resume, not the corresponding skillsets. The input data for training the system turned out to be mostly white men's resumes. The algorithm excluded resumes that contained words more commonly used by women. As a result, bias towards women in hiring was manifested [10].

Understanding these facts and a danger of AI biases several approaches have been proposed to reduce the risk of errors and mistakes of artificial systems.

The specialists of the London-based company DeepMind suggested as a defense against the influence of human prejudice use the hypothetical fairness method (counterfactual fairness). To formulate a fair and unbiased judgment about a citizen, AI forms a hypothetical situation in which a given citizen has the opposite characteristics: a woman turns into a man, a poor person turns into a rich person, an African American turns into a white one, etc. Thus, the real status does not affect the assessment of the citizen's actions. A judgment is formed in a hypothetical situation. Such a judgment is considered free from prejudice, and therefore fair [5].

The second approach is to improve the AI systems themselves, from the way data is used to design, implementation, and application processes, to prevent individual and societal biases from perpetuating or creating bias and related problems. Interdisciplinary collaboration aims to ensure the further development and implementation of technical improvements, working methods and ethical standards [25].

An important part of the fairness of AI systems is mandatory inclusion of direct human participation. While fairness statistics are certainly useful, they cannot consider the nuances of the social context in which the AI system is deployed and the potential problems associated with, for example, data collection [20].

VI. Results and Discussion

The paper shows that the research made in the direction ethical AI needs serious attention and there are several issues to be solved. All the proposed approaches for solving above mentioned problems are limited and do not assure the complete confidence of AI user that this technology will have only positive effects.

The research directions should answer the following questions:

- ✓ Where exactly and in what form is human judgment needed in the ethical development and operation of AI?
- ✓ Who decides when an AI system has already minimized bias and is safe to use?
- ✓ In what situations and contexts automated decision making is allowed and ethical?

No computer algorithm can answer these questions on its own, and they cannot be entrusted to any machine. They require human judgment and reflection from a variety of disciplines, including computer science, sociology, economics, psychology, law, and ethics. For this purpose, in each area and every location the trustworthy group of experts motivated to behave honestly and ethically should act to help a mankind ensure safe and ethical use of AI and digital technologies. This is an ultimate goal and success factor of our future.

References

[1] R. Banham, "Cybersecurity: Protective Measures Treasuries Should Be Taking." Treasury & Risk. 2018 Special Report, pp. 2-7.

[2] D.Beskow, K.Carley, "Social Cybersecurity: An Emerging National Security Requirement". Military Review. April 2019, Vol. 99 Issue 2, pp. 117-127. 11p.

[3] N. Bostrom, "Global Catastrophic Risks". Oxford: Oxford University Press, 2007.

[4] N. Bostrom. "The ethics of artificial intelligence." Cambridge Handbook of Artificial Intelligence, 2011. [Online]. Available: https://nickbostrom.com/ethics/artificial-intelligence.pdf [Accessed: March. 03, 2019].

[5] Chiappa S. "Path-Specific Counterfactual Fairness". [Online]. Available: https://csilviavr.github.io/assets/publications/silvia19path.pdf [Accessed: April. 1, 2021].

[6] Curry O., Mullins D., Whitehouse H. "Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies", Current Anthropology. 2019. Vol. 60, no 1. [Online]. Available: https://www.journals.uchicago.edu/doi/pdfplus/10.1086/701478 [Accessed: March. 25, 2021].

[7] R. Deibert, "Toward a Human-Centric Approach to Cybersecurity". Ethics & International Affairs. Winter 2018, Vol. 32 Issue 4, pp. 411-424.

[8] Gates S. W., Perry V. G., Zorn P. M. "Automated underwriting in mortgage lending: Good news for the underserved?" Housing Policy Debate. 2002. Vol. 13, no 2. P. 369‑391.

[9] D. Gunning, "Explainable Artificial Intelligence", DARPA project, 2018. [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence [Accessed: March. 07, 2019].

[10] Hamilton I. A. "Why It's Totally Unsurprising That Amazon's Recruitment AI Was Biased against Women." Business Insider. [Online]. Available: https://www.businessinsider.in/why-its-totally-unsurprising-that-amazons-recruitmentai-was-biased-against-women/articleshow/66192889.cms [Accessed: April. 1, 2021].

[11] S. Hawking, "Science in the next millenium", 1998. [Online]. Available:https://www.learnoutloud.com/Catalog/Science/Physics/Science-in-the-Next-Millennium/45223 [Accessed: March. 03, 2019].

[12] N.Hennig, "Privacy and Security Online: Best Practices for Cybersecurity". Library Technology Reports. April 2018, Vol. 54 Issue 3, pp. 1-37.

[13] A. Holzinger, "From Machine Learning to Explainable AI." World Symposium on Digital Intelligence for Systems and Machines August 2018. [Online]. Available: https://www.researchgate.net/publication/328309811_From_Machine_Learning_to_Explainable_AI [Accessed: Feb. 21, 2019].

[14] M. Kiss, C. Muha, "The cybersecurity capability aspects of smart government and industry 4.0 programmes." Interdisciplinary Description of Complex Systems. 2018, Vol. 16 Issue 3-A, pp. 313-319.

[15] Kleinberg J., Lakkaraju H., Leskovec J. et al. "Human decisions and machine predictions." The Quarterly Journal of Economics. 2018. Vol. 133, no 1. P. 237‑293.

[16] Lum K., Isaac W. "To predict and serve?" Significance. 2016. Vol. 13, no 5. P. 14‑19. URL: https://doi.org/10.1111/j.1740-9713.2016.00960.x

[17] Maconachy, S. and Ragsdale, W. (2001) "A Model for Information Assurance: An Integrated Approach." Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, West Point, pp. 308-310. [Online]. Available: http://it210.groups.et.byu.net/theitbok_files/msrwpaper_1.pdf [Accessed: March. 21, 2021].

[18] McCumber, John. "Information Systems Security: A Comprehensive Model". Proceedings 14th National Computer Security Conference. National Institute of Standards and Technology. Baltimore, MD. October 1991, pp. 328-337. [Online]. Available: https://csrc.nist.gov/csrc/media/publications/conference-paper/1991/10/01/proceedings-14th-national-computer-security-conference-1991/documents/1991-14th-ncsc-proceedings-vol-1.pdf [Accessed: March. 22, 2021].

[19] Merriam-Webster dictionary. Ethics. [Online]. Available: https://www.merriam-webster.com/dictionary/ethics?src=search-dict-box [Accessed: March. 21, 2021].

[20] Richardson R., Schultz J., Crawford K. "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice." New York University Law Review Online. 2019. March. [Online]. Available: https://www.benzevgreen.com/wp-content/uploads/2019/02/18-icmldebates.pdf [Accessed: April. 1, 2021].

[21] Ross C., Swetlitz I. "IBM's Watson supercomputer recommended «unsafe and incorrect» cancer treatments, internal documents show." [Online]. Available: https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watsonrecommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf [Accessed: March. 25, 2021].

[22] N. Sales, "Privatizing Cybersecurity". UCLA Law Review. April 2018, Vol. 65 Issue 3, pp. 620-688. 69p.

[23] W. Samek, T. Wegang, K. Muller. "Explainable artificial intelligence: understanding, Visualizing and interpreting deep learning models", Aug. 28, 2017. [Online]. Available: https://arxiv.org/abs/1708.08296 [Accessed: March. 19, 2019].

[24] Silberg J., Manyika J. "Tackling bias in artificial intelligence (and in humans)". McKinsey Global Institute. [Online]. Available: https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-andin-humans [Accessed: March. 23, 2021].

[25] Universal Declaration of Human Rights. [Online]. Available: https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=eng [Accessed: March. 25, 2021].

[26] R. Yampolskiy, "Artificial Superintelligence: a Futuristic Approach". New York: Chapman and Hall/CRC, 2015.

[27] Аналитический обзор мирового рынка робототехники. Сбербанк, 2019. [Online]. Available: http://www.sberbank.ru/common/img/uploaded/pdf/sberbank_robotics_review_2019_17.07.2019_m.pdf [Accessed: March. 23, 2021].

[28] "Этика и «Цифра»: Этические проблемы цифровых технологий. Аналитический доклад. РАНХиГС." [Online]. Available: https://ethics.cdto.center/ [Accessed: March. 25, 2021].