33

ETHICAL DILEMMAS OF ARTIFICIAL INTELLIGENCE SYSTEMS

Alexander Ageev, Russia[453]

33.1 Introduction

Currently[454], the replacement of humans by AI systems is taking place everywhere. Mostly modern AI ("weak AI") operate within

[453] The author Alexander Ageev is Professor of Economics at various universities in Moscow, with a PhD also in theology. He is Founder and Director of the Institute for Economic Strategies in Moscow, General Secretary of the International Research Institute for Advanced Studies IRIAS and Director of Globethics Russia.

[454] Article translated from Russian. First published in: Агеев А.И. *Этические дилеммы систем искусственного интеллекта*, в книге *Социогуманитарные аспекты цифровых трансформаций искусственного интеллекта* под редакцией В.Е. Лепского, А.Н. Райкова. Москва, Когито-Центр. 2022. Глава 3.5. © Globethics Publications, 2023 | DOI : 10.58863/20.500.12424/4276064 | CC BY-NC-ND 4.0 International.

knowledge and goal-oriented contours specified by the developers. Accordingly, posing the question of AI ethics at this level of its development is a purely declarative, though useful, exercise. The real question about the rules of conduct of AI arises when and if AI will be able to autonomously change the contour of knowledge used, learn from unlimited sources and types of information, adjust or even set its own goals. The second fundamental problem is related to the ethical situation in "ordinary" human society, where ethical relativism has formally prevailed, which in reality means the veiled dominance of quite certain ethical models. This creates the prerequisites for the inevitable translation of this situation into the space of development and application of AI.

33.2 Genuinely Aware Subjects and Artificial Intelligence Systems (Ethical Aspects)

Under certain conditions, it is possible to apply the definition of genuinely aware subjects (GAS), capable of selecting activity goals and arbitrarily working with different knowledge bases, to AI. In part, this reflects the dependence of AI on the goal-setting of its developers, who can put explicit or implicitly criminal decisions into AI. But to a greater extent, the emergence of GAS among AI depends on scientific and technological progress and its dissemination in everyday life.

GAS refers to living or quasi-living beings that have self-awareness and subjective experiences similar to those of humans or other highly evolved beings. Such objects could probably include AI in its strongest version . Strictly speaking, individuals with implanted "smart" prosthetic devices (from cardiac pacemakers to prostheses of the musculoskeletal system or individual sense organs) should also be classified as PIC. The intellectual component of such devices is developing rapidly. The number of people with such devices is growing and is measured in the tens of millions worldwide. It is easy to see the boundary where the operation of these devices collides with the problem of moral choice. In addition, as a “social creature,” man in today's world is in a number of living and virtually organized networks that can influence his daily, subject-situational, and even attitudinal choices. This means that he is influenced by supra-personal virtual systems, as well as by “big user data” collected independently of the individual's will.

Moral and ethical concerns stem not only from how arbitrary and/or harmful such a system can behave to the individual and human society, but also from the fact that society itself can inflict suffering on the artificial entities it creates when they reach the GAS level.

A similar moral dilemma has arisen in scientific and practical experiments on humans and animals. At the moment it has been resolved as follows: experiments on humans are forbidden without their consent, animals are considered less evolved and experiments on them are allowed. Recently, however, there has been a tendency to equate certain highly evolved animal species with humans, assigning them moral and legal rights. A precedent occurred in the United States: two primates participating in biomedical research at the State University of New York at Stony Brook were recognized as having some human rights.

When creating artificial beings and quasi-beings, it is not always possible to predict the outcome of an experiment and the extent to which a conscious being will eventually emerge, given the learning potential of such a system. In addition, it is impossible to ask the creature's consent to the experiment before it is created. This problem is also solvable: the creature does not require its consent to be born, but it receives certain rights and guarantees provided by law and custom immediately after birth, and in part before. Similarly, the creators of artificial beings, including SIs, can be required to preemptively minimize their “suffering” during experiments, also to keep these beings in proper conditions. Otherwise, when consequences cannot be predicted, such experiments should be prohibited.

An even more difficult question is whether, in the future, intelligent systems will be able to take on the risks of decision-making while solving some tasks on vital social problems. Is society ready to take responsibility for the activities of such a technical system? Because of the complexity of the tasks solved by the AI, the impossibility of full control over the machine, tracking and checking the decisions made, a person is already forced to shift part of the responsibility to the machine (program).

There is a certain degree of uncertainty in the development of AI, which is similar to human free will. But at the same time some principles must be laid down in the AI to be developed:

ethical self-restraint and knowledge of people's moral and ethical standards, imitation of the process of self-regulation of one's behavior, ability to empathize;

a mechanism for predicting the risks and consequences of one's own actions, limiting actions when certain risks occur;

the possibility of recognizing and correcting one's own mistake .

Further research in the field of AI ethics will undoubtedly lead to the creation of various standards and certification of rules for the design and operation of AI. In creating such standards, it is necessary to assess their impact on the further development of AI, eliminating the risks of both malicious development and inhibition of the development of intelligent systems.

33.3 Types of Mass Consciousness and the Challenge of Ethical Relativism

In the 20th century an unprecedented experience of manipulating consciousness, of purposefully forming mass consciousness defined by an initiator, usually the state or with support from state institutions, of personality types and parameters of their permissible (“ethical”, socially acceptable) behavior has been accumulated. This experience has largely been discredited in recent decades, giving rise to a crisis of basic ethical concepts that have developed over the last millennia of human history and have formed the dominant ethical relativism, the growth of multiculturalism along with the aggravation of conflicting identities. However, the discrediting of this experience does not mean either its disappearance or the loss of its potential to manifest social energy in the future. We are talking about a certain accumulated repertoire of social being, in which, depending on the combination of organizing and self-organized processes, a set of motivating and stimulating layers of culture for socialization, adaptation, subjectification of man and his communities on different grounds, reflecting biological and social reproduction and interaction with the outside world emerges. There is no doubt that the development and implementation of new technologies, including AI, will be influenced by the ethical state of society. There is also every reason to believe that advances in science and technology will increasingly influence collective and individual consciousness.

An ontological map of social-value orientations ("pictures of the world," "I (we)-concepts") can be formed on many grounds, in many ways similar to the traditional membership of philosophical schools (versions of idealism and materialism, gnoseological and existential concepts, etc.). Among these criteria are, for example, idealistic vs. materialistic, subjectivity vs. objectivity, generality-partiality, personalization-impersonalization, etc.

In the considered perspective the approach to the formation of the ontological map from the point of view of generalized personality types (GPT), taking into account the historical experience of the twentieth century, significant for social practices in the modern period, seems meaningful and useful. These issues are partly understood in a technocratic perspective as the formation of Industry 4.0 societies, in a sociotechnical perspective as "inclusive capitalism," and in a more complex perspective as the creation of Society 5.0.

In any case, the immense variety of vital problems, against the background of the increasing information flood and the expanding use of information technology for processing information arrays and flows (from the structure of the universe and its particles, road traffic, the regulatory framework to human brain research) provokes more and more people to turn to various techniques for ordering and simplifying this hectic and complex reality, including dexterity and complexity, for certainty in their identification and choice of behavior and life in general in this confusing world.

Under the conditions of "multiculturalism," ethical relativism is theoretically and de facto inevitable, manifesting itself as the existence of a certain, seemingly unlimited "menu" for the choice of personal identity. However, this does not mean that every line on this menu is equal. Among the entire set of ethical concepts offered or allowed to be developed "from scratch" (an option implicit in personal creativity that is almost entirely manipulative) there are both heavy and very light, situational, conjunctural, usually little conscious, but convenient fractions. "Heavy fractions" rely on conceptual, theoretical and/or religious solutions to deep questions of human existence in the world and society in particular. They tend to have a powerful system of reproduction, institutionalized in the social structure, in management, science, education, and anchored in architecture and art. Their gravitational pull is obviously higher than any of the newfangled attempts to create a "new ethics" and, as its political projection, an ideology. But the "heavy fractions" of ethical systems and constructs, representing the attractor of traditionalism, always experience difficulties in adapting to the new challenges of evolution formed by the progress of science, technology, lifestyles, political and economic struggles.

Technological changes and active project development have brought to life a diverse assortment of "new human" concepts ("digital humans," "GMO-humans," "transhumanism," "service people," "singulars," "new Europeans," "new nomads," etc.).

The ensemble of ideas fed by the concepts of noosphere, cosmism, "radiant humanity", socio-spiritual integrationism is also gaining strength. Behind it there is a substantial scientific basis going back to

V.I. Vernadsky and T. de Chardin, twentieth-century fantasts (H. Wells, I. Efremov, S. Lemm, Strugatsky brothers), philosophers and scientists (N. Fedorov, K. Tsiolkovsky), but more important - behind it practical triumphs of human development in the mid-20th century. - exploration of the Earth, near-Earth and outer space, the depths and expanses of the ocean, matter, human biology and psyche, etc. However, this set of ideas, for all their scientific validity, is complex, relying on new scientific paradigms that are poorly known in society. It is still far from being manifest as a new mass ideology, let alone an GPT. More popular are "geographical," "geopolitical," and "geoeconomic" versions of ideologies that raise to the flag the mechanical proximity of the residence and destiny of certain peoples and states and serve rather pragmatic interests.

All this diversity of GPTs affects scientific and technological solutions in the field of digitalization. Generally speaking, as AI reaches a level of development at which there is a need for certain ethical constraints on goal-setting, on the knowledge and behavior used, approaching AI at the GAS level, a "portfolio" of ethical principles and practices will be rapidly formed and regulated for the AI being developed. In this process, ensembles (sets) of ethical templates (matrices) developed at the national and civilizational levels will inevitably be projected onto artificial entities. With all the universality of technical and technological solutions it is digital technologies that have created a fundamental opportunity not only to imitate real individuals and processes in the form of their digital twins, but also to generate purposeful strategies to influence clusters of individuals, using the potential of identifying, observing and using their psychosemantic, biophysical and other properties. In other words, the possible future "battle of the machines" will in any case be a "battle" of quite human and humanoid ethics encrusted in the AI and the growing set of GASes.

33.4 Digital Society: Ethics and Trust in Artificial Intelligence Systems

The question should also be raised about the point at which society crosses a certain technological threshold, when there may be an irrevocable transformation of both society as a whole and the individual into what can conventionally be called a “digital society.” This bifurcation point appears to be associated with the formation of a digital stratification platform as a dominant feature of social structuring and management. AI will play a key role in this platform and the formation of appropriate social strata, based on the criteria set during the development. The necessary conditions for this include, first of all, the achievement of a technologically possible, virtually complete awareness of the critical parameters of the life of society and the entire set of individuals.

The experience of using digital technologies in the 2020-2021 pandemic has shown both the high potential of AI for social management and their many software and hardware and socio-psychological vulnerabilities. This is also evidenced by the experience of all kinds of digital platforms (from government services to marketplaces). Nevertheless, digitalization covers more and more spheres of social life. The prospect of comprehensive integration of the created surveillance systems, databases, data processing centers, and decision support systems is being seen. The possibilities of manipulating personal choice through digital personalized models and their clustering are outlined above.

The logic of the AI development shows that they increasingly implement the principle of self-construction not only in software, but also in hardware. For example, if so far, based on input and output of information in neural networks, it is possible to reconstruct the ways of its processing and, accordingly, to obtain human-readable rules and algorithms, then in the near future the situation will change. At the IARPA annual conference back in 2016, it was discussed that the intelligence community uses complexes with such deep neural networks that it is impossible to translate their algorithms into human-readable language in a reasonable amount of time. These are so far the first symptoms that a world is emerging where decision-making will be based on criteria closed to the decision makers.

In 2012, Human Rights International raised with the UN the issue of the need to ban autonomous combat systems (robots) that make their own decisions about the use of combat equipment in their possession. Despite strong support from a number of governments, no binding UN decision has been adopted to date.

A similar problem arises in the case of robotic vehicles. One way or another, all robotic vehicles will be controlled by an AI system of varying degrees of sophistication. Automakers approach this problem in different ways. For example, Mercedes issued a statement in 2016 that in the case of road rules, the company's duty is to protect passengers, not pedestrians. In any situation, if a robotic Mercedes decides it is following the rules, the choice will be made in favor of passengers, not pedestrians. Google has taken a fundamentally different stance. The new generation of Google cars has a software video filter. If the car's video sensors recognize a child on the road, regardless of whether the child is breaking the rules or not, the car will choose saving the child's life as its first priority.

The second condition for the transition to a “digital society” is determined by the real interests of modern society itself. First of all, it is socially heterogeneous almost everywhere, which predetermines the differences not only in access to digital technology, but also in their development, implementation and use. In this sphere there is already a fierce competition for leadership and control. However, practical management in society as a whole and at the level of integration, national, regional, sectoral formations is still determined not by digital imperatives, but by political national-state, corporate and private interests. They are by no means unconditionally favorable to the rapid development of “digital society”, while it requires the total coverage of all subjects, objects and processes in society. It is enough to note the risks of mass unemployment, the level of cybersecurity to be cautious about the most optimistic pace and stages of the formation of “digital society”.

The evolution of cyber-physical systems to the status of global awareness and successful management of the evolution of society is seen in the level of centralization-decentralization of the global awareness function and the combination of institutions of organization and self-organization of life activities, including the economy. Currently, various states have enacted regulations limiting the degree of centralization of personal data.

Nevertheless, the development of technological solutions for a new generation of public systems is underway in all leading states and corporations around the world. In the USA, one of the promising DARPA projects was the creation of a “dynamic virtual environment” in which the barriers that exist today (departmental, organizational, informational and technical) will be eliminated for effective and prompt joint work of representatives of various ministries and agencies engaged in crisis situations resolution in various areas of activity (political, military, economic and social). The decision-making tools being developed for territorially distributed groups of interests are intended to provide the fullest possible comprehension of complex situations and scenarios of their dynamics, the choice of optimal solutions based on all available information without fully studying it by the principle of "knowing without reading". The technology is based on methods of fuzzy structuring of arguments, three-dimensional color visualization and corporate memory.

The development of AI and digital transformation in general leads to the formation of collectives of autonomous agents of artificial and mixed genealogies, as well as complex constructions of information and regulatory environments with multiple possibilities and pathologies and increasing levels of uncertainty for managerial decision-making.

At the same time, there is an increasing phenomenon of “degradation of natural intelligence”. In particular, constant surfing of websites can lead to an erosion of the capacity for systemic and in-depth thinking. Medical research has conclusively proven that those who spend a lot of time on the Internet quickly develop two areas of the brain-the part responsible for short-term memory and the center responsible for making quick decisions. At the same time, those areas of the brain responsible for detailed analysis, deep thinking problems remain without load and gradually lose the skill to work intensively.

Today among the most acute issues of AI development is the problem of trustworthiness, which covers the problem of confidence of consumers, regulators and other stakeholders that the AI is able to perform its tasks with the required quality and safety level.

The world of social structures is projected to syndicate major social platforms and redefine standards for human interaction between 2020 and 2030. Mutual translation of neurodescriptions, social descriptions and descriptions of the semantics of the human Internet and the Internet of Things is being established. “Codes” of nervous systems and the brain will largely be described and used not only in medicine, but also for modeling similar processes in other substrata - economic and social systems, self-organization of “smart things” and artificial systems. By 2030 semantics of different types will be able to translate into each other, and this will be used in experimental settings. Mental modes will be described fully enough, including states of consciousness in relation to different types of activity. The structure of a person's consciousness can be easily reconstructed depending on the tasks facing him. Neuronet interfaces are absolutely invisible, transparent. A person does not work behind a keyboard and a screen - he works directly with data, with meanings, with people.

The subject of control at this stage is the human body, represented by a large amount of data from sensors of different types. This also includes collectives, of which a person is currently a part. There is also interaction with distributed systems of smart things, which are constantly redesigning themselves. One such subject of control is human life, education throughout the life cycle, and the life cycles of communities.

33.5 Conclusions

The question of the ethics of AI at this level of its development is purely philosophical and futurological. The real problem of the “ethics” of operation (behavior) of AI arises only when and if AI will be able to autonomously change the contour of the knowledge used, learn from unlimited sources and types of information, adjust or even set its own goals. This prospect can only be seen in the emergence and expansion of “strong AI”. Nevertheless, there are already a significant number of beings, including ordinary humans, equipped with various “smart” devices, whose behavior can pose ethical problems. This also affects artificial systems that obey the logic of the behavior of the “swarm” that affects the individual.

The second fundamental problem is related to the ethical situation in “ordinary” human society, where ethical relativism has formally prevailed. This creates the prerequisites for the inevitable translation of this situation into the space of development and application of AI. As a result, it is very likely that the “machine world” will reproduce the human world by transferring the existing ethical problems to the cyber-physical world.

33.6 Bibliography

Ageev A.I., Asanova E.A., Gribenko O.V., etc. Are you ready for the “digit”? Assessment of the adaptability of Russia's high-tech complex to the realities of the digital economy / Edited by Doctor of Economics, prof. A.I. Ageev. M.: INES, 2018. 60 p.

Approaches to the formation and launch of new industries in the context of the National Technological Initiative on the example of the sphere "Technologies and systems of digital reality and promising "human-computer" interfaces (in terms of neuroelectronics)": Analytical report URL: http://rusneuro.net/cambiodocs/media/files/analitijeskii-doklad-podhodyk-formirovaniu-i-zapusku-novyh-otraslei-promyhlennosti.pdf.

Cadwell et al, for monitoring everyday prosthesis use a systemanic review/ Journal of NeuroEngineering and Rehabilitation. 2020. https://qz.com/1577451/century-tech-signs-deal-to-put-ai-in-700-classrooms-in-belgium/

Internet surfing changes the user's brain, SecurityLab. 2010. 30 August. URL: https:// www.securitylab.ru/news/397247.php.

Jenny Anderson: A British start-up will put AI into 700 schools in Belgium //QUARTZ. 2019. 21 March. https://qz.com/1577451/century-tech-signs-deal-to-put-ai-in-700-classrooms-in-belgium/

Kukshev V.I. Digital Economy: Problems and Solutions, Economic strategies. 2020. № 5, pp. 51–57;

Kukshev V.I. Classification of Artificial Intelligence Systems, Economic strategies. 2020. № 6, pp. 58–67.

Razin A.A. Ethics of artificial intelligence, Philosophy and Society. 2019. Issue.1 (90), pp. 57–73. https://doi.org/10.30884/jfio/2019.01.04.

Report of the Deputy Director of the FBI, Head of CJIS Stephen L. Morris "Artificial intelligence: the FBI and the police against criminals", https://ordrf.ru/wp-content/uploads/2017/

10/Обзор-отдельных-вопросов-использования-больших-данных-и-искусственного-интеллекта.pdf#page=10&zoom=100,72,537

Shuravin A. The History of Artificial Intelligence, https://wiki.programstore.ru/istoriya-iskusstvennogo-intellekta/

Socio-economic aspects of the introduction of artificial intelligence, Under the scientific editorship of A.I. Ageev. M.: It-Service, 2020.

Van Dyk: U.S. Court Recognizes Chimpanzees as Legal Persons, BuzzFeedNews. 2015. 21 April. https://www.buzzfeednews.com/article/mbvd/us-court-recognizes-chimpanzees-as-legal-persons.