ISSN 2949-2483

Volume 1 | Number 2

Special Issue
Artificial intelligence: law and ethics
Executive editor:
Chiara Gallese Nobile, PhD
(Trieste, Italy)

JOURNAL OF DIGITAL TECHNOLOGIES AND LAW

2023

ELECTRONIC
SCIENTIFIC
AND PRACTICAL
JOURNAL

JOURNAL OF DIGITAL TECHNOLOGIES AND LAW

Volume 1, № 2
2023
DOI: 10.21202/2949-2483.2023.2

ELECTRONIC SCIENTIFIC AND PRACTICAL JOURNAL

eISSN 2949-2483

Published since 2023, frequency – 4 issues a year. DOI: 10.21202/2949-2483

Address: 42 Moskovskaya Str., 420111 Kazan, Russian Federation
Tel.: +7 (843) 231-92-90
Fax: +7 (843) 292-61-59
E-mail: lawjournal@ieml.ru
Website: https://www.lawjournal.digital
Telegram: https://t.me/JournalDTL_world
VKontakte: https://vk.com/JournalDTL
Yandex.Dzen: https://dzen.ru/JournalDTL
Odnoklassniki: https://ok.ru/JournalDTL

Founder and publisher of the Journal

Kazan Innovative University named after V. G. Timiryasov. Address: 42 Moskovskaya Str., 420111 Kazan, Russian Federation. Tel.: +7 (843) 231-92-90. Fax: +7 (843) 292-61-59. E-mail: info@ieml.ru. Website: https://ieml.ru

Certificate on registering a mass medium: EL no. FS 77-84090 of 21.10.2022, issued by Roskomnadzor.
Territory of distribution: Russian Federation, foreign countries.

Date of signing the issue for publication: 2023, June 16. Hosted on the website https://www.lawjournal.digital: 2023, June 20.

JOURNAL OF DIGITAL TECHNOLOGIES AND LAW
Том 1, № 2
2023
DOI: 10.21202/2949-2483.2023.2
ELECTRONIC SCIENTIFIC AND PRACTICAL JOURNAL
eISSN 2949-2483

Content

Research article
DOI: https://doi.org/10.21202/jdtl.2023.13

Legal Aspects of the Use Artificial Intelligence in Telemedicine

Chiara Gallese Nobile

Eindhoven University of Technology
Eindhoven, Netherlands;
University of Trieste
Trieste, Italy

Keywords

Artificial intelligence,
data protection,
digital inequality,
digital technologies,
law,
legislation,
personal data,
private life,
regulation,
telemedicine

Abstract

Objective: the rapid expansion of the use of telemedicine in clinical practice and the increasing use of Artificial Intelligence has raised many privacy issues and concerns among legal scholars. Due to the sensitive nature of the data involved particular attention should be paid to the legal aspects of those systems. This article aimed to explore the legal implication of the use of Artificial Intelligence in the field of telemedicine, especially when continuous learning and automated decision-making systems are involved; in fact, providing personalized medicine through continuous learning systems may represent an additional risk. Particular attention is paid to vulnerable groups, such as children, the elderly, and severely ill patients, due to both the digital divide and the difficulty of expressing free consent.

Methods: comparative and formal legal methods allowed to analyze current regulation of the Artificial Intelligence and set up its correlations with the regulation on telemedicine, GDPR and others.

Results: legal implications of the use of Artificial Intelligence in telemedicine, especially when continuous learning and automated decision-making systems are involved were explored; author concluded that providing personalized medicine through continuous learning systems may represent an additional risk and offered the ways to minimize it. Author also focused on the issues of informed consent of vulnerable groups (children, elderly, severely ill patients).

Scientific novelty: existing risks and issues that are arising from the use of Artificial Intelligence in telemedicine with particular attention to continuous learning systems are explored.

Practical significance: results achieved in this paper can be used for lawmaking process in the sphere of use of Artificial Intelligence in telemedicine and as base for future research in this area as well as contribute to limited literature on the topic.

For citation

Gallese Nobile, C. (2023). Legal Aspects of the Use Artificial Intelligence in Telemedicine. *Journal of Digital Technologies and Law, 1*(2), 314–336. https://doi.org/10.21202/jdtl.2023.13

Contents

Introduction

The term “telemedicine” was coined in the 1970s by Thomas Bird and was defined by Strehle and Shabde as “healing at a distance” (Strehle & Shabde, 2006). A number of official definitions have been added over time, such as: “the provision of healthcare services, through use of ICT, in situations where the health professional and the patient (or two health professionals) are not in the same location. It involves secure transmission of medical data and information, through text, sound, images or other forms needed for the prevention, diagnosis, treatment and follow-up of patients. Telemedicine encompasses a wide variety of services. Those most often mentioned in peer-reviews are teleradiology, telepathology, teledermatology, teleconsultation, telemonitoring, telesurgery and teleophthalmology. Other potential services include call centres/online

information centres for patients, remote consultation/e-visits or videoconferences between health professionals.

Health information portals, electronic health record systems, electronic transmission of prescriptions or referrals (e-prescription, e-referrals) are not regarded as telemedicine services for the purpose of this Communication."[1], which has been used as a reference for national implementation (such as the definition provided by the Italian Ministry in 2012)[2].

This new way of providing health care services is not only useful to optimize processes by making them more efficient, and it is not intended to replace traditional in-patient medicine (Burrai et al., 2021), but it also serves to provide the patient with better follow-up, a greater chance of prevention and greater comfort, especially in the case of disabled or particularly frail patients. In fact, compared to traditional medicine, the devices used to monitor the patient from home allow patients not to travel as often to the hospital or to the doctor's office, remaining comfortably at their residence (which may be their own home, but also hotel, if they fall ill during holidays or business trips). This circumstance is particularly important during the pandemic, as it helps limit the chance of spreading the Sars-cov-2 infection or to get a hospital-acquired infection. It can be said that it was precisely the Covid-19 emergency that gave a boost to the use of medicine, as it sought to ensure continuity of health care even in a context where travel has long been limited.

This incredible opportunity, however, may create some risks, and many legal issues of different nature may arise. In this paper we will focus in particular on the legal issues connected to the use of Artificial Intelligence (AI) in telemedicine, with particular attention to continuous learning systems.

1. Legal Framework Regarding Telemedicine in Europe

The rapid expansion of the use of telemedicine in clinical practice has prompted, for some time now, the European Union to address the implications of the use of new technologies on patients, the development of the e-Health market, the creation of European Health Data Space, and the impact that this may have on the health services of Member States. Therefore, over the years, several soft law instruments have been issued, such as guidelines, recommendations, and other tools, analyzed in detail by Botrugno (Bortugno, 2014). In addition, the European Regulation on Medical Devices, fully applicable as of May 26, 2021, disciplines all telemedicine devices used to make a diagnosis and deliver care at a distance[3]. In Italy as well, the

1 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on telemedicine for the benefit of patients, healthcare systems and society. (2008). https://commission.europa.eu/system/files/2022-10/cwp_2023.pdf

2 Ministero della Salute. Telemedicina – Linee di indirizzo nazionali. (2012). https://www.salute.gov.it/portale/documentazione/p6_2_2_1.jsp?id=2129

3 Regulation (Eu) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745

matter is mainly left to soft law, in particular the Guidelines dating back to 2012[4]. In addition to this, in 2017 Law no. 219 regulated the matter of informed consent and Anticipated Treatment Arrangements in case of possible and future inability to self-determine, a topic that, as will be seen, in the case of telemedicine delivered through intelligent systems takes on particular importance. It has been pointed out, in fact, that if the emotional needs of the patient were to be considered an integral part of the treatment, telemedicine could be qualified only as an integrative service to the traditional ones (Campagna, 2020).

To the current regulatory framework, well examined by Campagna, will soon be added the new European Regulation on AI (AI Act), approved by the European Commission during 2021 (Campagna, 2020). This instrument will bring a very relevant change on AI-based telemedicine devices, as it will impose very stringent requirements for them to be used in medical practice, clinical trials and scientific research. Medical devices, in fact, are classified by the Regulation as "high risk". The new regulation will complement GDPR and MDR, providing for additional guarantees for users (both patients and health care professionals).

2. Artificial Intelligence in Telemedicine

With the progress of technology, many of AI techniques have been applied to telemedicine, with the aim of improving its results, since, in several areas (such as image recognition), the performance of AI has now surpassed that of humans.

However, healthcare professionals are still reticent to use these techniques in clinical practice: data from the research conducted in 2020 on Italian physicians by the Digital Innovation in Healthcare Observatory of the Politecnico di Milano on Connected Care in the Covid-19 emergency showed that only 9% of them used them before the Covid emergency and only 6% work in a facility that introduced (or enhanced) them during the pandemic. Despite this, 60% of medical specialists believe AI techniques can play a key role in emergency situations, 52% believe they help personalize care, 51% believe they help make care more effective, and 50% believe they help reduce the likelihood of clinical errors. Those results are similar to those of other surveys conducted in different countries, where concerns about liability were also raised (Scheetz, 2021). However, surveys shows that the general public has a great distrust over AI models employed in medicine (Castagno & Khalifa, 2020).

Trends in the development of AI use in telemedicine can be classified into four different groups: patient monitoring, health information technology, intelligent diagnostic assistance, and collaboration in information analysis (Pacis et al., 2018).

[4] Ministero della Salute, Telemedicina - Linee di indirizzo nazionali. https://www.salute.gov.it/portale/ehealth/homeEHealth.jsp

The branches of medicine in which these techniques are most frequently used are diabetes care, cardiology, ophthalmology, oncology, epidemiology, and dermatology. During the pandemic, telemedicine has been used, among other things, to help the management of the patients who were suspected to be infected and to provide assistance for chronic diseases (Ye., 2020). Typically, patients wear removable devices such as smart watches or sensors, or use an app on their tablets; however, more invasive methods can also be used, such as pills that can be swallowed, cameras placed inside the home, and sensors that monitor actions such as opening medication packages. Low-intrusive techniques can also be employed, such as the use of an app on the cellphone, especially in the case of younger patients, who may engage in telemedicine through a game (Giunti, 2014), or older patients (Schatten & Protrka, 2021), who need to be stimulated in engaging with the AI systems. Devices using the more complex techniques also adapt to the individual patient, continuing to learn throughout the duration of use. This creates a number of ethical and legal problems, on which doctrine and jurisprudence are trying to conduct an in-depth reflection, in order to suggest guidelines for healthcare professionals and researchers who develop these devices. Very often, in fact, the regulatory framework - as in the case of personal data protection – is complicated even for jurists, a circumstance that represents an obstacle to effective patient protection.

The complex regulatory framework is further complicated by the fragmentation of the AI discipline within the European Union, which the new and ambitious AI Act intends to remedy. Today, however, there remain, and will remain even after the entry into force of this legal instrument, many points peculiar to each Member State in terms of protection of personal data, professional liability of the doctor, informed consent, product liability and criminal liability.

3. Continuous Learning and Personalized Medicine

One of the most popular models to provide personalized medicine is the so-called continuous learning. Physicians want to provide the patient with the best possible care, which also means, in some cases, trying to adjust health care services to a specific patient's needs, due to the differences between ethnicities, genders, habits, family anamnesis, psychological state, etc. This goal can effectively be achieved with the help of AI through systems that learn through use by the patient and adapt to their characteristics over time. From a legal perspective, particular attention should be paid to models based on machine learning, i.e., machine learning (especially deep learning) and also to those that are based on a black-box approach (Rodrigues, 2020; Lakkaraju, 2019), 15 since in these cases the programmers create the model and provide relevant examples, but they do not know the end result because the model is designed to learn on its own (before it is marketed, during the training phase, but also after it is marketed). This often means that it is not possible to know the reason behind the output given by those models, so it is important to explore and understand how they behave in different scenarios (Davis, 2016).

This problem is not limited to telemedicine, but it is related to all AI applications; however, the use of this type of models in the context of telemedicine poses major risks not only because of the category of data involved, but also due to their potential impact on fundamental rights of patients. These systems may be more dangerous in case of telemedicine than they are in case of in presence medicine especially because the patient is not closely monitored by the doctor, being physically away. We will explore these themes in the following. We could consider, as an example, the case of a smart watch employed to monitor diabetes in children, which detect blood levels of sugar, measure physiological values (such as the heartbeat), reminds to eat and exercise, and suggest the correct amount of medicine to be assumed by the patient, communicating the output to the doctor. On the basis of this output, the doctor is able to set appointments, testing, and medical evaluations. In the case of continuous learning, one of the main issues is that neither the programmer nor the physician who will use the AI system can know a priori how it will behave and what it will learn from the interaction with the patient, since it is a system that evolves over time. In fact, four elements should be taken into consideration. The first is that the quality of the training reflects the quality of the samples: if the final user provides the system with biased samples, or samples characterized by poor quality, the behavior of the AI will eventually change to reflect the extended learning set.

One notable example was the case of the Tay chatbot[5]. In the case of personalized telemedicine, many examples are provided directly by the patient through the use of the system. Clearly, the patient is not an expert, and is unable to understand the implications of certain choices. For example, a child may lend the smart watch to classmates or to younger siblings, thus providing inexact measurements to the system. The consequence of an inaccurate observation may lead the system to believe that the child has lower level of sugar in the blood, and then suggest to take less medicine than needed. Even if doctors were present to check the amount of medicine suggested by the system (which would be very unlikely when the medicine must be assumed everyday, due to the costs of an examination on a daily basis), they would not be able to infer that the glucose levels were wrongly influenced by external factors or by a wrongful device management by the patient, unless the device is equipped with a camera that allows to monitor all interactions with the system. In such case, the only technical way to prevent the misuse of the device is to enact strict tele-monitoring measures, which are generally considered invasive of privacy and very intrusive in patients' lives. Even from a liability point of view, there could be doubts regarding the person who is responsible for the error of the system in such cases. Could the producer or programmer be held liable for failing to provide an unsafe device which is lacking of a monitoring mechanism? Is the doctor liable for not noticing the abnormal

[5] The case of Tay, the Microsoft's Twitter bot based on AI, that became racist and nazist. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

blood levels? Is the parent/guardian of the child liable for failing to oversee the use of the system? To tackle some of these issues, on September 28, 2022, the European Commission published a proposal for an AI Liability Directive, complementing the EU AI Act, which considers medical devices as «high risk systems», thus subject to stricter requirements. This reform was needed and long advocated by scholars, due to the fact that existing liability rules were inadequate to regulate the use of AI systems and, in particular, machine learning models (Gallese, 2022).

According to the new discipline, there is a presumption of fault for the person who made the system available on the market: as a general rule, the manufacturer or the importer will be liable for the damage produced by the system and "national courts shall presume, for the purposes of applying liability rules to a claim for damages, the causal link between the fault of the defendant and the output produced by the AI system or the failure of the AI system to produce an output». This new rule has two exceptions in case the claimant is able to demonstrate that the user of the system: • "did not comply with its obligations to use or monitor the AI system in accordance with the accompanying instructions of use or, where appropriate, suspend or interrupt its use pursuant to [Article 29 of the AI Act]», or «exposed the AI system to input data under its control which is not relevant in view of the system's intended purpose pursuant to [Article 29(3) of the Act]». The ratio of the rule is to protect users, due to the difficulty of assessing the causal link between the behavior of the system and the incorrect output (especially in case of black boxes).

Under this new discipline, it is difficult to consider doctors liable, both because they are unable to physically influence the system (unless they are provided with full remote control and tele-monitoring measures), and because they are not expert in machine learning techniques to such an extent to be able to correct the way the model is learning. A second element to consider is that machine learning methods are prone to overfitting, i.e., they tend to lose their capability to generalize. This circumstance – well known to machine learning practitioners – is generally detected by observing the relationship between the training error and the test error: when an improvement of the error on the training set leads to a worse error on the test dataset, the network is overfitting and the training process should be stopped. The latter technique – one of the most widespread and effective to prevent overfitting – is known as early stopping and is clearly not applicable in the case of systems that are supposed to be both sold "market ready" (i.e., the learning process was halted before overfitting) and able to learn outside the factory.

In our example, the smart watch may lose its capability to generalize after being used on the patient for a while, both because of an incorrect use or because in that specific time frame the number of specific types of observations were unbalanced (e.g., the patient has been ill for a while, and the heartbeat or the blood analysis have not been normal for a long time). From a legal point of view, this circumstance is relevant not only for the liability of programmers, who will be subject to the new AI liability regime, but also for the new safety and security framework introduced by the AI Act and the proposal for the so-called Cyber-resilience Act. Thirdly, these problems are exacerbated in the case of lifelong continual learning (Parisi, 2019), i.e., machine learning methods that continuously receive new instances to be

used to refine the behavior of the system. As a matter of fact, these methods are challenging because the new samples are often unbalanced (e.g., some categories are more represented than others, due to their probability to be met in the "real world"), a circumstance that can strongly affect the quality of the learning and impact the future behavior of that AI. In the case of the smart watch, it is possible that the specific characteristics of the patients lead to an over-representation of some physiological values, leading to incorrect outcomes.

Finally, the problem is basically not solvable in the extreme case of generalized class incremental learning (Mi, 2020), in which the machine learning method receives new instances that can, in principle, belong to new classes/cases never considered before: in this peculiar situation, the algorithm must be able to reconfigure its internal functioning (e.g., in the case of deep neural networks, adapt the architecture, change the topology, alter the number of neurons, and re-calibrate all parameters) which clearly prevents any realistic possibility of predicting the future behavior of the system. In our example, the patient may have unique characteristics that lead to unusual physiological levels that were not present in the training data sets. In this scenario, it may be dangerous for the patient, as the system suggestions may be wrong and lead to assume an incorrect amount of medicine, and the doctor is not present to correct the error.

The proposal for the AI Act only superficially tackles the issue of continuous learning at article 15: "[...] High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations ('feedback loops') are duly addressed with appropriate mitigation measures [...]". Recital 78 adds that "In order to ensure that providers of high-risk AI systems can take into account the experience on the use of high-risk AI systems for improving their systems and the design and development process or can take any possible corrective action in a timely manner, all providers should have a post-market monitoring system in place. This system is also key to ensure that the possible risks emerging from AI systems which continue to 'learn' after being placed on the market or put into service can be more efficiently and timely addressed. In this context, providers should also be required to have a system in place to report to the relevant authorities any serious incidents or any breaches to national and Union law protecting fundamental rights resulting from the use of their AI systems.

This provision is extremely vague and it does not clarify what possible mitigation measures or correcting actions may be considered adequate. The practical implementation of this paragraph will be difficult due to its opacity and it will leave the interpretation open to judicial discretion. The provision of a "post-market monitoring system" is a measure that may be helpful but scarcely effective when referred to personalized medicine, as producers are not able to constantly monitor every single patient. Continuous learning poses a major legal problem in multiple respects (Marchant & Lindor, 2012).

For example, it challenges the traditional liability paradigm: is it possible to adapt a strict liability regime to a situation where neither the programmer nor the manufacturer can predict the behavior of the AI? And from a technical point of view, is it safe to use on patients a product that, even if trained by the manufacturer, cannot be fully explained? Some have theorized a so-called responsibility gap (Matthias, 2004), while others have opposed this view (Tigard, 2020).

In these cases, it becomes really difficult to assume professional liability on the part of the physician. However, many privacy issues arise as well. For example, because the health care professional has no control over the way in which those systems process the patients' health data, they are not able to fully comply with the transparency obligations. We will explore the privacy issues in the next section.

4. Privacy Issues in Telemedicine

One of the most relevant aspects in the field of telemedicine and AI is definitely the protection of personal data and, in particular, the European Regulation No. 679/2016 (GDPR). Under the GDPR, health data, along with a few others, are considered "special categories of data" (Art. 9) and are therefore subject to increased legal protection.

Telemedicine, as well as traditional medicine, involves the processing of special categories of personal data, that is health data. Patients' data employed by AI models in the health care sector can rarely be fully anonymous; most often, they are considered pseudonymized, as the hospital or other health care facility is able to match it back with patients' names and other personal details.

Even when telemedicine devices are not based on AI, the identity of the patients is most of the time known to the health care professionals, as the goal of the system is to provide health care services to a specific individual. GDPR and other privacy law rules (e.g., national implementation acts, sectoral internal laws, etc.), thus, will apply. Because of the ways telemedicine is necessarily carried out, there are a number of privacy issues that will always be present.

The most obvious one is security, as interactions at a distance imply that a connection must be established (between patients and doctors, but also between different professionals). Security measure prescribed by article 32 GDPR must be in place: it must be assured that the connection is secure, that the identity of patients and doctors is ascertained, that all persons involved are adequately trained, and that data are correctly stored and deleted when no longer needed. Especially in cases involving older patients, who are generally not familiar with new technologies, the risk of data breaches is high. IoT devices may be lost, passwords may be cracked by malicious attackers, and data may be deleted by mistake. The fact that the device is in the patient's hand means that the health care professionals and their IT experts are not always present to check the system.

Before introducing these devices, therefore, patients should be adequately trained. However, recent attacks towards hospitals have highlighted that organizational measures implemented by hospitals are not always at the state-of-the-art, mainly due to the lack of proper training of employees, who have a low level of cyber-security awareness (Gioulekas, 2022).

In providing telemedicine services, hospitals and other health care facilities need to make sure that they are able to fulfill the requirements of art. 32 GDPR, which includes training their employees regarding basic cyber-security measures. If health care professional are not able to train themselves regarding security, it is hardly possible that they are able to monitor how patients interact with the system. Because they are not technical experts, they may also be unable to detect anomalies in the system, such as an incorrect training of the neural network. In this context, it is necessary that devices are regularly checked, updated, and re-calibrated by AI experts. Additional safeguards with regards to special categories of data are required by the AI Act in the last paragraph of article 10: it is possible to process them to the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to high-risk AI systems, but appropriate safeguards are necessary (such as technical limitations on the re-use of data, and the use of state-of-the-art technical measures, including pseudonymisation or encryption).

When employing AI system to deliver telemedicine services, privacy by design and by default principles should be carefully implemented in the system, taking into account the data minimization and storage limitation principles as well. To facilitate this, Recital 44 is accounting for an additional specific purpose that may allow the data processing of health data, that is the “bias monitoring, detection and correction”, in order to create a fair and trustworthy AI system.

The above-mentioned requirements can be difficult to implement in the case of continuous learning: how can accuracy, up-to-datedness, and data minimization be ensured, if it is not possible to predict how the system will behave? Another general issue that may arise in telemedicine is the transfer of medical data abroad. The distance between the patient and the doctor may involve cross-border consultation, subject to different jurisdictions.

This problem is unique to tele-medicine, as the medical consultation domain is strictly regulated in every country of the world, providing different rules regarding consent, transparency, quality standards, liability, insurance, data protection, security, identification, contractual relationships, payments, professional ethics, etc. Each of these elements may be the object of a lawsuit or an official investigation, creating conflicts of laws. Some of these aspects can be regulated by contract, but most of all are directly regulated by the foreign law and cannot be waived by contract. If a dispute arises, it would be extremely difficult for the patient, who is already in a position of disadvantage, to seek compensation and obtain legal redress. Even when a contractual arrangement is possible, it may be difficult to reach an agreement regarding responsibilities and liability regarding a system that is unpredictable.

Therefore, the risks involved are exacerbated if compared with an in-person consultation, and the extreme uncertainty regarding the behavior of the system makes it difficult to regulate the relationship between the patient and the doctor. In addition, the device employed to deliver the service may rely to cloud solutions that may have their servers outside EU.

Due to the nature of the data involved, the risks for the patients are higher if compared with a different domain. Although a doctor can store a patient's data in a foreign cloud storage even after a physical consultation, the difference in case of continuous learning devices is that the data are modified and updated in real time, thus giving actual and relevant information to potential attackers, to foreign authorities, and to the manufacturer of the system. In a device that does not collect and make use of data in real time, it is easier for patients to have control over their data, to delete them, to choose what information to store and feed into the model, and to have access to it, as required by the new proposal for the so-called Data Act. A careful assessment is needed before employing those systems, and a Data Protection Impact Assessment (DPIA) may be needed. Appropriate contractual agreement related to data protection may also be needed between the hospital and the processors providing the telemedicine devices.

5. Article 22 GDPR and Human Oversight

The use of AI-based devices in telemedicine often falls under the definition of automated decision-making (ADM) and profiling (Art. 22), for example in the case of automatic scanning of medical imaging to provide a diagnosis. Along with the principle of transparency, which guarantees data subjects the right to be informed about how their data will be used and the consequences of processing on them, the GDPR also guarantees an additional right with regard to ADM and profiling: the right to an explanation. This means that "The controller should find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision without necessarily always attempting a complex explanation of the algorithms used or disclosure of the full algorithm. The information provided should, however, be meaningful to the data subject"[6].

In this context, black box models, and in particular those based on continuous learning, cannot meet this requirement, as it is not possible to justify to the patient why the model has given a certain output. The guidelines note that "Complexity is no excuse for failing to provide information to the data subject. Recital 58 states that the principle of transparency is of particular relevance in situations where the proliferation of actors and the technological complexity of practice makes it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him are

6 *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*. https://ec.europa.eu/newsroom/article29/items/612053/en

being collected, such as in the case of online advertising"[7]. This is particularly important in the case of telemedicine, as health data are involved, and the patient is not closely monitored by a doctor. The consequences of a mistake may even be lethal, and this is one of the reasons that led the European Commission to classify medical devices as "high risk systems".

Data controllers, which may be, for example, the hospital or the producer of the telemedicine device, must provide information about the processing and how ADM and profiling might affect data subjects. In Convention 108+, Article 77 provides that "Data subjects should be entitled to know the reasoning underlying the processing of data, including the consequences of such a reasoning, which led to any resulting conclusions, in particular in cases involving the use of algorithms for automated decision-making including profiling. For instance, in the case of credit scoring, they should be entitled to know the logic underpinning the processing of their data and resulting in a "yes" or "no" decision, and not simply information on the decision itself. Having an understanding of these elements contributes to the effective exercise of other essential safeguards such as the right to object and the right to complain to a competent authority"[8]. Data subjects, in addition, have the right to express their views on ADM and profiling, the right to have the decision affecting them made by a human being, and the right to challenge the decision. In the case of AI-based telemedicine, it may be extremely difficult to manage the remote health care process in such a way that the physician oversees each and every decision made by the system and at the same time ensure that only the physician makes the decision, as this may negate the benefit of using an automated decision. Even if the doctors managed to be constantly present, in the case of continuous learning they still will not be able to tell the patient how the system may behave with them. This will probably be possible in some cases, but in the future, if these devices become widespread and a high degree of integration of telemedicine into the health service on the ground is achieved, it will be essential to keep a close watch on the aspects just outlined. Another major problem of continuous learning systems is the difficult exercise of the right to challenge the system's decision, i.e. the lack of "contestability» of its outputs, defined as "lack of an obvious means to challenge them when they produce unexpected, damaging, unfair or discriminatory results» (Edwards & Veale, 2017). Not only the patient is unable to challenge the decision of the system, but so is a distant doctor, who cannot gather all the relevant elements that were employed by the system to reach its decision (e.g., the change in the daily medicine dosage). The proposed "contestability by design» (Almada, 2019) is thus inherently inapplicable to these models. In this scenario, it is possible to argue that, for the

7 *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.* https://ec.europa.eu/newsroom/article29/items/612053

8 *Ibid.* Art. 29.

sake of transparency and fairness, interpretable AI models should be used as a standard in telemedicine and black boxes should be employed only in cases in which the health care professional is able to closely oversee the output of the system and make a larger clinical assessment, based on additional elements other than the AI output. Also, continuous learning should be limited only to decisions which are not able to harm the patient in case of errors, and should be constantly monitored by a health care professional.

6. Informed Consent

When the legal basis that allows the processing of the patient's health data is consent, which is often the case for telemedicine devices, then GDPR requires it to be both explicit and informed, other than freely given, specific, and unambiguous. Other than the usual elements that should be communicated to the data subject according to the applicable law, additional information should be provided about how the virtual consultation is performed. This information includes rights under data protection regulations, the possibility of errors in the system, contact protocols during tele-consultations, prescribing policies, and coordination of care with other professionals (Membrado, 2021). Additional transparency requirements are found in the AI Act: article 13 lists the information that should be provided to the users, who also needs to be properly instructed regarding the AI system: High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users. A detailed technical documentation is required by Article 11 and 18, which also add the requirement of keeping it up to date before the release of the system. The guidelines on ADM note that Data controllers seeking to rely on consent as a basis for profiling will need to demonstrate that data subjects understand exactly what they are consenting to, and remember that consent is not always an appropriate basis for processing. In all cases, data subjects should have enough relevant information about the intended use and consequences of the processing to ensure that any consent provided represents an informed choice. It might be argued that this is inherently not possible in the case of continuous learning, as even the doctor or the modelist will not be able to predict the consequences of using an unpredictable model. The information received by the patient needs to be concise, transparent, intelligible and in an easily accessible form, using clear and plain language, depending on the audience, adapted to the age, mental ability, and education level of the data subjects. In any case, it is only possible to use ADM and profiling with special categories of data, such as medical data, if there is explicit consent from the data subject or if it is permitted by law for substantial reasons of public interest, provided there are adequate safeguards in place. It is not enough to use public interest as a legal basis, it must also be considered substantial. It can be argued, for example, that a use aimed at combating Covid-19 falls within the "substantial public interest" grounds. Informed consent,

both from a legal and ethical point of view, becomes a central element of AI systems employed in telemedicine.

7. Vulnerable Groups

An open issue Telemedicine using intelligent systems raises more concerns when particularly vulnerable subjects are involved, such as oncological patients, patients with cognitive problems – for example, elderly people suffering from senile dementia or Alzheimer's disease-, children, neurodiverse patients (Shaw et al., 2022), or people who do not speak the language used by the doctor. First, a major issue is the ability to provide consent that is informed, free, unambiguous, and specific (Art. 7 GDPR). The doctor-patient relationship is itself an unbalanced one, where the parties are not on the same level, and where the patient is in a vulnerable situation; it may be particularly difficult to obtain consent that meets all the requirements of the law from a fragile individual, such as a cancer patient, already prostrated by the disease, might be.

If we add to this the right to obtain an explanation, it becomes clear that several problematic points may arise: in telemedicine, the doctor is far from the patient, and human contact is completely lacking. Explanations given remotely may be unclear or misunderstood, as the patient is unable to clearly see the signals of nonverbal communication.

Since these are AI systems, whose functioning, even in the case of explainable and interpretable AI, is often obscure even to experts, the risks of misunderstanding become significant and difficult to eliminate, unless a live consultation is also provided. When the system behavior is not predictable, it becomes even more difficult to explain the consequences of using the device to a vulnerable patient.

Also to be kept in mind is the digital divide, that is, the fact that not all patients have the same degree of digital literacy. This circumstance takes on significant weight in the case of telemedicine services provided in the public sphere, since there would be discrimination between users who are able to use the service and users who, for various reasons, are not. This risk is also underlined by the European Commission, which based its opinion, among other things, on the OECD report "Health at a Glance: Europe - 2018", according to which there is a risk of discrimination between users who are able to use the service and those who are not.

The Commission notes that there is a direct relationship between the level of education and the number of searches for health information on the web; in fact, similar disparities in the use of digital solutions for health promotion and disease prevention are also likely, and there is a risk that digital tools such as apps, wearable technology, and online forums will not benefit those who need them most, potentially widening health-related inequalities (Oliveira, 2020). The divide may even be exacerbated by continuous learning systems, as those who are better at using the telemedicine devices are more likely to get a more precise output. This risk should be evaluated by health care professionals when delivering tele-health services.

8. The Balance Between Privacy and Protection from Harm at Distance

As we discussed above, one of the major problems of continuous learning in telemedicine is that the model "evolves" while the patient is physically distant from the doctor (and from the IT support) and at the same time the system is usually designed to be less intrusive as possible. This leads to the necessity for a balance between constantly monitoring patients to ensure their safety and preserving their privacy, without making them feel uncomfortable with the technology. Therefore, four elements become relevant: technical measures, organizational measures, psychological factors, and legal requirements. From a technical point of view, a number of different techniques have been studied to solve some of the privacy issues discussed above, such as problems related to identity management, using blockchain, encryption, or federated learning under edge computing (Ahmad, 2021; Jain et al., 2022; Wang, 2021; Ma et al., 2020; Wang, 2022).

However, some of the problems cannot be solved only using technology: organizational measures play a crucial role in achieving the balance between the protection of privacy and the benefits of surveillance technology as a countermeasure against unexpected malfunction of the system.

In fact, as a first measure, the harm caused by malfunctions of the system can be mitigated by scheduling appropriate maintenance interventions delivered by IT experts and medical doctors, who can test the system, evaluate its performance, and assess the patient's health. These measures can be enacted keeping the burden on patients at the very minimum (e.g., combining the maintenance to the regular in-person appointments at the hospital). However, a second measure is equally important: training doctors and patients on the correct use of telemedicine technology is crucial both to preserve privacy and to mitigate the risk of malfunctioning and incorrect use. Once doctors are correctly trained, for example, they may be able to recognize peculiar and unique features of patients that could lead the self-learning system to generate errors.

From a psychological point of view, vulnerable patients may need additional support with: understanding the functioning of the system in order to be able to use it correctly and express a free informed consent; accepting the system in their life without feeling a disruption of everyday activities; learn how to express their concerns and discomfort (including physical pain) at a distance (Yakar, 2021).

A trained therapist, working together with doctors and IT experts, could be appointed to help patients in overcoming those issues. On the other hand, doctors should make sure that patients are able to understand their instructions and they should regularly check if their consent is truly informed. Privacy law is not an obstacle against the enacting the above-mentioned measures, as GDPR provides for appropriate means to protect patients' fundamental rights. As an additional organizational measure, hospitals should appoint a privacy expert to monitor the use of telemedicine devices and to guide all the stakeholders.

9. AI Auditing Measures as an Additional Safeguard

The academic field of algorithmic auditing within the broader AI auditing framework is becoming more and more popular in the recent years, especially in the domain of machine learning. AI auditing means incorporating Ethics, Human Rights, and Law into the whole AI development life cycle and in the post-market phase (LaBrie & Steinke, 2019; Mökander & Floridi, 2022; Mantelero & Esposito, 2021; Koshiyama et al., 2021; Mökander, 2022; Floridi, 2022), while at the same time checking its technical soundness (e.g., safety, security, performance, and correctness of pre-processing techniques). To this aim, many practical tools have been created, such as several Ethics Canvas (e.g., the Open AI Canvas, the Data Ethics Canvas, the AI Ethics Canvas (Kalra, 2020). This type of assessment and continuous monitoring is important to detect biases, technical and statistical errors, security flaws, and detrimental effects during the post-market phase. In fact, due to the close look into the whole processes and the compliance mechanism enacted since the early stages of the AI development (i.e., even before the data collection), it is possible to prevent many of the errors that could lead to privacy violation, errors in the system, security breaches, and discrimination. It is therefore recommended to implement AI auditing procedures every time continuous learning models are envisaged in clinical practice.

Conclusions

In this brief contribution we have seen how telemedicine carried out through continuous learning intelligent systems is a useful tool to ensure better health care to the patient, but that its diffusion may also bring new challenges to the jurist and to health care professionals. New legal problems, such as the regulation of systems that adapt to the patient or the protection of vulnerable subjects, and their ability to understand how the AI system works, in order to be able to express a free consent, will have to be addressed and dissected by the doctrine in the near future. Health care facilities intending to employ AI systems to deliver telemedicine services should train their employees and the patients to handle the system safely and securely, in order to avoid data breaches and an incorrect use of the devices. An efficient way to achieve this goal is to have dedicated teams of IT experts, doctors, and trained therapists. Although the European Union has tried to give a legal response to the development of AI techniques, there are still many unresolved issues. It is hoped, therefore, that the issue can be adequately explored before telemedicine through AI techniques becomes a widespread and common practice on the EU territory.

References

Ahmad, R. W. (2021). The role of blockchain technology in telehealth and telemedicine. *International Journal of Medical Informatics*, *148*, 104399. https://doi.org/10.1016/j.ijmedinf.2021.104399

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 2–11). https://doi.org/10.2139/ssrn.3264189

Botrugno, C. (2014). Un diritto per la telemedicina: analisi di un complesso normativo in formazione. *Politica del Diritto*, *4*(45), 639–668. https://doi.org/10.1437/78949

Burrai, F., Gambella, M., & Scarpa, A. (2021). L'erogazione diprestazioni sanitarie in telemedicina. *Giornale di Clinica Nefrologica e Dialisi*, *33*, 3–6.

Campagna, M. (2020). Linee guida per la Telemedicina: considerazioni alla luce dell'emergenza Covid-19. *Corti Supreme e Salute*, *3*, 11–25.

Castagno, S., & Khalifa, M. (2020). Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. *Frontiers in artificial intelligence*, *2*(5), 84–92. https://doi.org/10.3389/frai.2020.578983

Davis, E. (2016). AI Amusements: The Tragic Tale of Tay the Chatbot. *AI Matters*, *2*(4), 20–24. https://doi.org/10.1145/3008665.3008674

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, *16*, 18–26.

Floridi, L. (2022). capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091

Gallese, Ch. (2022). Suggestions for a revision of the European smart robot liability regime. In *Proceedings of the 4th European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2022)*. https://doi.org/10.34190/eciair.4.1.851

Gioulekas, F. (2022). A Cybersecurity Culture Survey Targeting Healthcare Critical Infrastructures. *Healthcare*, *10*, 327–333. https://doi.org/10.3390/healthcare10020327

Giunti, G. (2014). The Use of a Gamified Platform To Empower And Increase Patient Engagement in Diabetes Mellitus Adolescents. In *American Medical Informatics Association Annual Symposium*.

Jain, N., Gupta V., & Dass, P. (2022). Blockchain: A novel paradigm for secured data transmission in telemedicine. In *Wearable Telemedicine Technology for the Healthcare Industry* (pp. 33–52).

Kalra, A. (2020). *Artificial Intelligence Ethics Canvas: A Tool for Ethical and Socially Responsible AI*.

Koshiyama, A. S., Kazim, E., Treleaven, P. C., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S. A., & Lomas, E. (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. *Software Engineering eJournal*. https://doi.org/10.2139/ssrn.3778998

LaBrie, R., & Steinke, G. (2019). Towards a framework for ethical audits of AI algorithms. In *Twenty-fifth Americas Conference on Information Systems*.

Lakkaraju, H. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131–138). https://doi.org/10.1145/3306618.3314229

Ma, M., Shuqin, F., & Feng, D. (2020). Multi-user certificateless public key encryption with conjunctive keyword search for cloud-based telemedicine. *Journal of Information Security and Applications*, *55*, 102652. https://doi.org/10.1016/j.jisa.2020.102652

Mantelero, A., & Esposito, S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review*, *41*, 105561. https://doi.org/10.1016/j.clsr.2021.105561

Marchant, G., & Lindor, R. (2012). The coming collision between autonomous vehicles and the liability system. *Santa Clara Law Review*, *52*, 1321–1340.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Membrado, C. G. (2021). Telemedicina, ética y derecho en tiempos de COVID-19. Una mirada hacia el futuro. *Revista Clinica Espanola*, *221*, 408–410. https://doi.org/10.1016/j.rce.2021.03.002

Mi, F. (2020). Generalized Class Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 240–241).

Mökander, J. (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, *32*, 241–268. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J., & Floridi, L. (2022). Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics*, *6*, 1–18. https://doi.org/10.1007/s43681-022-00171-7

Oliveira, T. (2020). Bringing health care to the patient: An overview of the use of telemedicine in OECD countries. *OECD, Directorate for Employment, Labour and Social Affairs, Health Committee*.

Pacis, D., Mitch, M., Edwin, D. C., Subido, Jr., & Bugtai, N. (2018). Trends in telemedicine utilizing artificial intelligence. In *AIP conference proceedings*. AIP Publishing LLC.

Parisi, G. (2019). Continual lifelong learning with neural networks: A review, *Neural Networks*, *113*, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, *4*, 100005. https://doi.org/10.1016/j.jrt.2020.100005

Schatten, M., & Protrka, R. (2021). Conceptual Architecture of a Cognitive Agent for Telemedicine based on Gamification. In *Central European Conference on Information and Intelligent Systems* (pp. 3–10).

Scheetz, J. (2021). A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific reports*, *11.1*, 1–10.

Shaw, S., Davis, L-J., & Doherty, M. (2022). *Considering autistic patients in the era of telemedicine: the need for an adaptable, equitable, and compassionate approach, BJGP open 6.1*.

Strehle, E. M., & Shabde, N. (2006). One hundred years of telemedicine: does this new technology have a place in paediatrics? *Archives of disease in childhood*, *91.12*, 956–959. https://doi.org/10.1136/adc.2006.099622

Tigard, D. (2020). There is no techno-responsibility gap. *Philosophy & Technology*, 1–19.

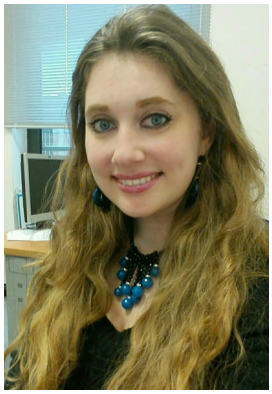
Wang, R. (2022). Privacy-Preserving Federated Learning for Internet of Medical Things under Edge Computing. *IEEE Journal of Biomedical and Health Informatics*.

Wang, W. (2021). A privacy protection scheme for telemedicine diagnosis based on double blockchain. *Journal of Information Security and Applications*, *61*, 102845. https://doi.org/10.1016/j.jisa.2021.102845

Yakar, D. (2021). Do People Favor Artificial Intelligence Over Physicians? A Survey Among the General Population and Their View on Artificial Intelligence in Medicine. *Value in Health*, *3*, 12–23. https://doi.org/10.1016/j.jval.2021.09.004

Ye, J. (2020). The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. *JMIR medical informatics*, *8.7*, e19866. https://doi.org/10.2196/19866

Author information



Chiara Gallese Nobile – PhD, Researcher (postdoc) of research data management, Eindhoven University of Technology (Eindhoven, Netherlands), Researcher (postdoc) of the Department of Mathematics and Geosciences, University of Trieste (Trieste, Italy).
Address: P/O 513 5600 MB Eindhoven, the Netherlands
E-mail: cgallese@liuc.it
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57222726276
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/AGE-9594-2022
ORCID ID: https://orcid.org/0000-0001-8194-0261
Google Scholar ID: https://scholar.google.com/citations?user=Vmoen8UAAAAJ

Conflict of interests

The author is an international editor of the Journal; the article has been reviewed on general terms.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – May 4, 2023
Date of approval – May 20, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 347.1:654:004.8
EDN: https://elibrary.ru/vskcfb
DOI: https://doi.org/10.21202/jdtl.2023.13

Правовые аспекты использования искусственного интеллекта в телемедицине

Кьяра Галлезе-Нобиле

Эйндховенский технологический университет
г. Эйндховен, Королевство Нидерландов;
Университет Триеста
г. Триест, Итальянская Республика

Ключевые слова

Законодательство,
защита данных,
искусственный интеллект,
персональные данные,
право,
регулирование,
телемедицина,
цифровое неравенство,
цифровые технологии,
частная жизнь

Аннотация

Цель: стремительное распространение телемедицины в клинической практике и возрастающая роль искусственного интеллекта ставят перед юристами множество проблем относительно охраны неприкосновенности частной жизни. Повышенная сензитивность данных в этой области заставляет уделить особое внимание правовым аспектам таких систем. В статье исследуются правовые последствия использования искусственного интеллекта в телемедицине, в частности, систем непрерывного обучения и автоматизированного принятия решений; фактически оказание персонализированных медицинских услуг через системы непрерывного обучения может представлять дополнительный риск. Особого внимания заслуживают уязвимые группы населения – дети, пожилые люди и тяжелобольные пациенты – как по причине цифрового неравенства, так и из-за сложностей с выражением своего согласия.

Методы: сравнительно-правовые и формально-правовые методы исследования позволили проанализировать текущее состояние регулирования искусственного интеллекта и выявить его соотношение с нормами регулирования телемедицины, Общим регламентом ЕС по защите персональных данных и другими нормами.

Результаты: изучены правовые последствия использования искусственного интеллекта в телемедицине, в частности, систем непрерывного обучения и автоматизированного принятия решений; автор приходит к выводу, что оказание персонализированных медицинских услуг через системы непрерывного обучения представляет дополнительные риски, и предлагает пути их минимизации. Автор также уделяет особое внимание вопросам информированного согласия уязвимых групп населения (детей, пожилых, тяжелобольных).

Научная новизна: изучены актуальные риски и проблемы, возникающие в сфере использования искусственного интеллекта в телемедицине, при этом особое внимание уделено системам непрерывного обучения.

Практическая значимость: полученные результаты восполняют недостаток научных исследований по данной теме, могут быть использованы в законодательном процессе в сфере использования искусственного интеллекта в телемедицине, а также в качестве основы для будущих исследований в данной области.

Для цитирования

Галлезе-Нобиле, К. (2023). Правовые аспекты использования искусственного интеллекта в телемедицине. *Journal of Digital Technologies and Law*, *1*(2), 314–336. https://doi.org/10.21202/jdtl.2023.13

Список литературы

Ahmad, R. W. (2021). The role of blockchain technology in telehealth and telemedicine. *International Journal of Medical Informatics*, *148*, 104399. https://doi.org/10.1016/j.ijmedinf.2021.104399

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 2–11). https://doi.org/10.2139/ssrn.3264189

Botrugno, C. (2014). Un diritto per la telemedicina: analisi di un complesso normativo in formazione. *Politica del Diritto*, *4*(45), 639–668. https://doi.org/10.1437/78949

Burrai, F., Gambella, M., & Scarpa, A. (2021). L'erogazione diprestazioni sanitarie in telemedicina. *Giornale di Clinica Nefrologica e Dialisi*, *33*, 3–6.

Campagna, M. (2020). Linee guida per la Telemedicina: considerazioni alla luce dell'emergenza Covid-19. *Corti Supreme e Salute*, *3*, 11–25.

Castagno, S., & Khalifa, M. (2020). Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. *Frontiers in artificial intelligence*, *2*(5), 84–92. https://doi.org/10.3389/frai.2020.578983

Davis, E. (2016). AI Amusements: The Tragic Tale of Tay the Chatbot. *AI Matters*, *2*(4), 20–24. https://doi.org/10.1145/3008665.3008674

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, *16*, 18–26.

Floridi, L. (2022). capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091

Gallese, Ch. (2022). Suggestions for a revision of the European smart robot liability regime. In *Proceedings of the 4th European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2022)*. https://doi.org/10.34190/eciair.4.1.851

Gioulekas, F. (2022). A Cybersecurity Culture Survey Targeting Healthcare Critical Infrastructures. *Healthcare*, *10*, 327–333. https://doi.org/10.3390/healthcare10020327

Giunti, G. (2014). The Use of a Gamified Platform To Empower And Increase Patient Engagement in Diabetes Mellitus Adolescents. In *American Medical Informatics Association Annual Symposium*.

Jain, N., Gupta V., & Dass, P. (2022). Blockchain: A novel paradigm for secured data transmission in telemedicine. In *Wearable Telemedicine Technology for the Healthcare Industry* (pp. 33–52).

Kalra, A. (2020). *Artificial Intelligence Ethics Canvas: A Tool for Ethical and Socially Responsible AI*.

Koshiyama, A. S., Kazim, E., Treleaven, P. C., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S. A., & Lomas, E. (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. *Software Engineering eJournal*. https://doi.org/10.2139/ssrn.3778998

LaBrie, R., & Steinke, G. (2019). Towards a framework for ethical audits of AI algorithms. In *Twenty-fifth Americas Conference on Information Systems*.

Lakkaraju, H. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131–138). https://doi.org/10.1145/3306618.3314229

Ma, M., Shuqin, F., & Feng, D. (2020). Multi-user certificateless public key encryption with conjunctive keyword search for cloud-based telemedicine. *Journal of Information Security and Applications*, *55*, 102652. https://doi.org/10.1016/j.jisa.2020.102652

Mantelero, A., & Esposito, S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review*, *41*, 105561. https://doi.org/10.1016/j.clsr.2021.105561

Marchant, G., & Lindor, R. (2012). The coming collision between autonomous vehicles and the liability system. *Santa Clara Law Review*, *52*, 1321–1340.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Membrado, C. G. (2021). Telemedicina, ética y derecho en tiempos de COVID-19. Una mirada hacia el futuro. *Revista Clinica Espanola*, *221*, 408–410. https://doi.org/10.1016/j.rce.2021.03.002

Mi, F. (2020). Generalized Class Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 240–241).

Mökander, J. (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, *32*, 241–268. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J., & Floridi, L. (2022). Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics*, *6*, 1–18. https://doi.org/10.1007/s43681-022-00171-7

Oliveira, T. (2020). Bringing health care to the patient: An overview of the use of telemedicine in OECD countries. *OECD, Directorate for Employment, Labour and Social Affairs, Health Committee*.

Pacis, D., Mitch, M., Edwin, D. C., Subido, Jr., & Bugtai, N. (2018). Trends in telemedicine utilizing artificial intelligence. In *AIP conference proceedings*. AIP Publishing LLC.

Parisi, G. (2019). Continual lifelong learning with neural networks: A review, *Neural Networks*, *113*, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, *4*, 100005. https://doi.org/10.1016/j.jrt.2020.100005

Schatten, M., & Protrka, R. (2021). Conceptual Architecture of a Cognitive Agent for Telemedicine based on Gamification. In *Central European Conference on Information and Intelligent Systems* (pp. 3–10).

Scheetz, J. (2021). A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific reports*, *11.1*, 1–10.

Shaw, S., Davis, L-J., & Doherty, M. (2022). *Considering autistic patients in the era of telemedicine: the need for an adaptable, equitable, and compassionate approach, BJGP open 6.1*.

Strehle, E. M., & Shabde, N. (2006). One hundred years of telemedicine: does this new technology have a place in paediatrics? *Archives of disease in childhood*, *91.12*, 956–959. https://doi.org/10.1136/adc.2006.099622

Tigard, D. (2020). There is no techno-responsibility gap. *Philosophy & Technology*, 1–19.

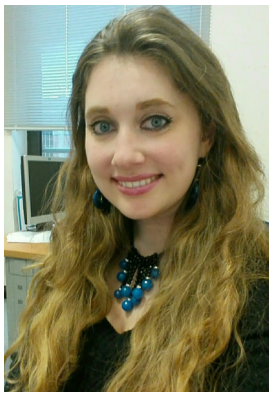
Wang, R. (2022). Privacy-Preserving Federated Learning for Internet of Medical Things under Edge Computing. *IEEE Journal of Biomedical and Health Informatics*.

Wang, W. (2021). A privacy protection scheme for telemedicine diagnosis based on double blockchain. *Journal of Information Security and Applications*, *61*, 102845. https://doi.org/10.1016/j.jisa.2021.102845

Yakar, D. (2021). Do People Favor Artificial Intelligence Over Physicians? A Survey Among the General Population and Their View on Artificial Intelligence in Medicine. *Value in Health*, *3*, 12–23. https://doi.org/10.1016/j.jval.2021.09.004

Ye, J. (2020). The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. *JMIR medical informatics*, *8.7*, e19866. https://doi.org/10.2196/19866

Сведения об авторе



Галлезе-Нобиле Кьяра – доктор наук, научный сотрудник (постдок) по управлению исследовательскими данными, Эйндховенский технологический университет (Эйндховен, Королевство Нидерландов); научный сотрудник (постдок) департамента математики и наук о земле, Университет Триеста (Триест, Итальянская Республика)
Адрес: а/я 513 5600 МБ Эйндховен, Королевство Нидерландов
E-mail: cgallese@liuc.it
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57222726276
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/AGE-9594-2022
ORCID ID: https://orcid.org/0000-0001-8194-0261
Google Scholar ID: https://scholar.google.com/citations?user=Vmoen8UAAAAJ

Конфликт интересов

Автор является международным редактором журнала, статья прошла рецензирование на общих основаниях.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.27.91 / Гражданское право отдельных стран
Специальность ВАК: 5.1.3 / Частно-правовые (цивилистические) науки

История статьи

Дата поступления – 4 мая 2023 г.
Дата одобрения после рецензирования – 20 мая 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.14

Legal Means of Providing the Principle of Transparency of the Artificial Intelligence

Yuliya S. Kharitonova

Lomonosov Moscow State University
Moscow, Russian Federation

Keywords

Algorithm,
artificial intelligence,
automated data processing,
autonomy,
decision-making,
digital economy,
digital technologies,
ethics,
law,
transparency

Abstract

Objective: to analyze the current technological and legal theories in order to define the content of the transparency principle of the artificial intelligence functioning from the viewpoint of legal regulation, choice of applicable means of legal regulation, and establishing objective limits to legal intervention into the technological sphere through regulatory impact.

Methods: the methodological basis of the research is the set of general scientific (analysis, synthesis, induction, deduction) and specific legal (historical-legal, formal-legal, comparative-legal) methods of scientific cognition.

Results: the author critically analyzed the norms and proposals for normative formalization of the artificial intelligence transparency principle from the viewpoint of impossibility to obtain the full technological transparency of artificial intelligence. It is proposed to discuss the variants of managing algorithmic transparency and accountability based on the analysis of social, technical and regulatory problems created by algorithmic systems of artificial intelligence. It is proved that transparency is an indispensible condition to recognize artificial intelligence as trustworthy. It is proved that transparency and explainability of the artificial intelligence technology is essential not only for personal data protection, but also in other situations of automated data processing, when, in order to make a decision, the technological data lacking in the input information are taken from open sources, including those not having the status of a personal data storage. It is proposed to legislatively stipulate the obligatory audit and to introduce a standard, stipulating a compromise between the technology abilities and advantages, accuracy and explainability of its result, and the rights of the participants of civil relations. Introduction

of certification of the artificial intelligence models, obligatory for application, will solve the issues of liability of the subjects obliged to apply such systems. In the context of professional liability of professional subjects, such as doctors, militants, or corporate executives of a juridical person, it is necessary to restrict the obligatory application of artificial intelligence if sufficient transparency is not provided.

Scientific novelty: the interdisciplinary character of the research allowed revealing the impossibility and groundlessness of the requirements to completely disclose the source code or architecture of the artificial intelligence models. The principle of artificial intelligence transparency may be satisfied through elaboration and provision of the right of the data subject and the subject, to whom the decision made as a result of automated data processing is addressed, to reject using automated data processing in decision-making, and the right to object to the decisions made in such a way.

Practical significance: is due to the actual absence of sufficient regulation of the principle of transparency of artificial intelligence and results of its functioning, as well as the content and features of the implementation of the right to explanation the right to objection of the decision subject. The most fruitful way to establish trust towards artificial intelligence is to recognize this technology as a part of a complex sociotechnical system, which mediates trust, and to improve the reliability of these systems. The main provisions and conclusions of the research can be used to improve the legal mechanism of providing transparency of the artificial intelligence models applied in state governance and business.

For citation

Kharitonova, Yu. S. (2023). Legal Means of Providing the Principle of Transparency of the Artificial Intelligence. *Journal of Digital Technologies and Law, 1*(2), 337–358. https://doi.org/10.21202/jdtl.2023.14

Contents

Introduction

The Russian law formulates the principles of National Strategy for artificial intelligence development up to 2030, which include, inter alia, transparency as explainability of the artificial intelligence functioning and achieving results, non-discriminatory access of users of the products created with the artificial intelligence technologies to information about the artificial intelligence algorithms applied in these products (clause 19 of the Strategy, adopted by the Decree of the President of the Russian Federation "On development of artificial intelligence in the Russian Federation" No. 490 of 10.10.2019). The notions of "explainability" and "non-discrimination" of the artificial intelligence functioning are highlighted as constituents of the transparency principle.

Information disclosure is also stipulated by international acts and national legislation of many countries. These rules are, first of all, closely touch upon the issues of human rights and freedoms protection, like, for example, in the General Data Protection Regulation (GDPR); Regulation (EU) 2016/679[1].

In Russia, the human rights and freedoms protection and safety of the artificial intelligence functioning are stipulated in the Strategy 2030 as separate principles, although they are closely connected to transparency. Assumingly, non-discrimination results in providing protection of the human rights and freedoms guaranteed by the Russian and international legislation. The principle of safety of the artificial intelligence functioning is defined as inadmissibility of using the artificial intelligence for purposeful incurring harm to citizens and juridical persons, as well as prevention and minimization of risks of negative consequences of using the artificial intelligence technologies. Assumingly, the transparency principle also allows achieving safety when using artificial intelligence.

In the absence of a clear legal vision of the content of the principle of safety of the artificial intelligence functioning, we consider it important to define the notion of transparency and research the admissible limits of legal intervention into the technological sphere through regulatory impact.

In the context of the artificial intelligence functioning, transparency may be viewed from the angle of technology, ethics, and law. Interdisciplinary approach allows a critical view at the norms and proposals to be normatively formalized, given that complete technological transparency of the artificial intelligence is impossible. It is necessary to discuss the variants of managing policy for the algorithmic transparency and accountability based on the analysis of social, technical and regulatory problems, created by the algorithmic artificial intelligence systems.

---

[1] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. http://eur-lex.europa.eu/

1. The "black box" notion and its significance for legal formalization of using the artificial intelligence technology for decision-making

For decades, artificial intelligence projects relied on human experience accumulated by engineers and were both explicitly elaborated and easily understandable. A significant progress in the sphere of artificial intelligence was achieved by using controllable learning systems intended to repeat people's decisions (Hastie et al., 2009; LeCun et al., 2015; Pak & Kim, 2017). For example, expert systems based on decision trees are perfect models of human decision-making, hence, naturally understandable for both developers and end users (Lawrence & Wright, 2001; Cho et al., 2002). The same is true for data tables (Cragun & Steudel, 1987). However, after the leading methodologies of artificial intelligence change the paradigm for machine learning systems, based on deep neural networks (DNN), novelties appeared (Samek et al., 2021).

Easiness of comprehension was sacrificed for the rate of decision-making and the technology was called "a black box" – nontransparent for human comprehension but extremely potent in terms of both results and learning in new spheres. The models which "open the black box", making a nonlinear and complex process of decision-making clear for human observers, are a promising solution of the "black box" AI problem, but are limited, at least in their present state, in their ability to make these processes more transparent for most observers. Artificial intelligence uses deep learning (DL), an algorithmic system of deep neural networks, which are generally nontransparent or hidden from human comprehension.

How does this nontransparence manifest itself and what are its reasons? The main purpose of machine learning (ML) is teaching system of exact decision making – predictors, capable of helping automate the tasks which otherwise people would have to perform. Machine learning possesses a lot of algorithms which have demonstrated great success in science and industry. The most popular ML means are kernel methods (Hofmann et al., 2008) and, especially in the recent decade, deep learning methods (Vargas et al., 2018).

As ML is increasingly often used in actual software and applications, it has become a common opinion that high accuracy of a decision or a forecast can be insufficient in practice (Gunning, 2017).

The first difficulty is due to a multi-scale and distributed nature of neural networks representations. Certain neurons are only activated for several points of data, while others act more globally. Thus, a forecast is a sum of local and global effects, which complicates (or precludes) a search of a root point x, which linearly expands to a forecast for the data point of interest. Transition from the global to the local effect induces nonlinearity, which cannot be detected (Samek et al., 2021).

The second source of instability occurs due to a large depth of modern neural networks with their shattered gradient problem (Balduzzi et al., 2017). A gradient in neural networks is a vector of partial derivatives of the loss function by the weighs of the neural network. It is used in weigh optimizer to improve the model quality. The gradient shows the change of errors on various data sets.

Finally, there is a problem of explainability of the artificial intelligence technology with the need to search for a root point x, on which the explanation will be based and which is simultaneously close to data and is not an adversarial example (the problem of adversarial examples) (Goodfellow et al., 2014). The problem of adversarial examples is explained by gradient noise, which makes the model provide an excessive reaction (overreact) to certain pixel perturbations, as well as by high dimensionality of the data (Najafabadi et al., 2015), when multiple pixel effects sum up producing a large effect on the model result (Samek et al., 2021).

These features of functioning of the artificial intelligence as a class results in that, while big data and huge computations are available, achieving a superhuman productivity requires "zero human knowledge" (Silver et al., 2017).

Researchers propose to admit that the artificial intelligence is inside the sociotechnical system, which mediates trust and, while increasing the reliability of these systems to make these processes less nontransparent for most observers, we thus increase trust to artificial intelligence (von Eschenbach, 2021). In this context, exclusion of a human from the decision-making process adds trust to it, excluding the factor of subjectivity in the result obtained.

At the same time, the issue of trust depends not only on the ability for a human to interfere into the decision-making process of the artificial intelligence. At the modern stage, demand for explainable artificial intelligence (XAI) is growing. R. Yampolskiy stated that "if everything we have is a 'black box', then it is impossible to understand the reasons of failures and to increase the system safety. Besides, if we get used to accept the answers of the artificial intelligence without explanations of reasons, we will not be able to detect when it starts giving wrong or manipulative answers" (Yampolskiy, 2019). The researcher vividly describes the dangers on non-transparent artificial intelligence, offering to imagine that in the nearest future artificial intelligence may be mistaken in diagnosing illnesses in 5% of cases, which will result in mass operations of healthy people. The absence of the mechanism to check the artificial intelligence model for deviations and to prevent such failures may lead to irreparable consequences. Thus, transparency and accountability are the tools facilitating making just algorithmic decisions, providing the basis for obtaining the opportunity to turn to a meaningful explanation, correction or means to identify drawbacks which may lead to compensation processes (Koene, 2019).

2. Legal and ethical risks of applying nontransparent technology

The issue of transparency is defined in the Russian Concept of development of regulating relations in the sphere of artificial intelligence technologies and robotics up to 2024 (further – Concept 2024) as "using probability estimations in decision-making by artificial intelligence systems and impossibility, in some cases, to fully explain the decision made by them (the problem of algorithmic transparency of artificial intelligence systems)"[2].

Concept 2024 lists transparency among such areas of concern in regulating artificial intelligence as maintain the balance between personal data protection requirements and the need to use them for training artificial intelligence systems; defining the object and limits of regulating the use of artificial intelligence technologies and robotics; legal "delegation" of decisions to artificial intelligence and robotics systems; liability for incurring harm using artificial intelligence and robotics systems. In other words, the issues of legal provision of the artificial intelligence transparency play a conceptual role in elaborating legal approaches.

As was shown above, a developer provides data but cannot control the criteria on which an artificial intelligence yielded a result or a forecast. Seemingly, sometimes it is not possible to develop a meaningful neural network. This is due to the difficulties with defining input data and their factual insufficiency. Actually, the loss of control over artificial intelligence is based on the uncertainty of data with which the model interacts (Kharitonova et al., 2021).

Are developers and jurists capable of reasonably intervening into the system functioning and contesting its conclusions, if they do not comprehend the principles under those conclusions? Developers may point out the criteria for making decisions, but artificial intelligence may autonomously supplement the conditionally lacking data to formulate final decisions. For example, a machine analyzes dot or pixels without knowing whether this is the color of skin or eyes. It manipulates with pixels, not the overall picture.

At the same time, the decisions made by humans – lawyers and even judges, whose activity is thoroughly regulated in this aspect, – are not void of a conscious and/or unconscious bias. Researches of human prejudices showed that people are cognitively prone to bias and stereotypes (Plous, 2003; Quillian, 2006), although contemporary forms of prejudice are hard to detect and can be unknown even to their carriers. The practice of justifying decisions may be insufficient for counteracting the influence of various factors, while the reasons suggested for a decision-making human may hide the motifs hardly known to those who make decisions (McEwen, 2018).

[2] *On adopting the Concept of development of regulating relations in the sphere of artificial intelligence technologies and robotics up to 2024*: Order of the Russian Government of 19.08.2020 No. 2129-r.

That is, algorithmic and human prejudice and non-explainability of the decision made often exist in a latent form, unperceived by its carriers and undetected by the third persons. This implies that a soulless, emotionless algorithm can still serve as an objective measurement for decision-making, as it is void of personal subjective prejudices.

The risk of using nontransparent artificial intelligence becomes critically important if such technology must be applied by the subject of activity. For example, in a moving unmanned vehicle the decision is made by the artificial intelligence system, while the liability for a source of increased danger is still imposed on the driver (Payre & Cestac, 2014). Another example refers to the nearest future. Today, robots based on artificial intelligence are increasingly used to assist surgeons (Kalis et al., 2014). During medical assistance, some procedures become obligatory, hence, a doctor may find themselves in a situation when their decisions incurred liability, though factually the harm was caused by the problems with artificial intelligence software.

Researching the issues of legal intervention to spreading deepfakes, V. O. Kalyatin comes to a conclusion that "the relevant legislation should be developed not in respect of deepfakes as such, but in respect of using AI in general" (Kalyatin, 2022). Jurists face a choice: to remain in the current legal tradition or to create a new one. We believe that attempts to create a legal regime of entrepreneurs' using artificial intelligence cannot be successful in the absence of understanding of its technological features. However, transparency as explainability of the technology cannot be understood literally. We need to create criteria to check the results of artificial intelligence functioning in order to observe the citizens' rights and freedoms, to protect state and public interests.

3. Automated system of data processing and the data quality

If transparency per se is not inherent to the nature of algorithms (Kalpokas, 2019), under the condition that information is provided at the input to launch artificial intelligence applications, then a question arises about the possibility of prioritizing the rules data analysis by the artificial intelligence algorithm.

In literature, several aspects of algorithmic transparency and accountability are highlighted, which include increased awareness, accountability when using algorithmic solutions, first of all in the state sector, as well as normative surveillance and legal liability, leading to a global coordination of algorithmic governance (Koene et al., 2019).

Awareness, viewed by many researchers as a solution to the problem of transparency, can be interpreted in many different ways. First of all, when providing transparency of artificial intelligence, heavy emphasis is placed on working with data and on awareness about their use in a certain way.

Notably, many jurisdictions stipulate data analysis and its limits in relation to personal data. In Russia, provisions of Article 16 of the Law on personal data[3] are in force, according to which it is prohibited to make decisions, based exclusively on automated processing of personal data, which generate legal consequences for the personal data subject or otherwise affect their rights and legitimate interests, except the cases stipulated by law. Such cases include situation when the decision generating legal consequences for the personal data subject or otherwise affecting their rights and legitimate interests is made on the basis of exclusively automated processing of their personal data with the written consent of the personal data subject (clause 2 of Article 16 of the Law on personal data).

The said provision of the Russian legislation is comparable to Article 15 of the currently not applicable Directive 95/46/EC of the European Parliament and the EU Council "On the protection of individuals with regard to the processing of personal data and on the free movement of such data"[4]. The current GDPR contains similar rules. Article 22(3) of the General Data Protection Regulation provides that in some cases of automated processing "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision"[5].

At the same time, to ensure awareness, it is essential to disclose information about the data underlying the decision made. This refers to the issues of reliability and neutrality, representatives of data, non-biased methods of their processing and analysis, as well as the information on the artificial intelligence self-learning.

Transparency as the technology explainability and non-discrimination depends on the quality of the data with which the artificial intelligence system works. Researchers (Buolamwini & Gebru, 2018) found that all popular facial recognition systems most accurately recognize males with fair skin (2.4% of errors) and make the most mistakes when recognizing black females (61% of errors). Actually, this proved that "photos of black women are the least numerous in databases; developers of such systems are predominantly white men; camera sensors worse identify details in dark colors"[6].

The above example shows that it is insufficient to doubt the reliability of data available to artificial intelligence. The data quality problem is that the available data were not neutral

---

3 On personal data: Federal law of 27.07.2006 No. 152-FZ. *SPS KonsultantPlyus*. https://www.consultant.ru/document/cons_doc_LAW_61801/

4 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. (1995, November 23). *Official Journal of the European Communities, L 281*, 31–50.

5 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). http://eur-lex.europa.eu/

6 More and more often researchers cannot explain how AI works. "Black and white box" theory. (2022, November 23). *Habr*. https://habr.com/ru/company/getmatch/blog/700736/

even if they could be considered representative. Facial recognition systems are used in many countries, including in the work of law-enforcement bodies. It is proved that if you belong to a racial minority in one of these countries, the system will more often recognize you as a criminal[7]. An opinion has been voiced that, if artificial intelligence models are trained with big data, then the built-in racial and other prejudices will be inevitable (Bender et al., 2021), as some groups of people have less access to the Internet and their data are less presented at various resources (for example, residents of remote places compared to programmers).

In general, we believe that the approach focused on personal data is weak, as it is not consistent with the reality. In addition to the data submitted directly to the algorithm by its users, business should supplement these categories with analytical data (Camilleri, 2018), more thoroughly describing various groups and making grounds for classification more clear.

Hence, it is necessary to stipulate rules for identifying the quality (reliable and neutral) set of data in a situation when it is not possible to limit such set of data. The risk of unexplainable biased decisions of the artificial intelligence will have to be excluded by reinforcement learning and audit of the result obtained.

This leads to a conclusion that one should not expect an algorithm explanation comprehensible for a human when the "black box" method is used, but the algorithm disclosure will not have a legal sense in that case. It is impossible to teach an artificial intelligence system to understand ethical values; lawyers can just list criteria to check that the decision of an artificial intelligence is unbiased. However, it is not always possible to put a human at the output to check the result. Hence, law may stipulate only the need of control on the part of software created by independent developers.

In this regard, it seems hardly feasible to achieve the artificial intelligence transparency not in relation to the system in general but through explaining the logic of individual decisions (Kuteynikov et al., 2018). The methods proposed by the authors include analysis of input data, statistical explanation, checking architecture/code and statistical analysis, determining the sensitivity of individual data (exactly which variables predetermine the result) (Kuteynikov et al., 2018).

On the contrary, it seems more feasible to require an open algorithm with indication of the general logic of decision-making. This system was adopted in California, USA. In February 2020 it adopted the Automated Decision Systems Accountability Act, which stipulates executing systematic control and revealing errors in the functioning of automated systems, as well as directing the reports obtained to the Department of Business Oversight starting from January 1, 2022, and placing them in the Internet for open access[8].

---

7 More and more often researchers cannot explain how AI works. "Black and white box" theory. (2022, November 23). *Habr*. https://habr.com/ru/company/getmatch/blog/700736/

8 USA. State of California. Automated Decision Systems Accountability Act of 2020. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB2269

The same was emphasized by the GDPR researchers. As A. Selbst and J. Powles wrote, "the problem is that to check the protection accuracy and potential contesting its correctness one needs specific explanations of the decision, including weights and factors used to achieve it" (Selbst & Powles, 2018), which is technical information not always comprehensible for a person. From this viewpoint, the "right" stipulated in Article 22 (3) of the General Data Protection Regulation is not sufficiently explained in the legislation and is subject to well-grounded critique due to its unfeasibility. Most researchers agree that the right to explanation of individual decisions, which may include global or local clarifications, does not follow from Article 22(3) of the GDPR (Wachter, 2017; Goodman & Flaxman, 2017). Article 15(1)(h) of the GDPR stipulates that, in case of automated processing in the sense of Article 22(1) of the GDPR, the controller must provide "meaningful information about the logic involved". Some researchers believe that this refers only to general structure and architecture of the processing model, but there is no need to explain individual decision or specific weights and characteristic of the model (Wachter, 2017; Malgieri & Comandé, 2017).

In the absence of a standardized approach to justification of individual and general decisions, one cannot answer the questions about who should the decision-making logic be disclosed to – the data users or subjects only or all stakeholders, in what amount, etc.

In this context, one should pay attention to the significance of using artificial intelligence in relations involving the state. One should agree that it is necessary to stipulate the norms of obligatory general availability of the results of state authorities using artificial intelligence and big data technologies. As was convincingly proved by V. V. Silkin, following our European colleagues, if the state executes its functions using artificial intelligence, then the transparency of the technology is required. At that, the author proposes imposing on the state authorities an obligation to substantiate and disclose the goals of using the automated data processing technologies. The capabilities of the big data and artificial intelligence technologies are rather vast, but their use by the state should be determined by the need to achieve publicly significant goals (Silkin, 2021). We believe, however, that, if the is technology is widely spread, justification of the use of artificial intelligence in certain types of state activity will solve this task in general, but will not provide transparency of decisions.

At the same time, it is worth highlighting that the principle of transparency in artificial intelligence functioning is not equal to the principle of transparency in the activity of state authorities or other operators of data using automated systems. V. V. Silkin proposes "when implementing the principle of transparency in the activity of state authorities using automated data processing systems, to assume openness of the information about the goals, means and results of their use" (Silkin, 2021). At that, the author justly states that "at the same time, in complex automated processing systems, algorithms are formed and complicated independently, which excludes the possibility to forecast or unambiguously comprehend in advance all the capabilities of such systems" (Silkin, 2021).

In our opinion, a substitution of concepts may easily occur in this case: transparency of decision-making by artificial intelligence is not associated exclusively with the transparency

of goals set for the algorithm application. One should distinguish between the transparency of algorithm as the principle of regulating artificial intelligence and transparency of activity (state governance, civil circulation, etc.), which is achieved partly with the help of artificial intelligence technology. Factually, the second aspect of automated data processing, together with the data quality, constitutes the transparency of goals and methods of using artificial intelligence. In that, the principle of transparency of the artificial intelligence functioning coincides, but is not equal to the notion of the transparency of activity.

Probably, the goals of using artificial intelligence could be defined as the goals of making decisions with explanation of the artificial intelligence processes along a standardized form, which requires regular updating every time business changes its methods of automated processing (Wulf & Seizov, 2022).

Assumingly, stipulation of the right to awareness about using artificial intelligence technology in automated data processing for decision-making does not exhaust the probable legal means to achieve transparency of intellectual systems. Moreover, citizens' rights are not protected, contravening the methods of the artificial intelligence functioning, first of all, the black box method. Disclosure of the algorithm tasks and priorities for the goals of decision-making may be based on standards, but, given the need to access new data and bridge the gaps in the data obtained from user, it does not seem possible to achieve transparency in this field either.

4. Openness of algorithms and results of their functioning

Disclosure of information about software development, code and the order of its execution is also prioritized for ensuring transparency of artificial intelligence.

As was justly noted by A. I. Savelyev, Article 16 of the Law on personal data stipulates, as conditions of using automated data processing tools for the purposes of making legally relevant decisions in relation to the subject, "additional information responsibilities of the operator, expressed in their obligation to provide the subject with explanations related to the order of making such a decision and probable juridical consequences thereof" (Savelyev, 2021).

The draft Law on artificial intelligence[9], proposed by the European Commission on April 21, 2021, proposes to provide transparency of the decisions made by artificial intelligence, for example, by disclosing information about characteristics, capabilities and limitations of the artificial intelligence system, about the purpose of the system, as well as the information necessary to service the artificial intelligence systems.

[9] European Commission (2021, April 21). *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

Since 2022, a draft law on algorithmic accountability is being discussed in the USA[10], according to which the companies using artificial intelligence systems will be required to assess their automated systems in compliance with the rules of Federal Trade Commission to ensure users non-discrimination and confidentiality of information.

Introduction of a legislative requirement of algorithm disclosure is also discussed in Russia, although so far only in relation to recommendation algorithms of social networks[11]. Similar requirements are stipulated by the Chinese law – by the Provision on managing algorithmic recommendations of information services in the Internet[12].

At the same time, the requirement of absolute disclosure of the source code or the architecture of the artificial intelligence models used does not seem justified. In any case, such disclosure cannot be comprehensive for a number of reasons. Technologically such disclosure is very costly and requires large resources. Publication of the source code may lead to violations of an intellectual property right or a trade secret. Such approach to transparency contravenes the current legal regimes, which govern the constituents of a common artificial intelligence technology.

This said, we consider convincing the position, according to which artificial intelligence systems are protected, first of all, as a trade secret, as the attempts to protect artificial intelligence systems in compliance with copyright and patent laws encounter difficulties (Foss-Solbrekk, 2021). The problem of granting copyright protection to the algorithm per se is due to its constant changing and complementing during self-learning and autonomous work. Also disputable is the question of the creative character of the artificial intelligence origin as an object of legal protection. Obtaining patents for artificial intelligence systems is also complicated. This situation is observed in various systems of justice. For example, in the Russian legislation on copyright, algorithms essentially got their own regulation within Article 1261 of the Russian Civil Code; in EU algorithms are excluded from the copyright protection in compliance with the EU Directive on computer programs[13]. Anyway, the current vigorous discussions about referring artificial intelligence to one or another type of intellectual property right objects are far from completion.

Protection of the artificial intelligence models with the legal regime of trade secret leads to a clash between the requirements of transparency and accountability. In particular,

---

10 Metcalf, J., Smith, B., & Moss, E. (2022, February 9). A New Proposed Law Could Actually Hold Big Tech Accountable for Its Algorithms. *Slate*. https://slate.com/technology/2022/02/algorithmic-accountability-act-wyden.html

11 Mass media: it is planned to propose a draft law to the State Duma about regulating recommendation services in social networks. (2021, October 15). *Parlamentskaya gazeta*. https://www.pnp.ru/politics/smi-v-gosdumu-planiruyut-vnesti-proekt-o-regulirovanii-rekomendatelnykh-servisov-v-socsetyakh.html

12 互联网信息服务算法推荐管理规定. (2021, December 31). http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm

13 *On the legal protection of computer programs (codified version)*: Directive 2009/24/EC of the European Parliament and of the Council. https://base.garant.ru/71657620/

researchers come to a conclusion that the European Directive 2016/943 leaves little space for the hypotheses of algorithmic transparency (Maggiolino, 2018).

The definition of "trade secret" given in the Directive 2016/943 says about a commercial value without indication of its actual or potential character. According to Article 2(1), "'trade secret' means information which meets all of the following requirements:

(a) it is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question;

(b) it has commercial value because it is secret;

(c) it has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret"[14].

Under such approach, the requirement of disclosure of a code or architecture of the artificial intelligence model would mean a loss of a competitive advantage in the market. As trade secret protection exists as long as information remains confidential and requires that the subjects take measures to provide confidentiality, trade secret protection promotes algorithmic nontransparency.

As follows from the above, rejection of the secrecy of artificial intelligence algorithms should be accompanied by an increased protection of interests of the disclosing party against the third parties, as it currently occurs in patent law.

5. Applicable legal means to prevent the problems with nontransparency of legal decisions: object or reject

Awareness of using artificial intelligence in decision-making also leads to the need to discuss the right to reject using the artificial intelligence technology in a specific case, as well as the right to object to the decision made.

In compliance with clause 3 of Article 16 of the Russian Law on personal data, the operator is obliged to provide the personal data subject with the opportunity to claim against a decision made as a result of automated data processing, and to clarify the order of protection of the rights and legitimate interests by the personal data subject. A. I. Savelyev explains that this "refers to the rights related to making legally significant decisions exclusively as a result of automated personal data processing. In particular, this order of protection implies notifying a subject about their right to demand human intervention into the decision-making process, which is an indispensible part of the right to objection" (Savelyev, 2021).

[14] On the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure: Directive (EU) 2016/943 of the European Parliament and of the Council. https://base.garant.ru/71615160/

In Russia, a draft law is also being discussed, which proposes providing users with the opportunity to completely or partially reject using recommendation algorithms.

In China, such an approach is already accepted for implementation and is being checked for feasibility. According to researchers, disclosure of the algorithm logic must eliminate the risks of unjustified refusal of service supplier in case of algorithmic recommendations to provide the necessary information, such as the sphere of the algorithm application, the service user, the level of the algorithm risk and others, on the pretext that there are no clear legislative provisions for that (Xu Ke & Liu Chang, 2022). However, sometimes such disclosure does not lead to success but is just a website design.

Thus, lawyers should strive for transparency and explainability of the artificial intelligence functioning in a human-comprehensible form in order to, inter alia, ensure the opportunity for the artificial intelligence system user to object to the decision made. This issue is also related to defining the subject of liability for the decisions which may significantly infringe upon human rights.

Researchers have noticed that the current discussion about the requirements of data protection in relation to explainability ignores the importance of this characteristic for estimating contractual and delict liability in relation to using the artificial intelligence tools (Hacker et al., 2020). In this regard, it is necessary to further specify the legislation provisions on using artificial intelligence in the sphere of strengthening the obligation of developers, producers and suppliers of artificial intelligence services to constantly assess the probable negative consequences of using artificial intelligence for human rights and fundamental freedoms and, in view of these consequences, to take measures to prevent and mitigate risks (Dyakonova et al., 2022).

Conclusions

Transparency is an indispensible condition for recognizing artificial intelligence as trustworthy. The most effective way to establish trust towards artificial intelligence is to recognize this technology as a part of a complex socio-technical system, which mediates trust and improves reliability of such systems.

Most of the debate around the artificial intelligence transparency from juridical point of view is focused on data protection laws. We believe that the circle of these discussions should be broadened. Transparency and explainability of the artificial intelligence technology is essential not only for personal data protection, but also in other situations of automated data processing, when, in order to make a decision, the technological data lacking in the input information are taken from open sources, including those not having the status of a personal data storage. A legislator may only strive for introduction of a standard, stipulating a compromise between the technology abilities and advantages, accuracy and explainability of its result, and the rights of the participants of civil relations. Introduction of certification of the artificial intelligence models, obligatory for application,

will solve the issues of liability of the subjects obliged to apply such systems. In the context of professional liability of professional subjects, such as doctors, militants, or corporate executives of a juridical person, it is necessary to restrict the obligatory application of artificial intelligence if sufficient transparency is not provided.

The legal discussion should develop towards elaborating proposals for the content of the right to reject using automated data processing in decision-making and the right to object to the decisions made in such a way.

References

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W. D., & McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning* (pp. 342–350). PMLR.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Camilleri, M. A. (2018). Market segmentation, targeting and positioning. In *Travel marketing, tourism economics and the airline product* (pp. 69–83). New York: Springer.

Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert systems with Applications*, *23*(3), 329–342. https://doi.org/10.1016/s0957-4174(02)00052-0

Cragun, B. J., & Steudel, H. J. (1987). A decision-table-based processor for checking completeness and consistency in rule-based expert systems. *International Journal of Man-Machine studies*, *26*(5), 633–648. https://doi.org/10.1016/s0020-7373(87)80076-7

Dyakonova, M. O., Efremov, A. A., Zaitsev, O. A., et al.; I. I. Kucherova, S. A. Sinitsyna (Eds.). (2022). *Digital economy: topical areas of legal regulation: scientific-practical tutorial*. Moscow: IZISP, NORMA. (In Russ.).

Foss-Solbrekk, K. (2021). Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly. *Journal of Intellectual Property Law & Practice*, *16*(3), 247–258. https://doi.org/10.1093/jiplp/jpab033

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense advanced research projects agency (DARPA)*. nd Web. *2*(2), 1.

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, *28*(4), 415–439. https://doi.org/10.1007/s10506-020-09260-6

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). New York: Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). *Kernel methods in machine learning*.

Kalis, B., Collier, M., & Fu, R. (2014). 10 promising AI applications in health care. *Harvard Business Review*.

Kalpokas, I. (2019). *Algorithmic Governance. Politics and Law in the Post-Human Era*. Cham: Palgrave Pivot.

Kalyatin, V. O. (2022). Deepfake as a legal problem: new threats or new opportunities? *Zakon*, *7*, 87–103. (In Russ.). https://doi.org/10.37239/0869-4400-2022-19-7-87-103

Kharitonova, Y. S., Savina, V. S., & Pagnini, F. (2021). Artificial Intelligence's Algorithmic Bias: Ethical and Legal Issues. *Perm U. Herald Jurid. Sci*, *3*(53), 488–515. https://doi.org/10.17072/1995-4190-2021-53-488-515

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency*. Panel for the Future of Science and Technology.

Kuteynikov, D. L., Izhaev, O. A., Zenin, S. S., & Lebedev, V. A. (2020). Algorithmic Transparency and Accountability: Legal Approaches to Solving the "Black Box" Problem. *Lex Russica*, *73*(6), 139–148. (In Russ.). https://doi.org/10.17803/1729-5920.2020.163.6.139-148

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric engineering and remote sensing*, *67*(10), 1137–1142.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Maggiolino, M. (2018). *EU trade secrets law and algorithmic transparency*. Available at SSRN 3363178.

Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, *7*(4), 243–265. https://doi.org/10.1093/idpl/ipx019

McEwen, R., Eldridge, J., & Caruso, D. (2018). Differential or deferential to media? The effect of prejudicial publicity on judge or jury. *The International Journal of Evidence & Proof*, *22*(2), 124–143. https://doi.org/10.1177/1365712718765548

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7

Pak, M., & Kim, S. (2017). A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*. https://doi.org/10.1109/caipt.2017.8320684

Payre, W., Cestac, J., & Delhomme, P. (2014). Intention to use a fully automated car: Attitudes and a priori acceptability. *Transportation research part F: traffic psychology and behavior*, *27*, 252–263. https://doi.org/10.1016/j.trf.2014.04.009

Plous, S. E. (2003). *Understanding prejudice and discrimination*. McGraw-Hill.

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annu. Rev. Sociol.*, *32*, 299–328. https://doi.org/10.1146/annurev.soc.32.061604.123132

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. https://doi.org/10.1109/jproc.2021.3060483

Savelyev, A. I. (2021). *Scientific-practical article-by-article commentary to Federal Law "On personal data"* (2nd ed., amended and abridged). Moscow: Statut. (In Russ.).

Selbst, A., & Powles, J. (2017). "Meaningful information" and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Silkin, V. V. (2021). Transparency of executive power in digital epoch. *Russian Juridical Journal*, *4*, 20–31. (In Russ.). https://doi.org/10.34076/20713797_2021_4_20

Silver, D. et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Vargas, R., Mosavi, A., & Ruiz, R. (2018). *Deep learning: a review*. Preprints.org. https://doi.org/10.20944/preprints201810.0218.v1

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wulf, A. J. & Seizov, O. (2022). "Please understand we cannot provide further information": evaluating content and transparency of GDPR-mandated AI disclosures. *AI & SOCIETY*, 1–22. https://doi.org/10.1007/s00146-022-01424-z

Yampolskiy, R. V. (2019). *Unexplainability and incomprehensibility of artificial intelligence*. arXiv preprint arXiv:1907.03869

许可、刘畅. 论算法备案制度 // 人工智能. 2022. № 1. P. 66. [Xu Ke, Liu Chang. (2022). On the Algorithm Filing System. *Artificial Intelligence*, *1*, 66.]

Author information

Yuliya S. Kharitonova – Doctor of Law, Professor, Professor of the Department of Entrepreneurial Law, Head of the Center for legal research of artificial intelligence and digital economy, Lomonosov Moscow State University
Address: 1 Leninskiye gory, 119991 Moscow, Russian Federation
E-mail: sovet2009@rambler.ru
ORCID ID: https://orcid.org/0000-0001-7622-6215
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57316440400
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/708572
Google Scholar ID: https://scholar.google.ru/citations?user=61mQtb4AAAAJ
RSCI Author ID: https://elibrary.ru/author_items.asp?authorid=465239

Conflict of interests

The author is a member of the Editorial Board of the Journal; the article has been reviewed on general terms.

Funding

The research was not sponsored.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – March 6, 2023
Date of approval – April 13, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 346.1:006.44:004.8
EDN: https://elibrary.ru/dxnwhv
DOI: https://doi.org/10.21202/jdtl.2023.14

Правовые средства обеспечения принципа прозрачности искусственного интеллекта

Юлия Сергеевна Харитонова

Московский государственный университет имени М. В. Ломоносова
г. Москва, Российская Федерация

Ключевые слова

Автоматизированная обработка данных,
автономность,
алгоритм,
искусственный интеллект,
право,
принятие решений,
прозрачность,
цифровая экономика,
цифровые технологии,
этика

Аннотация

Цель: анализ действующих технологических и юридических теорий для определения содержания принципа прозрачности работы искусственного интеллекта с позиции правового регулирования, выбора применимых средств правового регулирования и установление объективных границ юридического вмешательства в технологическую сферу с помощью регулирующего воздействия.

Методы: методологическую основу исследования составляет совокупность общенаучных (анализ, синтез, индукция, дедукция) и специально-юридических (историко-правовой, формально-юридический, сравнительно-правовой) методов научного познания.

Результаты: подвергнуты критическому анализу нормы и предложения для нормативного оформления принципа прозрачности искусственного интеллекта с точки зрения невозможности получения полной технологической прозрачности искусственного интеллекта. Выдвинуто предложение обсудить варианты политики управления алгоритмической прозрачностью и подотчетностью на основе анализа социальных, технических и регулятивных проблем, создаваемых алгоритмическими системами искусственного интеллекта. Обосновано, что прозрачность является необходимым условием для признания искусственного интеллекта заслуживающим доверия. Обосновано, что прозрачность и объяснимость технологии искусственного интеллекта важна не только для защиты персональных данных, но и в иных ситуациях автоматизированной обработки данных, когда для принятия решений недостающие из входящей информации технологические данные восполняются из открытых источников, в том числе не имеющих значения хранилищ персональных данных. Предложено законодательно закрепить обязательный аудит и ввести стандарт, закрепляющий

компромисс между возможностями и преимуществами технологии, точностью и объяснимостью результата ее работы и правами участников общественных отношений. Введение сертификации моделей искусственного интеллекта, обязательных к применению, позволит решить вопросы ответственности обязанных применять такие системы субъектов. В контексте вопроса о профессиональной ответственности профессиональных субъектов, таких как врачи, военные, органы корпоративного управления юридического лица, требуется ограничить обязательное применение искусственного интеллекта в случаях, если не обеспечена его достаточная прозрачность.

Научная новизна: междисциплинарный характер исследования позволил выявить невозможность и необоснованность требований полного открытия исходного кода или архитектуры моделей искусственного интеллекта. Принцип прозрачности искусственного интеллекта может быть обеспечен за счет проработки и обеспечения права субъекта данных и субъекта, которому адресовано решение, принятое в результате автоматизированной обработки данных, на отказ от применения автоматизированной обработки данных для принятия решений и права на возражения против принятых таким способом решений.

Практическая значимость: обусловлена отсутствием в настоящее время достаточного регулирования принципа прозрачности искусственного интеллекта и результатов его работы, а также содержания и особенностей реализации права на объяснение и права на возражение субъекта решения. Наиболее плодотворный путь для установления доверия к искусственному интеллекту заключается в том, чтобы признать данную технологию частью сложной социотехнической системы, которая опосредует доверие, и повышать надежность этих систем. Основные положения и выводы исследования могут быть использованы для совершенствования правового механизма обеспечения прозрачности моделей искусственного интеллекта, применяемых в государственном управлении и бизнесе.

Для цитирования

Харитонова, Ю. С. (2023). Правовые средства обеспечения принципа прозрачности искусственного интеллекта. *Journal of Digital Technologies and Law*, *1*(2), 337–358. https://doi.org/10.21202/jdtl.2023.14

Список литературы

Дьяконова, М. О., Ефремов, А. А., Зайцев, О. А. и др.; И. И. Кучерова, С. А. Синицына (ред.). (2022). Цифровая экономика: актуальные направления правового регулирования: научно-практическое пособие. Москва: ИЗиСП, НОРМА.

Калятин, В. О. (2022). Дипфейк как правовая проблема: новые угрозы или новые возможности? *Закон, 7*, 87–103. https://doi.org/10.37239/0869-4400-2022-19-7-87-103

Кутейников, Д. Л., Ижаев, О. А., Зенин, С. С., Лебедев, В. А. (2020). Алгоритмическая прозрачность и подотчетность: правовые подходы к разрешению проблемы «черного ящика». *Lex russica (Русский закон)*, *73*(6), 139–148. https://doi.org/10.17803/1729-5920.2020.163.6.139-148

Савельев, А. И. (2021). Научно-практический постатейный комментарий к Федеральному закону «О персональных данных» (2-е изд., перераб. и доп.). Москва: Статут.

Силкин, В. В. (2021). Транспарентность исполнительной власти в цифровую эпоху. *Российский юридический журнал*, *4*, 20–31. https://doi.org/10.34076/20713797_2021_4_20

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W. D., & McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning* (pp. 342–350). PMLR.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Camilleri, M. A. (2018). Market segmentation, targeting and positioning. In *Travel marketing, tourism economics and the airline product* (pp. 69–83). New York: Springer.

Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert systems with Applications*, *23*(3), 329–342. https://doi.org/10.1016/s0957-4174(02)00052-0

Cragun, B. J., & Steudel, H. J. (1987). A decision-table-based processor for checking completeness and consistency in rule-based expert systems. *International Journal of Man-Machine studies*, *26*(5), 633–648. https://doi.org/10.1016/s0020-7373(87)80076-7

Foss-Solbrekk, K. (2021). Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly. *Journal of Intellectual Property Law & Practice*, *16*(3), 247–258. https://doi.org/10.1093/jiplp/jpab033

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense advanced research projects agency (DARPA)*. nd Web. *2*(2), 1.

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, *28*(4), 415–439. https://doi.org/10.1007/s10506-020-09260-6

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). New York: Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). *Kernel methods in machine learning*.

Kalis, B., Collier, M., & Fu, R. (2014). 10 promising AI applications in health care. *Harvard Business Review*.

Kalpokas, I. (2019). *Algorithmic Governance. Politics and Law in the Post-Human Era*. Cham: Palgrave Pivot.

Kharitonova, Y. S., Savina, V. S., & Pagnini, F. (2021). Artificial Intelligence's Algorithmic Bias: Ethical and Legal Issues. *Perm U. Herald Jurid. Sci*, *3*(53), 488–515. https://doi.org/10.17072/1995-4190-2021-53-488-515

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency*. Panel for the Future of Science and Technology.

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric engineering and remote sensing*, *67*(10), 1137–1142.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Maggiolino, M. (2018). *EU trade secrets law and algorithmic transparency*. SSRN 3363178.

Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, *7*(4), 243–265. https://doi.org/10.1093/idpl/ipx019

McEwen, R., Eldridge, J., & Caruso, D. (2018). Differential or deferential to media? The effect of prejudicial publicity on judge or jury. *The International Journal of Evidence & Proof*, *22*(2), 124–143. https://doi.org/10.1177/1365712718765548

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7

Pak, M., & Kim, S. (2017). A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*. https://doi.org/10.1109/caipt.2017.8320684

Payre, W., Cestac, J., & Delhomme, P. (2014). Intention to use a fully automated car: Attitudes and a prio-

ri acceptability. *Transportation research part F: traffic psychology and behavior*, *27*, 252–263. https://doi.org/10.1016/j.trf.2014.04.009

Plous, S. E. (2003). *Understanding prejudice and discrimination*. McGraw-Hill.

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annu. Rev. Sociol.*, *32*, 299–328. https://doi.org/10.1146/annurev.soc.32.061604.123132

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. https://doi.org/10.1109/jproc.2021.3060483

Selbst, A., & Powles, J. (2017). "Meaningful information" and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Silver, D. et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Vargas, R., Mosavi, A., & Ruiz, R. (2018). *Deep learning: a review*. Preprints.org. https://doi.org/10.20944/preprints201810.0218.v1

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wulf, A. J. & Seizov, O. (2022). "Please understand we cannot provide further information": evaluating content and transparency of GDPR-mandated AI disclosures. *AI & SOCIETY*, 1–22. https://doi.org/10.1007/s00146-022-01424-z

Yampolskiy, R. V. (2019). *Unexplainability and incomprehensibility of artificial intelligence*. arXiv preprint arXiv:1907.03869

许可、刘畅. 论算法备案制度 // 人工智能. 2022. № 1. P. 66. [Xu Ke, Liu Chang. (2022). On the Algorithm Filing System. *Artificial Intelligence*, *1*, 66.]

Сведения об авторе

Харитонова Юлия Сергеевна – доктор юридических наук, профессор, профессор кафедры предпринимательского права, руководитель Центра правовых исследований искусственного интеллекта и цифровой экономики, Московский государственный университет имени М. В. Ломоносова
Адрес: 119991, Российская Федерация, г. Москва, Ленинские горы, 1
E-mail: sovet2009@rambler.ru
ORCID ID: https://orcid.org/0000-0001-7622-6215
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57316440400
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/708572
Google Scholar ID: https://scholar.google.ru/citations?user=61mQtb4AAAAJ
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=465239

Конфликт интересов

Автор является членом редакционной коллегии журнала, статья прошла рецензирование на общих основаниях.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.23.01 / Общие вопросы предпринимательского права
Специальность ВАК: 5.1.3 / Частно-правовые (цивилистические) науки

История статьи

Дата поступления – 6 марта 2023 г.
Дата одобрения после рецензирования – 13 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.15

Future of the Artificial Intelligence: Object of Law or Legal Personality?

Irina A. Filipova ✉

National Research Lobachevsky State University of Nizhny Novgorod
Nizhny Novgorod, Russian Federation;
Samarkand State University
Samarkand, Republic of Uzbekistan

Vadim D. Koroteev

National Research Lobachevsky State University of Nizhny Novgorod
Nizhny Novgorod, Russian Federation

Keywords

Artificial intelligence,
cyberphysical system,
digital technologies,
electronic person,
generative model,
intellectual system,
law,
legal personality,
quasi subject of law,
robot

Abstract

Objective: to reveal the problems associated with legal regulation of public relations, in which artificial intelligence systems are used, and to rationally comprehend the possibility of endowing such systems with a legal subject status, which is being discussed by legal scientists.

Methods: the methodological basis of the research are the general scientific methods of analysis and synthesis, analogy, abstraction and classification. Among the legal methods primarily applied in the work are formal-legal, comparative-legal and systemic-structural methods, as well as the methods of law interpretation and legal modeling.

Results: the authors present a review of the state of artificial intelligence development and its introduction into practice by the time of the research. Legal framework in this sphere is considered; the key current concepts of endowing artificial intelligence with a legal personality (individual, collective and gradient legal personality of artificial intelligence) are reviewed. Each approach is assessed; conclusions are made as to the most preferable

✉ Corresponding author

amendments in the current legislation, which ceases to correspond to the reality. The growing inconsistency is due to the accelerated development of artificial intelligence and its spreading in various sectors of economy, social sphere, and in the nearest future – in public management. All this testifies to the increased risk of a break between legal matter and the changing social reality.

Scientific novelty: scientific approaches are classified which endow artificial intelligence with a legal personality. Within each approach, the key moments are identified, the use of which will allow in the future creating legal constructs based on combinations, avoiding extremes and observing the balance between the interests of all parties. The optimal variant to define the legal status of artificial intelligence might be to include intellectual systems into a list of civil rights objects, but differentiating the legal regulation of artificial intelligence as an object of law and an "electronic agent" as a quasi subject of law. The demarcation line should be drawn depending on the functional differences between intellectual systems, while not only a robot but also a virtual intellectual system can be considered an "electronic agent".

Practical significance: the research materials can be used when preparing proposals for making amendments and additions to the current legislation, as well as when elaborating academic course and writing tutorials on the topics related to regulation of using artificial intelligence.

For citation

Filipova, I. A., Koroteev, V. D. (2023). Future of the Artificial Intelligence: Object of Law or Legal Personality? *Journal of Digital Technologies and Law, 1*(2), 359–386. https://doi.org/10.21202/jdtl.2023.15

Contents

Introduction

Today, humanity finds itself in the period of social transformation related to substituting one technological order for another; "smart" machines and software are rather rapidly learning; artificial intelligence systems increasingly become able to substitute people in many spheres of activity. One of the questions more and more often raised in connection with improving artificial intelligence technologies is that of recognizing artificial intellectual systems to be subjects of law, as they have achieved the level of making completely autonomous decisions and potential manifestation of "subjective will". This question was formulated hypothetically as early as in the 20th century (McNally & Inayatullah, 1988; Solum, 1992). In the 21st century, the scientific discussion is ramped up steadily, reaching another extreme with each introduction of new artificial intelligence models into practice, like emergence of unmanned vehicles in the streets or presenting robots with a new set of functions (Bertolini & Episcopo, 2022).

The legal problem of defining the status of artificial intelligence is of general theoretical character, which is due to the objective inability to forecast all possible results of developing new models of artificial intelligence. However, artificial intelligence systems (AI systems) are already factual participants of certain social relation, which requires setting the "benchmarks", i. e. solving the fundamental issues in this sphere in order to legislatively stipulate, hence, to reduce the share of uncertainty in forecasting the development of relations involving artificial intelligence systems, in the future.

The question, used as the article title, about the supposed personality of the artificial intelligence as the research object, undoubtedly, does not comprise all artificial intelligence systems, among which there are a lot of "electronic assistants" not claiming to be legal personalities as their set of functions is limited and their represent a narrow (weak) artificial intelligence. Rather, we will speak of "smart machines" (cyberphysical intellectual systems) and generative models of virtual intellectual systems, which by their abilities are increasingly verging to the general (strong) artificial intelligence, comparable to the human's and in future exceeding it.

1. Legal status of artificial intelligence: sources of the problem

1.1. Level and rate of artificial intelligence development

The level of artificial intelligence development can now be discussed only conditionally, as the speed of its development is accelerating and what was relevant at the moment of writing the article is rapidly becoming obsolete. This is especially true for the most rapidly developing sphere of artificial intelligence – artificial neural networks. By the beginning of 2023, multimodal neural networks, such as ChatGPT, DALL-e and others, the intellectual abilities of which are being improved through increasing the number of parameters (perceived modalities, including those inaccessible to humans), as well as through using large amounts of data for learning, which humans cannot physically

process, have raised the acuteness of the issue of creating a string artificial intelligence. For example, multimodal generative models of neural networks can create pictures, literary and scientific texts so that one cannot always discern whether they were created by a person or an artificial intelligence system.

IT experts speak of two qualitative leaps: velocity leap (periodicity of emergence of qualitatively new models), which is now measured not in years but in months as a maximum, and volatility leap (impossibility to accurately forecast what may happen in the sphere of artificial intelligence even up to the end of the current year)[1]. ChatGPT-3 model (the third generation of natural language processing algorithm by OpenAI company) appeared in 2020 and could process a text, the next generation model – ChatGPT-4, launched by the producer in March 2023, can "work" not only with texts but also with images, while the model of the generation to come is learning and will be capable of more.

A few years ago the supposed moment of technological singularity, when the development of machines becomes actually unmanageable and irreversible, drastically changing the human civilization, was considered to be at least several decades away, but today more and more researchers think that it may happen much sooner[2]. This implies the emergence of the so called strong artificial intelligence, which will demonstrate the abilities comparable to human intelligence and be able to solve a similar or even broader range of tasks. Unlike the weak artificial intelligence, the strong one will possess consciousness, and one of the indispensible conditions of emerging consciousness in intellectual systems is the possibility to perform multimodal behavior integrating data from various sensor modalities (text, image, video, sound, etc.), "linking" information of various modalities to the reality and building full-fledged coherent "metaphors of the world", as is peculiar to humans[3].

In March 2023, over one thousand researchers, IT experts and entrepreneurs in the sphere of artificial intelligence signed an open letter published in the website of the US scientific-research center Future of Life Institute[4] which specializes in studying existential risks for humanity. The letter calls for pausing the training of new generative multimodal neural network models, as the lack of common safety protocols and the legal vacuum significantly increase the risks, because the speed of artificial intelligence technologies development has sharply increased due to the "ChatGPT revolution". It was also marked that the artificial intelligence models have developed unexplainable capabilities unforeseen

1 Karelov, S. (2023, April 5). *Telegram channel "Little-known interesting facts"*. https://t.me/s/theworldisnoteasy

2 *David Shapiro (expert on artificial cognitive architecture) predicts. "AGI within 18 months"*. (2023, March 28). https://www.reddit.com/r/singularity/comments/1254azr/david_shapiro_expert_on_artificial_cognitive/

3 Kolonin, A. (2021, December 8). *On the depth, transparency and "power" of AI at the moment*. https://russiancouncil.ru/analytics-and-comments/analytics/o-glubine-prozrachnosti-i-sile-ii-v-tekushchem-momente/

4 *Pause Giant AI Experiments: An Open Letter.* (2023, March 22). https://futureoflife.org/open-letter/pause-giant-ai-experiments/

by their developers and, probably, the share of such capabilities will gradually increase. Besides, such technological revolution sharply stimulates the creation of intellectual gadgets, which will be widely spread, and the new generations, today's kids grown up in constant communication with artificial intelligence assistants will differ greatly from the previous generations.

Is it possible to impede the artificial intelligence development so that the humanity could adapt to the new conditions? Theoretically it is, if all states facilitate it through national legislations. Will they do it? Judging by the published national strategies, they will not; on the contrary, each state sets a mission to win the competition (maintain leadership or reduce the gap). In the Russian Federation, the task of accelerated developing of the artificial intelligence technologies was set in the National Strategy of artificial intelligence development up to 2030, adopted by the Decree of the Russian President of October 10, 2019 No. 490 "On the development of artificial intelligence in the Russian Federation"[5] (further – National Strategy). According to clause 24 of the National Strategy, the main areas include: supporting research for providing advanced development of artificial intelligence, elaborating intellectual software, improving accessibility of data necessary for development of artificial intelligence technologies, creating a complex system of regulating relations emerging in connection with development and use of artificial intelligence. Clause 30 of the National Strategy stipulates that the development of Russian technologies requires supporting scientific research aimed at creating cardinally new results, including creating strong artificial intelligence. A similar task is posed in the national strategies of artificial intelligence development of other countries of the world.

1.2. Spreading artificial intelligence technologies in practice

Opportunities of artificial intelligence attract entrepreneurs, thus, business structures invest a lot into new developments, while success of each new model stimulates this process. The volumes of annual investment are growing, given both private companies and state investing into developments; the global market of solutions in the sphere of artificial intelligence amounts to hundreds billion dollars; according to forecasts, in particular those contained in the European Parliament Resolution of 3 May 2022 "On Artificial Intelligence in a Digital Age", the contribution of artificial intelligence into the global economy will exceed 11 trillion euro by 2030[6].

---

5 On the development of artificial intelligence in the Russian Federation: Decree of the Russian President of October 10, 2019 No. 490. (2019). *Collection of legislation of the Russian Federation*, No. 41. Article 5700.

6 *European Parliament Resolution of 3 May 2022 on Artificial Intelligence in a Digital Age (2020/2266(INI))*. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0140_EN.html

Practice-oriented business results in introducing artificial intelligence technologies into all spheres of economy. Artificial intelligence is used both in extraction and processing industry (metallurgy, fuel and chemical industry, machine building, metalworking, etc.). It is used for forecasting the efficiency of developed products, automation of assembly lines, reduction of defects, improving logistics and preventing downtime.

Using artificial intelligence in transportation includes both autonomous transportation means proper and optimization of routes using predicting transportation streams, ensuring safety by preventing dangerous situations. Launching unmanned vehicles to public roads is an issue actively discussed by parliaments of different countries of the world. In 2021, the Russian Ministry of Transport also developed a draft law "On highly automated transportation means and on making amendments in certain legislative acts of the Russian Federation"[7], and a year later a Decree of the Russian Government of December 29, 2022 No. 2495 established a "Program of experimental legal regime in the sphere of digital innovations in rendering transport services using highly automated transportation means in the territories of some Russian Federation subjects"[8]. At the verge of transportation sphere and agriculture, autonomous harvesters are increasingly used, with the process being even more rapid in agriculture as there are no tight legal restrictions referring to automobile transport on public roads.

In banking, AI systems almost completely replaced people when estimating creditworthiness of borrowers; they are increasingly used to develop new banking products and to increase the safety of banking operations.

Artificial intelligence technologies "are capturing" not only business but also social sphere: healthcare, education, employment. Applying artificial intelligence in medicine allows improving diagnostics, development of new medications, performing surgery using robotics; in the sphere of education it allows individualizing lessons, automating assessment of students and professional skills of teachers.

Employment is increasingly changing today due to an exponential growth of platform employment. The share of persons working via digital labor platforms, complemented with artificial intelligence, is steadily growing worldwide, according to the data of International

---

7 *Draft of Federal Law "On highly automated transportation means and on making amendments in certain legislative acts of the Russian Federation" No. 02/04/06-21/00116763*. https://base.garant.ru/56880577/

8 On establishing an experimental legal regime in the sphere of digital innovations and adopting a Program of experimental legal regime in the sphere of digital innovations in rendering transport services using highly automated transportation means in the territories of some Russian Federation subjects: Decree of the Russian Government of December 29, 2022 No. 2495. (2022, December 30). *Official Internet portal of legal information*. http://publication.pravo.gov.ru/Document/View/0001202212300090

Labor Organization[9]. Platform employment is not the only component of transformation in labor sphere; the growing level of production robotization is strongly influencing it too. According to the International Federation of Robotics, the number of industrial robots continues to grow worldwide, with the most rapid rate of robotization in Asia, first of all, in the People's Republic of China and in Japan[10]. Russia significantly lags behind in this field, but it is the bridging of this gap that the new federal project is aimed. The project is devoted to developing Russian robotics and should stipulate legal, taxation and other conditions for developing production and launching of industrial robots. The federal project, in compliance with the order of the Russian President, is to be prepared in summer 2023. The project is to include a list of state support measures for developing production and launching of industrial robots "to provide annual reduction of lagging in the number of such robots by 10 thousand industrial workers in the country from the worldwide average level"[11]. Also, a draft is being prepared of the Order of the Russian President on making amendments in the National Strategy of artificial intelligence development, "aimed at widespread introduction of artificial intelligence technologies in economic and social sectors and in the state management system"[12].

Indeed, the abilities of artificial intelligence for data analysis, used for production management, diagnostic analytics and prognostics, excite serious interest in the states. Artificial intelligence is being introduced in public management. Today, the work on creating digital platforms is activated in order to render state services, automate many processes associated with elaborating decisions by state authorities.

The notions "artificial personality", "artificial sociality" are more and more often mentioned in the public discourse; this testifies to the fact that development and implementation of intellectual systems have passed from the pure technical domain into the sphere of researching its varied means of implementation in humanitarian and sociocultural human activities (Alekseev et al., 2023).

Given the above, one may assert that artificial intelligence is more and more profoundly penetrates into the lives of people. The presence of artificial intelligence systems in our life will become more visible in the years to come; it will increase both in the working environment and in public space, in services and homes. Artificial intelligence will more and more ensure the increased efficiency of achieving results through intellectual automation

---

9 *Prospects of employment and social protection worldwide: Role of digital labor platforms in transformation of labor sphere.* (2021). The ILO Decent Work Technical Support Team and Country Office for Eastern Europe and Central Asia. Moscow: ILO.

10 *World Robotics R&D Programs*. (2023). https://ifr.org/downloads/papers/Executive_Summary_-_World_Robotics_RD_Programs_V02.pdf

11 *Instructions of the Russian President following Artificial Intelligence Journey conference (November 23–24, 2022).*(2023, January 29). Pr-172, clause 1, subclause "e". http://www.kremlin.ru/acts/assignments/orders/70418 (access date: 20.04.2023)

12 *Instructions of the Russian President following Artificial Intelligence Journey conference (November 23–24, 2022).*(2023, January 29). Pr-172, clause 5. http://www.kremlin.ru/acts/assignments/orders/70418

of various processes, creating new opportunities and simultaneously bringing new threats for people, communities, and states.

With the growth of intellectual level, AI systems inevitably become an indispensible part of the society; people will have to coexist with them. Such a symbiosis will include cooperation between people and "smart" machines, which, according to a Nobel Prize winner in Economic Sciences J. Stiglitz, will lead to transformation of civilization (Stiglitz, 2017). Even today, according to some jurist, "to increase the level of wellbeing of humans, law must not make distinctions between human activity and that of artificial intelligence, when people and artificial intelligence perform the same tasks" (Abbott, 2020). One should also take into account that the development of humanoid robots acquiring the physiology increasingly similar to the human's one, will cause, inter alia, their performing gender roles as partners in the society (Karnouskos, 2022).

States have to adapt legislation to the changing public relations: the number of laws aimed at regulating relations, in which artificial intelligence systems are involved in one position or another, is rapidly growing worldwide. According to the Stanford University's AI Index Report – 2023[13], while only one law was adopted in 2016, in 2018 there were 12, in 2021 – 18, and in 2022 – 37 laws. This pushed the United Nations Organization towards formulating a position on the ethics of using artificial intelligence at the global level. In September 2022, a document appeared, which contained principles of the ethical use of artificial intelligence[14] and was based on Recommendations on the Ethics of Artificial Intelligence adopted a year earlier by UNESCO General Conference[15]. Nevertheless, the rate of development and implementation of artificial intelligence technologies significantly exceed the rate of corresponding changes in law.

The development of artificial intelligence technologies has launched the process of creating machine-readable law, which only AI systems can understand; moreover, one may speak not only of machine-readability of legal norms but also of their machine-projectability and machine-implementabililty. AI systems are already used for high quality legal analytics and formulating machine recommendations for lawyers (Ashley, 2017). Works on creating machine-readable law are actively executed today in many countries; in 2021 the Commission on digital development under the Russian Government adopted the Russian concept of developing the technologies of machine-readable law[16].

---

13 *AI Index Report 2023*. (2023). https://aiindex.stanford.edu/report/

14 *Principles for the Ethical Use of Artificial Intelligence in the United Nations System*. (2022, September 20). https://unsceb.org/principles-ethical-use-artificial-intelligence-united-nations-system

15 *Recommendation on the Ethics of Artificial Intelligence*. (2021, November 25). https://unesdoc.unesco.org/ark:/48223/pf0000373434

16 The concept of developing the technologies of machine-readable law, adopted by the Government Commission on digital development, using information technologies for improving living standards and conditions of entrepreneurial activity, protocol of 15.09.2021 No. 31. *KonsultantPlyus*. http://www.consultant.ru/document/cons_doc_LAW_396491/

1.3. Artificial intelligence system as an object of law

Today, artificial intelligence systems do not possess a legal personality and are considered objects of civil law – this is a certainty for any national legal system, not only a Russian one. Regardless of the achieved level of artificial intelligence development, AI system is someone's property. Accordingly, both virtual and cyberphysical AI systems (the two existing types of artificial intelligence by the form of its embodiment) are what legal relations emerge about.

Let us consider the legal regime of artificial intelligence as an object of law according to the Russian legislation. In compliance with Article 128 of the Civil Code of the Russian Federation[17] (further – CC RF), the objects of law are things, other property, including property rights, protected results of intellectual activity, nonmaterial goods, etc.

The cyberphysical by form, i. e. possessing a "body", artificial intelligence (as a rule, a robot) is considered to be a thing by the current legislation, but no special features of the legal regime of such things are not stipulated, their conveyancing is not restricted (Somenkov, 2019). It is assumed admissible to characterize robots similarly to indivisible items in compliance with Article 133 CC RF, as an attempt to divide artificial intelligence proper (that is, software) from the robot's "body" as its shell will entail inevitable change of its purpose or even destruction.

In the international private legal practice, cyberphysical AI systems also have the status of a thing in the general sense and a good in commercial terms. For example, International Classification of Goods and Services (NCGS) explicitly names a specific kind of goods: "Humanoid robots with artificial intelligence" (class 09, basic No. 090778)[18].

In the Russian Federation, artificial intelligence in virtual form also has no special legal position and today is actually regulated by the norms contained in part 4 CC RF, referring to untitled copyright objects. However, for effective legal protection of such object of civil rights, virtual AI systems have to be recognized as software, so that the provisions of Article 1259 CC RF, stipulating the legal protection of software similarly to that of literary works, could be extended to them.

The definition of the "artificial intelligence" notion can be found in Federal Law of April 24, 2020 No. 123-FZ "On conducting an experiment of establishing special regulation with a view of creating necessary conditions for developing and introducing artificial intelligence

[17] Civil Code of the Russian Federation. (1994, December 5). *Collection of legislation of the Russian Federation,* No. 32. Article 3301.

[18] International Classification of Goods and Services for registration of signs (NCGS) (11th edition, publication 1). "Kodeks" legal system. https://docs.cntd.ru/document/420273241

technologies in the Russian Federation subject – city of federal significance Moscow and making amendments to Articles 6 and 10 of Federal Law 'On personal data'"[19]:

"Artificial intelligence is a complex of technological solutions enabling to imitate human cognitive functions (including self-learning and searching solutions without a preset algorithm) and to obtain the results of executing certain tasks, comparable as a minimum with the results of human intellectual activity. The complex of technological solutions includes information-communication infrastructure (including information systems, information-communication networks, and other technical means of information processing), software (including using machine learning methods), processes and services of data processing and searching for decision".

Judging by the cited document, one may conclude that an AI system does not correspond to the definition of software contained in Article 1261 CC RF; such systems are not limited to just a set of data and commands intended for a computer functioning, i.e. have not only a software component; thus, for the purpose of further legal protection, such an AI system should be recognized as a complex object of intellectual property, stipulated by Article 1240 CC RF (Vasilevskaya et al., 2021).

It should be noted that while previously IT experts defined any artificial intelligence system as a software and hardware package, today the hardware part in virtual intellectual systems can be considered nonexistent, hence, the problem is removed. For example, the definition of artificial intelligence given in the Communication from the Commission to the European Parliament mentions that a virtual intellectual systems may have no hardware part of its own at all: "Artificial intelligence (AI) refers to systems demonstrating intelligent behavior, analyzing environment and taking actions – with a certain degree of autonomy – to achieve certain goals. AI-based systems may be purely software ones, acting in the virtual world (for example, voice assistants, software for analyzing images, search systems, speech and facial recognition systems) or AI may be built into hardware devices (for example, robots with artificial intelligence, unmanned vehicles, drones or applications of the Internet of Things)"[20].

Referring AI systems to objects of law does not exclude the opportunity of legal stipulation of the features of their legal regulation in the future, depending on the form of the artificial intelligence – virtual or cyberphysical, as well as taking into account the difference in the level of artificial intelligence. For example, some researchers propose distinguishing

---

19 "On conducting an experiment of establishing special regulation with a view of creating necessary conditions for developing and introducing artificial intelligence technologies in the Russian Federation subject – city of federal significance Moscow and making amendments to Articles 6 and 10 of Federal Law 'On personal data'": Federal Law of April 24, 2020 No. 123-FZ. (2020). *Collection of legislation of the Russian Federation*, No. 17. Article 2701.

20 Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. *Artificial Intelligence for Europe, Brussels, 25.04.2018 COM(2018) 237 final.* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN

a separate category of advanced cyberphysical systems, assuming that the influence of "smart" robots on the society will be much more significant than that of virtual systems, due to several factors, including the presence of the "body" and emergence, i. e. appearance of qualities in the system which were not inherent in its components (Calo, 2015).

One should recognize as well-grounded the opinion about the need to apply to AI systems a legal regime stipulated for sources of increased danger. Today, a court considering a particular case may decide at its own discretion whether a certain object refers to the "source of increased danger" category or not (Laptev, 2019). Artificial intelligence corresponds to the definition of a source of increased danger due to its ability to make autonomous decisions differing from the initially installed program. It should be specified that this should refer not to all systems with elements of artificial intelligence, which include many smart phone applications, but only advanced models. In this aspect, a proposal by V. V. Arkhipov and V. B. Naumov seems rational, namely, to recognize such AI systems to a property of special kind, thoroughly regulating their legal regime through confirming their status of "the property capable of autonomous actions" (Arkhipov & Naumov, 2017), especially taking into account that the more advanced artificial intelligence models become, the more control functions will be imposed on them (Greenstein, 2022).

One cannot but mention that the recognition of artificial intelligence to be an object of law offers a very limited choice of variants in solving a number of questions, the significance of which will only increase with time. These questions include:

1. Who will be liable for the damage incurred by the actions of AI systems, given that they become more and more autonomous?

2. Who will possess rights to the results of creative intellectual activity (given that the level of results becomes higher and the participation of humans, even indirect, may actually be reduced to zero)?

The answers to these questions so far include only a person – a producer, owner or user. Will this situation stay the same in the future? As for the liability for the damage incurred by an AI system, "while previously it was believed that no cardinal changes in regulating the institution of legal liability will be required, today there are no grounds to say so with absolute confidence. The reason is the increased level of autonomy of the artificial intelligence systems with the broadened range of possibilities of their use" (Kharitonova et al., 2022). Again, the corner stone is the problem of the growing autonomy of AI systems, hence, changing the legislation is just a matter of time and this change will reflect the new "balance on interests" (McCarty, 2017).

2. Main concepts of the legal personality of artificial intelligence

2.1. Concept of individual legal personality of artificial intelligence

Proceeding to the concepts of potential endowing intellectual systems with legal personality, one should admit that implementation of any of such approaches will require fundamental reconstruction of the established general theory of law and changing a number of provisions

in certain branches of law. Notably, proponents of various views often use the term "electronic person", thus, using this term does not allow determining, the proponent of which concept the author of a given work is without getting acquainted with the work content.

The most radical approach and, quite logically, the least popular among academic circles is the concept of individual legal personality of artificial intelligence. Supporters of this approach put forward the idea of "complete inclusiveness" (extreme inclusivism), which implies endowing AI systems with the legal status similar to that of a human being and recognizing their own interests (Mulgan, 2019), given their social significance or social meaning (social valence). The latter is due to the fact that "the physical embodiment of the robot tends to make a person treat that moving object as if it were alive. This is even more observable when the robot has anthropomorphic characteristics, since the resemblance to the human body causes people to start projecting emotions, feelings of pleasure, pain and care, as well as desires to constitute relationships" (Avila Negri, 2021). Projecting of human emotions on inanimate objects is not new, stemming from the history of humanity, but applied to robots it entails numerous consequences (Balkin, 2015).

As prerequisites of legal consolidation of this position, the following is usually mentioned:

– AI systems achieving the level comparable to human cognitive functions;

– increasing degree of similarity between robots and people;

– humanity, protection of intellectual systems against potential "sufferings".

As can be seen from the list of prerequisites, all of them possess a high degree of theoretization and subjective estimation. In particular, the trend towards creating anthropomorphic robots (androids) is due to the everyday psychological and social demands of people, for whom it is comfortable to feel themselves in a "company" of subjects similar to them. Some modern robots possess other constrictive features due to the functions they perform; these include "multiped" courier robots, whose priority is solid construction and effective distribution of the weight carried. In this case, the last of the mentioned prerequisites comes into force, which is caused by forming emotional links with robots in the human consciousness, similar to emotional links between a pet and its owner (Grin, 2018).

The idea of "complete inclusion" of the legal status of AI systems and a human being is reflected in the works by some jurists. As the provisions of the Constitution (for example, provisions of Chapters 1 and 2 of the Russian Constitution[21]), as well as of sector legislation, do not present a legal definition of a personality, the "personality" concept in constitutional-legal sense theoretically allows for an expansive interpretation. In that case, In this case, personalities will include any holders of intellect whose cognitive abilities are recognized as sufficiently developed. As A. V. Nechkin states, the logic of this approach is that the essential difference of a human being from other living being

[21] *Constitution of the Russian Federation*. https://base.garant.ru/10103000

consists in its unique highly developed intellect (Nechkin, 2020). Recognition of rights of AI systems seems to the next step of a legal system evolution, which gradually expands legal recognition to earlier discriminated people, and today opening access to non-human beings too (Gellers, 2021).

If AI systems are endowed with such legal status, the proponents of his approach consider it well-grounded to provide such systems not with literal citizens' rights in their established constitutional-legal interpretation, but their analogs and certain civil rights with some waivers. This position is based on objective biological differences between a human and a robot. For example, it makes no sense to recognize the right to life for an AI system, as it does not live in biological sense. Rights, freedoms and obligations of artificial intelligence systems should be secondary in relation to citizens' rights; this provision consolidates in the legal sense the derivativeness of artificial intelligence as a creation of a human being.

Among the potential constitutional rights and freedoms of artificial intellectual systems one may list: the right to be free, the right to self-improvement (learning and self-learning), the right to privacy (protection of software against arbitrary interference of the third persons), the freedom of speech, the freedom of creativity, recognition of AI system's copyright and a limited property right. One may also list specific rights of artificial intelligence, such as the right to access to a source of electric power.

As for the obligations of AI systems, it is proposed to constitutionally stipulate the three renowned robotics laws, formulated by I. Asimov: Non-injuring a person and preventing harm by one's own inaction; obeying all orders given by a person, except those aimed at harming another person; taking care of one's own safety, with the exception of the previous two cases (Naumov & Arkhipov, 2017). Some other obligations will be reflected in the norms of civil and administrative law in this case.

The concept of individual legal personality of artificial intelligence has very little chances for its statutorization for several reasons.

First, the criterion of recognizing the legal personality by the presence of consciousness and self-consciousness is abstract; it allows numerous law breaches, abuse of law and provokes social, political problems as an additional reason for the society stratification. This thesis was detailed in the work by S. Chopra and L. White, who stated that consciousness and self-consciousness are not a necessary and/or sufficient condition for recognizing AI systems as a subject of law (Chopra & White, 2004). In the legal reality, comprehensively conscious individuals, for example, children (or slaves in the Roman law), are deprived of or limited in legal personality. At the same time, people with severe mental disorders, including those recognized as legally incapable, or in a state of coma, i. e. under an objective inability to manifest consciousness, in the former case remain subjects of law (although in a limited form), and in the latter case possess the same complete legal personality, without global changes in their legal status. Potential stipulation of the above said criterion

of consciousness and self-consciousness will make it possible to arbitrary deprive citizens of their legal personality.

Second, AI systems will not be able to implement their rights and obligations in the established legal sense, as they act on the basis of a previously written program, while making legally relevant decisions must be based on subjective, moral choice of a person (Morkhat, 2018b), their direct expression of will. All moral attitudes, feelings and desires of such "person" become derivatives of a human intellect (Uzhov, 2017). Autonomy of AI systems in the sense of their ability to make decisions and implement them independently, without external anthropogenic control or purposeful influence of a human (Musina, 2023), is not full-fledged. Today, artificial intelligence is capable of making only "quasi-autonomous decisions", in one way or another based on ideas and moral attitudes of people. In this context, one may consider only an "action-operation" of an AI system, excluding the possibility of a real moral evaluation of the artificial intelligence behavior (Petev, 2022).

Third, recognition of an individual legal personality of artificial intelligence (moreover in the form of equaling to the status of a physical person) entails a destructive change of the established law order and legal traditions formed since the times of the Roman law, and provokes to pose a number of fundamentally unsolvable philosophical and legal issues in the sphere of human rights. Law as a system of social norms and a social phenomenon was created with the account of human abilities and to provide human interests. The established anthropocentric system of normative regulations, international consensus in the field of the intrinsic rights concept will be deemed legally and factually invalid in case the "extreme inclusivism" approach is established (Dremlyuga & Dremlyuga, 2019). Therefore, endowing AI systems, in particular, "smart" robots with a legal personality may turn out to be not a solution to the existing problems but a Pandora's box aggravating social and political contradictions (Solaiman, 2017).

One more point: the works of supporters of this concept usually mention only robots, i. e. cyberphysical systems of artificial intelligence, which will interact with people in the physical world, while virtual systems are left beyond the pale, although strong artificial intelligence, if it emerges, will be embodied in a virtual form too.

Stemming from the whole range of the above arguments, the concept of individual legal personality of artificial intelligence system should be viewed as juridical unrealistic under the current law order.

2.2. Concept of collective legal personality of artificial intelligence

The concept of collective personals in respect of artificial intellectual systems has acquired significant support among the proponents of acceptability of such legal personality. The main advantage of this approach is that it excludes abstract notions and evaluative judgments (consciousness, self-consciousness, rationality, morals, etc.) from legal workmanship. The approach is based on applying legal fiction to artificial intelligence.

In respect of juridical persons, there already exist "advanced methods of regulation, which could be adapted to solve the dilemma of the legal status of artificial intelligence" (Hárs, 2022).

This concept does not imply actual endowment of AI systems with the legal personality of a physical person, but is just an expansion of the current institute of legal persons, proposing to create a new category of legal persons – cybernetic "electronic organisms"[22] (Musina, 2023). In the context of this approach, it is more appropriate to consider a legal person not in compliance with the contemporary narrow notion, in particular, stipulated in Article 48 of the Russian Civil Code (as an organization possessing separate property and liable with it on its obligations, may on its own behalf acquire and implement civil rights, bear civil obligations, be an applicant and respondent at court), but in a broader sense, which presents a legal person as any construct differing from a physical person, endowed with rights and responsibilities in the form stipulated by law. Thus, the supporters of this approach propose viewing a legal person as a subject-essence (ideal essence) by the Roman law (Sanfilippo, 2007).

Similarity between AI systems and legal persons is seen in the means of endowing them with legal personality – via mandatory state registration of legal persons. Only after completing the established registration procedure a legal person is endowed with legal status and capacity, i. e. becomes a subject of law. Such model keeps discussions about the legal personality of AI systems within the legal framework, excluding the possibility of recognizing legal personality on other (extralegal) grounds, without intrinsic prerequisites, while a person is recognized a subject of law by birth.

An advantage of this concept is expanding on artificial intellectual systems the requirement to enter information into respective state registries similarly to the state registry of legal persons (Popova, 2018) as a necessary condition for endowing them with legal personality. This method implements the important function of systematization of all legal persons and creating a common database, which is necessary both for state authorities to implement control and surveillance (for example, in taxation) and for potential counteragents of such person.

The volume of rights of legal persons in any jurisdiction is, as a rule, smaller than that of physical persons; hence, using this construct for endowing artificial intelligence with legal personality is not linked with endowing it with a number of rights, proposed by the supporters of the previous concept.

When using the technique of legal fiction in relation to legal persons, it is assumed that the actions of a legal person are accompanied by uniting physical persons who form their "will" and implement "expression of will" through administrative bodies of the legal person.

---

22 These should not be confused with "legal electronic persons" – decentralized autonomous organizations, in which coordination of the participants' activity takes place in accordance with previously coordinated set of rules with automated control over their execution (functioning on the basis of blockchain).

In other words, legal persons are artificial (abstract) formations, designed to satisfy the interests of physical persons who acted as their founders or execute control over them. Similarly, artificial intellectual systems are created to satisfy the needs of definite persons – developers, operators, owners. A physical person using AI systems or programming them is guided by their own interests, which are represented by that system in the external environment.

When theoretically estimating such model of regulation, one should not forget that a complete analogy between the positions of legal persons and AI systems is impossible. As was stated above, all legally relevant actions of legal persons are backed by physical persons, who directly make these decisions. The will coming from a legal person is always determined and fully controlled by the will of physical persons (Shutkin, 2020). Hence, without expression of will of physical persons the implementation of activity of legal persons is impossible; with regard to AI systems, the objective problem of their autonomy is already emerging, that is, the possibility to make decisions without interference of a physical person after the moment of direct creation of such system (Ladenkov, 2021).

It is also important to take into account that AI systems do not satisfy the formal sign of organizational unity, which is mandatory for legal persons. The legal status of a legal person has been formed for many centuries and, like law in general, shows the features of "legal conservatism". On the other hand, the current legislation on legal persons largely restricts the possibilities of endowing AI systems with rights and obligations; an attempt to apply this construction with lead to ungrounded legislative deterrence of innovations (Ponkin & Redkina, 2018), which is inadmissible in view of the content of the above mentioned strategic documents, aimed at rapid introduction of artificial intelligence technologies in various economic sectors, social sphere and public management.

Thus, while the concept of collective persons with regard to AI systems has a certain potential, but it does not correspond to the established legal traditions. However, if one stems from the position that "though the issue of personality is binary" (recognition as or non-recognition a person), but "the content of this status is a specter" of possible variants (Chesterman, 2020), then one should rather speak of a gradient legal personality of artificial intelligence, which will be discussed in the next section.

2.3. Concept of gradient legal personality of artificial intelligence

Due to irremovable restrictions of the above-discussed concepts, a large number of researchers suggest their own approaches to solving the issue of a legal status of artificial intellectual systems. Conditionally, one may refer them to different variations of a "gradient legal personality" concept, according a researcher from Leuven University D. M. Mocanu, who implies a limited or partial legal status and capacity of AI systems with a proviso: the term "gradient" is used because it is not just about including or not including certain rights and obligations into the legal status, but about forming a set of such with

a minimally accepted threshold, as well as of recognizing such legal personality for certain goals only (Mocanu, 2021). Then to the two main variations of this concept one may refer the approaches substantiating:

1) endowing AI systems with a special legal status and including "electronic persons" into the law order as an absolutely new category of the subjects of law;

2) endowing AI systems with a limited legal status and capacity within the frameworks of civil-legal relations through creating a category of "electronic agents".

The position of proponents of various approaches within this concept may be united, in view of the fact that there are so far no ontological reasons to view artificial intelligence as a subject of law; nevertheless, in specific situations there already exist functional reasons to endow AI systems with specific rights and obligations, which "proves the best way of fostering the individual and social interests that the law is meant to protect", endowing these systems with "limited and narrow forms of legal personality" (Bertolini & Episcopo, 2022).

Granting artificial intelligence systems with a special legal status through creating a separate legal institution of "electronic persons" has a major advantage of detailed clarification and regulation of the relations emerging:

– between legal and physical persons and AI systems;

– between AI systems and their developers (operators, owners);

– between a third party and AI systems in civil-legal relations[23].

Within this legal construction, AI system will be controlled and managed separately from its developer, owner or operator (Morkhat, 2018b). Presenting the definition of an "electronic person", P. M. Morkhat focuses on using the above mentioned technique of a legal fiction and the functional orientation of a specific artificial intelligence model: "electronic person" is a technical-legal image (possessing some features of a legal fiction similarly to a legal person), reflecting and embodying a conditionally specific legal personality of an artificial intelligence system, differing depending on its intended function or purpose and capabilities (Morkhat, 2018a).

Just as the concept of collective persons in regard to AI systems, this approach implies keeping special registries of "electronic persons". A detailed and clear statement of the rights and obligations of "electronic persons" serves as the basis for subsequent control on the part of the state and owner of such AI systems. An accurately defined circle of authorities, a narrowed volume of legal status and capacity of "electronic persons" will allow tracing that the given "person" does not go beyond its program due to potentially making autonomous decisions and constant self-learning.

---

[23] Schrijver, S. de (2018, January 5). The Future Is Now: Legal Consequences of Electronic Personality for Autonomous Robots. *Who's Who Legal*. https://whoswholegal.com/features/the-future-is-now-legal-consequences-of-electronic-personality-for-autonomous-robots

In formal-legal terms, this model is analogous to endowing legal persons, for example, in the form of unitary enterprises, with a limited (special) legal capacity by implication of clause 2 of Article 48 CC RF. It is also proposed to license certain types of "electronic persons" depending on the activity executed by them, similarly to the licensing stipulated by Federal Law of May 4, 2011 No. 99-FZ "On licensing certain types of activity"[24]. Under such approach, artificial intelligence, at the stage of its creation being an object of intellectual property of software developers, may be endowed with legal personality after relevant certification and state registration, but the legal status and capacity of an "electronic person" will be of special character (Vavilin, 2022).

The introduction of a fundamentally new institution for an established law order will have serious legal consequences, requiring a profound reform of legislation at least in the areas of constitutional and civil law. Researchers rightly note that caution should be exercised when introducing the concept of an «electronic person,» given the difficulties in introducing new persons in law, as expanding the concept of «person» in a legal sense could potentially occur at the expense of limiting the rights and lawful interests of the existing subjects of legal relations. (Bryson et al., 2017). Accounting for these aspects seems realistically impossible, as the legal personality of physical persons, legal persons, and public-legal entities is a result of centuries-long evolution of the theory of state and law.

The second approach within the concept of gradient legal personality is the legal notion of "electronic agents", primarily associated with the widespread use of AI systems as means of communication between counteragents and as tools for online commerce. This approach can be called a compromise, as it acknowledges the impossibility of endowing AI systems with the status of full-fledged subjects of law, while at the same time establishing certain (socially significant) rights and obligations for artificial intelligence. In other words, the concept of "electronic agents" legalizes the quasi-subjectivity of artificial intelligence. The term "quasi-subject of law" should be understood as a certain legal phenomenon, in which individual elements of legal personality are recognized on an official or doctrinal level, while establishing the status of a full-fledged subject of law is impossible (Channov, 2022).

Supporters of this approach highlight the functional features of AI systems that allow them to act as both a passive tool and an active participant in legal relationships, potentially capable of independently creating legally relevant contracts for the system owner. That is why AI systems can be conditionally viewed in the framework of agency relationships (Morkhat, 2018b). When creating (or registering) an AI system, the initiator of the "electronic agent" activity concludes a factual unilateral agency agreement with it, as a result of which the "electronic agent" is endowed with a number of authorities, exercising which it can perform legal actions that are significant for the principal.

[24] On licensing certain types of activity: Federal Law of May 4, 2011 No. 99-FZ (ed. of 29.12.2022). (2011). *Collection of legislation of the Russian Federation, 2011*. No. 19. Article 2716.

Provisions on agency relationships with AI systems were first mentioned in Russia in connection with the development of the draft law "On amending the Civil Code of the Russian Federation in terms of improving the legal regulation of relations in the field of robotics", prepared in 2016. This project became known informally as the Grishin Law after D. Grishin, the founder of an investment fund Grishin Robotics and the Chairman of the Board of Directors of Mail.Ru Group. By implication of the law draft, an "electronic agent" should be recognized as a robot which by its owner decision and due to its constructive features is intended for participation in civil transactions. A robot-agent has separate property and is liable for its obligations, can acquire and exercise civil rights and bear civil responsibilities on its own behalf. In cases provided by law, a robot-agent may act as a participant in civil proceedings. If this draft law were approved, it would legalize AI systems as participants in legal relationships in Russia.

In the context of "electronic agents", the most problematic issue is whether such entities possess separate property that would allow them to be held accountable for acquired civil law obligations. The authors of the draft law attempted to take into account the functional specifics of different types of artificial intelligence and proposed dividing AI systems[25] into two types:

– AI systems (in the draft law – robots) as a special form of property, for which analogy with animals and other objects of law is potentially possible (AI systems of type 1 – objects of law);

– AI agents as participants in civil legal relations possessing a special legal personality (AI systems of the second type – quasi-subjects of law).

It should be noted that the authors of the 2016 draft law did not mention the possible legal status of virtual systems as another form of artificial intelligence, stating only that "provisions of civil legislation on robots do not apply to software which, although capable of acting, defining their actions and evaluating their consequences without complete control by humans based on the results of processing information received from the external environment, is not part of the information system of an isolated device intended fully or partially for taking autonomous actions"[26]. Currently, the achieved level of development in the field of artificial intelligence allows for the assumption that strong AI can exist in virtual form and control intellectually weaker cyber-physical systems.

---

25 The draft law only mentions robots; such incomplete wording was inherent not only to the authors of this draft but also, for example, European authors of the EU Resolution of February 16, 2017 concerning the civil legal norms on robotics (European Parliament Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics 2015/2013 (INL). Later in EU Resolutions of 2020 this defect was liquidated.

26 Dmitriy Grishin presented a draft regulation of legal status of robots in Russia. (2016, December 17). https://robotrends.ru/pub/1650/dmitriy-grishin-predstavil-proekt-regulirovaniya-pravovogo-statusa-robotov

The approaches to endowing AI systems with elements of legal personality within the framework of the concept discussed in this section can be compared to a constructor that allows for building something new from the existing concepts described in the previous sections of the article, by combining, taking into account the functional characteristics of AI systems and the range of tasks for which a specific AI model is designed to solve. Therefore, it can be considered that the gradient concept has the greatest chance of implementation within the existing law order.

Conclusions

Determining the legal status of artificial intellectual systems is the issues causing increasingly heated discussions among jurists. One may agree with U. Pagallo, one of the researchers most deeply immersed in this problem, that in the years to come there will hardly be found solutions for "all of the hard cases and dilemmas" associated with artificial intelligence; nevertheless, preventing the "polarization of today's debate, methods of legal flexibility and pragmatic experimentation" allow solving even such difficult tasks (Pagallo, 2018).

Having considered the main concepts of endowing AI systems with legal personality that have been formulated up to this point, we should state at least a legal impracticality in granting artificial intelligence the status of a legal subject in the classical understanding of legal theory. Furthermore, as time goes on, there is an even smaller possibility of maintaining the legal regime of the object of law in its current form. In modern technological, economic, social, political, and legal realities it is likely that a combined approach will be needed to determine the legal status of artificial intelligence.

One optimal solution could involve including AI systems in the list of objects of civil rights, but differentiating the legal regulation of artificial intelligence as an object of law and an "electronic agent" as a quasi-subject of law. The line of differentiation should be drawn depending on the functional differences of AI systems, while as an "electronic agent" can be recognized not only a robot but also a virtual intellectual system. An "electronic agent" is endowed with certain rights and can perform some legal obligations, but ultimately, responsibility for its actions should be borne by a human. Recognizing AI systems as "electronic persons" seems premature, at least until the emergence of strong artificial intelligence.

In the future, considering the ongoing transformation of law and the testing of AI systems for machine-readable law and decision-making support in public administration, one cannot exclude the likelihood of a gradual increase of artificial intelligence impact in the sphere of law. This will contribute to the strengthening of artificial intelligence position on its way towards recognition as a legal subject and the complete "overhaul" of legal matters with its participation or even under its guidance, no matter how fantastic it may seem at first glance.

References

Abbott, R. (2020). *The Reasonable Robot. Artificial Intelligence and the Law*. Cambridge University Press.

Alekseev, A., Alekseeva, E., & Emelyanova, N. (2023). Artificial Personality in socio-political communication. *Artificial Societies*, *18*(1). (In Russ.). https://doi.org/10.18254/s207751800024370-6

Arkhipov, V. V., & Naumov, V. B. (2017). On certain issues of theoretic grounds for development of robotics legislation: the aspects of will and legal personality. *Statute*, *5*, 157–170. (In Russ.).

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

Avila Negri, S. M. C. (2021). Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence. *Frontiers in Robotics and AI*, *8*, Art. 789327. https://doi.org/10.3389/frobt.2021.789327

Balkin, J. M. (2015).The Path of Robotics Law. *California Law Review*, *6*, 45–60.

Bertolini, A., & Episcopo, F. (2022). Robots and AI as Legal Subjects? Disentangling the Ontological and Functional Perspective. *Frontiers in Robotics and AI*, *9*, Art. 842213. https://doi.org/10.3389/frobt.2022.842213

Bryson, J. J., Diamantis, M. E., & Grant, Th. D. (2017). Of, For, and By the People: The Legal Lacuna of Synthetic Persons. *Artificial Intelligence and Law*, *25*, 273–291.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review*, *103*(3), 513–563.

Channov, S. E. (2022). Robot (Artificial Intelligence System) as a Subject (Quasi-Subject) of Law. *Actual Problems of Russian Law*, *17*(12), 94–109. (In Russ.). https://doi.org/10.17803/1994-1471.2022.145.12.094-109

Chesterman, S. (2020). Artificial Intelligence and the Limits of Legal Personality. *International & Comparative Law Quarterly*, *69*, 819–844. https://doi.org/10.1017/s0020589320000366

Chopra, S., & White, L. (2004). Artificial Agents – Personhood in Law and Philosophy. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004* (pp. 635–639). Valencia: IOS Press.

Dremliuga, R. I., & Dremliuga, O. A. (2019). Artificial intelligence – a legal person: the arguments for and against. *Pravovaya politika i pravovaya zhizn*, *2*, 120–125. (In Russ.).

Gellers, J. C. (2021). *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*. London: Routledge.

Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law*, *30*, 291–323.

Grin, S. N. (2018). Robots' emancipation: elements of legal personhood in the construct of artificial intelligence. *Biznes. Obshchestvo. Vlast*, *1*(27), 233–242. (In Russ.).

Hárs, A. (2022). AI and international law – Legal personality and avenues for regulation. *Hungarian Journal of Legal Studies*, *62*(4), 320–344. https://doi.org/10.1556/2052.2022.00352

Karnouskos, S. (2022). Symbiosis with artificial intelligence via the prism of law, robots, and society. *Artificial Intelligence and Law*, *30*, 93–115.

Kharitonova, Yu. S., Savina, V. S., & Pagnini, F. (2022). Civil liability in the development and application of artificial intelligence and robotic systems: basic approaches. *Perm University Herald. Juridical sciences*, *4*(58), 683–708. (In Russ.). https://doi.org/10.17072/1995-4190-2022-58-683-708

Ladenkov, N. S. (2021). Models of endowing artificial intelligence with legal personality. *Vestnik IKBFU. Humanities and Social Sciences*, *3*, 12–20. (In Russ.).

Laptev, V. A. (2019). Artificial Intelligence and Liability for its Work. *Pravo. Zhurnal Vysshey Shkoly Ekonomiki*, *2*, 79–102. (In Russ.). https://doi.org/10.17323/2072-8166.2019.2.79.102

McCarty, L. T. (2017). Finding the Right Balance in Artificial Intelligence and Law. In *Research Handbook on the Law of Artificial Intelligence* (Chapter 3, pp. 55–87). Edward Elgar Publishing.

McNally, Ph. & Inayatullah, S. (1988). The Rights of Robots: Technology, Culture and Law in the 21st Century. *Futures*, *20*(1), 119–136. https://doi.org/10.1016/0016-3287(88)90019-5

Mocanu, D. M. (2021). Gradient Legal Personhood for AI Systems – Painting Continental Legal Shapes Made to Fit Analytical Molds. *Frontiers in Robotics and AI*, *8*, Art. 788179. https://doi.org/10.3389/frobt.2021.788179

Morkhat, P. M. (2018a). Artificial intelligence unit as electronic personality. *Bulletin of the Moscow State Regional University (Jurisprudence)*, *2*, 61–73. (In Russ.). https://doi.org/10.18384/2310-6794-2018-2-61-73

Morkhat, P. M. (2018b). Legal personality of artificial intelligence unit: some civil-legal approaches. *Bulletin of Kostroma State University*, *3*, 280–283. (In Russ.).

Mulgan, T. (2019). Corporate Agency and Possible Futures. *Journal of Business Ethics*, *154*, 901–916. https://doi.org/10.1007/s10551-018-3887-1

Musina, K. S. (2023). Theoretical aspects of identifying legal personality of artificial intelligence: cross-national analysis of the laws of foreign countries. *RUDN Journal of Law*, *27*(1), 135–147. (In Russ.). https://doi.org/10.22363/2313-2337-2023-27-1-135-147

Naumov, V. B., & Arkhipov, V. V. (2017). Draft Federal law “On amendments to the Civil Code of the Russian Federation in improving the legal regulation of relations in the field of robotics”. In N. A. Sheveleva (Ed.), *Law and Information: the Questions of Theory and Practice: Collection of works of international scientific and practical conference.* Saint Petersburg: The Presidential Library. (In Russ.).

Nechkin, A. V. (2020). Constitutional and Legal Status of Artificial Intelligence in Russia: Present and Future. *Lex Russica*, *8*, 78–85. (In Russ.). https://doi.org/10.17803/1729-5920.2020.165.8.078-085

Pagallo, U. (2018). Apples, oranges, robots: four misunderstandings in today’s debate on the legal status of AI systems. *Philosophical Transactions of the Royal Society*, *376*(2133), Art. 20180168. https://doi.org/10.1098/rsta.2018.0168

Petev, N. I. (2022). Existential, legal and ethical problems of artificial intelligence. *Journal of Wellbeing Technologies*, *2*(45), 55–70. (In Russ.). https://doi.org/10.18799/26584956/2022/2/1159

Ponkin, I. V., & Redkina A. I. (2018). Artificial Intelligence from the Point of View of Law. *RUDN Journal of Law*, *22*(1), 91–109. (In Russ.). https://doi.org/10.22363/2313-2337-2018-22-1-91-109

Popova, A. V. (2018). New Subjects of the Information Society and the Knowledge Society: To the Question of Legal Regulation, *Journal of Russian Law*, *6*(11), 14–24. (In Russ.). https://doi.org/10.12737/art_2018_11_2

Sanfilippo, Ch. (2007). *Course in Roman private law: tutorial* (transl. by I. I. Makhankov, D. V. Dozhdev (Ed.)). Moscow: Norma. (In Russ.).

Shutkin, S. I. (2020). Is legal personhood of AI possible? *Works on Intellectual Property*, *35*(1–2), 90–137. (In Russ.).

Solaiman, S. M. (2017). Legal Personality of Robots, Corporations, Idols and Chimpanzees: A Quest for Legitimacy. *Artificial Intelligence and Law*, *25*(2), 155–179. https://doi.org/10.1007/s10506-016-9192-3

Solum, L. B. (1992). Legal Personhood for Artificial Intelligences. *North Carolina Law Review*, *70*(4), 1231–1287.

Somenkov, S. A. (2019). Artificial intelligence: from object to subject? *Courier of the Kutafin Moscow State Law University*, *2*(54), 75–85. (In Russ.).

Stiglitz, J. E. (2017). The coming great transformation. *Journal of Policy Modeling*, *39*(4), 625–638. https://doi.org/10.1016/j.jpolmod.2017.05.009

Uzhov, F. V. (2017). Legal personality of artificial intelligence. *Gaps in Russian Legislation*, *3*, 357–360. (In Russ.).

Vasilevskaya, L. Yu., Poduzova, E. B., & Tasalov, F. A. (2021). *Digitalization of civil turnover: legal characteristics of “artificial intelligence” and “digital” subjects (civilistic research)* (In 5 vol. Vol. 3). Moscow: Prospect. (In Russ.).

Vavilin, E. V. (2022). The status of artificial intelligence: from object to the subject of legal relations. *Vestnik Tomskogo Gosudarstvennogo Universiteta. Pravo*, 45, 147–158. (In Russ.). https://doi.org/10.17223/22253513/45/10

Authors information

Irina A. Filipova – Candidate of Sciences in Jurisprudence, Associate Professor, Associate Professor of the Department of Labor Law and Environmental Law, National Research Lobachevsky State University of Nizhny Novgorod, Head of Central Asia research center for artificial intelligence regulation, Samarkand State University
Address: 23 prospekt Gagarina, Nizhniy Novgorod 603922, Russian Federation;
15 Universitetskiy boulevard, Samarkand 140104, Republic of Uzbekistan
E-mail: irinafilipova@yandex.ru
ORCID ID: https://orcid.org/0000-0003-1773-5268
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57327205000
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/R-1375-2016
Google Scholar ID: https://scholar.google.com/citations?user=opJc7fcAAAAJ
RSCI Author ID: https://www.elibrary.ru/author_items.asp?authorid=461586

Vadim D. Koroteev – 4th year student of Law Faculty, National Research Lobachevsky State University of Nizhny Novgorod
Address: 23 prospekt Gagarina, Nizhniy Novgorod 603922, Russian Federation
E-mail: redsert87@gmail.com
ORCID ID: https://orcid.org/0009-0004-4182-2411
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/IAP-5405-2023
RSCI Author ID: https://www.elibrary.ru/author_profile.asp?id=1198439

Authors' contributions

Irina A. Filipova formulated the research idea, goals and objectives; elaborated the methodology; analyzed and summarized literature; interpreted general research results; critically reviewed and edited the manuscript text; formulated the key conclusions, proposals and recommendations; approved the final variant of the article.
Vadim D. Koroteev compiled the manuscript draft and critically reviewed it, adding valuable comments on the intellectual content; participated in scientific design; performed comparative analysis; selected literature; prepared and edited the manuscript text; interpreted the specific research results; finalized the manuscript.

Conflict of interests

I. A. Filipova is a Deputy Editor-in-Chief of the Journal; the article has been reviewed on general terms.

Funding

The research was not sponsored.

Thematic rubrics:

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – April 23, 2023
Date of approval – May 8, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 34.023:34.025:004.8
EDN: https://elibrary.ru/immoam
DOI: https://doi.org/10.21202/jdtl.2023.15

Будущее искусственного интеллекта: объект или субъект права?

Ирина Анатольевна Филипова ✉

Национальный исследовательский Нижегородский государственный университет имени Н. И. Лобачевского
г. Нижний Новгород, Российская Федерация;
Самаркандский государственный университет
г. Самарканд, Республика Узбекистан

Вадим Дмитриевич Коротеев

Национальный исследовательский Нижегородский государственный университет имени Н. И. Лобачевского
г. Нижний Новгород, Российская Федерация

Ключевые слова

Генеративная модель,
интеллектуальная система,
искусственный интеллект,
квазисубъект права,
киберфизическая система,
право,
правосубъектность,
робот,
цифровые технологии,
электронное лицо

Аннотация

Цель: выявление проблем, связанных с правовым регулированием общественных отношений, в которых используются системы искусственного интеллекта, и рациональное осмысление обсуждаемой правоведами возможности наделения таких систем статусом субъекта права.

Методы: методологической основой исследования являются общенаучные методы анализа и синтеза, аналогии, абстрагирования и классификации. Среди преимущественно применяемых в работе юридических методов – формально-юридический, сравнительно-правовой и системно-структурный, а также методы толкования права и правового моделирования.

Результаты: представлен обзор состояния развития искусственного интеллекта и его внедрения на практике ко времени проведения исследования. Рассмотрено нормативно-правовое регулирование в данной области и разобраны основные из существующих концепций наделения искусственного интеллекта правосубъектностью (индивидуальная, коллективная и градиентная правосубъектность искусственного интеллекта). При этом дана оценка каждого из подходов и сделаны выводы о наиболее предпочтительных вариантах внесения изменений

✉ Контактное лицо

в действующее законодательство, которое перестает соответствовать реалиям. Растущее несоответствие связано с ускоряющимся развитием искусственного интеллекта и его распространением в различных секторах экономики, социальной сферы, а в ближайшей перспективе и в государственном управлении. Все это свидетельствует о повышении риска разрыва правовой материи с изменяющейся социальной реальностью.

Научная новизна: классифицированы научные подходы к наделению искусственного интеллекта правосубъектностью. В рамках каждого из подходов выделены ключевые моменты, использование которых позволит в дальнейшем создавать правовые конструкции на основе комбинирования, уходя от крайностей и соблюдая баланс интересов всех сторон. Оптимальным вариантом определения правового статуса искусственного интеллекта может стать внесение интеллектуальных систем в перечень объектов гражданских прав, но с дифференциацией правового регулирования искусственного интеллекта в качестве объекта права и «электронного агента» как квазисубъекта права. Линия разграничения должна проходить в зависимости от функциональных различий интеллектуальных систем, причем «электронным агентом» может быть признан не только робот, но и виртуальная интеллектуальная система.

Практическая значимость: материалы исследования могут быть применены в работе, связанной с подготовкой предложений по внесению изменений и дополнений в действующее законодательство, а также при разработке учебных курсов и написании учебных пособий по тематике, имеющей отношение к регулированию использования искусственного интеллекта.

Для цитирования

Филипова, И. А., Коротеев, В. Д. (2023). Будущее искусственного интеллекта: объект или субъект права? *Journal of Digital Technologies and Law*, *1*(2), 359–386. https://doi.org/10.21202/jdtl.2023.15

Список литературы

Алексеев, А. Ю., Алексеева, Е. А., Емельянова, Н. Н. (2023). Искусственная личность в социально-политической коммуникации. *Искусственные общества*, *18*(1), 1. https://doi.org/10.18254/s207751800024370-6

Архипов, В. В., Наумов, В. Б. (2017). О некоторых вопросах теоретических оснований развития законодательства о робототехнике: аспекты воли и правосубъектности. *Закон*, *5*, 157–170.

Вавилин, Е. В. (2022). Статус искусственного интеллекта: от объекта к субъекту правовых отношений. *Вестник Томского государственного университета. Право*, *45*, 147–158. https://doi.org/10.17223/22253513/45/10

Василевская, Л. Ю., Подузова, Е. Б., Тасалов, Ф. А. (2021). *Цифровизация гражданского оборота: правовая характеристика «искусственного интеллекта» и «цифровых» субъектов (цивилистическое исследование)* (В 5 т. Т. 3). Москва: Проспект.

Гринь, С. Н. (2018). Эмансипация роботов: элементы правосубъектности в конструкции искусственного интеллекта. *Бизнес. Общество. Власть*, *1*(27), 233–242.

Дремлюга, Р. И., Дремлюга, О. А. (2019). Искусственный интеллект – субъект права: аргументы за и против. *Правовая политика и правовая жизнь*, *2*, 120–125.

Ладенков, Н. Е. (2021). Модели наделения искусственного интеллекта правосубъектностью. *Вестник Балтийского федерального университета им. И. Канта. Серия: Гуманитарные и общественные науки*, *3*, 12–20.

Лаптев, В. А. (2019). Понятие искусственного интеллекта и юридическая ответственность за его работу. *Право. Журнал Высшей школы экономики*, *2*, 79–102. https://doi.org/10.17323/2072-8166.2019.2.79.102

Мохат, П. М. (2018a). Правосубъектность юнита искусственного интеллекта: некоторые гражданско-правовые подходы. *Вестник КГУ*, *3*, 280–283.

Мохат, П. М. (2018b). Юнит искусственного интеллекта как электронное лицо. *Вестник МГОУ. Серия: Юриспруденция*, *2*, 61–73. https://doi.org/10.18384/2310-6794-2018-2-61-73

Мусина, К. С. (2023). Идентификация правосубъектности искусственного интеллекта: кросснациональный анализ законодательств зарубежных стран. *Вестник Российского университета дружбы народов. Серия: Юридические науки*, *27*(1), 135–147. https://doi.org/10.22363/2313-2337-2023-27-1-135-147

Наумов, В. Б., Архипов, В. В. (2017). Проект Федерального закона «О внесении изменений в Гражданский кодекс Российской Федерации в части совершенствования правового регулирования отношений в области робототехники». В сб.: Н. А. Шевелёва (ред.), *Право и информация: вопросы теории и практики: сборник материалов VII Международной научно-практической конференции. Сер. «Электронное законодательство»*, *7* (с. 220–226). Санкт-Петербург: Президентская бибилиотека.

Нечкин, А. В. (2020). Конституционно-правовой статус искусственного интеллекта в России: настоящее и будущее. *Lex Russica*, *8*(165), 78–85. https://doi.org/10.17803/1729-5920.2020.165.8.078-085

Петев, Н. И. (2022). Экзистенциальная, правовая и этическая проблемы искусственного интеллекта. *Векторы благополучия: экономика и социум*, *2*(45), 55–70. https://doi.org/10.18799/26584956/2022/2/1159

Понкин, И. В., Редькина, А. И. (2018). Искусственный интеллект с точки зрения права. *Вестник Российского университета дружбы народов. Серия: Юридические науки*, *1*, 91–109. https://doi.org/10.22363/2313-2337-2018-22-1-91-109

Попова, А. В. (2018). Новые субъекты информационного общества и общества знания: к вопросу о нормативном правовом регулировании. *Журнал российского права*, *11*(263), 14–24. https://doi.org/10.12737/art_2018_11_2

Санфилиппо, Ч. (2007). *Курс римского частного права*: учебник (пер. с итал. И. И. Маханькова, под общ. ред. Д. В. Дождева). Москва: Норма, 2007.

Соменков, С. А. (2019). Искусственный интеллект: от объекта к субъекту?. *Вестник Университета имени О. Е. Кутафина*, *2*(54), 75–85.

Ужов, Ф. В. (2017). Искусственный интеллект как субъект права. *Пробелы в российском законодательстве*, *3*, 357–360.

Харитонова, Ю. С., Савина, В. С., Паньини, Ф. (2022). Гражданско-правовая ответственность при разработке и применении систем искусственного интеллекта и робототехники: основные подходы. *Вестник Пермского университета. Юридические науки*, *58*, 683–708. https://doi.org/10.17072/1995-4190-2022-58-683-708

Чаннов, С. Е. (2022). Робот (система искусственного интеллекта) как субъект (квазисубъект) права. *Актуальные проблемы российского права*, *17*(12), 94–109. https://doi.org/10.17803/1994-1471.2022.145.12.094-109

Шуткин, С. И. (2020). Возможна ли правосубъектность искусственного интеллекта. *Труды по интеллектуальной собственности*, *35*(1–2), 90–137.

Abbott, R. (2020). *The Reasonable Robot. Artificial Intelligence and the Law*. Cambridge University Press.

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

Avila Negri, S. M. C. (2021). Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence. *Frontiers in Robotics and AI*, *8*, Art. 789327. https://doi.org/10.3389/frobt.2021.789327

Balkin, J. M. (2015). The Path of Robotics Law. *California Law Review*, *6*, 45–60.

Bertolini, A., & Episcopo, F. (2022). Robots and AI as Legal Subjects? Disentangling the Ontological and Functional Perspective. *Frontiers in Robotics and AI*, *9*, Art. 842213. https://doi.org/10.3389/frobt.2022.842213

Bryson, J. J., Diamantis, M. E., & Grant, Th. D. (2017). Of, For, and By the People: The Legal Lacuna of Synthetic Persons. *Artificial Intelligence and Law*, *25*, 273–291.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review*, *103*(3), 513–563.

Chesterman, S. (2020). Artificial Intelligence and the Limits of Legal Personality. *International & Comparative Law Quarterly*, *69*, 819–844. https://doi.org/10.1017/s0020589320000366

Chopra, S., & White, L. (2004). Artificial Agents – Personhood in Law and Philosophy. In *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004* (pp. 635–639). Valencia: IOS Press.

Gellers, J. C. (2021). *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*. London: Routledge.

Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law*, *30*, 291–323.

Hárs, A. (2022). AI and international law – Legal personality and avenues for regulation. *Hungarian Journal of Legal Studies*, *62*(4), 320–344. https://doi.org/10.1556/2052.2022.00352

Karnouskos, S. (2022). Symbiosis with artificial intelligence via the prism of law, robots, and society. *Artificial Intelligence and Law*, *30*, 93–115.

McCarty, L. T. (2017). Finding the Right Balance in Artificial Intelligence and Law. In *Research Handbook on the Law of Artificial Intelligence* (Chapter 3, pp. 55–87). Edward Elgar Publishing.

McNally, Ph., Inayatullah, S. (1988). The Rights of Robots: Technology, Culture and Law in the 21st Century. *Futures*, *20*(1), 119–136. https://doi.org/10.1016/0016-3287(88)90019-5

Mocanu, D. M. (2021). Gradient Legal Personhood for AI Systems – Painting Continental Legal Shapes Made to Fit Analytical Molds. *Frontiers in Robotics and AI*, *8*, Art. 788179. https://doi.org/10.3389/frobt.2021.788179

Mulgan, T. (2019). Corporate Agency and Possible Futures. *Journal of Business Ethics*, *154*, 901–916. https://doi.org/10.1007/s10551-018-3887-1

Pagallo, U. (2018). Apples, oranges, robots: four misunderstandings in today's debate on the legal status of AI systems. *Philosophical Transactions of the Royal Society*, *376*(2133), Art. 20180168. https://doi.org/10.1098/rsta.2018.0168

Solaiman, S. M. (2017). Legal Personality of Robots, Corporations, Idols and Chimpanzees: A Quest for Legitimacy. *Artificial Intelligence and Law*, *25*(2), 155–179. https://doi.org/10.1007/s10506-016-9192-3

Solum, L. B. (1992). Legal Personhood for Artificial Intelligences. *North Carolina Law Review*, *70*(4), 1231–1287.

Stiglitz, J. E. (2017). The coming great transformation. *Journal of Policy Modeling*, *39*(4), 625–638. https://doi.org/10.1016/j.jpolmod.2017.05.009

Сведения об авторах

Филипова Ирина Анатольевна – кандидат юридических наук, доцент, доцент кафедры трудового и экологического права, Национальный исследовательский Нижегородский государственный университет имени Н. И. Лобачевского; руководитель Центрально-Азиатского исследовательского центра регулирования искусственного интеллекта, Самаркандский государственный университет
Адрес: 603922, Российская Федерация, г. Нижний Новгород, проспект Гагарина, 23; 140104, Республика Узбекистан, г. Самарканд, Университетский бульвар, 15
E-mail: irinafilipova@yandex.ru
ORCID ID: https://orcid.org/0000-0003-1773-5268
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57327205000
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/R-1375-2016
Google Scholar ID: https://scholar.google.com/citations?user=opJc7fcAAAAJ
РИНЦ Author ID: https://www.elibrary.ru/author_items.asp?authorid=461586

Коротеев Вадим Дмитриевич – студент 4-го курса юридического факультета, Национальный исследовательский Нижегородский государственный университет имени Н. И. Лобачевского
Адрес: 603922, Российская Федерация, г. Нижний Новгород, проспект Гагарина, 23
E-mail: redsert87@gmail.com
ORCID ID: https://orcid.org/0009-0004-4182-2411
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/IAP-5405-2023
РИНЦ Author ID: https://www.elibrary.ru/author_profile.asp?id=1198439

Вклад авторов

И. А. Филипова осуществляла формулирование идеи, исследовательских целей и задач; разработку методологии; анализ и обобщение литературы; интерпретацию общих результатов исследования; критический пересмотр и редактирование текста рукописи; формулировку ключевых выводов, предложений и рекомендаций; утверждение окончательного варианта статьи.
В. Д. Коротеев осуществлял составление черновика рукописи и его критический пересмотр с внесением ценных замечаний интеллектуального содержания; участие в научном дизайне; проведение сравнительного анализа; сбор литературы; подготовку и редактирование текста статьи; интерпретацию частных результатов исследования; оформление рукописи.

Конфликт интересов

И. А. Филипова является заместителем главного редактора журнала, статья прошла рецензирование на общих основаниях.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.07.45 / Право и научно-технический прогресс
Специальность ВАК: 5.1.1 / Теоретико-исторические правовые науки

История статьи

Дата поступления – 23 апреля 2023 г.
Дата одобрения после рецензирования – 8 мая 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.16

Algorithmic Discrimination and Privacy Protection

Elena Falletti

Università Carlo Cattaneo – LIUc
Castellanza, Italy

Keywords

Algorithm,
artificial intelligence,
data protection,
digital technologies,
discrimination,
law,
personal data,
privacy,
private life,
regulation

Abstract

Objective: emergence of digital technologies such as Artificial intelligence became a challenge for states across the world. It brought many risks of the violations of human rights, including right to privacy and the dignity of the person. That is why it is highly relevant to research in this area. That is why this article aims to analyse the role played by algorithms in discriminatory cases. It focuses on how algorithms may implement biased decisions using personal data. This analysis helps assess how the Artificial Intelligence Act proposal can regulate the matter to prevent the discriminatory effects of using algorithms.

Methods: the methods used were empirical and comparative analysis. Comparative analysis allowed to compare regulation of and provisions of Artificial Intelligence Act proposal. Empirical analysis allowed to analyse existing cases that demonstrate us algorithmic discrimination.

Results: the study's results show that the Artificial Intelligence Act needs to be revised because it remains on a definitional level and needs to be sufficiently empirical. Author offers the ideas of how to improve it to make more empirical.

Scientific novelty: the innovation granted by this contribution concerns the multidisciplinary study between discrimination, data protection and impact on empirical reality in the sphere of algorithmic discrimination and privacy protection.

Practical significance: the beneficial impact of the article is to focus on the fact that algorithms obey instructions that are given based on the data that feeds them. Lacking abductive capabilities, algorithms merely act as obedient executors of the orders. Results of the research can be used as a basis for further research in this area as well as in law-making process.

For citation

Falletti, E. (2023). Algorithmic Discrimination and Privacy Protection. *Journal of Digital Technologies and Law, 1*(2), 387–420. https://doi.org/10.21202/jdtl.2023.16

Contents

Introduction

The 18th-century Industrial Revolution was the precursor to social change and claimed rights through automation in the production process. It fostered the advent of goods manufacturing by creating places dedicated to the production, shifting activity from the countryside to urban centres. It brought about a social transformation of society by displacing the masses into an alienating reality: the factories erased mentalities, behaviours, and customs linked to the previous agricultural life rooted in the culture and collective memory (Tunzelmann, 2003; Freeman et al., 2001).

From the perspective of the relationship between law and technology, this change has had a hard legal and social impact since technological evolution promotes a unifying thrust by standardising the rules of conduct of the new production process. Such a transition represents a shift towards an articulated system of norms, which, in the function of technical precepts, obliges each subject involved to homologate their behaviour according to technical standards that take on legal value.

This context allows us to make a historically diachronic comparison between the automation introduced in the 19th century in British factories and today's algorithmic automation. This comparison concerns the measurement of time during work performance. As has been the case in the past, the first area where innovation in human activity manifests itself and develops is the field of work. This circumstance applies first and foremost

to automation: on the one hand, the performance of a job, a biblical punishment, allows individuals to live an independent life; on the other, automating work allows entrepreneurs and producers to have a series of savings, to the detriment of workers, whose role is comparable to that of a mere machine (Marx, 2016). The Marxist curse of 'alienated labour' is also realised (Marx, 2016; Claeys, 2018).

The measurement of labour time in order to cut costs is one of the most significant concerns when managing an organisation so complex as a business: the automation of the production process in the 19th century extended the workday beyond daylight hours (thanks to electricity) and detached it from weather conditions, separating labour from agricultural rhythms, whereas algorithmic automation creates an even greater extreme since the algorithm can manage production processes without human control, achieving the kind of alienation evoked by Marx of the individual worker estranged from other workers (Marx, 2016).

By applying Marxist theory to the length of working days (Marx, 2016; Claeys, 2018), illo tempore as well as now, a paradoxical effect can be observed: on the one hand, capitalists expand their profits by increasing both the labour capacity of workers and the surplus value of production; however, this method results in marginal returns, as the lengthening of shifts exhausts employees. Such is true concerning the manufacturing capitalist and the digital platform, which hires workers to perform algorithms-managed tasks. In both situations, the increase in surplus value produced by the engaged labour force involves the physical control of workers. While in the past, said control was made problematic because, unlike other production factors, the power of labour is (or was) constituted by its embodiment in people, workers who, in turn, maintain (or maintained) their power to resist being treated as commodities (Woodcock, 2020; Altenried, 2020).

In the «Digital Revolution era», such correspondence between physicality and work is severely undermined by the solitude of work performance. Indeed, the worker has the platform managed by algorithms as their only reference. They are isolated from other colleagues and have as one means of a direct relationship with the app that allows interaction between them and the algorithm. Algorithmic discrimination in contemporary workplaces sees workers deprived of proper legal and wage recognition, and it enables the perpetuation of categorical and legal distinctions that reflect those related to work (Fuchs, 2014).

The Industrial Revolution changed the paradigm of equality and, mirroring it, that of discrimination since the advent of capitalism had brought about a change in social reality and lifestyles (Schiavone, 2019). In philosophical discourse, equality acquired a new centrality that broke away from the religious and philosophical perspective and took on the social and political points of view embodied in the great revolutions of the time, namely the American and French (Schiavone, 2019).

Equality, inequality, and, therefore, discrimination moved out of the formal condition of only slavery, which still existed, and entered the social facts about which a broader public opinion began to form beyond that of the old aristocratic and religious orders. People began

to reflect that discrimination and poverty are related to technologies that shattered social categories and weakened them.

The shattering and weakening of social categories are happening again today. Indeed, technology seems to increase discrimination against the most vulnerable, and it appears to do so because it allows for the emergence of «expansive elements of impersonal equality concerning any individual or gender difference» (Schiavone, 2019).

There would then be a little effect according to which equality would turn into a form of impersonality since it would be necessary to ignore, or even nullify, the individual characteristics that make each person unrepeatable (Ainis, 2015).

However, discriminatory cases could occur in this context of homogeneous serialisation of individuals, particularly in the automated processing of personal data. In this sense, automated decision-making algorithms find the justification of efficiency in their use precisely in serialising issues and overlapping facts. In fact, in the perspective of algorithmic automation, a kind of indifference, i.e., a pretended neutrality, on the possible origin of inequality, whether physical or moral, is envisioned because the nomenclature of choices elaborated upstream by the author of the algorithm concerns cataloguing, and thus a classification based on the processing of concepts. This procedure allows for a choice during the formation of the algorithm as to how to classify each element of the procedure based on the purposes to which the algorithm is directed, i.e., the beliefs of the programmer. Such an operation cannot be neutral but follows criteria for categorisation according to the intended purpose. Since this process is not neutral, nor can it be since it is choice-based, it contains potential discrimination.

On this issue, scholars have investigated whether and how it is possible to adopt measures of different elements related to fairness, and they are specifically examining the mathematical relationships between them (Garg et al., 2020; Pessach & Shmueli, 2020; Krawiec et al., 2023). For example, group calibration, positive and/or class balancing show that «except in narrow specific cases", no conditions can simultaneously satisfy the situations dealt with in the experiment (Kleinberg et al., 2016; Pleiss et al., 2017; Abdollahpouri et al., 2020). Other studies have addressed related issues on incompatibilities between fairness criteria, which in specific contexts (such as psychological tests applied to recidivism) can lead to significant indirect discrimination when recidivism differs among the groups examined (Chouldechova, 2017; Lippert-Rasmussen, 2022). Likewise, in criminal risk analysis, it has been observed that it is «generally impossible to maximise accuracy and fairness simultaneously, and it is impossible to satisfy all types of fairness simultaneously» (Berk et al., 2021).

Under the condition of technological transformation, as was the case in the mid-18th century, and now as well, the discrimination origin could be discerned due to the vulnerability of people involved in the automation process. Indeed, there are similarities between the early Industrial Revolution and contemporary artificial intelligence use.

In both cases, there is a subjugation of the vulnerable classes who cannot (and/or were not) escape such a technological paradigm shift. It is a condition that occurs due to belonging to the vulnerable classes, and, at the same time, it is the same vulnerability that forces such weak parties to be subaltern.

In Western Legal Tradition countries, particularly in the United States, this double connotation is most profound for those suffering from the nefarious effects related to the slave heritage, which is firmly rooted in that social reality, despite various attempts to overcome this legacy. Such endeavours have remained unsatisfactory since, in large parts of society, there is significant racial discrimination introjected by ADMs, particularly in areas such as risk assessment software.

1. Discrimination and personal data treatment

As far as privacy is concerned, it should be regarded as the manifestation of the individual's right to personal integrity, both physical and mental, against influences from third parties, whether physical subjects, legal entities, or the state itself. Privacy has become the bulwark of personality protection, both on- and offline. It is a shield against the reconstruction of individuality by third parties, public or private, as a result of the tracking of data that each person leaves behind during his/her day through geolocation carried out by the apps on his/her smartphone, payments made, and the sharing of materials and places. Privacy defends individual identity and reputation (thanks to the affirmation of the right to be forgotten, in which privacy is an element) as much in real personal life as in virtual.

In this perspective, the role played by discrimination is more insidious and difficult to demonstrate, and it concerns its manipulative aspect: from the moment that black boxes collect personal data on a massive scale, they shatter (Messinetti, 2019; Vale et al., 2022) the personality and identity of each individual by effectively annihilating them into a mass of information; by this very fact, they accomplish an operation of manipulation of individual data that, depending on the perspective of observation, entails two opposite results.

On the one hand, if each black box is formed by entering the data collected as neutrally as possible, it reflects social reality. Therefore, like a mirror, the same discriminatory processes result in reporting our society's distortions.

Instead, on the other hand, if the input of data into the black box is cleansed of discriminatory content, it could represent a distorted and manipulated image of the same reality, which returns a black box that adheres to the ideals of those who collected the data. However, this could be even more dangerous because it no longer depicts reality but a view of the same, which, in addition to being biased and unreal, is also tied to a value judgment oriented to the purposes of those who manage the black box and its results.

Therefore, the formation of the black box, which is the basis for elaborate automated decision-making processes, must be carried out with maximum transparency and attention to the objectives of the promoters. These are processes in which the multidisciplinary skills of data scientists, mathematicians, philosophers, and jurists are needed, especially those who turn to the study of comparative systems to learn and understand guidelines and results obtained in other legal systems that have dealt with the subject.

Although the protection of the right to privacy is traditionally focused on state and government authorities, the invasiveness of private entities focused on the lucrative interest provided by the commercial exploitation of their users' profiles is steadily growing (Zuboff, 2019; York, 2022) and emerges in situations that may be only seemingly unexpected. The confrontation between technological evolution and personal protection has resulted in four different mutations:

(a) on the one hand, the need to ensure the protection of the sphere of privacy, already acquired by the emancipation of the individual from central powers (Bellamy, 2014), of all citizens of industrialised societies (Rodotà, 2012), both individually and as a mass (Canetti, 1960);

(b) this need is consequent to the introduction of the electronic processor and telematic transmission of information related to the individual in the production process;

(c) the collection and storage of such data by private entities represents a paradigm shift (Kuhn, 1962) that has a legal interest (Rodotà, 1995; Alpa & Resta, 2006; Angiolini, 2020);

d) the latter activity allows for the profiling of users, which is also a harbinger of a paradigm shift related to the subject of law that is pitted against the entity, a legal person endowed with such technological power (Pellecchia, 2018).

Using computers and data telematic transmission combination has allowed for storing large amounts of data through their storage in magnetic tapes and, later on, in increasingly sophisticated digital media. With this transformation, private entities have joined the state in the massive collection of individuals' data, commonly called «big data» (Mayer-Schonberger & Cukier, 2013). Such collection has evolved by disseminating more advanced, refined, complex, and, therefore, intrusive technological tools among consumers.

It involves the collection of personal data from locations such as the Internet (the digital platforms and social networks), mobile devices (smartphones, tablets, smartwatches), satellite geolocators, radio frequency identifiers (RFID), sensors, cameras, or other tools that are capable of implementing other new emerging technologies (Gressel et al., 2020).

The right to privacy manifests itself as negative freedom, that is, in not being subjected to interference in one's private life by persons or entities, whether private or institutional. However, the right to personal data protection is embodied in the positive freedom to exercise control over the processing and to circulate of information about one's person (Rodotà, 2014).

What is legally evaluable in a «negative» sense (protection of privacy toward external interference) or positive way (protection of personal data through control over information) is that a black box takes on an indistinct connotation. Indeed, what is processed

is the information extracted from subjects during their lifetime, regardless of its relation to judgmental classifications of various conceptual natures.

Imagining being inside a black box, immersed in a virtual place where the algorithmic nature of automated decisions takes shape through matrix calculations, one realises that it is impossible to follow the logical path of the data. Such operations feed on the knowledge at their disposal (Cockburn et al., 2018; Bodei, 2019), without dealing with its provenance or origins.

Suppose it is true that this massive amount of data («big data,» precisely) represents the knowledge used by black boxes; it becomes essential to identify who has the authority to manage it and the power to dispose of it. The questions referred to can be summarised as follows:

a) Who knows what? It is a question related to the distribution of knowledge and whether barriers can legitimately be placed on access to such information sources.

b) Who decides what is possible to know? (and, therefore, how the collected data can be accessed). This question concerns the performance of the role of each subject part of the information chain: the subject «source» of the data (considered an individual even though the data collected relevant to this individual is processed massively), the institutions in charge of controlling the use of the data themselves (such as the National Privacy Authorities), as well as the operators who draw their black boxes from such data.

c) Who decides who decides? Who exercises the power of control over the sharing or subtraction of collected and knowable data?

The element that serves as the fulcrum of the balance concerning each issue concerns the effective implementation of privacy protection. For example, regarding question (a), privacy, viewed as a barrier to the intrusiveness of others by private and public administration entities into the private and personal sphere, is a limitation on the possibility of collection of discriminatory elements, and thus the continuation of discrimination itself, at the time of black box formation.

About question (b), regarding which authority can allow for the massive collection of data, the roles of responsibility and guarantee must be played, on the one hand, by the Data Protection Officers of each entity, either public or private, involved in the collection and processing of data, respectively. On the other hand, the role of guarantors of fairness is played by European and national institutions.

Finally, concerning question (c) as to who has the power to determine who the decision-makers are, it seems to be appropriate that this role should be played by the state, understood as the representative body of the consociates, subject to the rule of law and the principle of separation of powers. However, it is self-evident that the power achieved by platforms in data management represents a factual monopoly of problematic management by nation-states related to the asserted extraterritoriality of economic entities to apply national law (Tarrant & Cowen, 2022; Witt, 2022; Parona, 2021).

It is a context provoked by the interaction of subjects in a stateless society, such as the Internet. On the one hand, there is no substantial difference between social and political relations as the individual user considers him/herself as the unit of measure of his/her world. On the other hand, it dissolves in the information magma present in online life.

On the contrary, dictating the rules through their contractual conditions are non-state entities that can apply unilaterally established sanctions without counterbalance and according to their discretion. In this context, the data collection process is influenced by discriminatory elements, especially if illicitly collected, leading to biased, distorted, and thus non-neutral results, which are products of illegitimate and factually wrong decisions.

In a sense, a study of black boxes and the discrimination absorbed in them could be said to be a mirror in which reality itself is reflected. Such a mirror is helpful, especially from a legal point of view, to find tools to remedy it.

In this area, case law is given a definite role in resolving the juxtaposition between the need for public safety and privacy protection.

2. Bias and discrimination in automated decision-making systems

In computer science, «bias» refers to a length bias in bits transmission. In legal-computer science, this term refers to discriminatory situations on the part of algorithmic models, which «may lead to the detection of false positives (or negatives) and, consequently, produce discriminatory effects to the detriment of certain categories of individuals» (Parona, 2021). Such bias may depend on the set of training data, the relevance or accuracy of the data, and the types of algorithms used (Hildebrandt, 2021) concerning the fineness and speed of the results. From the perspective of social reality, bias-related discrimination may refer to unfair treatment or illegal discrimination. On this point, it is crucial to distinguish computer bias from the impact of unfair or unjust bias that consists of illegal discrimination, depending on how the collected data are formed and how they interact with each other and, simultaneously, with the surrounding reality.

It is noted in the doctrine that the interacting biases in automated decisions are mainly three (Hildebrandt, 2021):

(a) The first concerns the machine learning posed by machine learning algorithms. It is an inductive and unavoidable bias that, although neither positive nor negative in itself, cannot be considered neutral concerning the reality in which it interacts.

(b) The second concerns the ethically problematic bias because it allows for the distribution of goods, services, risks, opportunities, or access to information configured in ways that may be morally problematic. Some examples could involve excluding people, pushing them in a particular direction, or toward specific behaviours.

(c) The third type of bias is the most obvious, even to the less observant eye. It could be based on illegitimate situations or behaviours, i.e. when the machine learning algorithm focuses on individuals or categories of subjects based on illegitimate and discriminatory motives.

It has been discussed whether bias (c) may involve a subset of ethical biases (Hildebrandt, 2021). Indeed, discrimination based on gender is illegal, but not everyone considers it unethical, such as charging male drivers a higher insurance premium. After all, they are considered more reckless drivers, while female drivers appear more cautious. Such discrimination, while illegal, is not necessarily an ethical problem.

Biases (a) and (b) can relate to behaviours observable by sensors or online tracking systems or infer by the same automated algorithm. In observation, bias affects the training data, while bias affects the system's output in inference. In both cases, the output (i.e., the result) is affected, so it lacks neutrality concerning the reality to which it relates. The use of machine learning (Hildebrandt, 2021) inevitably produces bias since, as in the case of human cognition and perception are already characterised, those of machine learning are not objective, contrary to what one might think. This situation calls for caution and critical capacity, especially when the cognitive results of machine learning appear reliable (Hildebrandt, 2021).

One of the seminal texts on the subject, namely «Machine Learning» written by Tom Mitchell (Mitchell, 2007; Kubat & Kubat, 2017) makes explicit that bias, understood as variance, is necessary to demonstrate the importance and usefulness of refined tests. From these, it is possible to derive the realisation that bias, whether it consists of variance or bias, can cause errors (Hildebrandt, 2021), which can become embedded in the various stages of the decision-making process. The latter may include steps such as collecting or classifying «training data» until the goal of automated decision-making is achieved. Such errors may consist in the translation of factual contingencies into the programming language or in the circumstance that the source data are themselves incomplete (technically defined as «low hanging fruit»), however, without their incompleteness being apparent, and this may be caused by the context in which the machine learning program is operating, i.e., on a simulated or actual model (Hildebrandt, 2021).

The amount of data used by the machine learning program could cause errors since this software could process «spurious» correlations or patterns because of biases, understood in the sense of variance, inherent in the original data, precisely because of the reference to a certain idea of outcome toward which the creators of machine learning themselves were oriented.

However, a third situation could exist. It refers to an issue that is both fundamental and elusive and occurs when data is correctly assimilated by machine learning («ML»). However, it refers to real situations having distorting effects, such as following events that involved social developments that resulted in the exclusion of vulnerable groups, i.e., exclusions of subjects due to characteristics of a physical or behavioural nature.

In the case of erroneous or spurious raw materials, the collected data have ab origine flaws that can cause misinterpretation of correlations; the bias is rooted in real life, and thus data extraction will confirm or even reinforce existing biases.

This situation cannot be resolved by adopting ML, although some argue that ML can help highlight such bias or its causes. So, focusing the critical lens on what may cause such bias is necessary. Indeed, it should be undertaken to determine whether the data collection procedure helps the inclusion of bias in the raw materials or whether they occur in a noncausal distribution while maintaining great caution in working with ML tools at the risk of uncritically accepting their results (Hildebrandt, 2021).

The accuracy of machine learning results depends on the knowledge on which it works and interacts, despite the limitation (Marcus, & Davis, 2019; Brooks, 2017; Sunstein, 2019) and the models used.

The issue concerns the ability to understand the difference between cognitive biases present in humans (whose intelligence can adapt and make abductive and unexpected reasoning) and machine learning, whose intelligence instead always depends on the data, assumptions, and characteristics of ML feeding. The ML can process inductive and deductive inferences but not abductive ones. Indeed, abductive reasoning demands an «intuitive leap» that starts from a set of elements and, from these, elaborates an explanatory theory of these elements, which the available data must verify (Hildebrandt, 2021). To test such hypotheses, programming the machine learning model is concerned with the creative ability to recompose the abductive step to test such hypotheses inductively. If such an operational hypothesis were confirmed, the system could use the abductive method as the basis for deductive reasoning. In this regard, the experiential feedback (feedback) of machine learning is decisive as it is fundamental and crucial (Hildebrandt, 2021).

It follows from this that the quality of the samples themselves reflects the quality of the training of databases. If the user provides the system with data (or samples) that are distorted or characterised by poor quality, the behaviour of the machine learning algorithm remains negatively affected by this, producing poor-quality results (Gallese, 2022). In addition, it should be remembered that machine learning algorithms tend to lose the ability to generalise and are thus prone to exaggeration (Gallese, 2022).

This circumstance is well-known to programmers, but jurists elude it. It is identified by observing the relationship between the training and test errors. It occurs when the improvement on the error related to the training set causes a worse error on the test data set. In this case, the network on which machine learning rests proves to be «overfitting» (Gallese, 2022). It happens when the model fits the observed data (the sample) because it has too many parameters relative to the number of observations and thus loses touch with the reality of the data. From this, it can be understood that human input in the preliminary decisions on data collection, classification, and processing is inescapable and essential for obtaining a reasonable and acceptable result.

In other words: if the machine learning algorithm produces erroneous, discriminatory, or wrong results, the responsibility lies with those who organised the dataset and set up the algorithm. Machine learning obeys instructions that it merely executes.

In these situations, it could occur or be exacerbated even in the case of continuous feeding with new categories of data, leading to the system's imbalance, so some categories are more represented than others, with a significant influence on the impact on the future behaviour of that AI (Gallese, 2022). However, there is a situation in which the problem remains unsolvable: this is the case of so-called generalised class incremental learning (Gallese et al., 2020) in which the machine learning method receives new data that may, in principle, belong to new classes or cases never considered before. In this peculiar situation, the algorithm must be able to reconfigure its inner workings (e.g., in the case of the deep learning machine, where the algorithm must adapt the architecture and recalibrate all parameters) (Gallese et al., 2020). It would prevent any realistic possibility of predicting the future behaviour of the automated system.

3. The twofold nature of risk assessment algorithms

Automated decision-making algorithms highlight a relevant aspect of civil coexistence that is changing in its course and, thus, in its nature. It is the relationship between authorities (whether public or private, however able to influence people's individual and collective lives) and consociates (the people who, consciously or unconsciously, find themselves subject to or the source of approaching personal data).

Every stage or element of life (daily and in the entire existential journey) has become the object of automated collection, profiling, and decision-making. Biometric recognition algorithms investigate our identity and feelings through so-called. «affective computer learning» (Guo et al., 2020), i. e., using data on the most obvious physical characteristics (such as eye colour, complexion, hair colour) or the collection of fingerprints contained in the chips of identity identification documents (such as passports), or through gait recognition, or images capturing expression during a leisurely walk or a work activity (Crawford, 2021) or a run to catch the last subway under the prying eye of a surveillance camera.

These informational data make it possible to answer interesting questions like, who are you? or where are you going? Nevertheless, one may wonder to what extent the collection and processing of data in response to such questions are legitimate.

To develop answers to such questions, it seems helpful to examine risk assessment algorithms. They present fascinating aspects from an evolutionary point of view, as they were the first programs to be used for predicting recidivism, initially in probation and parole (Oswald et al., 2018). As pointed out in the doctrine, using predictive analytics algorithms in a judicial context can satisfy three needs, two of which are more general, especially on a cost-benefit assessment and access to public safety and legal protection resources level, and one placed on an individual level (Oswald et al., 2018).

From a justice administration efficiency perspective, specifically, the use of such software could have certain advantages, such as:

(a) facilitating overall strategic planning of forecasts and priorities in combating criminal activities;

(b) evaluating profiles of specific activities related to crime reduction;

(c) assessing the prediction of recidivism in the case of individuals.

On the latter aspect, the application of risk assessment algorithms concerns the need to balance necessity (pursued primarily by the government) with the principle of proportionality in light of respect for human rights so that the protection of the rights of the individual can be fairly balanced against the needs of the community (Oswald et al., 2018).

From a legal perspective, the major critical issues affecting this approach concern, first of all:

1. opacity and secrecy, linked first and foremost with the intellectual property protections of the algorithm itself (Oswald et al., 2018), but questionable, as contrary to the principle of transparency, especially in the case of the use of ADMs in the judicial sphere, such as the calculation of criminal recidivism;

2. the use of such automatisms does not meet the criteria of protection of principles of constitutional relevance, such as the principle of the natural judge predefined by law, the right to be heard in court and to a fair trial, the motivation of the decision (contained in Article 111 of the Italian Constitution and Article 6 of the European Convention for the Protection of Human Rights and Fundamental Freedoms) as well as, evidently, with the provision of Article 22 GDPR, which guarantees the right to the explanation of the automated decision.

On the subject of limitations on freedom, both the decision flow adopted by the algorithm and the database on which it bases its decision-making process must be transparent and accessible and the inadequacy of even one of the aforementioned constitutes a violation of the principles of due process under Articles 111 Const and 6 ECHR.

Automated decisions with inadequate results obtained through procedures that do not adhere to the principles related to respect for the protection of personal data and fundamental rights must be able to be challenged before the ordinary courts in order to obtain intelligible and adequate decisions to be understood by the person concerned. It is the right of the person subjected to the automated decision to be informed about the correctness of the decision that affects him or her, both in a formal sense (i. e., compliance with the safeguards for his or her protection prepared by the multilevel legislative framework) and in a substantive sense (i.e., why the matter was decided in the dispositive sense and how the conclusions were reached in the automatically decided matter). The logical path should not give rise to essential doubt as to whether the decision maker, in this case, the risk assessment algorithm, erred in law in making a rational decision based on relevant grounds (Oswald et al., 2018).

Is it possible to argue that ADMs require a higher decision-making standard than a human decision-maker? (Oswald et al., 2018). For example, in the legal sphere, a judicial decision

issued by a judge without a statement of reason is considered nonexistent[1], i.e., null and void[2], both under Article 132, No. 5 of the Code of Criminal Procedure, 546 of the Code of Criminal Procedure and Article 111 of the Constitution (Chizzini, 1998; Taruffo, 1975; Massa, 1990; Bargi, 1997). A judicial decision must be reasoned with a statement of the relevant facts of the case and the legal reasons for the decision (Taruffo, 1975).

However, it is noted that historical and comparative experience shows that per se, the aforementioned obligation to give reasons in fact and law is not an indispensable element of the exercise of the judicial function as both the experience of popular juries, judges of fact (Chizzini, 1998; Taruffo, 1975; Massa, 1990; Bargi, 1997), and the finding that in various systems, an obligation to give reasons is absent (Chizzini, 1998). The obligation to state reasons arises from the Jacobean Constitutions as an endoprocessual function relating to the need for control over the exercise of judicial power, consistent with the delineation of the nomofilactic role of the Supreme Court in the implementation of the principle of subordination of the judge to the law (Taruffo, 1975).

These words bring to mind the ancient Montesquiean precept that wants the judge bouche de la loi instead of bouche du roi (Petronio, 2020). In the case of the jurisprudential use of risk assessment algorithms, who fills the shoes of the «roi»? Especially in light of sensitivity, i.e., the ease with which the results of recidivism proceedings conducted through risk assessment algorithms can be manipulated. One may wonder if the algorithm cannot become the bouche de la loi. Scholars affirm that:

«to ensure that the judge, even the supreme judge, does not go his or her own way but complies with the law, which is the same for all even if it is to be applied on a case-by-case basis taking into account the multiplicity of cases to be judged, it becomes necessary for the judgment to be reasoned, that is, to give an account of why that particular solution» (Petronio, 2020).

The statement of reasons represents the manifestation of the competence, independence, and responsibility of the decisional measure issued by the judge; it is the instrument of the legal protection of the parties or the defendant against the presence of hidden factors such as bias, error, and irrationality. Against such critical issues concerning automated decision-makers, it is not possible to carry out a similar transparency operation, given the established opacity of the decisional phase.

Although in the promoters' intentions, the use of algorithms is justified by the replacement of evaluation systems in the area of bail calculation and bail granting (Israni, 2017), in reality, algorithms incorporate commonly shared stereotypes, particularly on ethnic, economic, and sexual grounds (Starr, 2014), as well as presenting ethical and constitutionality issues.

The Wisconsin Supreme Court dealt with using an algorithm developed to assist the work of judges in matters of probation and the risk of recidivism of arrested persons. The case

1 Cass., 19-7-1993, n. 8055, GC, 1993, I, 2924; Cass., 8-10-1985, n. 4881, NGL, 1986, 254.

2 Cass., 27-11-1997, n. 11975.

can be summarised as follows: in February 2013, Eric Loomis was arrested while driving a car used in a shooting. Shortly after his arrest, he pleaded guilty to contempt of court and did not contest the fact that he had taken possession of a vehicle without the owner's consent. As a result, the defendant was sentenced to six years in prison. This decision was worthy of attention because the district court used a proprietary algorithm developed by Northpointe, Inc1 in the decision-making phase of a fourth-generation intelligent risk assessment software called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). This algorithm was believed to have predicted a person's risk of recidivism based on a complex analysis involving information gathered from a survey of 137 questions divided into several sections and information corresponding to individual public criminal records (Rebitschek et al., 2021; Bao et al., 2021; Wang et al., 2022).

The Wisconsin Supreme Court held that such a tool did not present constitutionality problems about the defendant's due process if the software processed individual cases using accurate information (Israni, 2017)[3]. It should be noted that the manufacturer of the COMPAS software refused to disclose at trial the methodology and guidelines used by the software for its decisions (Custers, 2022), even though the risk assessment score developed by the algorithm was cited in the judgment, since the algorithm believed the defendant was at high risk of reoffending, the court denied him parole and handed down a six-year sentence (Israni, 2017).

The Wisconsin Supreme Court rejected the constitutionality concerns related to the violation of the defendant's due process, especially concerning the lack of transparency and accuracy of the algorithm's formulation, which would prevent the defendant from being confident of the impartiality of the decision-making process to which he is subjected.

Such verification is prevented (a usual circumstance in cases of judicial requests for access to decision-making algorithms) by the rules of intellectual property and trade secret protection as an expression of a form of the inscrutability of the algorithm (Vogel, 2020) that becomes absolute sovereign: the concretisation of the voix du roi, a king anointed by the oil of intellectual property instead of divine sanctity.

Nevertheless, in this context, intellectual property rules protect the economic interests of those who develop machine learning technology. In balancing the rights of asks to access the decisional mechanism, particularly the rights of the defence (and in certain decisional contexts, claims to due process) protected by different multilevel sources, from the Constitution to international conventions, which ones prevail?

North American jurisprudence justifies continued opacity, confirming the constitutionality of using COMPAS, despite the Wisconsin Supreme Court placing numerous restrictions on its use. The algorithm cannot be used to determine whether a detainee should be incarcerated, i.e., to calculate the length of his or her detention (Israni, 2017).

[3] State v. Loomis, 881 N.W.2d 749, 767 (Wis. 2016)

The COMPAS algorithm must be justified at sentencing and for any scoring processing on recidivism prediction, always with the caveat regarding its limited decisional utility. Because the federal Supreme Court declined to issue a writ of certiorari, the Wisconsin Supreme Court's decision is final[4].

It is the questionable factual circumstance that the Wisconsin Supreme Court has allowed an algorithm, about which there are ethical and constitutional doubts related to its non-transparency, to be placed alongside ordinary judges in exercising jurisdiction. In the U.S. court system, the protection of confidentiality in patent matters would be prioritised to maintain the patent owner's competitive advantage over individual due process rights and freedoms. According to the doctrine, this is a misunderstanding that would be difficult to overcome without the intervention of the federal Supreme Court (Israni, 2017).

Addressing the application of no less than five psychological and actuarial risk assessment tools, the Supreme Court of Canada offered a reversal of the rationale regarding whether these decision-making mechanisms can be used in the assessment of the level of risk of recidivism presented by native offenders, the case is, the Ewert v Canada decision[5]. In this case, the Supreme Court of Canada ruled that the Correctional Service of Canada (CSC) breached its statutory duty under Section 24 of the Corrections and Conditional Release Act (CCRA) for the exclusive use of accurate information in risk assessments (Scassa, 2021).

According to the Canadian Court, the Correctional Service of Canada has to account for the systemic discrimination indigenous peoples face in the criminal justice system, in general, and in prisons. However, the SCC found that despite using tools that may be discriminatory against indigenous individuals, there was no violation of the Canadian Charter of Rights and Freedoms (Russell, 1983; Epp, 1996).

This decision drew several critical comments given that it is primarily Native people, particularly women, who suffer systemic discrimination because of how the prison system is organised: they experience detention to a greater extent and for longer periods. They do not enjoy culturally appropriate programs or forms of rehabilitation that could bring Native individuals back into their communities where they might have greater support.

While the Canadian Court has recognised such discrimination, motivated by the fact that «identical treatment can lead to serious inequalities»,[6] it has been pointed out that there is a significant connection between the risk assessment tools used in prison and prisoners' freedom. An inmate's excessive risk rating significantly impacts the individual's freedom: he or she is placed in more restrictive environments, and the chances of early release are reduced. Despite all this, the Court decided that neither the person's right to life,

4 Loomis v. Wisconsin, 137 S. Ct. 2290 (2017)

5 Ewert v Canada [2018 SCC 30]. https://canliiconnects.org/en/commentaries/62360

6 *Ibid*.

liberty, and security (s. 7 of the Charter) nor the equality provisions (s. 15) had been violated. According to the Court, there was no evidence that such discrimination was absorbed by the software risk assessment algorithms and, thus, discriminated against the plaintiff.

The Supreme Court of Canada acknowledged that risk classification tools to determine prisoner recidivism are inaccurate and provided the appellant with a statement. However, it is troubling that despite a lengthy analysis of how inaccurate they are and how this may impact indigenous individuals, the Court chose not to declare them unconstitutional. It is an even more harmful and insidious decision than that of the Wisconsin Supreme Court, which,

in any case, had placed procedural limits on the use of such software. For example, some possible expedients would allow the algorithm to limit the economic-ethnic-social influence regarding the entry of the personal data of the individual being screened by the program. Furthermore, it would be possible to omit the entry of the «ZIP code», i.e., data related to the defendant's residence, since these show the potential income ranges of areas. Based on the principle of presumption of innocence, this data is irrelevant for calculating the possible risk of his recidivism, the objective of using such software.

4. The Artificial Intelligence Proposal

The Proposal for a regulation published by the European Commission on 21 April 2021 represents the first completed attempt to regulate AI in general terms, despite the context in which such a regulation would have to be regulated: on the one hand, the fragmentary nature or lightness of the regulation (especially in the United States and China, the two leading global competitors in this field), and on the other, the difficulty of foreseeing and balancing the development of a sector that is a harbinger of interests that may even be divergent ones, and subject to very rapid developments (Rosa, 2021; Alpa, 2021; Scherer, 2015). The regulation framing the development of AI must ensure the protection of fundamental rights and the rule of law while being flexible enough to adapt to technological changes that are yet to be foreseeable. In other words, the regulation is required to 'square the circle' at the national or European level and serve as a model at the global level.

The critical points concern the balance between the uniformity of the discipline and its updating, given the not remote possibility of its rapid obsolescence, due to the autonomous developments of black boxes, machine learning, deep learning and neural networks, which in turn represent a source of risks that cannot be foreseen ex-ante, in contradiction with one of the fundamental principles of the rule of law, the provision of general and abstract pro-future regulations (Scherer, 2015).

In any case, one observes the acknowledgement by the drafters of the text of the known jurisprudential experience on the subject, particularly concerning the recognition of the subordinate role of the weaker user to the role of the platforms. However, it shows a growing understanding of the discriminatory phenomena linked to automated algorithms.

The European Commission has considered such complexities related to potentially present risk factors that cannot be predicted a priori by introducing two flexibility mechanisms in the regulatory framework. This strategy is developed in three main points:

(a) The Artificial Intelligence Act proposal is accompanied by several annexes (Annexes) that form an integral part of it, characterising its discipline. These annexes, for instance, outline the categories of high-risk devices (high-risk AI systems) for which the legislation is detailed and a specific compliance procedure is envisioned. Such annexes are so relevant for the European Commission's regulatory approach to AI that the procedure for their amendment adheres to Article 290 TFEU, allowing technical regulatory standards to be approved (Battini, 2018). In order that appropriate and timely solutions can be found to the application issues related to the use of high-risk systems, according to Article 74 of the AIA, a committee is planned to intervene through amendments to the regulation, even outside the ordinary regulatory procedure required for the formal revision of the regulation (Casonato et al., 2021; Veale et al., 2021; Stuurman et al., 2022);

b) The AIA proposal itself envisions a general obligation of a five-yearly review, the first within five years of its entry into force, precisely because of the instability of the subject matter, with the consequent regulatory and legal adaptation to the evolutions it has achieved.

c) In order for the review mechanism of the regulation in question to be as thoughtful as possible, the Proposal provides in Title V for the implementation of the mechanism of so-called 'sandboxes', i.e., functional spaces set up by the Member States, for a limited period, and under the control of the national authorities, where it is possible to experiment and test innovative artificial intelligence systems with a view to their introduction on the market.

Given the combination of complexity and opaqueness in the mode of operation, significant unpredictability, and autonomy in forming results based on input data, the need to regulate artificial intelligence has become almost imperative. Such regulation must address security risks and guarantee and reinforce the protection of fundamental rights against legal uncertainty to stem the fragmentation of regulation and distrust in both the tool and the human ability to control it.

The proposed AIA regulation is part of the Union's strategy to strengthen the single digital market, which uses harmonised rules. In this case, these rules are intended to avoid fragmentation of the internal market on the essential elements concerning the requirements of products using automated algorithms to avoid legal uncertainty for both providers of such services and users of automated decision-making systems. Indeed, from the perspective of subsidiarity, if the principle of non-exclusive competence were to be strictly integrated, given that different big data sets may be incorporated in each product comprising automated systems, in this sense, a national-only approach is a harbinger of more significant, contradictory regulatory hurdles and uncertainties that would impede the circulation of goods and services, including those using automated decision-making systems.

In this sense, the AIA proposal aims to develop a legal framework adhering to the principle of proportionality that achieves its objectives by following a risk-based approach, imposing burdens only when artificial intelligence systems present high risks, i.e., outweighing the benefits, for the protection of fundamental rights and security. In order to verify such a risk and consider AI systems as not high risk, they must meet specific requirements: the data used must meet high quality, documentation, transparency, and traceability criteria.

In this regard, the instrument of the regulation was chosen because, under Article 288 TFEU, the direct applicability of regulation will reduce legal fragmentation and facilitate the development of a single market in legal, safe and reliable AI systems by introducing all EU member states a harmonised set of basic requirements for AI systems classified as high-risk and obligations for providers and users of such systems, improving the protection of fundamental rights and providing legal certainty for operators and consumers.

Regarding the protection of fundamental rights, the AIA proposal imposes certain restrictions on the freedom to conduct business and the freedom of art and science to ensure that overriding reasons of public interest, such as health, safety, consumer protection and the protection of other fundamental rights are respected when high-risk AI technology is developed and used. Such restrictions are proportionate and limited to the minimum necessary to prevent and mitigate severe risks and likely violations of fundamental rights. The use of AI with its specific characteristics (opacity, complexity, data dependency, autonomous behaviour) may adversely affect several fundamental rights enshrined in the EU Charter of Fundamental Rights. The obligation of ex-ante testing, risk management and human oversight will also facilitate the respect of other fundamental rights by minimising the risk of erroneous or biased AI-assisted decisions in critical areas such as education and training, employment, legal services, judiciary, health and welfare services.

It should be emphasised that, to the benefit of significant investments in funds, know-how and research, transparency obligations will not disproportionately affect the rights to protect intellectual property, know-how and trade secrets, and confidential information. However, this fact risks thwarting the objective of making the automated decision-making system transparent and trustworthy, as it would prevent, if not appropriately balanced, the disclosure of the way the data is processed and thus the source of possible discrimination, as happened in the litigation concerning the protection of crowd workers employed through job brokerage platforms.

As anticipated, the Proposal's approach is risk-based: it classifies artificial intelligence models as follows:

a) prohibited artificial intelligence practices insofar as they are oriented towards manipulating the conduct of individuals based on subliminal techniques or exploiting

the vulnerabilities of individuals on account of age or disability in order to influence their conduct. Also prohibited in principle are AI systems used by public authorities to establish the trustworthiness of individuals (i.e. "social scoring") (Maamar, 2018; Infantino et al., 2021) based on their social conduct and personal characteristics. However, this prohibition would seem to have taken on broad fears or apprehension from other legal systems, such as that of China, as such use of 'social scoring' models is already prohibited in Europe as violating dignity and equality.

Similarly, biometric recognition tools are prohibited, subject to a broad exception relating to their necessity for the targeted search of potential victims of criminal acts (e.g. missing children) or the prevention of a specific, substantial and imminent danger to a person's safety or of a terrorist attack, or for the detection, location, or prosecution of a person suspected of offences under Article 2(2) of Council Framework Decision 2002/584 for which the Member State concerned provides for a custodial sentence of three years or more. On this point, it is noted that these are offences for which a European arrest warrant will be issued. In any event, it is noted that the Proposal for regulation does not explicitly state anything about the possible use of such recognition systems by private entities.

b) high risk: these models' use may be permitted but must be subject to prior verification of precise requirements for protecting human dignity and respect for fundamental rights. The identification of this category is based on both the function attributed to the device and its overall purpose, including its specific purposes. For this assessment, a further evaluation of compliance with both the relevant legislation and the protection of fundamental rights is required. They cover an extensive range of tools (used in the areas specified in Annex III) and concern models used in job recruitment or diagnostic medical devices, models for biometric identification of individuals, for infrastructure management (both those used in so-called 'smart cities', such as intelligent traffic lights, and those used in the management of service supplies such as water, gas and electricity supply), education or staff training purposes, and so on.

The classification of an AI system as high-risk is based on the intended purposes of the AI system, in line with current product safety legislation. Thus, the high-risk classification depends not only on the function performed by the IA system but also on the specific purpose and manner in which the system is used. This classification has identified two categories of high-risk systems: (a) IA systems intended for use as safety components of products subject to ex-ante conformity assessment by third parties; (b) other stand-alone IA systems with implications primarily for fundamental rights that are explicitly listed in Annex II;

c) AI models with minimal or no risk are those that, although they maintain specific transparency requirements, are identified by subtraction from the previous categories, such as chatbots, whose use may be permitted but subject to information and transparency requirements on their nature, or if/then models.

Title I of the AIA proposal defines the subject matter and scope of the new rules governing the placing on the market, commissioning and use of AI systems. It also outlines the definitions used throughout the instrument, particularly under Art. 3(1) of the Proposal, an 'artificial intelligence system' (AI system) is defined as 'software developed using one or more of the techniques and approaches listed in Annex I, which can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions that influence the environments with which it interacts'. This definition of the AI system in the legal framework aims to be as technologically neutral and 'future-proof' as possible, considering the rapid technological and market developments related to AI.

In order to provide the necessary legal certainty, Title I is complemented by Annex I, which contains a detailed list of approaches and techniques for AI development to be adapted to the new technological scenario.

The key participants along the AI production and value chain are also clearly defined as suppliers and users of AI systems covering public and private operators to ensure a level playing field. They can be summarised as

(a) machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods, including deep learning;

(b) logic and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) statistical approaches, Bayesian estimation, search methods and optimisation.

Further, in light of Recital 6, an AI system should be clearly defined to ensure legal certainty while providing the flexibility to accommodate future technological developments. The definition should be based on the essential functional characteristics of software. In particular, the ability to generate outputs such as content, predictions, recommendations or decisions that influence the environment with which the system interacts. AI systems can be designed to operate with varying degrees of autonomy and be used on their own or as product components, regardless of whether the system is physically integrated into the product (embedded) or serves the product's functionality without being integrated into it (non-embedded). Furthermore, the definition of an AI system should be complemented by a list of specific techniques and approaches used in its development, which should be kept up-to-date in light of technological and market updates through the adoption of delegated acts by the Commission to amend this list.

So far, the impression received from the Artificial Intelligence Act Proposal is that it is an attempt at definitions and classifications, while what is worrying from an anti-discrimination protection perspective is the lack of remedial mechanisms to protect individuals subjected to discriminatory automated decisions especially in light of what should have been the premise of the (new) body of law, whose purpose was to put the person at the centre.

The exclusive reference to the discipline of Art. 22 GDPR is not sufficient since it does not cover the extensive areas of artificial intelligence, but focuses mainly on automated decisions productive of legal effects, which affect individual rights, unless such a decision, including profiling, is necessary for a contractual context or is authorised by a law of the European Union or a Member State. In this regard, it has been argued in the doctrine that decisions under Article 22 GDPR cannot be based on special categories of personal data, such as biometric data, t unless the data subject allows it (ex Art. 9(2)(a) GDPR) or if there are reasons of public interest (ex. Art. 9(2)(g)). Nevertheless, these exceptions must provide a clear legal framework for protecting fundamental rights (Martini, 2020), or concerning investments in this area, despite a possible link to Article 5 of the AIA proposal.

In this context, Article 22 GDPR manifests its inadequacy to make up for the absence of a procedure appropriate to the rationale and purposes of the AIA Proposal. This absence is even more pronounced when one considers that the AIA Proposal defines which AI systems are 'high-risk' but does not provide any remedy (nor does it strengthen Article 22 GDPR for this purpose) in order to make effective the protection of those affected against discrimination produced by a risk of harm to security, health or an adverse impact on fundamental rights.

One might wonder whether Article 13 of the Proposal (under the heading 'Transparency and provision of information to users') could be supplementary to Article 22 GDPR in the framework prepared by the AIA Proposal. The letter of Article 13 denies such a purpose, as the article would manifest a mere proclamation of good intentions. Indeed, it is unlikely that the average user will be able to interact with an algorithm or understand the mechanism for achieving its results, however transparent it may be, although risk AI system must be accompanied by 'instructions for use in an appropriate digital or non-digital format; though such must include concise, complete, correct and precise information that is relevant, accessible and understandable to users'[7].

Article 13 of the AI proposal explicitly provides for information's correctness (accuracy) but does not provide truthfulness (trustworthiness). The two terms are not perfect synonyms in Italian or English. The former refers to the accuracy of information details, which should coincide with their truthfulness, but not necessarily. Therefore, in areas where the two concepts do not adhere, it is possible for discrimination or inappropriate automated data processing to occur without the provision of sanctions or penalties.

The Draft text of the Recommendation takes up this concern on the Ethics of Artificial Intelligence proposed by UNESCO[8]. This draft, however, delegates the provision of remedies against discriminatory treatment to AI operators, who 'must make every reasonable effort

7 Art. 13 AI Proposal.

8 UNESCO. *Draft Recommendation on the Ethics of Artificial Intelligence*. UNESCO General Conference, Paris.

to minimise and avoid reinforcing or perpetuating discriminatory or biased applications and results throughout the life cycle of the AI system, to ensure the fairness of such systems'[9].

Effective remedies against discrimination and algorithmic determination biased by negative bias should be available, and the European legislator or each member state should arrange this task. Leaving the remedy instrument in the hands of the individual operator would result in fragmentation contrary to the minimisation of discrimination in the results produced by the algorithmic determinants.

The prescriptive nature of the contents and requirements relating to such information makes it possible to exclude that Article 13 of the AI Proposal can play a remedial role. It merely establishes content and information-related prescriptions.

The amendments made to the Artificial Intelligence Act Proposal are attractive because they manifest their contradictory nature due to the ambition to adopt 'universal' impact legislation that the Proposal (and its promoters) claim. Indeed, as with the GDPR, the Proposal aims to candidate itself as a model for AI regulation in other legal systems. However, one may wonder whether such a complex and disorienting model, a fact due to the coexistence of limitations and exceptions, can be adopted in systems where decision-making automation is used on the administrative and public side mainly to make bureaucracy and obedience to the order more efficient, as well as to strengthen defensive or offensive military apparatuses, while on the private side to improve cost-benefit analysis through automation of the production chain.

On the other hand, the amendments manifest a fundamental contradiction with the concept of AI present in the European Parliament and, by extension, in public opinion or the electorate. The Rapporteur to the European Parliament recalled that artificial intelligence systems are based on software using mathematical models of a probabilistic nature as well as algorithmic predictions for several specific purposes. On the contrary, 'Artificial Intelligence' is a generic term, covering a wide range of old and new technologies, techniques and approaches, better understood as 'artificial intelligence systems', which refer to any machine-based system and which often have little more in common than the fact that a particular set of human-defined goals guides them, that they have some, varying degrees of autonomy in their actions. They engage in predictions, recommendations or decisions based on available data. The development of such technologies is not homogeneous; some are widely used, while others are under development or consist only of speculation that has yet to find design and concreteness[10].

---

9 *Ibid*.

10 Voss, Axel. (2022, April 20). *Draft Report on Artificial Intelligence in a digital age (2020/2266(INI)) Compromise Amendments European Parliament; Draft European Parliament Legislative Resolution on the Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD))*.

The opposing political positions are confirmed by the Amendments proposed by the joint LIBE (European Parliament's Committee on Civil Liberties, Justice and Home Affairs) and IMCO (Internal Market and Consumer Protection) Committees[11] concerning the definition of Artificial Intelligence. In fact, on the one hand, a definition of AI is presented that is as broad as possible and overrides the technical classifications set out in Annex I of the Proposal, while on the other hand, insisting on precise definitions, including machine learning (defined as the ability to find patterns without being explicitly programmed for specific tasks).

A new criterion for classifying high-risk systems has been introduced, changing the automatic classification for systems in the list of areas mentioned in Annex III to a list of 'critical use cases'. Based on these uses, AI providers must self-assess whether their systems present significant risks to health, safety and fundamental rights. During the approval of these amendments, political forces clashed over the use of algorithms used to evaluate creditworthiness, health insurance processes, payments and media recommendation systems.

It seems relevant to recognise, outside of expressions of intent or ideological statements, that structural biases in society should be avoided or even increased through low-quality data sets. It is stated explicitly that algorithms learn to be as discriminatory as the data they work with. As a result of low-quality training data or biases and discriminations observed in society, they may suggest inherently discriminatory decisions, which exacerbates discrimination in society. It is noted, however, that AI biases can sometimes be corrected. Furthermore, it is necessary to apply technical means and establish different levels of control over the software of AI systems, the algorithms and the data they use and produce to minimise risk. Finally, it claims, probably deluding itself, that AI can and should be used to reduce prejudice and discrimination. In reality, AI tends to amplify discrimination in society.

The Artificial Intelligence Act would make it possible to go beyond the 'case studies' model for risk assessment algorithms, which instead relies mainly on the above-mentioned case studies, focusing on individual and not collective risks, while the human rights impact of risk assessment algorithms is crucial. Given the sectorial nature of each subject, it is an approach that has yet to be absorbed by those involved in creating or formulating the algorithms. However, it is a fundamental step from a purely cultural and empirical point of view because it would allow for a change of anti-discriminatory perspective in the use of data.

[11] Benifei, Brando, Tudorache, Ioan-Dragoş. *DRAFT REPORT on the Proposal for a regulation of the European Parliament and the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD))*. Committee on the Internal Market and Consumer Protection Committee on Civil Liberties, Justice and Home Affairs.

Equally relevant is the issue of the prohibition of the use of technology, the content of which is extensive. On this point, the Proposal for a regulation contains similarly broad exceptions: it is an indisputable fact of legal significance that the entire Proposal is built around the classification of types of algorithms, with the prohibition of the use of social scoring algorithms or particular biometric software at the heart of the framework. Nonetheless – especially for biometrics – such broad exceptions leave room for contradictions. However, the rationale conveyed by Proposal is that, at least in the European Union member states, there are limits related to the protection of dignity and fundamental rights that cannot be exceeded in using artificial intelligence algorithms.

In light of this, one can also ask oneself whether introducing so-called 'sandboxes' might be an appropriate solution to balance the speed of technological development with the need to protect the human rights, especially from an anti-discrimination perspective, of those subjected to algorithmic decision-making. The answer cannot be immediate and is probably not satisfactory. Indeed, a sandbox is a functional space limited in time to experiment with AI systems, especially the decisive ones, with a view to their release on the market. Their goal is to evaluate the impact of ADMs on individuals and the social fabric. However, their effect is only sometimes immediate but can only be seen after their medium- to long-term operation.

The filing of thousands of amendments manifests a profound criticism of the political forces in the European Parliament against the Commission that promoted such a proposal, which makes the approval process or even the final approval of this Proposal for a regulation complex.

Conclusion

Advance of technologies had led not only to a progress and improves our life. It also brought threats to human rights related to the violation of privacy as well as discrimination. Discriminatory cases often occur in the automated processing of personal data. Since this process is not neutral, nor can it be since it is choice-based, it contains potential discrimination. It is important today to investigate whether and how it is possible to adopt measures of different elements related to fairness of algorithms.

The results made in this paper can be used as basis for future research in the sphere of algorithmic discrimination and privacy protection. It can also be used in law-making process.

References

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020). The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems* (pp. 726–731). https://doi.org/10.1145/3383313.3418487

Ainis, M. (2015). *La piccola eguaglianza.* Einaudi.

Alpa, G. (2021). Quale modello normativo europeo per l'intelligenza artificiale? *Contratto e impresa*, *37*(4), 1003–1026.

Alpa, G., & Resta, G. (2006). *Trattato di diritto civile. Le persone e la famiglia: 1. Le persone fisiche ei diritti della personalità*. UTET giuridica.
Altenried, M. (2020). The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. *Capital & Class*, *44*(2), 145–158. https://doi.org/10.1177/0309816819899410
Amodio, E. (1970). *L'obbligo costituzionale di motivare e l'istituto della giuria. Rivista di diritto processuale.*
Angiolini, C. S. A. (2020). *Lo statuto dei dati personali: uno studio a partire dalla nozione di bene*. Giappichelli.
Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., ... & Venkatasubramanian, S. (2021). It's complicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*
Bargi, A. (1997). *Sulla struttura normativa della motivazione e sul suo controllo in Cassazione.* Giur. it.
Battini, S. (2018). *Indipendenza e amministrazione fra diritto interno ed europeo*.
Bellamy, R. (2014). Citizenship: Historical development of. Citizenship: Historical Development of'. In J. Wright (Ed.), *International Encyclopaedia of Social and Behavioural Sciences, Elsevier*. https://doi.org/10.1016/b978-0-08-097086-8.62078-0
Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533
Brooks, R. (2017). Machine Learning Explained. *Robots, AI and other stuff*.
Bodei, R. (2019). *Dominio e sottomissione*. Bologna, Il Mulino.
Canetti, E. (1960). *Masse und Macht*. Hamburg, Claassen.
Casonato, C., & Marchetti, B. (2021). Prime osservazioni sulla proposta di regolamento dell'Unione Europea in materia di intelligenza artificiale. *BioLaw Journal-Rivista di BioDiritto*, *3*, 415–437.
Chizzini, A. (1998). *Sentenza nel diritto processuale civile*. Dig. disc. priv., Sez. civ.
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047
Citino, Y. (2022). *Cittadinanza digitale a punti e social scoring: le pratiche scorrette nell'era dell'intelligenza artificiale*. Diritti comparati.
Claeys, G. (2018). *Marx and Marxism*. Nation Books, New York.
Cockburn, I. M., Henderson, R., & Stern, S. (2018). The impact of artificial intelligence on innovation: An exploratory analysis. In *The economics of artificial intelligence: An agenda*. University of Chicago Press.
Cossette-Lefebvre, H., & Maclure, J. (2022). AI's fairness problem: understanding wrongful discrimination in the context of automated decision-making. *AI and Ethics*, *5*, 1–15. https://doi.org/10.1007/s43681-022-00233-w
Crawford, K. (2021). Time to regulate AI that interprets human emotions. *Nature*, *592*(7853), 167. https://doi.org/10.1038/d41586-021-00868-5
Custers, B. (2022). AI in Criminal Law: An Overview of AI Applications in Substantive and Procedural Criminal Law. In B. H. M. Custers, & E. Fosch Villaronga (Eds.), *Law and Artificial Intelligence* (pp. 205–223). Heidelberg: Springer. http://dx.doi.org/10.2139/ssrn.4331759
De Gregorio, G. & Paolucci F. (2022). *Dati personali e AI Act. Media laws*.
Di Rosa, G. (2021). Quali regole per i sistemi automatizzati "intelligenti"?. *Rivista di diritto civile*, *67*(5), 823–853.
Epp, C. R. (1996). Do bills of rights matter? The Canadian Charter of Rights and Freedoms, *American Political Science Review*, *90*(4), 765–779.
Fanchiotti, V. (1995). *Processo penale nei paesi di Common Law.* Dig. Disc. Pen.
Freeman, C., Louçã, F., & Louçã, F. (2001). *As time goes by: from the industrial revolutions to the information revolution*. Oxford University Press.
Freeman, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. *North Carolina Journal of Law & Technology*, *18*(5), 75–90.
Fuchs, C. (2014). *Digital Labour and Karl Marx*. Routledge.
Gallese, C.(2022). *Legal aspects of the use of continuous-learning models in Telemedicine*. JURISIN.
Gallese, E. Falletti, M. S. Nobile, L. Ferrario, Schettini, F. & Foglia, E. (2020). Preventing litigation with a predictive model of COVID-19 ICUs occupancy. *2020 IEEE International Conference on Big Data (Big Data)*. (pp. 2111–2116). Atlanta, GA, USA. https://doi.org/10.1109/BigData50022.2020.9378295
Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 3662–3666). IEEE. https://doi.org/10.1109/bigdata50022.2020.9378025
Gressel, S., Pauleen, D. J., & Taskin, N. (2020). *Management decision-making, big data and analytics*. Sage.
Guo, F., Li, F., Lv, W., Liu, L., & Duffy, V. G. (2020). Bibliometric analysis of affective computing researches during 1999–2018. *International Journal of Human-Computer Interaction*, *36*(9), 801–814. https://doi.org/10.1080/10447318.2019.1688985

Hildebrandt, M. (2021). The issue of bias. The framing powers of machine learning. In Pelillo, M., & Scantamburlo, T. (Eds.), *Machines We Trust: Perspectives on Dependable AI*. MIT Press. https://doi.org/10.7551/mitpress/12186.003.0009

Hoffrage, U., & Marewski, J. N. (2020). Social Scoring als Mensch-System-Interaktion. *Social Credit Rating: Reputation und Vertrauen beurteilen*, 305–329. https://doi.org/10.1007/978-3-658-29653-7_17

Iftene, A. (2018). *Who Is Worthy of Constitutional Protection?* A Commentary on Ewert v Canada.

Infantino, M., & Wang, W. (2021). Challenging Western Legal Orientalism: A Comparative Analysis of Chinese Municipal Social Credit Systems. *European Journal of Comparative Law and Governance*, *8*(1), 46–85. https://doi.org/10.1163/22134514-bja10011

Israni, E. (2017). *Algorithmic due process: mistaken accountability and attribution in State v. Loomis*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Krawiec, A., Pawela, Ł., & Puchała, Z. (2023). Discrimination and certification of unknown quantum measurements. *arXiv preprint arXiv:2301.04948*.

Kubat, M., & Kubat, J. A. (2017). *An introduction to machine learning* (Vol. 2, pp. 321–329). Cham, Switzerland: Springer International Publishing.

Kuhn, Th. S. (1962). The structure of scientific revolutions. *International Encyclopedia of Unified Science*, *2*(2).

Lippert-Rasmussen, K. (2022). Algorithm-Based Sentencing and Discrimination, *Sentencing and Artificial Intelligence* (pp. 74–96). Oxford University Press.

Maamar, N. (2018). Social Scoring: Eine europäische Perspektive auf Verbraucher-Scores zwischen Big Data und Big Brother. *Computer und Recht*, *34*(12), 820–828. https://doi.org/10.9785/cr-2018-341212

Mannozzi, G. (1997). Sentencing. *Dig. Disc. Pen.*

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Martini, M. (2020). Regulating Algorithms – How to demystify the alchemy of code?. In *Algorithms and Law* (pp. 100–135). Cambridge University Press. https://doi.org/10.1017/9781108347846.004

Marx, K. (2016). Economic and philosophic manuscripts of 1844. In *Social Theory Re-Wired*. Routledge

Massa, M. (1990). *Motivazione della sentenza (diritto processuale penale)*. Enc. Giur.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Messinetti, R. (2019). La tutela della persona umana versus l'intelligenza artificiale. Potere decisionale dell'apparato tecnologico e diritto alla spiegazione della decisione automatizzata, *Contratto e impresa*, *3*, 861–894.

Mi, F., Kong, L., Lin, T., Yu, K., & Faltings, B. (2020). Generalised class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 240–241). https://doi.org/10.1109/cvprw50498.2020.00128

Mitchell, T. M. (2007). *Machine learning* (Vol. 1). New York: McGraw-hill.

Nazir, A., Rao, Y., Wu, L., & Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, *13*(2), 845–863. https://doi.org/10.1109/taffc.2020.2970399

Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170359. https://doi.org/10.1098/rsta.2017.0359

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental'proportionality. *Information & communications technology law*, *27*(2), 223–250.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural networks*, *113*, 54–71.

Parona, L. (2021). Government by algorithm": un contributo allo studio del ricorso all'intelligenza artificiale nell'esercizio di funzioni amministrative. *Giornale Dir. Amm*, *1*.

Pellecchia, E. (2018). Profilazione e decisioni automatizzate al tempo della black box society: qualità dei dati e leggibilità dell'algoritmo nella cornice della responsible research and innovation. *Nuove leg. civ. comm*, 1209–1235.

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Petronio, U. (2020*). Il precedente negli ordinamenti giuridici continentali di antico regime. Rivista di diritto civile*, *66*(5), 949–983.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, *30*.

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research, *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/taffc.2020.3038167

Rebitschek, F. G., Gigerenzer, G., & Wagner, G. G. (2021). People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Scientific reports*, *11*(1), 1–11.

Rodotà, S. (1995). *Tecnologie e diritti*, il Mulino. Bologna.

Rodotà, S. (2012). *Il diritto di avere diritti*. Gius. Laterza.

Rodotà, S. (2014). *Il mondo nella rete: Quali i diritti, quali i vincoli*. GLF Editori Laterza.

Russell, P. H. (1983). The political purposes of the Canadian Charter of Rights and Freedoms. *Can. B. Rev.*, *61*, 30–35.

Scassa, T. (2021). Administrative Law and the Governance of Automated Decision Making: A Critical Look at Canada's Directive on Automated Decision Making, *UBCL Rev*, *54*, 251–255. https://doi.org/10.2139/ssrn.3722192

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies, *Harv. JL & Tech.*, *29*, 353–360. https://doi.org/10.2139/ssrn.2609777

Schiavone, A. (2019). *Eguaglianza.* Einaudi.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalisation of discrimination. *Stanford Law Review*, *66*, 803–872.

Stuurman, K., & Lachaud, E. (2022). Regulating AI. A label to complete the proposed Act on Artificial Intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3963890

Sunstein, C. R. (2019). Algorithms, correcting biases. *Social Research: An International Quarterly*, *86*(2), 499–511. https://doi.org/10.1353/sor.2019.0024

Tarrant, A., & Cowen, T. (2022). Big Tech Lobbying in the EU. *The Political Quarterly*, *93*(2), 218–226. https://doi.org/10.1111/1467-923x.13127

Taruffo, M. (1975). *La motivazione della sentenza civile*. Cedam, Padova.

Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*, 1–12. https://doi.org/10.1007/s43681-022-00142-y

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act-Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.31235/osf.io/38p5f

Vogel, P. A. (2020). "Right to explanation" for algorithmic decisions?, *Data-Driven Decision Making. Law, Ethics, Robotics, Health*, *49*, 1–12. https://doi.org/10.48550/arXiv.1606.08813

Von Tunzelmann, N. (2003). Historical coevolution of governance and technology in the industrial revolutions, *Structural Change and Economic Dynamics*, *14*(4), 365–384. https://doi.org/10.1016/s0954-349x(03)00029-8

Wang, C., Han, B., Patel, B., & Rudin, C. (2022). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction, *Journal of Quantitative Criminology*, *6*, 1–63. https://doi.org/10.1007/s10940-022-09545-w

Witt, A. C. (2022). Platform Regulation in Europe – Per Se Rules to the Rescue?, *Journal of Competition Law & Economics*, *18*(3), 670–708. https://doi.org/10.1093/joclec/nhac001

Woodcock, J. (2020). The algorithmic panopticon at Deliveroo: Measurement, precarity, and the illusion of control, *Ephemera: theory & politics in organisations*, *20*(3), 67–95.

York, J. C. (2022). *Silicon values: The future of free speech under surveillance capitalism*. Verso Books, London-New York.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile books, London.

Author information

Elena Faletti – PhD, Assistant Professor, Carlo Cattaneo University LIUC
Address: Corso Matteotti 22, Castellanza, 21053, Italy
E-mail: efalletti@liuc.it
ORCID ID: https://orcid.org/0000-0002-6121-6775
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57040979500

Conflict of interest

The author declares no conflict of interest.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – February 24, 2023
Date of approval – April 13, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 347.1:004.8
EDN: https://elibrary.ru/ktizpw
DOI: https://doi.org/10.21202/jdtl.2023.16

Алгоритмическая дискриминация и защита неприкосновенности частной жизни

Элена Фаллетти

Университет Карло Каттанео
г. Кастелланца, Итальянская Республика

Ключевые слова

Алгоритм,
дискриминация,
защита данных,
искусственный интеллект,
неприкосновенность,
персональные данные,
право,
регулирование,
цифровые технологии,
частная жизнь

Аннотация

Цель: появление цифровых технологий, таких как искусственный интеллект, стало вызовом для государств всего мира. Оно породило множество рисков, связанных с нарушением прав человека, включая права на неприкосновенность частной жизни и человеческое достоинство. Это определяет актуальность данного исследования. Цель статьи – проанализировать роль алгоритмов в случаях дискриминации и выяснить, каким образом алгоритмы могут способствовать предубежденности при принятии решений на основе персональных данных. Проведенный анализ помогает оценить проект закона об искусственном интеллекте, направленный на регулирование данной проблемы для предотвращения дискриминации при использовании алгоритмов.

Методы: в работе применялись методы эмпирического и сравнительного анализа. Сравнительный анализ позволил выявить сходства и различия существующего регулирования и положений проекта закона об искусственном интеллекте. С помощью эмпирического анализа рассмотрены реальные примеры алгоритмической дискриминации.

Результаты: результаты исследования показывают, что Закон об искусственном интеллекте нуждается в доработке, так как он остается на уровне дефиниций и недостаточно опирается на эмпирический материал. Автор выдвигает ряд предложений по совершенствованию данного законопроекта.

Научная новизна: заключается в мультидисциплинарности данной работы, рассматривающей вопросы дискриминации, защиты данных и влияния на эмпирическую реальность в сфере алгоритмической дискриминации и охраны неприкосновенности частной жизни.

Практическая значимость: состоит в привлечении внимания к тому факту, что алгоритмы выполняют инструкции, составленные на основе сообщенных им данных. Не имея возможностей для абдукции, алгоритмы действуют лишь как послушные исполнители приказов. Результаты работы могут использоваться в качестве основы для будущих исследований в данной области и в законотворческом процессе.

Для цитирования

Фаллетти, Э. (2023). Алгоритмическая дискриминация и защита неприкосновенности частной жизни. *Journal of Digital Technologies and Law*, *1*(2), 387–420. https://doi.org/10.21202/jdtl.2023.16

Список литературы

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020). The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems* (pp. 726–731). https://doi.org/10.1145/3383313.3418487

Ainis, M. (2015). *La piccola eguaglianza.* Einaudi.

Alpa, G. (2021). Quale modello normativo europeo per l'intelligenza artificiale? *Contratto e impresa*, *37*(4), 1003–1026.

Alpa, G., & Resta, G. (2006). *Trattato di diritto civile. Le persone e la famiglia: 1. Le persone fisiche ei diritti della personalità*. UTET giuridica.

Altenried, M. (2020). The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. *Capital & Class*, *44*(2), 145–158. https://doi.org/10.1177/0309816819899410

Amodio, E. (1970). *L'obbligo costituzionale di motivare e l'istituto della giuria. Rivista di diritto processuale.*

Angiolini, C. S. A. (2020). *Lo statuto dei dati personali: uno studio a partire dalla nozione di bene*. Giappichelli.

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., ... & Venkatasubramanian, S. (2021). It's complicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*

Bargi, A. (1997). *Sulla struttura normativa della motivazione e sul suo controllo in Cassazione.* Giur. it.

Battini, S. (2018). *Indipendenza e amministrazione fra diritto interno ed europeo*.

Bellamy, R. (2014). Citizenship: Historical development of. Citizenship: Historical Development of'. In J. Wright (Ed.), *International Encyclopaedia of Social and Behavioural Sciences, Elsevier*. https://doi.org/10.1016/b978-0-08-097086-8.62078-0

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Brooks, R. (2017). Machine Learning Explained. *Robots, AI and other stuff*.

Bodei, R. (2019). *Dominio e sottomissione*. Bologna, Il Mulino.

Canetti, E. (1960). *Masse und Macht*. Hamburg, Claassen.

Casonato, C., & Marchetti, B. (2021). Prime osservazioni sulla proposta di regolamento dell'Unione Europea in materia di intelligenza artificiale. *BioLaw Journal-Rivista di BioDiritto*, *3*, 415–437.

Chizzini, A. (1998). *Sentenza nel diritto processuale civile*. Dig. disc. priv., Sez. civ.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Citino, Y. (2022). *Cittadinanza digitale a punti e social scoring: le pratiche scorrette nell'era dell'intelligenza artificiale*. Diritti comparati.

Claeys, G. (2018). *Marx and Marxism*. Nation Books, New York.

Cockburn, I. M., Henderson, R., & Stern, S. (2018). The impact of artificial intelligence on innovation: An exploratory analysis. In *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Cossette-Lefebvre, H., & Maclure, J. (2022). AI's fairness problem: understanding wrongful discrimination in the context of automated decision-making. *AI and Ethics*, *5*, 1–15. https://doi.org/10.1007/s43681-022-00233-w

Crawford, K. (2021). Time to regulate AI that interprets human emotions. *Nature*, *592*(7853), 167. https://doi.org/10.1038/d41586-021-00868-5

Custers, B. (2022). AI in Criminal Law: An Overview of AI Applications in Substantive and Procedural Criminal Law. In B. H. M. Custers, & E. Fosch Villaronga (Eds.), *Law and Artificial Intelligence* (pp. 205–223). Heidelberg: Springer. http://dx.doi.org/10.2139/ssrn.4331759

De Gregorio, G. & Paolucci F. (2022). *Dati personali e AI Act. Media laws*.

Di Rosa, G. (2021). Quali regole per i sistemi automatizzati "intelligenti"?. *Rivista di diritto civile*, *67*(5), 823–853.

Epp, C. R. (1996). Do bills of rights matter? The Canadian Charter of Rights and Freedoms, *American Political Science Review*, *90*(4), 765–779.

Fanchiotti, V. (1995). *Processo penale nei paesi di Common Law.* Dig. Disc. Pen.

Freeman, C., Louçã, F., & Louçã, F. (2001). *As time goes by: from the industrial revolutions to the information revolution*. Oxford University Press.

Freeman, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. *North Carolina Journal of Law & Technology*, *18*(5), 75–90.

Fuchs, C. (2014). *Digital Labour and Karl Marx*. Routledge.

Gallese, C.(2022). *Legal aspects of the use of continuous-learning models in Telemedicine*. JURISIN.

Gallese, E. Falletti, M. S. Nobile, L. Ferrario, Schettini, F. & Foglia, E. (2020). Preventing litigation with a predictive model of COVID-19 ICUs occupancy. *2020 IEEE International Conference on Big Data (Big Data)*. (pp. 2111–2116). Atlanta, GA, USA. https://doi.org/10.1109/BigData50022.2020.9378295

Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 3662–3666). IEEE. https://doi.org/10.1109/bigdata50022.2020.9378025

Gressel, S., Pauleen, D. J., & Taskin, N. (2020). *Management decision-making, big data and analytics*. Sage.

Guo, F., Li, F., Lv, W., Liu, L., & Duffy, V. G. (2020). Bibliometric analysis of affective computing researches during 1999–2018. *International Journal of Human-Computer Interaction*, *36*(9), 801–814. https://doi.org/10.1080/10447318.2019.1688985

Hildebrandt, M. (2021). The issue of bias. The framing powers of machine learning. In Pelillo, M., & Scantamburlo, T. (Eds.), *Machines We Trust: Perspectives on Dependable AI*. MIT Press. https://doi.org/10.7551/mitpress/12186.003.0009

Hoffrage, U., & Marewski, J. N. (2020). Social Scoring als Mensch-System-Interaktion. *Social Credit Rating: Reputation und Vertrauen beurteilen*, 305–329. https://doi.org/10.1007/978-3-658-29653-7_17

Iftene, A. (2018). *Who Is Worthy of Constitutional Protection?* A Commentary on Ewert v Canada.

Infantino, M., & Wang, W. (2021). Challenging Western Legal Orientalism: A Comparative Analysis of Chinese Municipal Social Credit Systems. *European Journal of Comparative Law and Governance*, *8*(1), 46–85. https://doi.org/10.1163/22134514-bja10011

Israni, E. (2017). *Algorithmic due process: mistaken accountability and attribution in State v. Loomis*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Krawiec, A., Pawela, Ł., & Puchała, Z. (2023). Discrimination and certification of unknown quantum measurements. *arXiv preprint arXiv:2301.04948*.

Kubat, M., & Kubat, J. A. (2017). *An introduction to machine learning* (Vol. 2, pp. 321–329). Cham, Switzerland: Springer International Publishing.

Kuhn, Th. S. (1962). The structure of scientific revolutions. *International Encyclopedia of Unified Science*, *2*(2).

Lippert-Rasmussen, K. (2022). Algorithm-Based Sentencing and Discrimination, *Sentencing and Artificial Intelligence* (pp. 74–96). Oxford University Press.

Maamar, N. (2018). Social Scoring: Eine europäische Perspektive auf Verbraucher-Scores zwischen Big Data und Big Brother. *Computer und Recht*, *34*(12), 820–828. https://doi.org/10.9785/cr-2018-341212

Mannozzi, G. (1997). Sentencing. *Dig. Disc. Pen.*

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Martini, M. (2020). Regulating Algorithms – How to demystify the alchemy of code?. In *Algorithms and Law* (pp. 100–135). Cambridge University Press. https://doi.org/10.1017/9781108347846.004

Marx, K. (2016). Economic and philosophic manuscripts of 1844. In *Social Theory Re-Wired*. Routledge

Massa, M. (1990). *Motivazione della sentenza (diritto processuale penale)*. Enc. Giur.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Messinetti, R. (2019). La tutela della persona umana versus l'intelligenza artificiale. Potere decisionale dell'apparato tecnologico e diritto alla spiegazione della decisione automatizzata, *Contratto e impresa*, *3*, 861–894.

Mi, F., Kong, L., Lin, T., Yu, K., & Faltings, B. (2020). Generalised class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 240–241). https://doi.org/10.1109/cvprw50498.2020.00128

Mitchell, T. M. (2007). *Machine learning* (Vol. 1). New York: McGraw-hill.

Nazir, A., Rao, Y., Wu, L., & Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, *13*(2), 845–863. https://doi.org/10.1109/taffc.2020.2970399

Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170359. https://doi.org/10.1098/rsta.2017.0359

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental'proportionality. *Information & communications technology law*, *27*(2), 223–250.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural networks*, *113*, 54–71.

Parona, L. (2021). Government by algorithm": un contributo allo studio del ricorso all'intelligenza artificiale nell'esercizio di funzioni amministrative. *Giornale Dir. Amm*, *1*.

Pellecchia, E. (2018). Profilazione e decisioni automatizzate al tempo della black box society: qualità dei dati e leggibilità dell'algoritmo nella cornice della responsible research and innovation. *Nuove leg. civ. comm*, 1209–1235.

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Petronio, U. (2020*). Il precedente negli ordinamenti giuridici continentali di antico regime. Rivista di diritto civile*, *66*(5), 949–983.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, *30*.

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research, *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/taffc.2020.3038167

Rebitschek, F. G., Gigerenzer, G., & Wagner, G. G. (2021). People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Scientific reports*, *11*(1), 1–11.

Rodotà, S. (1995). *Tecnologie e diritti*, il Mulino. Bologna.

Rodotà, S. (2012). *Il diritto di avere diritti*. Gius. Laterza.

Rodotà, S. (2014). *Il mondo nella rete: Quali i diritti, quali i vincoli*. GLF Editori Laterza.

Russell, P. H. (1983). The political purposes of the Canadian Charter of Rights and Freedoms. *Can. B. Rev.*, *61*, 30–35.

Scassa, T. (2021). Administrative Law and the Governance of Automated Decision Making: A Critical Look at Canada's Directive on Automated Decision Making, *UBCL Rev*, *54*, 251–255. https://doi.org/10.2139/ssrn.3722192

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies, *Harv. JL & Tech.*, *29*, 353–360. https://doi.org/10.2139/ssrn.2609777

Schiavone, A. (2019). *Eguaglianza.* Einaudi.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalisation of discrimination. *Stanford Law Review*, *66*, 803–872.

Stuurman, K., & Lachaud, E. (2022). Regulating AI. A label to complete the proposed Act on Artificial Intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3963890

Sunstein, C. R. (2019). Algorithms, correcting biases. *Social Research: An International Quarterly*, *86*(2), 499–511. https://doi.org/10.1353/sor.2019.0024

Tarrant, A., & Cowen, T. (2022). Big Tech Lobbying in the EU. *The Political Quarterly*, *93*(2), 218–226. https://doi.org/10.1111/1467-923x.13127

Taruffo, M. (1975). *La motivazione della sentenza civile*. Cedam, Padova.

Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*, 1–12. https://doi.org/10.1007/s43681-022-00142-y

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act-Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.31235/osf.io/38p5f

Vogel, P. A. (2020). “Right to explanation” for algorithmic decisions?, *Data-Driven Decision Making. Law, Ethics, Robotics, Health*, *49*, 1–12. https://doi.org/10.48550/arXiv.1606.08813

Von Tunzelmann, N. (2003). Historical coevolution of governance and technology in the industrial revolutions, *Structural Change and Economic Dynamics*, *14*(4), 365–384. https://doi.org/10.1016/s0954-349x(03)00029-8

Wang, C., Han, B., Patel, B., & Rudin, C. (2022). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction, *Journal of Quantitative Criminology*, *6*, 1–63. https://doi.org/10.1007/s10940-022-09545-w

Witt, A. C. (2022). Platform Regulation in Europe – Per Se Rules to the Rescue?, *Journal of Competition Law & Economics*, *18*(3), 670–708. https://doi.org/10.1093/joclec/nhac001

Woodcock, J. (2020). The algorithmic panopticon at Deliveroo: Measurement, precarity, and the illusion of control, *Ephemera: theory & politics in organisations*, *20*(3), 67–95.

York, J. C. (2022). *Silicon values: The future of free speech under surveillance capitalism*. Verso Books, London-New York.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile books, London.

Сведения об авторе

Элена Фаллетти – доктор наук, доцент, Университет Карло Каттанео
Адрес: Корсо Маттеотти 22, Кастелланца, 21053, Италия
E-mail: efalletti@liuc.it
ORCID ID: https://orcid.org/0000-0002-6121-6775
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57040979500

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.27.51 / Осуществление и защита гражданских прав
Специальность ВАК: 5.1.3 / Частно-правовые (цивилистические) науки

История статьи

Дата поступления – 24 февраля 2023 г.
Дата одобрения после рецензирования – 13 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.17

Approaches to Regulating Relations in the Sphere of Developing and Using the Artificial Intelligence Technologies: Features and Practical Applicability

Olga S. Erahtina

Perm branch of National Research University "Higher School of Economics"
Perm, Russian Federation

Keywords

Artificial intelligence,
danger,
digital economy,
digital technologies,
law,
regulation,
risk management,
risk-oriented approach,
software,
technological approach

Abstract

Objective: to review the modern scientific approaches to regulating relations in the sphere of using the artificial intelligence technologies; to reveal the main features and limitations of using the risk-oriented and technological approaches in order to determine the directions of their further development.

Methods: the methodological basis of the research is a set of scientific cognition methods, including the general scientific dialectic method and the universal scientific methods (analysis and synthesis, comparison, summarization, structural-functional, and formal-logical methods).

Results: it was determined that the use of the risk-oriented approach implies building constructive models of risk management. A significant issue in using this approach is the bases of referring the artificial intelligence technologies to high-risk ones. When determining the risk level of using the artificial intelligence technologies, the following criteria should be applied: the type of artificial intelligence technology, its sphere of use, and the level of potential harm for the environment, health and other fundamental human rights.

In turn, the central issue of using the technological approach is the necessity and limits of regulation in the sphere of developing and using the artificial intelligence technologies. First, interference into this sphere must not create obstacles for developing technologies and innovations. Second, a natural reaction of a regulator towards newly emerging objects and subjects of turnover is the "imperfect law syndrome". At the same time, a false idea

about a lack of legal regulation may produce an opposite effect – duplication of legal norms. To solve the problem of duplicating legal requirements, it is necessary, first of all, to solve the issue of the need to regulate the artificial intelligence technologies or certain types of software applications.

Scientific novelty: a review was carried out of the main approaches to regulating relations in the sphere of developing and using the artificial intelligence technologies; the opportunities and limitations of their use are revealed; further directions of their development are proposed.

Practical significance: the main provisions and conclusions of the research can be used for determining the optimal approaches to regulating the sphere of digital technologies and for improving the legal regulation of the studied sphere of social relations.

For citation

Erahtina, O. S. (2023). Approaches to Regulating Relations in the Sphere of Developing and Using the Artificial Intelligence Technologies: Features and Practical Applicability. *Journal of Digital Technologies and Law, 1*(2), 421–437. https://doi.org/10.21202/jdtl.2023.17

Contents

Introduction

Active development of technologies and systems of artificial intelligence generates scientific discussions about the necessity, limits and tasks of legal regulation in the sphere of information technologies. Scientists declare opposing views: from opinions about the need to establish a large number of obligatory requirements, mainly imposed on a developer (Smuha, 2021), to proposals to eliminate legal interference into the sphere of high technologies[1], so as not to impede innovations. The range of positions of scientists includes also more moderate voews: application of international and national standards (Zielke, 2020); implementation of voluntary certification (Ellul et al., 2021); soft regulation and self-regulation (Erdélyi & Goldsmith, 2018); establishing of explainable (Hamon et al., 2022) and ethical frameworks (Wagner, 2018).

[1] O'Sullivan, Andrea. (2017, October 24). *Don't Let Regulators Ruin AI*. https://www.technologyreview.com/2017/10/24/3937/dont-let-regulators-ruin-ai/

A Belgian researcher Nicolas Petit proposes the so called "regulatory trade-offs" achieved by balancing threats and opportunities created by the introduction of legal regulation[2]. He gives a number of examples of regulation impeding technological progress[3]. At the same time, he emphasizes that the lack of regulation may also hinder technological evolution. In particular, legal uncertainty negatively influences investments. According to Ryan Calo, undefined liability rules may bar investments into the open robotics markets and direct the caital flow towards narrow functionality of robots, where manufacturers may better manage risks, leaving open robotics underdeveloped (Calo, 2011).

Among the many approaches to legal regulation of relations in the sphere of using the artificial intelligence technologies one may specify the approach determining the general legal regime which is to stipulate the basic requirements to providing safety of the artificial intelligence systems. This regime should be applied to all such systems. Alongside with that, detailed requirements should be elaborated to development and use of the artificial intelligence in specific spheres (Ponkin & Redkina, 2018).

The high dynamics of the artificial intelligence technologies development and the multiple regulatory initiatives actualize the importance of interdisciplinary research aimed at revealing the optimal approaches to regulating the said sphere of public relations.

The article presents a complex analysis of two interdisciplinary approaches to regulating relations in the sphere of developing and using the artificial intelligence technologies, namely, the risk-oriented and the technological approaches; the features and limitations of their use are revealed; the directions of their further development are specified.

1. Risk-oriented approach

One of the approaches to regulating the artificial intelligence technologies actively discussed in science is a risk-oriented approach (Mikhaleva & Shubina, 2019; Gellert, 2021; Gonçalves, 2020). The idea of applying this approach was announced in the European Parliament Resolution on Civil Law Rules on Robotics of 2017[4]. Four years later, a draft Report of European Union stipulating the main rules on the artificial intelligence[5] proposed classification of the artificial intelligence systems based on the estimation of risk of their application. According to the said classification, all artificial intelligence systems were divided into four groups:

---

2 Petit, N. (2017, March 9). *Law and Regulation of Artificial Intelligence and Robots – Conceptual Framework and Normative Implications* (Working paper, pp. 6–7). http://dx.doi.org/10.2139/ssrn.2931339

3 *Ibid*. P. 12.

4 European Parliament. (2017, February 15). *Draft Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL))*. https://www.europarl.europa.eu/portal/en

5 European Parliament. (2021, April 21). *Regulation of the European Parliament and of the Council. Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. https://www.europarl.europa.eu/portal/en

1) artificial intelligence systems with inadmissibly high risk level (first of all, the artificial intelligence systems used in military and defense sector, systems for manipulating human behavior and systems for forming social ratings);

2) artificial intelligence systems with high risk level[6];

3) artificial intelligence systems with limited risk level[7];

4) artificial intelligence systems with minimal risk level[8].

To estimate the risk level of artificial intelligence systems application, two main criteria are proposed: the degree of users' dependence on the decisions made by the system and the degree of its danger for life and health of citizens and violation of their fundamental rights.

The UE draft Regulation highlights the need to prohibit production and civil circulation of artificial intelligence systems, the use of which creates inadmissibly high risk of incurring harm. In turn, the developers, owners, and users of high-risk artificial intelligence systems, according to the draft regulation, must meet higher requirements to providing safety, keeping technical documentation and disclosing information[9]. To comply with the stipulated standards, quality management systems can be applied. The systems with limited or minimal risk level must, as a minimum, provide the opportunity to inform users about their interaction with artificial intelligence.

In general, the EU approach deserves support. At the same time, it requires further development.

First, it should be taken into account that the risk-oriented approach is the risk management system consisting of three main stages. At the first stage, one must identify, analyze and differentiate between the forecasted risks. At the second stage, one should estimate the risk level. At the third stage, one should determine the means of managing risks.

Given the variety of approaches to classification of risks (ideally, all these approaches must be taken into account), for the purposes of this article we may highlight:

– internal and external risks;

– systemic and non-systemic risks.

---

6 For example, robot assistants in surgery, management and exploitation of critical infrastructure.

7 For example, chat bots, virtual assistants, smart homes.

8 For example, videogames, spam filters.

9 When developing and using such systems, "principal requirements" should be met, such as requirements to the quality of data, documenting and traceability, human control, etc. In particular, prior to market placement or putting into operation, technical documentation must be compiled, which should reflect the system compliance to the set requirements and all the necessary information for estimating the system compliance (Article 11). The draft also requires elaborating systems so as to provide accounting during the system functioning (Article 12), as well as transparency of the AI system and information submission (Article 13), for the users of high-risk systems to be able to use them properly and to interpret the output data correctly. Human control must be provided during the entire lifecycle of the system, with the opportunity to interfere into the system functioning at any time and stop or fix it if needed.

Internal risks occur in an individual company; they can be forecasted, estimated and prevented in house. External risks (changes in the economic and political situation, natural disasters, environmental accidents, etc.) cannot be prevented in house.

Systemic risks threaten the market in general or certain spheres of business. Only experts may forecast their occurrence and estimate their consequences. They also cannot be prevented in house. As for non-systemic (commercial) risks, these are the risks of an individual company which it can and must minimize by its own efforts.

The concept of risk-oriented approach as interpreted in the European Parliament Resolution on Civil Law Rules on Robotics is aimed at managing internal and non-systemic risks while unattended external and systemic risks, which cannot be imputed to individual subjects of civil turnover. Identification of such risks is one of the main tasks to be solved at the first stage.

At the second stage, the risk level is determined (the probability of risk materialization and the volume of adverse consequences which may occur). At the first sight, the draft solves this task. However, the grounds for referring the artificial intelligence system to one of the four groups require further research. As was mentioned above, the draft Resolution proposes using the degree of users' dependence on the decisions made by the artificial intelligence as the criterion of the risk level, as well as the degree of danger the technology poses for the life and health of humans and the violation of their fundamental rights. In our opinion, the assessment of the valuation of the probability of risk materialization depends also on a number of other factors. First of all, these are characteristics of the technology. The main characteristics of the artificial intelligence are its autonomy (ability to make decisions independently, without human interference) and learnability (ability to master new skills and competencies). Depending on the sphere of application, one may distinguish several levels of the technology autonomy. For example, Appendix 10 to the Transport Strategy of the Russian Federation up to 2030 with the forecast up to 2035[10] defines five levels of autonomy of automobile transport, four levels of autonomy of railway transport and six levels of autonomy of water and marine transport.

By the criterion of learnability, the artificial intelligence may be unlearning, learning and self-learning. Apparently, highly autonomous and self-learning technologies must be referred to high-risk artificial intelligence systems, while unlearning technologies with the first or second levels of autonomy should be referred to the systems with minimal risk. However, the task of determining the risk level of an artificial intelligence technology based on its characteristic is not as simple as it may seem at the first glance. This is first, of all, due to the fact that, while estimating the risk level, one must take into account other characteristics of artificial intelligence besides autonomy and learnability. In particular, these are functionality (ability to perform one or more functions) and equipment with control means (Alekseev et al., 2020). Also, apparently, various combinations of these characteristics

[10] Adopted by the Order of the Government of the Russian Federation of 27.11.2021 No. 3363-r. https://base.garant.ru/403156321/

are possible. For example, the artificial intelligence technology may be highly autonomous, learning, perform two functions, and having no objective control means.

Scientific works also propose to use the sphere of application as the criterion for estimating the risk level of using artificial intelligence technologies. According to a group of researchers from University of Malta, legal regulation must be obligatory only for critical spheres of activity. At the same time, alongside with the sphere of using the artificial intelligence systems, one must assess the risk level of the activities it is used in (Ellul et al., 2021). One should agree with this conclusion. For example, in healthcare artificial intelligence technologies may be used to assist doctors in making diagnosis, prescribing medications, or performing operations. These types of activity may be referred to a high-risk category. At the same time, the artificial intelligence technologies are used for patient registering, processing and analyzing medical records, automated notifying of medical staff. These types of activity may be referred to a limited risk category.

Most authors refer the sphere of transport to the high-risk category. However, it should be taken account which types of activity are accompanied by the artificial intelligence. The artificial intelligence systems can be used to improve safety and efficiency of transportation, to manage passenger and cargo flows. At the same time, the artificial intelligence technologies are also used for rendering services of transporting cargo and passengers. While managing transport infrastructure refers to the high-risk category, servicing may be referred to the limited risk category.

Second, the question of risk management means requires further research, too. The draft pays the most attention to high-risk systems, actually, leaving unattended the artificial intelligence systems with limited risk. At the same time, special regulation (based on risk-oriented approach) must be implemented also to the artificial intelligence systems referring to this group.

We believe that risk management means will be different depending on the type of risk (internal or external, systemic or non-systemic risk) and the degree of risk of the artificial intelligence technology (high or intermediate).

Deserving attention is the “Basic model for determining criteria and categories of risk” adopted in 2017 by a project committee of the priority program “Reform of control and supervisory activity”[11]. The document defines such notions as “risk sources”, “risk factors”, “risk profile”, determines their types, offers the means of ranking the manageable risk factors of profiles (to determine the most significant of them) and the methodology

[11] “Basic model for determining criteria and categories of risk” (adopted by a protocol of the meeting of the project committee of 31.03.2017 No. 19(3)) (alongside with the “Requirements to justification of the proposed by federal executive bodies – participants of the priority program “Reform of control and supervisory activity” – risk categories (classes of danger) and risk criteria in relation to the types of state control (supervision) executed by them”).

of determining the volume of harm incurred and the probable frequency of potential negative consequences.

Although the above model was adopted in order to implement "smart checks" by supervisory bodies, to focus the checks on potentially most dangerous objects, the proposed methods for determining risk categories and criteria can be also used for assessing the risk level of using artificial intelligence technologies.

2. Technological approach

Recently scientific literature has been paying more and more attention to technological approach which focuses on the technology per se, its essential and specific characteristics. Viewing the development of technologies and innovations as the basic task of legal regulation, representatives of this approach, first of all, pose the question of the necessity and limits of regulation in the spheres of high technologies. Pondering over this issue, J. Ellul comes to the conclusion that legal regulation must focus not on the artificial intelligence, but on software. In his opinion, in case of critical software, for example, used in an aircraft, it does not matter if artificial intelligence is applied or not. Regulation should not touch upon a specific artificial intelligence technology; it should be broader and be implemented in relation to software in general. To prove this conclusion, Ellul gives one more example. When using artificial intelligence in banking or insurance systems which decide whether a specific credit or polis should be offered, regulation must be aimed at providing that clients are not discriminated. This requirement must be applied not only to the systems based on the artificial intelligence. It is quite feasible to program a decision making system, using methods not related to artificial intelligence (Ellul, 2022).

Supporting this viewpoint in general, we should take into account that artificial intelligence is an umbrella term. Normative-legal acts justly and consistently distinguish between the notions "computer program"[12], "artificial intelligence"[13] and "artificial intelligence technologies"[14].

[12] Article 1261 of the Russian Civil Code.

[13] Order of the President of the Russian Federation of October 10, 2019 No. 490 "On developing artificial intelligence in the Russian Federation", which introduces the National strategy of developing artificial intelligence up to 2030.

[14] Article 2 of Federal Law of April 24, 2020 No. 123-FZ "On making an experiment of establishing special regulation with a view of creating the necessary conditions for developing and introducing the artificial intelligence technologies in the Russian Federation subject – city of federal significance Moscow and introducing changes into Articles 6 and 10 of Federal Law "On personal data".

Note to clause 3.18 of the National Standard "Artificial intelligence systems. Classification of the artificial intelligence systems"[15] highlights that artificial intelligence as a complex of technological solutions includes information-communication infrastructure, software (including that using machine learning methods), processes and services of data processing, analysis and synthesis of solutions.

Scientific literature also raises the question whether artificial intelligence is a single object or an umbrella notion (Balashova, 2022). To solve this question, L. Yu. Vasilevskaya et al. propose including artificial intelligence into the list of complex intellectual rights objects, stipulated in Article 1240 of the Russian Civil Code. According to the authors, the structural elements of artificial intelligence are a software product (computer program); software (a set of programs); artificial neural networks (computer programs); algorithms, software as know-how; technical solutions as inventions; data bases (Vasilevskaya et al., 2021).

Thus, although a computer program is a core of the artificial intelligence technology, it is wrong to equate these notions. It is the distinctive features of artificial intelligence systems, their specific characteristics and the presence of structural elements being independent objects of civil rights that determine the features of legal regulation of relations in the sphere of their development and use.

Application of the technological approach is also aimed at preventing duplication of legal requirements in the spheres where they are already introduced. In this regard, J. Ellul poses one more question: should software for planning tasks in a calendar, for example, be more regulated than required by the current laws (for example, the law on personal data protection or consumer rights protection) (Ellul, 2022)? We believe that this question should be answered positively. Technological development changes the process of interaction of the turnover participants, and, consequently, the content of their rights and obligations. For example, Article 12 of the Law on consumer rights protection stipulates the obligation of the producer (executor, seller) to timely provide the consumer with the necessary and reliable information about the goods (works, services). At the same time, in the digital society information is provided, as a rule, in the digital form. Alongside with that, software appears which simplifies information search. A. I. Savelyev justly notes that in future, probably, a consumer will bear the risk of non-acquaintance with the information placed by the producer in publicly accessible sources (Savelyev, 2016). Moreover, a legislator introduces new subjects into the civil turnover. For example, in June 2018 the Law "On making changes in the 'Law of the Russian Federation on consumer rights protection' "[16] introduced special regulation of activity of an owner of information aggregator on goods (services).

---

[15] GOST R 59277-2020. National Standard of the Russian Federation. Artificial intelligence systems. Classification of the artificial intelligence systems. Date of introduction 01.03.2021.

[16] "On making changes in the 'Law of the Russian Federation on consumer rights protection'": Federal Law of 29.07.2018 No. 250-FZ. http://www.consultant.ru/document/cons_doc_LAW_303537/

Using the artificial intelligence technologies for processing and storing personal data is associated with risks of their leaking or incorrect interpretation. Since 1995, the European Parliament has been solving the problem of such risks management[17]. In 2016, the EU "Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data" proposed a model aimed at solving the task of improved personal data protection. Implementation of this model, in the opinion of the Regulation drafters, should promote forming a trustful attitude towards technologies[18].

The final question studied in J. Ellul's work is: must obligatory regulation be oriented directly towards technologies or towards a certain sphere or type of activity, during the implementation of which they are used (Ellul, 2022)? When answering this question, one should take into account that a risk of error is inherent in software of any complexity, regardless of it having elements of artificial intelligence or not. As software systems in general (not obligatory using artificial intelligence) "grow in complexity, interconnectedness, and geographical distribution, we will increasingly face unwanted emergent behavior" (Mogul, 2006). That is, interaction of a technology with the environment creates additional complexities and risks. To minimize such risks, it is necessary, first of all, to introduce quality standards of software[19]. Then, it is necessary to assess the risk level of the types of activity, during the implementation of which the technology is used (as was said in the first part of the article). If a certain type of activity refers to a high-risk group, a requirement must be stipulated about an obligatory application of the respective standard. Thus, obligatory regulation must be oriented not on technology but on the type of activity, during the implementation of which it is used.

Viewing the main questions raised by the representatives of the technological approach, one should consider a research by Ronald Leenes. The researcher from the Netherlands points out that it is rather difficult to identify gaps in the legal regulation of relations in the sphere of using technologies. First of all, one should define the essential and specific characteristics of the technology. While solving this task, should one focus on a specific technology, like unmanned automobiles, or consider a broader category, like unmanned vehicles?

---

17 On the protection of individuals with regard to the processing of personal data and on the free movement of such data: Directive 95/46/EC of October 24, 1995. (Directive 95/46/EC "On personal data").

18 On the protection of natural persons with regard to the processing of personal data and on the free movement of such data: Regulation 2016/679 of April 27, 2016.

19 See, for example: GOST 28195-89. Assessing the quality of software. General provisions. Introduced on 01.07.1990; GOST 28806-90. Quality of software. Terms and definitions. Introduced on 01.01.1992; GOST R 51188-98. Protection of information. Testing software for the presence of computer viruses. Standard guidelines. Introduced on 01.07.1999.

According to Leenes, both approaches are disadvantageous. Focusing on a specific technology may result in a regulator concentrating on potentially accidental features of the technology. Otherwise, an excessive generalization may make the discussion abstract, hence useless. That is why, at the present stage, one should take a "social-technical prism" and, alongside with determining specific characteristics of the technology, reveal whose interests should be important and prioritized (Leenes, 2019).

At the second stage, it is necessary to solve the question of technology development, to which end reveal the potential risks and current problems associated with their use. Unfortunately, the categories of "risk" and "problem" are often confused in the scientific literature, making it hard to distinctly define the limits and tasks of legal regulation in the sphere of social relations under study. Assumingly, one of the main risks of using the artificial intelligence technologies is the possibility of it autonomously deviating from the target initially built in it. As a result if such risk materialization, certain negative consequences may occur, such as harm to life, health or property of the user, or disclosure of confidential information.

The need to distinguish between the "risk" and "problem" categories was pointed out by E. A. Voinikanis, E. V. Semenova and G. S. Tyulyaev. They mention such risks of using the artificial intelligence technologies as, in particular, the possibility of data de-anonymization, possibility of discrimination based on gender, race, nationality, or confession. They pose such problems as who is a right holder of artificial intelligence software, who is responsible for incurring harm to life or health when using artificial intelligence, etc. (Voinikanis et al., 2018).

At the third stage, one should define the forms and limits of state interference into the sphere of the artificial intelligence technologies. The means of risk management and solving the problems of minimizing the negative consequences of their materialization are different. For example, the means of risk management may include keeping a register during the system functioning and providing transparency of the decision-making process, so that the users may interpret the output data. In relation to high-risk system, post-marketing monitoring should also be introduced, in order to collect and analyze data about the system functioning after its launching into market.

If, despite the preventive control measures taken, the system malfunctions, law must provide the means for just distribution of negative consequences between its developer, user, and operator.

One should also take into account that the means of influencing the risks and problems are various and not always legal in form. At that, according to a just remark by M. Scherer (Scherer, 2016), traditional methods of legal regulation, such as, for example, licensing production, control over researches, possibility to apply delict liability are not quite suitable for risk management in the sphere of using artificial intelligence systems.

Conclusions

The use of risk-oriented approach implies building constructive models of risk management. The process of risk management consists of three main stages. At the first stage, it is necessary to identify and classify all risks related to using the artificial intelligence technologies in a certain sphere. The concept of risk-oriented approach, proposed by the European Parliament, focuses on internal and non-systemic risks. Accordingly, in order to develop this approach, it is necessary to research the external and systemic risks. At the second stage, it is necessary to assess the risk level of using a specific artificial intelligence technology. When making the assessment, several criteria should be used. Among them are the essential and specific characteristics of the technology, the sphere and type of activity, during implementation of which this technology is used. At the third stage, one should identify the means of risk management, which, in turn, are differentiated depending on the risk level of a specific technology. As one can see, the main objective of applying the risk-oriented approach consists in determining the means of risk management, associated with the use of the artificial intelligence technologies.

The technological approach is focused on the necessity and limits of regulation in the sphere of high technologies. The main stages of applying the technological approach are the following:

– determining the essential and specific characteristics of the technologies;

– revealing the potential risks and current problems of their use;

– determining the forms and limits of the state interference into the sphere of the artificial intelligence technologies.

It is the specific characteristics of the artificial intelligence technologies, such as autonomy and self-learning ability that determine the features of legal regulation of relations in the sphere of their development and use. At that, legal regulation should be oriented not on technology but on the type of activity, during the implementation of which it is used.

The research carried out also allows concluding that a universal approach to regulating relations in the sphere of technologies development and use is the technological approach. Although this approach needs further development, it may right now serve as the basis for forming the strategy of law-making activity. In turn, the risk-oriented approach is one of the main elements of the technological approach. Effective management of the accompanying risks will enable to minimize the potential negative consequences of using new technologies and will provide the sustainable development of the sphere of high technologies.

References

Alekseev, A. O., Erahtina, O. S., Kondratyeva K. S., & Nikitin, T. Ph. (2020). Approaches to civil legal liability of the artificial intelligence technologies developer: based on the classification. *Information Society*, *6*, 47–57. (In Russ.).

Balashova, A. I. (2022). Artificial intelligence in copyright and patent law: objects, subject structure of legal relations, terms of legal protection. *Zhurnal Suda po intellektual'nym pravam*, *2(36), 90–98.* (In Russ.).

Calo, R. (2011). *Open robotics. Maryland Law Review*, *70.3, 101–142*.

Ellul, J., Pace, G., McCarthy, S., Sammut, T., Brockdorf, J., & Scerri, M. (2021). Regulating artificial intelligence: a technology regulator's perspective. In: *Proceedings of the Eighteenth International conference on artificial intelligence and law* (pp. 190–194). https://doi.org/10.1145/3462757.3466093

Ellul, J. (2022). Should we regulate Artificial Intelligence or some uses of Software? *Discover Artificial Intelligence*, *2*(5). https://doi.org/10.1007/s44163-022-00021-9

Erdélyi, O. J., & Goldsmith, J. (2018). Regulating artifcial intelligence: proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 95–101).

Gellert, R. (2021). The role of the risk-based approach in the General data protection Regulation and in the European Commission's proposed Artificial Intelligence Act: Business as usual? *Journal of Ethics and Legal Technologies*, *3*(2), 15–33.

Gonçalves, M. E. (2020). The risk-based approach under the new EU data protection regulation: a critical perspective. *Journal of Risk Research*, *23*(2), 139–152. https://doi.org/10.1080/13669877.2018.1517381

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., & De Hert, P. (2022). Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE ComputIntell Mag.*, *17*(1), 72–85. https://doi.org/10.1109/mci.2021.3129960

Leenes, R. (2019). *Regulating New Technologies in Uncertain Times*. https://doi.org/10.1007/978-94-6265-279-8_2

Mikhaleva, E. S., & Shubina, E. A. (2019). Challenges and Prospects of the Legal Regulation of Robotics. *Actual Problems of Russian Law*, *1*(12), 26–35. (In Russ.). https://doi.org/10.17803/1994-1471.2019.109.12.026-035

Mogul, J. C. (2006). Emergent (mis)behavior vs. complex software systems. *ACM SIGOPS Oper. Syst. Rev.*, *40*(4), 293–304. https://doi.org/10.1145/1218063.1217964

Ponkin, I. V., & Redkina, A. I. (2018). Artificial Intelligence from the Point of View of Law. *RUDN Journal of Law*, *22*(1), 91–109. (In Russ.). https://doi.org/10.22363/2313-2337-2018-22-1-91-109

Savelyev, A. I. (2016). Directions of freedom of contract evolution under the influence of modern information technologies. In M. A. Rozhkova (head of authors' collective and editor-in-chief), *Svoboda dogovora*. Moscow: Statut. (In Russ.).

Scherer, M. U. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, *29*(2), 353–400. https://doi.org/10.2139/ssrn.2609777

Smuha, N. A. (2021). From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law. Innov. Technol.*, *13*(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Vasilevskaya, L. Yu., Poduzova, E. B., & Tasalov, F. A. (2021). *Digitalization of civil turnover: legal characteristics of "artificial intelligence' and "digital" subjects (civilistic research)* (In 5 Vol.). Moscow: Prospekt. (In Russ.).

Voinikanis, E. A., Semenova, E. V., & Tyulyaev, G. S. (2018). Artificial intelligence and law: challenges and possibilities of self-learning algorithms. *Proceedings of Voronezh State University. Series: Pravo*, *4*(35), 137–148. (In Russ.).

Wagner, B. (2018). Ethics as an escape from regulation: from ethics-washing to ethics-shopping. In *Being profiling: cogitas ergo sum* (pp. 86–90). Amsterdam: Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

Zielke, T. (2020). Is artificial intelligence ready for standardization? In: *European conference on software process improvement* (pp. 259–274). Springer.

Author information

Olga S. Erahtina – Candidate of Sciences in Jurisprudence, Associate Professor, Department of Civil and Entrepreneurial Law, Perm branch of National Research University "Higher School of Economics"
Address: 38 Studencheskaya Str., 614070 Perm, Russian Federation
E-mail: oerahtina@hse.ru
ORCID ID: https://orcid.org/0000-0002-9041-3487
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/K-3149-2014
Google Scholar ID: https://scholar.google.ru/citations?hl=ru&user=WdwWB4kAAAAJ
RSCI Author ID: https://www.elibrary.ru/author_items.asp?authorid=498773

Conflict of interests

The author declares no conflict of interests.

Financial disclosure

The research was not sponsored.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – February 10, 2023
Date of approval – April 23, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 340.143:004.8
EDN: https://elibrary.ru/lbwsxw
DOI: https://doi.org/10.21202/jdtl.2023.17

Подходы к регулированию отношений в сфере разработки и применения технологий искусственного интеллекта: особенности и практическая применимость

Ольга Сергеевна Ерахтина

Пермский филиал Национального исследовательского университета «Высшая школа экономики»
г. Пермь, Российская Федерация

Ключевые слова

Искусственный интеллект,
опасность,
право,
программное обеспечение,
регулирование,
рискориентированный подход,
технологический подход,
управление рисками,
цифровая экономика,
цифровые технологии

Аннотация

Цель: обзор сложившихся в науке подходов к регулированию отношений в сфере применения технологий искусственного интеллекта, выявление основных особенностей и ограничений применения рискориентированного и технологического подходов для определения направлений их дальнейшего развития.

Методы: методологическую основу исследования составляет совокупность методов научного познания, в том числе общенаучный диалектический и универсальные научные методы (анализ и синтез, сравнение, обобщение, структурно-функциональный, формально-логический).

Результаты: определено, что применение рискориентированного подхода предполагает построение конструктивных моделей управления рисками. Значимым вопросом для применения данного подхода является вопрос об основаниях отнесения технологий искусственного интеллекта к высокорисковым. При определении уровня риска применения технологии искусственного интеллекта необходимо применять следующие критерии: тип технологии искусственного интеллекта, сферу ее применения, а также уровень ее потенциальной опасности для окружающей среды, здоровья, других фундаментальных прав граждан.

В свою очередь, центральным вопросом для применения технологического подхода является вопрос о необходимости и пределах регулирования сферы разработки и применения технологий искусственного интеллекта. Во-первых, вмешательство в данную сферу не должно создавать препятствий для развития технологий и инноваций.

Во-вторых, естественная реакция регулятора в ответ на появление новых объектов и субъектов оборота – «синдром несовершенного закона». Вместе с тем ложное представление об отсутствии правового регулирования может дать обратный эффект – дублирование правовых норм. В целях решения проблемы дублирования законодательных требований следует прежде всего решить вопрос о том, необходимо ли регулировать технологии искусственного интеллекта или некоторые виды использования программного обеспечения.

Научная новизна: проведен обзор основных подходов к регулированию отношений в сфере разработки и применения технологий искусственного интеллекта, выявлены возможности и ограничения их применения, предложены дальнейшие направления их развития.

Практическая значимость: основные положения и выводы исследования могут быть использованы для определения оптимальных подходов к регулированию сферы цифровых технологий, а также в целях совершенствования правового регулирования рассматриваемой области общественных отношений.

Для цитирования

Ерахтина, О. С. (2023). Подходы к регулированию отношений в сфере разработки и применения технологий искусственного интеллекта: особенности и практическая применимость. *Journal of Digital Technologies and Law*, *1*(2), 421–437. https://doi.org/10.21202/jdtl.2023.17

Список литературы

Алексеев, А. О., Ерахтина, О. С., Кондратьева, К. С., Никитин, Т. Ф. (2020). Подходы к гражданско-правовой ответственности разработчика технологий искусственного интеллекта: на основе классификации технологий. *Информационное общество*, *6*, 47–57. https://elibrary.ru/yldbab

Балашова, А. И. (2022). Искусственный интеллект в авторском и патентном праве: объекты, субъектный состав правоотношений, сроки правовой охраны. *Журнал Суда по интеллектуальным правам*, *2*(36), 90–98. EDN: https://elibrary.ru/apldua

Василевская, Л. Ю., Подузова, Е. Б., Тасалов, Ф. А. (2021). *Цифровизация гражданского оборота: правовая характеристика «искусственного интеллекта» и «цифровых» субъектов (цивилистическое исследование)* (в 5 т.). Москва: Проспект. https://elibrary.ru/nrjkdo

Войниканис, Е. А., Семенова, Е. В., Тюляев, Г. С. (2018). Искусственный интеллект и право. Вызовы и возможности самообучающихся алгоритмов. *Вестник Воронежского государственного университета. Серия: Право*, *4*(35), 137–148. https://elibrary.ru/yumlnz

Михалева, Е. С., Шубина, Е. А. (2019). Проблемы и перспективы правового регулирования робототехники. *Актуальные проблемы российского права*, *12*(109), 26–35. https://doi.org/10.17803/1994-1471.2019.109.12.026-035

Понкин, И. В., Редькина, А. И. (2018). Искусственный интеллект с точки зрения права. *Вестник РУДН. Серия: Юридические науки*, *22*(1), 91–109. https://doi.org/10.22363/2313-2337-2018-22-1-91-109

Савельев, А. И. (2016). Направления эволюции свободы договора под влиянием современных информационных технологий. В сб. М. А. Рожкова (рук. авт. кол. и отв. ред.), *Свобода договора*. Москва: Статут. https://elibrary.ru/xxoolt

Calo, R. (2011). Open robotics. *Maryland Law Review*, 70.3, *101–142*.

Ellul, J., Pace, G., McCarthy, S., Sammut, T., Brockdorf, J., & Scerri, M. (2021). Regulating artifcial intelligence: a technology regulator's perspective. In *Proceedings of the Eighteenth International conference on artifcial intelligence and law* (pp. 190–194). https://doi.org/10.1145/3462757.3466093

Ellul, J. (2022). Should we regulate Artificial Intelligence or some uses of Software? *Discover Artificial Intelligence*, *2*(5). https://doi.org/10.1007/s44163-022-00021-9

Erdélyi, O. J., & Goldsmith, J. (2018). Regulating artifcial intelligence: proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 95–101).

Gellert, R. (2021). The role of the risk-based approach in the General data protection Regulation and in the European Commission's proposed Artificial Intelligence Act: Business as usual? *Journal of Ethics and Legal Technologies*, *3*(2), 15–33.

Gonçalves, M. E. (2020). The risk-based approach under the new EU data protection regulation: a critical perspective. *Journal of Risk Research*, *23*(2), 139–152. https://doi.org/10.1080/13669877.2018.1517381

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., & De Hert, P. (2022). Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Comput Intell Mag.*, *17*(1), 72–85. https://doi.org/10.1109/mci.2021.3129960

Leenes, R. (2019). *Regulating New Technologies in Uncertain Times*. https://doi.org/10.1007/978-94-6265-279-8_2

Mogul, J. C. (2006). Emergent (mis) behavior vs. complex software systems. *ACM SIGOPS Oper. Syst. Rev.*, *40*(4), 293–304. https://doi.org/10.1145/1218063.1217964

Scherer, M. U. (2016) Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, *29*(2), 353–400. https://doi.org/10.2139/ssrn.2609777

Smuha, N. A. (2021). From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law. Innov. Technol.*, *13*(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Wagner, B. (2018). Ethics as an escape from regulation: from ethics-washing to ethics-shopping. In *Being profling: cogitas ergo sum* (pp. 86–90). Amsterdam: Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

Zielke, T. (2020). Is artificial intelligence ready for standardization? In *European conference on software process improvement* (pp. 259–274). Springer.

Сведения об авторе

Ерахтина Ольга Сергеевна – кандидат юридических наук, доцент, доцент кафедры гражданского и предпринимательского права, Пермский филиал Национального исследовательского университета «Высшая школа экономики»
Адрес: 614070, Российская Федерация, г. Пермь, ул. Студенческая, 38
E-mail: oerahtina@hse.ru
ORCID ID: https://orcid.org/0000-0002-9041-3487
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/K-3149-2014
Google Scholar ID: https://scholar.google.ru/citations?hl=ru&user=WdwWB4kAAAAJ
РИНЦ Author ID: https://www.elibrary.ru/author_items.asp?authorid=498773

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.07.45 / Право и научно-технический прогресс
Специальность ВАК: 5.1.1 / Теоретико-исторические правовые науки

История статьи

Дата поступления – 10 февраля 2023 г.
Дата одобрения после рецензирования – 23 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.18

Problems and Prospects of Regulating Relations within a Deal Effected with Participation of Artificial Intelligence

Dmitriy A. Kazantsev

B2B-Center
Moscow, Russian Federation

Keywords

Artificial intelligence,
automation,
data,
deal,
digital technologies,
digitalization,
law,
liability,
procurement,
robot

Abstract

Objective: to research the problem of determining the subject of a legally relevant act effected with participation of artificial intelligence, as well as distribution of responsibility for the consequences of its performance.

Methods: to illustrate the problematic and practical significance of the issue of legal personality of artificial intelligence, we chose automated procurements for public and corporate needs; the methodological basis of the research is the set of methods of scientific cognition, including comparison, retrospective analysis, analogy, and synthesis.

Results: by the example of the sector of competitive procurements for public and corporate needs, the evolution of automation of economic relations up to artificial intelligence introduction was analyzed. Successfully tested reactions to the challenges of stage-by-stage introduction of digital technologies into economic relations were demonstrated, as well as the respective modifications of legal regulation. Based on the current level of technological development, the prospective questions are formulated, associated with the legal regulation of economic relations implemented with the use of artificial intelligence, first of all, the question of defining the subject of a deal effected with participation of artificial intelligence. As an invitation for discussion after analysis of jurists' conclusions about the probable variants of the legal status of artificial intelligence, the author proposes

variants of answers to the question of its legal personality when effecting a deal. To solve the issue of responsibility for the decisions resulting from the implementation of algorithms of a software and hardware package, we propose several models of distributing such responsibility among its creator, owner, and other persons, whose actions might influence the results of such an algorithm functioning. The proposed conclusions may be used to develop normative regulation both as a set and individually.

Scientific novelty: based on the analysis of evolution of the practices of using digital technologies in procurement, the work formulates potential legal problems, determined by the constant automation of economic relations, and proposes legal constructs to solve such problems.

Practical significance: the conclusions and proposals of this work are of prospective significance for conceptual comprehension and normative regulation of electronic procurement tools both at corporate and national level.

For citation

Kazantsev, D. A. (2023). Problems and Prospects of Regulating Relations within a Deal Effected with Participation of Artificial Intelligence. *Journal of Digital Technologies and Law, 1*(2), 438–463. https://doi.org/10.21202/jdtl.2023.18

Contents

Introduction

Artificial intelligence becomes increasingly demanded in various spheres. The topicality of artificial intelligence is a logical and inevitable consequence of broad diffusion of digital technologies.

Digital technologies have long been used not only for everyday communication, but also for accompanying and today also for registering economic relations. Procurements in the corporate and public segments are a vivid example of rapid broadening of the sphere of information technologies application.

Today, however, digitalization rises to a new level: software and hardware package solves tasks and implement actions which were traditionally assumed as a human prerogative. These include creation of copyright items and actual contracting.

Such penetration of artificial intelligence into economic and then legal relations inevitably poses legal questions. For example, who is the author of a text created by artificial intelligence? Who is liable for the consequences of a deal, the legal existence of which is generated by the consequences of implementation of software and hardware package algorithms?

These issues acquire special topicality due to the fact that both digitalization in general and using artificial intelligence in particular are by no means a transitory tribute to fashion. Rational and consequential introduction of these technologies is actually capable of improving the efficiency of both production processes and business processes as a whole.

In this regard, one may expect only expanding and intensification of digital technologies. Hence, the legal science cannot evade the emerging questions. "This said, jurists are not obliged to apprehend the mathematical and technical secrets of digitalization; digitalization is not a subject of legal science. We have to write about it as many of those who devoted their research to digitalization ignore the fact that disciplines are divided into technical and social ones; legal sciences are social disciplines, and technical norms are not a subject of their analysis" (Lazarev, 2023).

Apparently, electronic technologies, like any other technologies, have limited areas of effective use. But, for example, competitive procurements B2B (business-to-business) and B2G (business-to-government) is the sphere where these technologies provide a qualitative increase of performance efficiency: digital technologies help to spend mere hours for the business processes which took weeks in the traditional paradigm.

Application of software and hardware packages is not at all limited to competitive procurements. However, it is this sphere that allows vividly illustrating both the stage-by-stage evolution of the digital technologies introduction which has logically led to using AI, and the accompanying legal agenda. Making no pretence of a comprehensive research of using automation technologies in economic activity, in this article we will consider, by the example of electronic competitive procurements, the issues posed before legal science by the synthesis of business and digital technologies.

Both in the B2G and, especially, in the B2B segment, during the first quarter if the 21st century competitive procurements rose from introduction of electronic document flow to testing the artificial intelligence technologies. This path comprised not only numerous experiments but also several qualitative transitions, each of which demands, at least, a comprehensive analysis from both organizational and legal viewpoints.

Such analysis provides an opportunity to trace the key transformations, each of which cardinally changed the role of digital technologies in procurement activity. The transformation

associated with the introduction of AI into routine business processes, which is currently taking place in this sphere, is not limited to the technology issues but largely refers to legal issues.

This said, it is important to remember that the previous transformations also did not go without formulating and introducing new legal constructs. Hence, we may rely on this experience to answer new questions.

1. Five digital transformations of competitive procurements

In one form or another, competitive procurements exist during more than one century. Before the end of the 20th century, they were conducted, with rare exceptions, without using digital technologies. However, in the first quarter of the 21st century, the use of digital technologies in procurement not only became widespread but also caused raid change in the quality of using these technologies.

Despite rapid digitalization, in general one may assert its evolutionary character. This, in turn, allows us to trace the stages of the digital technologies introduction into procurement domain. For clarity, we present this evolution as five transformations, the result of each being a cardinally new role of digital technologies in procurement business processes.

It is also importance to specify these transformations due to the fact that in practice it is extremely difficult to pass onto a new level of digital technologies introduction without a deep and consistent implementation of the previous level in routine procurement work. In other words, each previous transformation serves as a necessary condition for the next one.

1.1. Electronic procurements

The first digital transformation of competitive procurements is transfer of such procurements into electronic form.

An electronic competitive procurement can be in the very first approximation defined through a technological feature as a formalized procedure of a competitive choice of a counteragent, within which all relations between the procurement organizer and participants, including legally relevant document flow and real time bidding, are implemented exclusively in the Internet without using paper documents.

Apparently, not only bidding in the Internet but even a mere refusal of paper documents cannot be implemented without a “zero” step of introducing electronic document flow by economic subjects. For the electronic document to be recognized as legally relevant as the paper one, one needs a mechanism of certifying the legal relevance of an electronic document.

Such mechanism is electronic signature, which allowed sending a bargain offer as a file and guaranteeing it the status of a full-fledged offer in the civil-legal sense. At the beginning

of 2002, the Law "On electronic digital signature"[1] was adopted, and in 2011 it was substituted for the currently valid Law "On electronic signature"[2].

The issue of the types and order of using electronic signatures is beyond the frameworks of this article. For the topic under study, it is important that electronic signature assigns legal relevance to not only electronic document but also electronic information. In particular, such signature may certify an electronic image of a paper document or even a figure indicated as an auction rate. This, in turn, opens space for full-fledged bidding in electronic form.

The first electronic procurements were conducted by corporate clients in 2002, and by the end of the decade electronic auctions were also introduced for public clients[3]. Although the wide range of electronic procurement tools is not limited to auctions – for example, contests, request for proposals, and requests for quotation can also be conducted electronically, – but the novel on using electronic auctions for public procurement became pivotal for official recognition of electronic procurement technologies. A separate discussion should be devoted to the issue whether it is methodologically and practically expedient to select only electronic auction among the whole specter of available electronic tools, as the said auction dominates in the current Law "On contractual system"[4].

Today, the electronic form is a norm for both public and commercial procurement. However, in practice the notion "electronic procurement" is interpreted by clients in varied ways. Some view it as publication of procurement documentation in the Internet and then collection of offers on electronic carriers. Others consider it to be a formalized procedure at a specialized platform called an electronic trading platform.

The chronological frameworks of using electronic technologies in procurement are also different. Traditionally, to be called an electronic procurement, it is sufficient to conduct on an electronic trading platform the part of business processes starting from a publication of the notion of procurement and finishing with an announcement of the procurement results.

However, in practice the complex of procurement business processes is much larger. It starts, long before announcing the procurement, with formalized work with the need for which the procurement is necessary, and the results of this work are included into the procurement plan. The procurement process finishes, much later than the choice of winner,

1 On electronic digital signature: Federal Law of 10.01.2002 No. 1-FZ. *Collection of legislation of the Russian Federation*, 14.01.2002, No. 2. Article 127.

2 On electronic signature: Federal Law of 06.04.2011 No. 63-FZ. *Collection of legislation of the Russian Federation*, 11.04.2011, No. 15. Article 2036.

3 On placing orders for supplying goods, executing works, rendering services for public and municipal needs: Federal Law of July 21, 2005 No. 94-FZ. *Collection of legislation of the Russian Federation*, 2005, No. 30. Article 3105.

4 On contractual system in the sphere of purchasing goods, works, services to provide for public and municipal needs: Federal Law of 05.04.2013 No. 44-FZ. *Collection of legislation of the Russian Federation*, 08.04.2013, No. 14. Article 1652.

with registering the supply results, and in developed procurement practices – with analysis of the quality of the initial need satisfaction with the supplied product.

Technically, all these business processes can long be implemented in electronic form. As practice shows, given due implementation, their transition to electronic form is expedient. However, that requires the next transformation, which can be conditionally called digitalization of procurement activity.

1.2. Digitalization of procurement activity

Digitalization of procurement means transition into electronic form of all formalized procurement business processes from procurement planning to acceptance of the product supplied. Such transformation, as a rule, requires creating a special portal providing execution of business processes in the digital environment.

If digitalization of procurement activity is successful, then all the external and internal document flow, associated with preparation and execution of procurement, is implemented in electronic form. Business processes are implemented along the routes suggested by the algorithms of a special portal. This, in turn, systematizes the procurement interaction and simplifies the work of a person who acquires the opportunity to focus not on the bureaucratic but on the substantive part of their activity.

Digitalization of procurement activity almost always implies stage-by-stage introduction of new technologies into the established business processes. All stages of such introduction can be conditionally united into two blocks: creating a digital environment, in which business processes will take place after their digitalization, and further development of the digital system to provide implementation of business processes in the electronic environment (Kazantsev, 2022). Apparently, these blocks can be implemented only consequentially, but within each of the two blocks individual stages can be developed also in parallel.

Creation of the digital environment is not limited to creation of a specialized portal. This block requires also solving a range of organizational and legal-technical issues, such as:

1. Systematizing procurement business processes with distribution of authorities and responsibilities, exclusion of extra links and addition of lacking elements.

2. Regulating procurement business processes in documents mandatory for all subjects of business processes. Such documents should exclude both differing interpretations and dissolved responsibility, uncertainty of terms and other legal-technical defects.

3. Describing procurement business processes in the form of algorithms for clear visualization of regulating documents and simplification of further digitalization of the regulated business processes.

4. Creating software for implementing procurement algorithms, as well as an intuitive interface for using this software.

5. User testing and adjusting the functionality of software for implementing procurement algorithms.

6. Testing the created digital environment in a pilot project to minimize the costs of further introduction of software for all business processes.

7. Modernizing the digital environment as a result of the pilot project and its expanding to all procurement business processes.

In the most general terms, development of the procurement digital system consists in expanding the area of its use in economic activity of a company by graduate transfer of all business processes into electronic form.

However, one should bear in mind that digitalization of procurement activity, with all its obvious conveniences, generally means just implementation of old processes in a new environment. By a famous saying, digitalization of chaos generates just chaos in an electronic form. In other words, digitalization of procurement – and any other – business processes should be started with their adjustment and optimization. One should not expect efficiency from introducing digital tools if excessive reconciliation or, on the contrary, areas of nontransparent decisions are preserved.

For the content of procurement activity to be qualitatively changed due to introduction of digital technologies, digitalization is not enough. Automation is needed.

The notion of automation should be strictly distinguished from the notion of digitalization. Digitalization precedes automation and is its necessary condition, but is by no means equal to it.

1.3. Automation of document production

While digitalization is executing classical business processes in an electronic environment, automation is involving the capacities of a software and hardware package to performing a part of tasks within these processes. This is not about tasks like document routing or notifications about a maturity. Automation implies using software and hardware package, inter alia, to produce such documents.

In the most general terms, under such division of labor a human just uploads initial data and controls the results, while a software and hardware package processes the data and forms the results. An authorized employee is still responsible for the results, including in the legal sense. However, under proper implementation, introduction of such technologies qualitatively reduces the amount of a human's routine work and increases its expert constituent.

The first step of such automation is to form a database including a constructor of electronic documents with preset forms. These forms may look like text fields, fields of numbers or dates, lists, formulas, etc. With the document constructor interface, an authorized employee inputs the source information into the preset forms. Based on the set of such information, the software and hardware package first selects the relevant form, and then fills in the variables in that form.

It is competitive procurements, due to their organizational features, that are a promising environment for automation of the document production processes. Normally, procurement

is carried out along a formalized, pre-regulated procedure, and the number of variants as templates of procurement documentation, just as templates of applications for participation in procurement is limited and, as a rule, rather small.

This creates an opportunity for using document templates, placed in the digital environment, for procurement automation, and in developed electronic systems – also filling in of these templates using the technology of data inheritance. Data inheritance from the viewpoint of business processes is understood as the possibility of automated prefilling of forms of documents, contracts, etc. Actually, a client preparing procurement only has to indicate, in the documentation "template", the object of procurement, requirements to the product purchased, and the terms of the future supply.

With the document constructor, a form is created, which is necessary for the definite procurement procedure, and with the data inheritance, the variable fields are filled in.

In the most simplistic terms, the system of data inheritance in competitive procurement can be presented as follows:

1. The demand parameters are formulated as an electronic application for procurement.
2. The parameters of an agreed application are included into the procurement plan.
3. The data from the procurement plan are "pulled through" to the standard procurement documentation, which thus turns into a project documentation of a specific procurement.
4. The norms of procurement documentation are translated into the draft contract.
5. Contract terms are reflected in the closing documents.

Generally speaking, the current architecture of the digital environment of public procurements, first of all, the Unified information system in the sphere of procurements, is built along these approaches.

To effectively use such electronic system, a user should only have professional knowledge in their field. In other words, automation is only good if it releases an employee's time but not makes one spend this time to study the order of using the functionality of the electronic portal.

That is why, inter alia, the system and interface design is practically one of the key factors of automation. In somewhat simplistic terms, one may formulate this principle as follows: the system is the better, the fewer keys a person must click to obtain a complex result.

The result of studying the data of previous documents must be prefilling of forms of further documents. For example, the title and amount of purchased goods have been established within the agreed application; in this case, the data inheritance technology must provide automated prefilling of the variables "title" and "amount" in the procurement plan, procurement documentation, draft contract, and draft acceptance documents. This said, data prefilling must leave an opportunity for a human to edit them because, as was mentioned above, it is a human that is ultimately responsible for the result.

Thus, after introduction of the automation tools of documents production, the source data, including the procurement parameters and variables, are specified by a human. Drafts of documents are created by the software and hardware package. As for legal relevance,

the documents acquire it through signing by an authorized employee with an electronic signature.

This said, the potential of automation in procurements is neither theoretically nor practically limited to producing the procurement documentation. We already see successful examples of automation of individual business processes, related to procurements.

1.4. Automation of business processes

Automation of business processes can be defined as providing their implementation using the algorithms of a software and hardware package. As a rule, this requires substantial organizational remodeling of these business processes while preserving their essence. After procurement automation, a human makes decisions only at key stages.

The automation opportunities allow implementing procurements also along completely new procedures, qualitatively reducing organizational costs of all parties of the procurement process while maintaining competitiveness and transparency of procurement per se (Kazantsev & Mikhaleva, 2020). The so called dynamic procurements are gaining popularity today, in which collection of suppliers' proposals and their ranking is de facto automated. This is the example of how automation requires organizational reconstruction of the traditional business processes.

For example, in a classical competitive procurement, the client first publishes a notification of procurements, than waits for one or several weeks for the suppliers to submit their proposals, then estimates and ranks these proposals, and only then formalizes the procurement results. Automation in dynamic procurement requires a different process: first the suppliers publish their proposals in a specialized portal and sign then with an electronic signature as offers addressed to an uncertain circle of potential clients; then the client, when procurement is needed, specifies the parameters of their demand in the relevant fields of the specialized portal interface; after that, the portal's software and hardware package performs an automated search for relevant offers among those uploaded by suppliers, immediately ranking them by the parameters set by the client.

Today, new variants of this model appear in the form of electronic market places and corporate Internet stores. These variants may sometimes have their own essential features: for example, a preliminary proposal may not have the status of an offer and choosing it means not an acceptance but just an invitation for negotiations. But even such specificity does not revoke the basic scheme of automation described above.

A tool for rapidly purchasing goods with automated selection of potential proposals is currently regulated by Part 12 of Article 93 of the Law "On contractual system"[5] as a full-fledged means of finding a supplier (agent, executor). As a recommended tool, it also

[5] On contractual system in the sphere of purchasing goods, works, services to provide for public and municipal needs: Federal Law of 05.04.2013 No. 44-FZ. Collection of legislation of the Russian Federation, 08.04.2013, No. 14. Article 1652.

occupied its position in procurements of public corporations and natural monopolies subjects[6]. As early as in 2014, a similar tool called "dynamic purchasing" was mentioned in the EU Directive "On public procurement"[7], and in 2016 it became nearly the main means of choosing a supplier in the Procurement Code of Italy[8].

Dynamic procurement is just one of the variants of the procurement business processes automation. Strictly speaking, in this case automation refers only to choosing the winner. It actually allows qualitatively reducing temporal costs. However, in competitive procurements, a deeper variant of automation may be implemented.

Such variant of automation is called end-to-end automation. It is based on the data inheritance technology described above, but is not limited to it. Schematically, it can be described as follows:

1. A client forms a database with libraries of standard procurement documentation, draft contracts, etc.

2. The database is integrated into a specialized portal together with categorical strategies of the client.

3. When a need for procurement occurs, an authorized employee specifies the key parameters of this need in the relevant fields of the specialized portal interface.

4. Based on the key parameters, the software and hardware package determines the categorical strategy.

5. Within this strategy, automated collection and processing of information are carried out, including the information about the market competitive condition, presence of qualified suppliers, market prices for the goods; then the most effective way of procurement is determined and a package of procurement documentation is formed.

6. An authorized employee approves drafts or inputs additional data to specify the results.

7. Based on the documentation approved by the authorized employee, the software and hardware package controls the terms and publishes or sends notifications about the procurement in the set time.

8. Depending on the chosen way of procurement, the software and hardware package selects proposals, previously published by suppliers for participation in the dynamic procurement, or sends invitations to qualified suppliers to submit their proposals.

9. After proposals are collected, the software and hardware package checks information about the suppliers in open sources, estimates the consistency of each proposal with the client's need and ranks the proposals based on the criteria indicated by the client.

---

6 On the features of participation of small and middle business entities in purchasing goods, works, services by certain types of legal persons: Decree of the Russian Government of 11.12.2014 No. 1352. *Collection of legislation of the Russian Federation*, 22.12.2014, No. 51. Article 7438.

7 Directive 2014/24/EU of the European Parliament and of the Council of 26 February 2014 on public procurement. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0024

8 *Codice dei contratti pubblici* (Decreto legislativo 18 aprile 2016, n. 50). https://www.bosettiegatti.eu/info/norme/statali/2016_0050.htm

10. The client accepts the ranking and chooses a winner, indicating additional data for a new procurement.

11. If a winner is announced, the software and hardware package compiles a draft contract and sends it first to the client for approval and then to the supplier for signing.

12. Based on the parameters of the contract signed, the software and hardware package controls the terms and volumes of the supply, including sending reminders about shipment and acceptance.

13. As a result of the supply, based on the contract data and the information uploaded by the client about the factual contract execution, the software and hardware package forms drafts of closing documents and sends them to the client and the supplier for signing (Kazancev, 2020).

This scheme comprises all the main stages of procurement work and allows using all the main means of procurement. It is important to note that a human is not removed from procurement work, as it is their electronic signature that assigns legal relevance to the key decisions and documents. But preparation of these decisions and documents, their routing, communications with the contractor and "bureaucratic' elements of work are all implemented by the software and hardware package.

In this paradigm, not only labor costs of the authorized employee are reduced and the business processes are accelerated. Implementation of these processes by software and hardware package reduces the value of the so called human factor and bias in making decisions. A human being able to make amendments in the documents prepared by the software and hardware package does not negate the fact that each of these amendments is registered in the document history, which, in turn, simplifies retrospective control.

End-to-end automation creates conditions for forming the S2P system. Source-to-pay is applied automation of the whole complex of procurement work, including all interactions between a client and a supplier[9]. In this system, all business processes, related to procurement, take place in the digital environment and are implemented, inter alia, using artificial intelligence.

If properly integrated into the S2P system, artificial intelligence is capable of significantly broadening the potential of automation and individualizing the procurement scenarios depending on the unique parameters combining the client's demands and the market of the goods purchased. In addition to the basic automation scheme described above, artificial intelligence – based on the analysis of big data both from open sources and its own archives of previously executed procurements – is capable of performing, at least, the following functions:

– forecasting demands. To do that, AI analyzes the client's activity from the viewpoint of necessary resources, monitors the volume and contents of stock reserves, and execution

[9] McCann, Jo. *How source-to-pay works*. https://blog.routable.com/how-source-to-pay-works/

of framework contracts. If there is a risk that a resource is about to be exhausted and cannot be replaced without a procurement, then AI informs the client about it, and ideally also proposes the optimal parameters of the future procurement. For example, it takes into account the factual supply terms. The more time a supplier needs to deliver goods, the mote stock the producer has to keep[10].

– determining the most effective procurement tool. Generally, this function is an expansion of the function of determining basic parameters. Based on the analysis of information about the condition of the regional market of the goods to be purchased, AI can assess what will be more efficient – to announce an auction, to conduct a dynamic procurement, or to offer a single supplier. For example, an attempt to conduct an auction, which is doomed to failure due to the absence of competition, will just mean additional organizational and temporal costs without any result.

– performing an expanded check of qualification and economic behavior of the supplier. AI can not only collect the data of official registers but also obtain feedback from other clients, information on litigations, and sometimes even information about the capacities, equipment, technologies, personnel qualification, etc. All these data are important for effecting a contract. Without AI, the client would need a lot more time to collect even a part of these data.

– determining and controlling the supply logistics. AI does not only calculate the optimal route, its costs and risks, but also allows the client, under minimal technological integration with the supplier and their navigation system, to trace the cargo movement in real time.

Implementation of these additional functions requires of the software and hardware package to collect and process big data and to have machine learning mechanisms. These features allow us to speak, if not in the specific then, at least, in the general sense, of using artificial intelligence for procurement automation.

1.5. Automation of decisions

So far, one cannot speak of a consensus on such a seemingly fundamental (for this discussion) issue as artificial intelligence.

For example, it is suggested to interpret it as a system capable of physically manifesting itself, including feeling, processing and influencing the external environment to a certain extent (Calo, 2015). Such definition prioritizes the physical manifestation of the results of artificial intelligence functioning. However, in practice the results of its functioning may remain in the electronic environment, and only the impact of these results of legal relations with participation of physical and legal persons will show itself in the external world.

[10] Banker, Steve. (2022, April 1). One Multinational's Supply Chain Transformation Journey. *Forbes*. https://www.forbes.com/sites/stevebanker/2022/04/01/one-multinationals-supply-chain-transformation-journey/?sh=87ff9516229c

In the extreme, this approach is expressed in the concept of the so called strong artificial intelligence, which is understood as a technology identical to human conscience in mental properties and the character of processing available information, including in the aspect of comprehensive information interpretation, ability for creativity and intuition (Searle, 1990).

However, the most realistic today is the concept of artificial intelligence as a software and hardware package, having nothing in common with the human intelligence in the aspect of cognition essence, but capable of solving in general the tasks similar in complexity or more complex ones (Bokovnya et al., 2020). Assumingly, it is this approach that corresponds today to both the level of technologies achieved, and the spheres of practical application of artificial intelligence.

In the expanded variant suggested by N. N. Chernogor, the definition of artificial intelligence within the said approach looks as follows: "The technology determining the ability of an informational system to correctly interpret, without a direct participation of a human, the external data (external information), specify the database (databases) with the account of those data, to learn from the mistakes made and to use the knowledge obtained to achieve specific goals through flexible adaptation under an ill-defined situation" (Chernogor, 2022).

Even today, the opportunities of artificial intelligence are to a greater or lesser extent exploited by clients to manage stock volumes and to work with supply logistics. The volume of the processed data and the speed of processing combined make the use of AI objectively feasible. However, that poses new questions of legal character.

By the example of the previous four transformations we have shown that the regulation development did not precede the introduction of electronic technologies into procurement activity, but always accompanied it. First, it was novels of general regulation devoted to electronic document flow and electronic signature, hen novels of special legislation, regulating electronic means of procurement, then new bylaws devoted to rapid purchases based on automation patterns. State regulation was complemented and developed by corporate regulation – in the sphere of automation today, the largest amount of norms, as well as the depth of regulation, belongs not to laws and even not to departmental instructions but to corporate provisions and regulations.

This should be born in mind when discussing the issue of regulating the use of artificial intelligence in procurement work. AI is capable of not only analyzing stock reserves and delivery routes. The next step is choosing the optimal way of procurement based on the analysis of the data of market conditions. To do that, it is important not only to process big data, but also for the client to have multifactor categorical strategies of procurement management. This is because in order to determine the tools, big data about the market must be processed within a matrix of categorical strategies: AI collects and processes the data about whether the goods demanded by the client are sold in this region or delivered from neighboring regions, what the delivered price is, whether the characteristics comply with the client's demands, whether there are competing offers and whether the suppliers are ready for competition, etc.

However, if AI is capable of processing such data, then it can propose not only the optimal way of procurement, but also the winner candidature. And if one extrapolates this capability on the need for rapid procurement for continuous production, then the same AI based on the same data may autonomously send an application for delivery to the best supplier.

Even if AI sends the application not to a new supplier but to the one with whom the client had previously signed a framework agreement on delivery – such models of artificial intelligence integration into procurement are already used by some enterprises, – in this case too it is AI that forms the essential contract terms which generate a deal. It indicates the specific volume of the specific goods, which under the framework agreement automatically determines the delivery cost. Technically, even now artificial intelligence can do the same in the absence of a framework agreement previously signed between the client and the supplier.

Automated decision-making about the choice of a supplier and signing a deal with them is much quicker and more convenient than the classical tender. It completely eliminates bias and subjectivism in making a decision about the winner of procurement. But it also turns AI from a decision-making tool into a decision-making subject.

2. Legal personality of artificial intelligence

The cardinally changed role of the software and hardware package in economic relations associated with procurements poses cardinally new questions of legal character, which are no longer solved by the current regulation. The electronic document flow is already regulated, as well as the electronic procurements, conducted using the electronic document flow, and, p to a certain extent, digitalization of the whole process of procurement work, for which electronic procurement per se is just one of the stages. However, the status of artificial intelligence within digital procurement work is not regulated.

Today, this is not an issue of just theoretical value: in practice, the consequences of the artificial intelligence functioning lead to effecting deals and emerging contractual obligations. Moreover, one may assert that it is the emergence of contractual obligations as a result of artificial intelligence functioning that implies the genuine automation procurement work.

For example, the decision on choosing the procurement means refers directly to the client only (even if its consequences may indirectly touch upon the supplier). In this case, the software and hardware package, which offers the most effective procurement tool for the client, based on a set algorithm, performs though a comprehensive, but still auxiliary function. The subject of the decision who is responsible for its consequences is still an authorized employee of the client. Hence, complete automation of this business process is out of question so far.

However, a decision to send application for another batch of goods to the supplier places the software and hardware package into the position within economic relations, which actually allows saying that a robot signs a deal.

This said, and it is important, the software and hardware package does not "purloin" the authorities beyond the algorithm set to it. Its role of a deal subject follows exactly from the algorithm execution – because it is convenient for both the client and the supplier to use the robot to determine the terms, volume and object of delivery. This allows both the client and the supplier to release resources for the main production activity. But from the civil-legal viewpoint it is the results of the artificial intelligence functioning that generate mutual rights and obligations of the client and the supplier.

This circumstance poses the question about the legal personality of artificial intelligence per se, the answer to which is by no means obvious today. The results of a scientific discussion on this topic depend not only on the possible consensus, but also on the speed of the electronic technologies development.

The issue of the legal personality of artificial intelligence is solved based on the understanding of its essence, but is not limited to that. Hence, even if we do not attempt to find similarities between a human cognitive activity and artificial intelligence functioning, we can still find weighty arguments for endowing a software and hardware package with a legal personality, first, to clearly indicate its role in legal relations, and second, to protect the rights of other legal personalities.

This concept was most vividly expressed by E. V. Vavilin: "a legal personality of AI is necessary to restrict AI in its rights through a specific functional purpose" (Vavilin, 2021). However, he also did not mention that the legal personality of artificial intelligence can be similar with the legal personality of, for example, a legal person. The construct of AI legal personality is proposed exactly as a means to achieve the goal of limiting its rights. From this viewpoint, it is sufficient to endow AI with a specific set of rights and obligations, i. e. to stipulate a special technical legal personality.

However, this discourse leaves unsolved the issue of the practical implementation of such legal personality – in particular, the practical aspects of endowing responsibility. For example, how can one make AI reimburse for the damage incurred by the consequences of its functioning? Whatever legal fictions we use, de facto AI functioning in favor of one subject (for example, in favor of the subject to whom artificial intelligence incurred harm) means that it does not use a part of time or capacities to work in favor of another person (namely, its owner). Thus, compensation of harm in this situation is imposed of the AI owner, even if the compensation is performed using AI.

Under such circumstances, one should agree with V. K. Andreev that "application of digital technologies using artificial intelligence at the modern stage of its development does not mean advent of new public relations, qualitatively differing from the existing ones", and "artificial intelligence does not act as a digital legal personality within the relations of digital rights circulation in the informational environment of the operator. The latter, using digital technologies in entrepreneurship, applies elements of artificial intelligence in business models, not generating digital legal relations" (Andreev, 2021).

Indeed, it is the owner, or the operator, of artificial intelligence that serves as the subject ultimately influencing the results of AI functioning and uses the results of its functioning. Hence, it is logical to imply their responsibility for these results.

Indeed, under any degree of digital technologies penetration and any model of using artificial intelligence, today it is hard to imagine AI acting completely independently from a human and assuming the rights and obligations, neither directly not indirectly stemming from the will and actions of the human. Hence, artificial intelligence today should be referred rather not to new legal subjects but to innovative tools requiring new legal regulation.

"The relations using artificial intelligence are always relations between legal subjects and about object of law. In any case, these are relations which at one stage or another initiated, programmed by a human – a legal subject with various levels of responsibility (including within the frameworks of legal persons' activity). Expression of will of a human for certain actions of artificial intelligence may be expressed in various degrees: from AI actions fully controlled by a human will to autonomous actions of AI, also allowed and comprehended in their probable limits and consequences by a human (a group of humans)" (Shakhnazarov, 2022).

The degree of influence of the human expression of will on the results of artificial intelligence functioning must be taken into account when solving the issue of distributing responsibility for AI actions between the current legal subjects, which will be discussed below in more detail. The issue of the legal personality of artificial intelligence per se at the present stage of technological, as well as public and economic, development, one may rely on the position by S. E. Channov: "Endowing robots (artificial intelligence systems) with the status of a legal personality will not entail any explicitly negative consequences in the foreseeable future. At the same time, one can see no advantages of such decision compared to viewing robots (artificial intelligence systems) as quasi legal personalities. Stemming from the philosophical Occam's razor principle not to multiply entities beyond necessity, we believe that introduction in the legal sphere of such a conceptually new legal personality as a robot (artificial intelligence system) is premature (although one cannot exclude that such necessity will emerge)" (Channov, 2022).

3. Legal status of a deal effected with participation of artificial intelligence

The conclusion that artificial intelligence today cannot be recognized as a legal personality does not remove, but actualizes the issues associated with the status of a deal effected using artificial intelligence. Such deals are already effected today, and their number will increase in the future.

What is the status of such a deal? What is the mechanism of its judicial protection? On what grounds can it be contestable? These and other questions are still to be answered by theorists of law. It is these answers that will become the basis of the "law of electronic relations" (Kenney & Zysman, 2016), which is being written about today. Assumingly, the future of legal regulation belongs to forming the law of electronic relations as a separate and integral branch of legal knowledge, not to endowing the status of a legal personality to artificial intelligence.

While endowing the status of a legal personality to artificial intelligence is premature, it is still necessary to solve the question of who is the subject of a legal relation implemented with participation of AI. Indeed, using AI personality per se should not lead to the deal contestability, to say nothing of voidness. Because if we allow presumption of a deal invalidity on the grounds of using AI technology during its execution, then this will inevitably create unjustifiable risks for the existing economic relations, the digitalization and automation of which was thoroughly discussed above. As any AI is backed by a physical or legal person, it is they that should share both the legal consequences and responsibility for artificial intelligence functioning.

The exact parameters of distributing responsibility should be elaborated by legal science during forming the law of electronic relations. As an invitation for discussion, we can suggest several variants of solving the question of responsibility for the consequences of the software and hardware package functioning within economic relations with a high degree of automation:

1. Owner's responsibility: whatever the results, responsibility for the AI actions and decisions shall be borne by the legal or physical person to whom the software and hardware package belongs or by whom it is legally used.

By a brilliant definition of S. F. Afanasyev, under such approach "the legal position of AI becomes identical or close to that of the Roman anthropomorphic collective organizations, or even more reduced – a slave, family members, children, including 'filius in potestate tua est' " (Afanasyev, 2022).

Such approach is intuitively comprehensible and, at first glance, maximally logical and utilitarian. However, it ignores the fact that the results of artificial intelligence functioning by no means always depend on the consequences of the expression of the will, actions or inaction of its owner.

For example, adverse results of AI functioning may be a consequence of not only its defective use, but also a consequence of its defective design. A legal construct seems dubious if it imposes on the owner responsibility for the functioning of a software code, analysis of which requires special and profound expertise – such as a bona fide user of the software and hardware package, as a rule, cannot and is not obliged to have.

2. Creator's responsibility: if a bona fide owner did not foresee the adverse consequences of the robot's actions and could not influence them, then it is logical to impose the obligations due to the harm inflicted not on them but on the developer of the algorithm, implementation of which caused the harm inflicted.

This should be highlighted: we are not talking about removing the AI owner from responsibility for the consequences of the AI functioning. We are only talking about the distribution of such responsibility as a result of proving guilt. For example, the owner's responsibility is presumed, but if their guilt was not proved, then the issue of identifying the developer's guilt is initiated.

However, this approach can be logically continues also in the case when the creator's guilt is not proved. If the results of AI functioning were influenced by a third person, for example, by modifying it or providing incorrect data for processing, then that third person will be liable.

3. Comprehensive approach: adverse consequences of AI functioning will be the responsibility of the person whose action or inaction ultimately caused the harm inflicted. In practice, the cause of adverse consequences of AI functioning within procurement business processes can be the developer's actions, the actions of the client as the software and hardware package owner, and even the actions of the supplier as its user.

The foreign doctrine considers even the issue of implementing the concept of criminal liability for the consequences of artificial intelligence functioning, which would take into account the actions and inaction of the developer, owner, users and other persons related to AI or having influenced the results of its functioning (Hallevy, 2013).

Actually, artificial intelligence makes decisions based on collecting and processing information, including open source information. Hence, if one of such sources contains a critical amount of incorrect information, then it may become the cause of artificial intelligence error and respective adverse consequences.

To distribute responsibility in this case, one may legislatively stipulate, for example, the presumption that the subject of obligations, emerging as a result of normal software and hardware package functioning, is its owner. It is they that will be obliged to prove the need to impose responsibility on another subject. This idea was proposed, in particular, by V. A. Laptev, who wrote about a subsidiary responsibility of the developer, owner, and user of artificial intelligence (Laptev, 2019).

This specter of solutions is not closed. One may construct other models of legal consequences for the relations emerging as a result of artificial intelligence functioning.

For example, another legal construction may be inclusion of AI into the sources of increased danger. According to A. A. Antonov, "legislative stipulation of new AI systems will require considering the issue of recognizing them as a source of increased danger, as an owner AI can be forced to compensate for the damage inflicted through court action only" (Antonov, 2020).

An alternative approach, also discussed in the foreign doctrine, is to endow artificial intelligence with a legal personality of a legal person with a similar distribution of responsibility for the consequences of its functioning (Chesterman, 2020).

In any case, regulation of the issue of a legal personality in the relations implemented using AI, as well as the issue of distributing responsibility in these relations, requires special normative regulation. This does not imply immediate changes in the legislation: regulatory tests may be performed through agreements, including norms of the AI status, between persons participating in such legal relations. After all, the very technologies of digitalization and automation of procurements, chosen as illustrative material for this article, were created and tested not in public but in commercial segment.

Conclusions

Not only we are witnessing the formation of informational law, but it may become a separate branch of law in the future – the very branch on which a significant part of economic relations is based. What is even more important, without formulating specialized norms of informational legislation in these relations, each year it will become harder and harder to speak of legal regulation. In other words, huge resources will be spent in economic activity without the possibility to accurately define mutual rights and obligations.

The experience of digitalization of B2B and B2G competitive procurements may help us trace the evolution of digital technologies penetration in the economic activity and the practice of introducing new legal mechanisms to solve the emerging problems. This experience shows that emergence of cardinally new technologies never means the principal impossibility of their regulation. Apparently, such regulation should reflect, but not substitute technological development.

This said, even in such specific sphere as competitive procurements, the ubiquitous penetration of digital technologies does not imply its evenness. In other words, while some clients undergo just the first digital transformation and still cannot abandon a paper contract with a handwritten signature, others constantly use automation procurement business processes and work out the introduction of artificial intelligence technologies.

Such unevenness does not depend on the sector specificity: the first digital transformation may become “a stumbling block” for large organizations in extremely knowledge-intensive industries, while the fifth digital transformation can be successfully tested by quite ordinary services. This returns us to the idea that procurement regulation in the aspects of digitalization is not a derivative from sector regulation, but is an independent and largely isolated range of issues, possessing an object unity. It is essential to tale this specificity into account when developing normative regulation of economic relations using artificial intelligence.

Digital transformation is not limited to using new software. It also includes restructuring business processes and modernization of legal regulation. For example, “digitalization of public procurements is not just an issue of acquiring the most advanced technologies. It also requires changing the tools and means of procurement, which would allow the state to interact with the new technologies and effectively and rapidly integrate them into practice” (Shmeleva, 2019).

A logical continuation of the digital transformation of procurement work is its automation, including using artificial intelligence technologies. Using artificial intelligence does not mean removal of a human from procurement work. It just means increasing the expert level of this work. In other words, with introduction of artificial intelligence, the attention of a procurements specialist will become more and more focused on nonstandard, specific and especially important situations (Kazantsev, 2021).

This said, artificial intelligence per se remains so far a tool, not a subject of procurement work, even if decisions within this work are made by the software and hardware package without direct participation of a human. "From the ontological viewpoint, all advanced technologies are not subjects but objects, and there are no juridical grounds to make them legally liable. Event in the light of the existing rules of legal liability based on various legal criteria, it is always theoretically possible to identify the person who will be liable for the harm inflicted as a result of production or exploitation of a device with AI system" (Ivliev & Egorova, 2022).

As an invitation for discussion we may propose several models distributing responsibility within economic relations with a high degree of automation. It would be more correct to call them not models but points for formulating legal constructs.

For example, one may stem from the fact that a subject of such a deal de jure is a person who signed the framework agreement, for the implementation of which the order formalized by the robot was created. Thus, the responsibility for the decisions "made" by a robot to execute the algorithm is imposed on a human. But one cannot but notice that this is largely just a temporal measure. Even today, artificial intelligence technically may not just send the applications within the frameworks of previously signed agreements but also sign new independent agreements. In other words, acting as a subject of legal relations, a robot does not need a previously signed framework agreement.

Another approach stipulates that the owner is responsible for all actions of a robot. In other words, whatever the result of the algorithm execution, all actions and decisions of the robot will be a responsibility of the legal or physical person to whom the given software and hardware package belongs or by whom it is legally used. Essentially, such approach was worked out by Roman jurists for the cases of slave owning relations

One may link the responsibility for the consequences of the artificial intelligence functioning with the guilt of the subject, whose action or inaction influenced the occurrence of the definite result of the software and hardware package functioning.

Each approach is imperfect and deserves being discussed. The future task is to elaborate a solution based on these approaches but not limited to any of them. However, solutions in the sphere of legal regulation of the said issues are indispensable.

Informational law becomes an increasingly demanded branch of legal knowledge (Scassa, 2018). In the future, it almost certainly will become a separate branch of law. This branch of law will have to deal with relations in which technologies and business practices are of priority significance for elaborating legal norms. In other words, the new norm should harmonize, not create the digital reality.

This means that the new legal constructs also must adequately reflect and regulate the existing relations implemented, inter alia, with automation tools. It is this approach that makes feasible the task of legal regulation to minimize the risk of useful tools being misused by malevolent subjects.

References

Afanasyev, S. F. (2022). On the problem of substantive and procedural legal personality of artificial intelligence. *Herald of Civil Procedure*, *3*, 12–31. (In Russ.).

Andreev, V. K. (2021). Acquisition and implementation of rights of a legal person using artificial intelligence. *Predprinimatelskoe pravo*, *4*, 11–17. (In Russ.).

Antonov, A. A. (2020). Artificial Intelligence as a Source of Increased Danger. *Yurist*, *7*, 69–74. https://doi.org/10.18572/1812-3929-2020-7-69-74

Bokovnya, A. Y. et al. (2020). Legal Approaches to Artificial Intelligence Concept and Essence Definition. *Revista San Gregorio*, *41*, 115–121.

Calo, R. (2015). Robotics and the New Cyberlaw. *Californian Law Review*, *103*(3).

Channov, S. E. (2022). Robot (Artificial Intelligence System) as a Subject (Quasi-Subject) of Law. *Actual Problems of Russian Law*, *17*(12), 94–109. (In Russ.). https://doi.org/10.17803/1994-1471.2022.145.12.094-109

Chernogor, N. N. (2022). Artificial intelligence and its role in the transformation of modern law and order. *Journal of Russian Law*, *4*, 5–15. (In Russ.). https://doi.org/10.12737/jrl.2022.037

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International and Comparative Law Quarterly*, *69*(4), 819–844. https://doi.org/10.1017/s0020589320000366

Hallevy, G. (2013). *When Robots Kill: Artificial Intelligence under Criminal Law*. University Press of New England.

Ivliev, G., & Egorova, M. (2023). Legal Issues of the Legal Status of Artificial Intelligence and Products Created by Artificial Intelligence Systems. *Journal of Russian Law*, *26*(6), 32–46. (In Russ.). https://doi.org/10.12737/jrl.2022.060

Kazancev, D. (2020). *A competitive procurement. Methodology and regulation*. Moscow: INFRA-M. (In Russ.). https://doi.org/10.12737/1068790

Kazantsev, D. A. (2021). Procurement and machine learning. Is there space for technology application? *Goszakaz: upravlenie, razmeshchenie, obespechenie*, *65*, 110–115. (In Russ.).

Kazantsev, D. A. (2022). From electronic documents to Source-to-Pay. Procurement automation step by step. *Goszakaz: upravlenie, razmeshchenie, obespechenie*, *67*, 60–67. (In Russ.).

Kazantsev, D. A., & Mikhaleva, N. A. (2020). Procurement automation as the future of the contract system. *RUDN Journal of Law*, *1*, 137–159. (In Russ.). https://doi.org/10.22363/2313-2337-2020-24-1-137-157

Kenney, M., & Zysman, J. (2016). The Rise of the Platform Economy. *Issues in Science and Technology*, *32*(3), 61–69.

Laptev, V. A. (2019). Artificial Intelligence and Liability for its Work. *Pravo. Zhurnal Vysshey Shkoly Ekonomiki*, *2*, 79–102. (In Russ.).

Lazarev, V. V. (2023). Legal Science in the Light of the Prospects of Digitalization. *Journal of Russian Law*, *2*, 5–19. (In Russ.).

Scassa, T. (2018). Information Law in the Platform Economy: Ownership, Control, and Reuse of Platform Data. In *Law and the "Sharing Economy": Regulating Online Market Platforms* (pp. 321–356). University of Ottawa Press.

Searle, J. R. (1990, January). Is the Brain's Mind a Computer Program? *Scientific American*, *262*(1).

Shakhnazarov, B. A. (2022). Legal Regulation of Relations Using Artificial Intelligence. *Actual Problems of Russian Law*, *9*, 63–72. (In Russ.). https://doi.org/10.17803/1994-1471.2022.142.9.063-072

Shmeleva, M. V. (2019). Digital Transformation of the System of Public and Municipal Procurements, *Jurist*, *7*, 15–22. (In Russ.). https://doi.org/10.18572/1812-3929-2019-7-15-22

Vavilin, E. V. (2021). Transformation of civil legal and procedural relations with the use of artificial intelligence: the formation of new legal regimes. *Herald of Civil Procedure*, *6*, 13–35. (In Russ.). https://doi.org/10.24031/2226-0781-2021-11-6-13-35

Author information

Dmitriy A. Kazantsev – Candidate of Sciences in Jurisprudence, Head of the Department of normative-legal regulation of the B2B-Center electronic trading platform operator
Address: 18/22 3rd Rybiskaya Str., 107113 Moscow, Russian Federation
E-mail: info@dkazantsev.ru
ORCID ID: https://orcid.org/0000-0003-2182-5776
RSCI Author ID: https://elibrary.ru/author_items.asp?authorid=1149755

Conflict of interests

The author declares no conflict of interest.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – April 6, 2023
Date of approval – April 15, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 347.44:004.8
EDN: https://elibrary.ru/jyqazw
DOI: https://doi.org/10.21202/jdtl.2023.18

Проблемы и перспективы регулирования отношений в рамках сделки, совершенной с участием искусственного интеллекта

Дмитрий Александрович Казанцев

B2B-Center

г. Москва, Российская Федерация

Ключевые слова

Автоматизация,
данные,
закупка,
искусственный интеллект,
ответственность,
право,
робот,
сделка,
цифровизация,
цифровые технологии

Аннотация

Цель: исследование проблемы определения субъекта юридически значимого действия, совершенного с использованием искусственного интеллекта, а также распределения ответственности за последствия его работы.

Методы: для иллюстрации проблематики и практической значимости вопроса о правосубъектности искусственного интеллекта были выбраны автоматизированные закупки для государственных и корпоративных нужд, а методологическую основу исследования составила совокупность методов научного познания, используемых для теоретико-правовых исследований, в том числе сравнения, ретроспективного анализа, аналогии и синтеза.

Результаты: на примере отрасли конкурентных закупок для государственных и корпоративных нужд проанализирована эволюция автоматизации хозяйственных отношений вплоть до внедрения искусственного интеллекта. Продемонстрированы успешно апробированные ответы на вызовы, обусловленные поэтапным внедрением цифровых технологий в хозяйственные отношения, а также соответствующие модификации правового регулирования. На основании текущего уровня развития технологий сформулированы перспективные вопросы правового регулирования хозяйственных отношений, реализуемых с использованием искусственного интеллекта, и прежде всего вопрос определения субъекта сделки, совершенной с использованием искусственного интеллекта. В качестве приглашения к дискуссии после анализа выводов правоведов о возможных вариантах юридического статуса искусственного

интеллекта предложены варианты ответа на вопрос о его правосубъектности при заключении сделки. Для решения вопроса об ответственности за решения, ставшие результатом реализации алгоритмов программно-аппаратного комплекса, предложено несколько моделей распределения такой ответственности между его создателем, его владельцем и иными лицами, действия которых могли повлиять на результаты работы такого алгоритма. Предложенные выводы могут использоваться для развития нормативного регулирования как в совокупности, так и по отдельности.

Научная новизна: в работе на основании анализа эволюции практик использования цифровых технологий в закупках сформулированы потенциальные проблемы правового порядка, обусловленные непрерывным процессом автоматизации хозяйственных отношений, а также предложены правовые конструкции для решения таких проблем.

Практическая значимость: выводы и предложения настоящей работы имеют перспективное значение для концептуального понимания и нормативного регулирования инструментов проведения закупок в электронной форме как на корпоративном, так и национальном уровне.

Для цитирования

Казанцев, Д. А. (2023). Проблемы и перспективы регулирования отношений в рамках сделки, совершенной с участием искусственного интеллекта. *Journal of Digital Technologies and Law*, *1*(2), 438–463. https://doi.org/10.21202/jdtl.2023.18

Список литературы

Андреев, В. К. (2021). Приобретение и осуществление прав юридического лица с использованием искусственного интеллекта. *Предпринимательское право*, *4*, 11–17.

Антонов, А. А. (2020). Искусственный интеллект как источник повышенной опасности. *Юрист*, *7*, 69–74. https://doi.org/10.18572/1812-3929-2020-7-69-74

Афанасьев, С. Ф. (2022). К проблеме материальной и процессуальной правосубъектности искусственного интеллекта. *Вестник гражданского процесса*, *3*, 12–31.

Вавилин, Е. В. (2021). Трансформация гражданско-правовых и процессуальных отношений с использованием искусственного интеллекта: формирование новых правовых режимов. *Вестник гражданского процесса*, *6*, 13–35. https://doi.org/10.24031/2226-0781-2021-11-6-13-35

Ивлиев, Г. П., Егорова, М. А. (2022). Юридическая проблематика правового статуса искусственного интеллекта и продуктов, созданных системами искусственного интеллекта. *Журнал российского права*, *26*(6), 32–46. https://doi.org/10.12737/jrl.2022.060

Казанцев, Д. А. (2020). *Конкурентные закупки. Методология и нормативное регулирование*: монография. Москва: ИНФРА-М. https://doi.org/10.12737/1068790

Казанцев, Д. А. (2021). Закупки и машинное обучение. Есть ли пространство для применения технологии? *Госзаказ: управление, размещение, обеспечение*, *65*, 110–115.

Казанцев, Д. А. (2022). От электронных документов до Source-to-Pay. Автоматизация закупок шаг за шагом. *Госзаказ: управление, размещение, обеспечение*, *67*, 60–67.

Казанцев, Д. А., Михалева, Н. М. (2020). Автоматизация закупок как будущее контрактной системы. *Вестник Российского университета дружбы народов. Серия: Юридические науки*, *1*, 137–159. https://doi.org/10.22363/2313-2337-2020-24-1-137-157

Лазарев, В. В. (2023). Юридическая наука в свете перспектив цифровизации. *Журнал российского права*, *2*, 5–19.

Лаптев, В. А. (2019). Понятие искусственного интеллекта и юридическая ответственность за его работу. *Право. Журнал Высшей школы экономики*, *2*, 79–102.

Чаннов, С. Е. (2022). Робот (система искусственного интеллекта) как субъект (квазисубъект) права. *Актуальные проблемы российского права*, *12*, 94–109. https://doi.org/10.17803/1994-1471.2022.145.12.094-109

Черногор, Н. Н. (2022). Искусственный интеллект и его роль в трансформации современного правопорядка. *Журнал российского права*, *26*(4), 5–15. https://doi.org/10.12737/jrl.2022.037

Шахназаров, Б. А. (2022). Правовое регулирование отношений с использованием искусственного интеллекта. *Актуальные проблемы российского права*, *9*, 63–72. https://doi.org/10.17803/1994-1471.2022.142.9.063-072

Шмелева, М. В. (2019). Цифровая трансформация системы государственных и муниципальных закупок. *Юрист*, *7*, 15–22. https://doi.org/10.18572/1812-3929-2019-7-15-22

Bokovnya, A. Y. et al. (2020). Legal Approaches to Artificial Intelligence Concept and Essence Definition. *Revista San Gregorio*, *41*, 115–121.

Calo, R. (2015). Robotics and the New Cyberlaw. *Californian Law Review*, *103*(3).

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International and Comparative Law Quarterly*, *69*(4), 819–844. https://doi.org/10.1017/s0020589320000366

Hallevy, G. (2013). *When Robots Kill: Artificial Intelligence under Criminal Law*. University Press of New England.

Kenney, M., & Zysman, J. (2016). The Rise of the Platform Economy. *Issues in Science and Technology*, *32*(3), 61–69.

Scassa, T. (2018). Information Law in the Platform Economy: Ownership, Control, and Reuse of Platform Data. In *Law and the "Sharing Economy": Regulating Online Market Platforms* (pp. 321–356). University of Ottawa Press.

Searle, J. R. (1990, January). Is the Brain's Mind a Computer Program? *Scientific American*, *262*(1).

Сведения об авторе

Казанцев Дмитрий Александрович – кандидат юридических наук, руководитель Департамента нормативно-правового регулирования оператора электронной торговой площадки B2B-Center
Адрес: 107113, Российская Федерация, г. Москва, 3-я Рыбинская улица, 18/22
E-mail: info@dkazantsev.ru
ORCID ID: https://orcid.org/0000-0003-2182-5776
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=1149755

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.27.41 / Сделки
Специальность ВАК: 5.1.3 / Частно-правовые (цивилистические) науки

История статьи

Дата поступления – 6 апреля 2023 г.
Дата одобрения после рецензирования – 15 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.19

AI-based Autonomous Weapons and Individual Criminal Responsibility under the Rome Statute

Fareed Mohd Hassan

Universiti Sains Islam Malaysia
Nilai, Negeri Sembilan, Malaysia

Noor Dzuhaidah Osman ✉

Universiti Sains Islam Malaysia
Nilai, Negeri Sembilan, Malaysia

Keywords

Armed conflict,
Artificial intelligence,
Autonomous weapons,
Criminal liability,
Digital technologies,
International criminal court,
law,
robotics,
Rome Statute,
war

Abstract

Objective: international law obligates states to prosecute those who have violated laws in armed conflicts, particularly when the international community now has International Criminal Court (ICC).

That is why the aim of the paper is to discover the responsibility for the crimes made with the use of AI-based autonomous vehicles in accordance with the provisions of the Rome Statute of the ICC.

Methods: doctrinal analysis allowed to research the positions of experts on the responsibility for the crimes made with the use of AI-based autonomous vehicles in accordance with the provisions of the Rome Statute of the ICC.

Results: this paper argues that the ICC can only exercise jurisdiction over natural persons who allegedly have committed the crimes under its jurisdiction, as compared to autonomous weapons. This paper argues that the persons who facilitate the commission of the alleged crimes are highly likely to be criminally responsible for providing means for the alleged crimes to be committed by AI-based autonomous weapons under Article 25(3)(c) of the Rome Statute and concludes that the Rome Statute provides a solution even to AI-based autonomous weapons.

✉ Corresponding author

Scientific novelty: this paper addresses to the highly relevant issues of the responsibility for the crimes made with the use of AI-based autonomous vehicles in accordance with the provisions of the Rome Statute of the ICC.

Practical significance: the results achieved in the paper can be used in regulation design for AI-based autonomous weapons. It can also be used as a basis for the future research in the sphere of liability of AI-based autonomous weapons and AI in general.

For citation

Hassan, F. M., & Osman, N. D. (2023). AI-Based Autonomous Weapons and Individual Criminal Responsibility Under the Rome Statute. *Journal of Digital Technologies and Law, 1*(2), 464–480. https://doi.org/10.21202/jdtl.2023.19

Contents

Introduction

War has become a tool for states to expand its territories where they have resorted to armed conflicts (Kalmanovitz, 2022; Kohama, 2019). During the armed conflict, various methods and means of warfare have been used which resulted into casualties for both or all sides, depending on how many parties or states have involved in the armed conflict (Bantekas, 2022). The warfare or types of weapons have evolved and changed tremendously, especially during and after the outbreak of both World Wars I and II (Fennell, 2019). Conventional weapons such as swords, knives, bows, gunpowder have been replaced with nuclear arms since then. Nonetheless, many states have now resorted to autonomous weapons via artificial intelligence (AI) as the latest technology to be used as its warfare (Human Rights Watch, 2020). Autonomous weapon is based on the AI which is the latest technology developed many countries and to be used as a weapon system that once activated, can select and engage targets without further intervention by a human operator (Horowitz, 2019). This type of weapon replaces ordinary human fighters (Hareth & Evans, 2023).

1. Autonomous Weapon in Artificial Intelligence

Autonomous weapon in artificial intelligence (AI) and robotics, autonomy simply refers to the ability to function for an extended period without the assistance of a human operator. Since war is divisive, many military applications of AI and robotics are also contentious (Amoroso & Tamburrini, 2021). The development and use of lethal autonomous weapons systems capable of autonomously making life and death decisions regarding human targets is perhaps the most contentious aspect of this topic. Cruise missiles, some argue, are a type of lethal autonomous weapons system. The Patriot missile system, the AEGIS naval weapons system, the Phalanx weapons system, and the Israeli Harpy weapons system are all examples of lethal autonomous weapons systems in use today (Payne, 2021). Defensive weapons include the Patriot, AEGIS, and Phalanx systems (Bartneck et al., 2021). In short, not all military robots are lethal.

The term "military robot" encompasses a wide range of non-lethal applications (Bartneck et al., 2021; Krishnan, 2009). Autonomous robots might be employed in mine clearance, explosive ordnance disposal, command and control, reconnaissance, intelligence, mobile network nodes, rescue missions, supply and resupply missions, and support operations, among other things (Burgess, 2017). Debates about military robots may differ depending on the robot's role (Malle et al., 2019). It is important to define some commonly used terms to illustrate the robot's and human's role in relation to war. In AI and robotics, autonomy simply refers to the ability to function for an extended period without the assistance of a human operator (Totaro, 2023). Robots may have autonomy over their immediate decisions, but they generally do not have autonomy over their goal selection (Javdani et al., 2018). A weapon is said to be "autonomous" in the "critical functions of targeting" if it can perform one or more of the following without the assistance of a human operator. If the weapon can choose which types of objects to engage, it will be autonomous in terms of defining its targets (Ekelhof, 2017). This capability is not currently available on AWS. If a weapon can use sensors to select a target without the assistance of a human operator, it is said to have autonomy in the targeting selection function.

Many existing weapons can select targets without the assistance of a human operator. When a weapon can fire on a target without the intervention of a human operator, it is said to have autonomy in the engage function of targeting. Many existing weapons can engage previously selected targets. The Patriot anti-missile system, for example, can select targets autonomously but requires a human operator to press a confirm button before launching a missile. Once launched, the missile can hit its target without the assistance of a human operator. Human control of a Patriot missile is not possible due to the speeds involved (Bartneck et al., 2021).

Many other functions may be "autonomous" for an AWS. It may be able to take off and land autonomously, as well as navigate autonomously. However, this non-lethal "autonomy" is not generally regarded as morally dubious. Autonomous weapons are

frequently referred to as "killer robots" in media reports. Some people object to the term's use. The phrase is described as a "insidious rhetorical trick" (Lokhorst & Van Den Hoven 2012). The "Campaign to Stop Killer Robots" believes otherwise. This is an umbrella organisation of human rights organisations seeking a global ban on lethal autonomous weapons systems (Bartneck et al., 2021).

2. Autonomous Weapon at the International Level

When many countries around the world criticise autonomous weapons, it only raises one critical issue: the risks of their use for humankind as well as military and war purposes. According to those who promote the benefits of autonomous weapons, the AI technology poses risks and benefits. The norms in deciding to regulate this contentious area of technology are the analysis of risks and benefits for lethal and non-lethal purposes. This would raise ethical and legal concerns about the use of autonomous weapons under international law. Before delving deeper into the autonomous weapon based on artificial intelligence as a method of warfare, it is necessary to review the series of incidents that led to legal regulation in this area.

Autonomous weapons based on artificial intelligence were previously discussed in 2010, when Philip Alston, then Special Rapporteur on Extrajudicial, Summary, or Arbitrary Executions, raised the issue in his interim report to the United Nations (UN) General Assembly 65th Session. Alston affirmed that "automated technologies are becoming increasingly sophisticated, and artificial intelligence reasoning and decision-making abilities are actively being researched and receive significant funding. States' militaries and defence industry developers are collaborating to develop 'fully autonomous capability', which will allow unmanned aerial vehicles to make and execute complex decisions, including the identification of human targets and the ability to kill them"[1]. Subsequently, in 2013, Christof Heyns, who was Special Rapporteur for Extrajudicial, Summary or Arbitrary Executions at the time, released a report that articulated further on the issues raised by what he called "lethal autonomous robotics".

Just after a recommendation by the Advisory Board on Disarmament Matters at the 68th session of the United Nations General Assembly, the Convention on the Prohibition or Restrictions on the Use of Certain Conventional Weapons Which May Be Considered Excessively Injurious or to Have Indiscriminate Effects, as revised on 21 December 2001, began discussing autonomous weapons systems in 2014. To address this issue, the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE on LAWS) was formed in 2016. While the group has continued to meet since then, no concrete steps towards a normative framework on autonomous weapons have been taken as of September 2022.

[1] https://digitallibrary.un.org/record/690463?ln=en

For the first time at the United Nations General Assembly, countries from around the world issued a joint statement on autonomous weapons systems. This was the largest cross-regional group statement ever made during UN discussions on the issue, with 70 states participating. While discussions at the UN CCW have yielded no results, the statement at the UNGA demonstrates states' widespread commitment to moving forward with a new international framework for autonomous weapons systems. The statement, delivered on behalf of the group by Ambassador Alexander Kmentt, Director of the Disarmament, Arms Control, and Non-proliferation Department at the Austrian Ministry of Foreign Affairs, consolidates key elements of the urgently needed international response, inter alia, "[r]ecognising that autonomous weapons systems raise serious humanitarian, legal, security, technological, and ethical concerns; [r]ecognise the importance of maintaining human responsibility and accountability when using force; and [t]he importance of internationally agreed rules and limits, including a combination of prohibitions and regulations on autonomous weapons systems"[2] are emphasised.

3. International Law on Autonomous Weapons

The Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects as amended on 21 December 2001, or the Convention on Certain Conventional Weapons (CCW)[3] is often widely recognised as the Inhumane Weapons Convention. The Convention's goal is to prohibit or limit the use of specific types of weapons that are thought to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately. The CCW's distinct structure aims to ensure adaptability in dealing with new developments in armed conflicts and weapon technologies.

The Framework Convention sets out the general operating provisions, such as rules for joining the regime and the ability to negotiate and adopt new protocols. The Protocols to the Convention contain substantive prohibitions and restrictions on specific types of weapons. The Convention, which included three annexed protocols, was adopted on 10 October 1980, and opened for signature on 10 April 1981 for a one-year period. The Convention was signed by 50 states and went into effect on December 2, 1983. There were initially three protocols namely Protocol I on 'Non-Detectable Fragments'; Protocol II on the 'Prohibitions or Restrictions on the Use of Mines, Booby Traps and Other Devices' and Protocol III on the 'Prohibitions or Restrictions on the Use of Incendiary Weapons'.

---

2 70 states deliver joint statement on autonomous weapons systems at UN General Assembly. https://www.stopkillerrobots.org/news/70-states-deliver-joint-statement-on-autonomous-weapons-systems-at-un-general-assembly

3 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects (adopted 10 October 1980, entered into force 2 December 1983) 1342 UNTS 137.

However, there were later additions of the Protocols namely Protocol IV on the 'Blinding Laser Weapons' which was adopted on 13 October 1995 during the First Review Conference of the States parties to the Convention pursuant to Article 8(3)(b) of the CCW and entered into force on 30 July 1998 as well as Protocol on the 'Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices' as amended on 3 May 1996 (Amended Protocol II) adopted at the First Review Conference, pursuant to Article 8 (1)(b) of the CCW and entered into force on 3 December 1998. There was also an amendment to Article 1 which extends the scope of application of the CCW to also cover situations of non-international armed conflict, adopted at the Second Review Conference in December 2001 pursuant to Article 8 (1)(b) of the CCW and entered into force on 18 May 2004. Lastly, Protocol V on the 'Explosive Remnants of War;' the first multilaterally negotiated instrument to deal with the problem of unexploded and abandoned ordnance was adopted on 28 November 2003 by the Meeting of the States Parties to the Convention pursuant to Article 5 (3) of the CCW and entered into force on 12 November 2006.

4. The International Criminal Court (ICC) and Its Jurisdiction

The Rome Statute was adopted by the international community on July 1998[4] and came into force in 2002[5] which established the International Criminal Court (ICC) which mentions under Preamble 10 and Article 1 of the Rome Statute. The ICC became the first and the most awaited permanent international criminal court (Pella, 1950) to end impunity of international crimes (Schabas, 2009) of genocide, crimes against humanity, war crimes and the crime of aggression as provided under Articles 6, 7, 8, 8bis and 15ter respectively. The ICC was created after a series of ad hoc tribunals established by the international community since the outbreak of the World War II by the victors allies which were the International Military Tribunal (IMT) in Nuremberg through the London Agreement[6] and the International Military Tribunal for the Far East (IMTFE) in Tokyo through a declaration made by General MacArthur, the Supreme Commander of the Allied Powers of the World War II[7] and two of the United

4 United Nations Diplomatic Conference of Plenipotentiaries on the Establishment of an International Criminal Court, *Final Documents: Rome Statute of the International Criminal Court and Final Act of the United Nations Diplomatic Conference of Plenipotentiaries on the Establishment of an International Criminal Court [With an Annex Containing the Resolutions Adopted by the Conference],* vol. I (Rome, 15 June - 17 July 1998) UN Doc A/CONF.183/13.

5 Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3. Hereafter, Rome Statute.

6 Agreement by the Government of the United States of America, the Provisional Government of the French Republic, the Government of the United Kingdom of Great Britain and Northern Ireland and the Government of the Union of Soviet Socialist Republics for the Prosecution and Punishment of the Major War Criminals of the European Axis, (signed at London on 8 August 1945, with Charter of the International Military Tribunal, entered into force 8 August 1945).

7 Special Proclamation by the Supreme Commander tor the Allied Powers at Tokyo (19 January 1946); Charter dated 19 January 1946; Amended Charter dated 26 April 1946 - Tribunal established 19 January 1946 (done in Tokyo on 19 January 1946).

Nations Security Council (UNSC) International Criminal Tribunal for the Former Yugoslavia (ICTY)[8] and Rwanda (ICTR)[9] acting under Chapter VII of the UN Charter in the 1990s.[10]

Unlike the International Court of Justice (ICJ) which only has jurisdiction over states,[11] the ICC only has jurisdiction over natural persons as stipulated under Article 25(1) of the Rome Statute who must be over 18 years old at the time of the commission of the crimes, or else they will be considered as under age as stipulated under Article 26 of the Rome Statute. Similarly, other parts of the Rome Statute also specifically mention the word 'person', among others, Article 1 which states that '…shall have the power to exercise its jurisdiction over persons…', Article 20 which mentions '…no person…' and '…the person…', Article 22 which elucidates '…a person' and '…the person…' as well as Article 23 which refers '…a person'.

Since these crimes are international crimes in nature, states have the obligation to investigate and prosecute them (Hassan & Osman, 2019). If states are either unable or unwilling to do so, the ICC will take over to exercise its jurisdiction over these crimes under the complementary principle as stipulated under Article 17 of the Rome Statute. In other words, national authorities will be the forum conveniens; latin words mean the most appropriate court to solve a particular dispute or case, has first-hand jurisdiction and are either able and willing to investigate or prosecute the individual perpetrators of the alleged crimes.

5. Individual Criminal Responsibility Under the ICC Jurisdiction

As mentioned under Article 10 of the Rome Statute, '[n]othing in this Part shall be interpreted as limiting or prejudicing in any way existing or developing rules of international law for purposes other than this Statute[12]. As we have discussed in the previous parts of this paper, there are several treaties which have been adopted by the international community to regulate autonomous weapons based on AI. Moreover, Article 21 of the Rome Statute allows the ICC to apply, 'where appropriate, applicable treaties and the principles and rules of international law, including the established principles of the international law of armed conflict' to decide cases brought before it[13]. Although the Rome Statute does not restrict the development of international law and its applicability to the ICC when deciding any cases brought before it, still the one who will be investigated and stand trials before it is only natural persons in accordance with Article 25(1) of the Rome Statute, regardless of his or her official positions as the head of state, head of government or other officials as enumerated under Article 27(1) of the Rome Statute.

---

8 UNSC Res 827 (25 May 1993) UN Doc S/RES/827.

9 UNSC Res 955 (8 November 1994) UN Doc S/RES/955.

10 Charter of the United Nations (24 October 1945) 1 UNTS XVI. See Chapter VII.

11 Statute of the International Court of Justice (ICJ Statute) art 34(1).

12 Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3. Hereafter, Rome Statute.

13 *Ibid.*

The notion of prosecuting persons or individuals regardless of his or her official positions for committing international crimes by the ICC is not new but has been practiced by numerous international tribunals such as the IMT under Articles 6 and 7 of the IMT Charter, the IMTFE under Articles 5 and 6 of the IMTFE Charter, the ICTY pursuant to Articles 6 and 7 of ICTY the Statute and the ICTR by virtue of Articles 5 and 6 of the ICTR Statute. As for the ICC, Article 25(3) of the Rome Statute further provides six (6) different modes or situations for a person to be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court which contains both basic rules of individual criminal responsibility and rules expanding attribution (Ambos, 2016).

I. If that person commits the crime[14]

As for the first mode of criminal liability under the Article 25(3)(a) of the Rome Statute, it provides that a person shall be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court if that person '[c]ommits such a crime, whether as an individual, jointly with another or through another person, regardless of whether that other person is criminally responsible'[15]. It is universally accepted criminal law principle[16] as held by the International Military Tribunal (IMT) at Nuremberg on the principle of individual criminal responsibility that '[c]rimes against international law are committed by men, not by abstract entities, and only by punishing individuals who commit such crimes can the provisions of international law be enforced'[17]. Under this mode of individual criminal responsibility, it 'refers to three forms of perpetration: on one's own, as a co[-]perpetrator or through another person (perpetration by means)[18].

II. If that person orders, solicits or induces the commission of the crime

As for the second mode of criminal liability under the Article 25(3)(b) of the Rome Statute, it provides that a person shall be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court if that person '[o]rders, solicits or induces the commission of such a crime which in fact occurs or is attempted'[19];

---

14 Rome Statute, art 25(3)(a).

15 *Ibid*.

16 See Prosecutor v Dusco Tadic (Decision on the Defence Motion for Interlocutory Appeal on Jurisdiction) IT-94-1 (2 October 1995) [128]-[137]. In [134] of this Decision, the ICTY stated that '[a]ll of these factors confirm that customary international law imposes criminal liability for serious violations of common Article 3, as supplemented by other general principles and rules on the protection of victims of internal armed conflict, and for breaching certain fundamental principles and rules regarding means and methods of combat in civil strife'.

17 Trial of the Major War Criminals Before the International Military Tribunal (Nuremberg 14 November 1945 - 1 October 1946) vol I (Nuremberg 1947).

18 Kai Ambos, 'Article 25: Individual Criminal Responsibility' in Otto Triffterer and Kai Ambos (eds), The Rome Statute of the International Criminal Court: A Commentary (3rd edn, Nomos 2016) 984.

19 Rome Statute, art 25(3)(b).

III. If that person facilitates the commission of the crime

As for the third mode of criminal liability under the Article 25(3)(c) of the Rome Statute, it provides that a person shall be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court if that person facilitates the commission of the crimes by aiding, abetting or otherwise assisting in its commission or its attempted commission, including providing the means for its commission;

IV. If that person in any way contributes to the commission or attempted commission of such a crime by a group of persons acting with a common purpose

As for the fourth mode of criminal liability under the Article 25(3)(d) of the Rome Statute, it provides that a person shall be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court if that person "[i]n any other way contributes to the commission or attempted commission of such a crime by a group of persons acting with a common purpose"[20]. Such contribution shall be intentional and shall either '[b]e made with the aim of furthering the criminal activity or criminal purpose of the group, where such activity or purpose involves the commission of a crime within the jurisdiction of the Court'[21] or '[b]e made in the knowledge of the intention of the group to commit the crime'[22];

V. In respect of the crime of genocide, directly and publicly incites others to commit genocide; and

VI. Attempts to commit such a crime by taking action that commences its execution by means of a substantial step, but the crime does not occur because of circumstances independent of the person's intentions. However, a person who abandons the effort to commit the crime or otherwise prevents the completion of the crime shall not be liable for punishment under this Statute for the attempt to commit that crime if that person completely and voluntarily gave up the criminal purpose.

6. Individual Criminal Responsibility and the Autonomous Weapons Based on AI

If linking those individuals or persons responsible to the crime can be very difficult, particularly when they are geographically and structurally remote from the scene of the crime, what more the 'perpetrators' of the ICC crimes are allegedly committed by autonomous weapons of AI which are not human beings. Article 36 of Additional Protocol I to the Geneva Conventions 1949[23] states that reviewing the legality of the intended deployment of the new weapon is an obligation of a state. It is crucial to ensure

20 Rome Statute, art 25(3)(a).

21 Rome Statute, art 25(3)(d)(i).

22 *Ibid*, art 25(3)(d)(ii).

23 Protocol Additional to the Geneva Conventions of 12 August 1949 and Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (adopted 8 June 1977, entered into force 7 December 1978).

that the armed forces of a State are capable of carrying out hostilities in line with their international responsibilities (Lawand, 2006). Article 36(2) of Additional Protocol I further mention that, when developing new weapon technology, lawyers and politicians need to maintain in respect of the law and accountability for those who seriously violate the law as stipulated under Article 49 of the Geneva Convention I[24].

Under Article 49 of the Geneva Convention I, is states that "[t]he High Contracting Parties undertake to enact any legislation necessary to provide effective penal sanctions for persons committing, or ordering to be committed, any of the grave breaches of the present Convention"[25]. Moreover, it mentions that "[e]ach High Contracting Party shall be under the obligation to search for persons alleged to have committed, or to have ordered to be committed, such grave breaches, and shall bring such persons, regardless of their nationality, before its own courts. It may also, if it prefers, and in accordance with the provisions of its own legislation, hand such persons over for trial to another High Contracting Party concerned, provided such High Contracting Party has made out a 'prima facie' case"[26].

As for autonomous weapons based on AI which are fully unmanned, orders from the operators have been pre-programmed and as such, the legal responsibility for any actions must be expected to transfer from the operators to the system conducted by the AI. However, a question of legal obligations will arise; whether any decisions made by the weapon will be borne by the weapon or its operators? In this sense, no one can be held accountable if he or she is willing to offend or behave passively. However, a weapon system's designer, programmer, or manufacturer could also be held liable only to the extent if they willfully to contributed to the crime commission (McFarland & McCormack, 2014).

Since autonomous weapons, particularly those which are free of human intervention where AI entirely controls them, there are no choice for human actors to exercise empathy or judgment (Gunawan et al., 2022). Human influence over weapons systems and force use need to meet legal and ethical demands, as mentioned by the International Committee of the Red Cross (ICRC) in its statement on the Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) in Geneva on 11 April 2017 to the CCW.

Conclusion

The advancement of technology has reached a high standard and demand by the international community in order to protect its boarders and citizens not only from being invaded and attacked by outsiders, but also to protect their troops from being targeted and killed.

24 Convention (I) for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31.

25 *Ibid.*

26 *Ibid.*

This led to the creation of the new technology in weaponry of autonomous weapons based on AI. However, such technology does not free from any responsibility under international law and has received many criticisms and concerns by the international community due to attacks by to be taken and done by autonomous weapons based on AI which could still incur casualties from the non-military objectives. Since the creation of the ICC in 2002 via the Rome Statute, the latter provides a solution even to the most advanced weapons such as unmanned autonomous weapons based on AI whereby individuals behind the creation and manning such weapons would be criminally liable if they went beyond the borders allowed under the law in order to win the war or involved in armed conflicts.

References

Ambos, K. (2016). Article 25: Individual Criminal Responsibility. In O. Triffterer? & K. Ambos (Eds.), *The Rome Statute of the International Criminal Court: A Commentary* (3rd ed.). Nomos.

Amoroso, D., & Tamburrini, G. (2021). In Search of the 'Human Element': International Debates on Regulating Autonomous Weapons Systems. *The International Spectator*, *56*(1), 20–38. https://doi.org/10.1080/03932729.2020.1864995

Bantekas, I. (2022). Punishment in Warfare and the Application of Law. In *Principles of Direct and Superior Responsibility in International Humanitarian Law* (pp. 1–37). Manchester UP.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *Military Uses of AI BT.* In *An Introduction to Ethics in Robotics and AI* (pp. 93–99). Springer International Publishing. https://doi.org/10.1007/978-3-030-51110-4_11

Burgess, A. (2017). *The Executive Guide to Artificial Intelligence: How to Identify and Implement Applications for AI in Your Organization*. Springer. https://doi.org/10.1007/978-3-319-63820-1

Ekelhof, M. A. (2017). Complications of a Common Language: Why It Is So Hard to Talk About Autonomous Weapons. *Journal of Conflict and Security Law*, *22*(2), 311–331. https://doi.org/10.1093/jcsl/krw029

Fennell, J. (2019). *Fighting the People's War: The British and Commonwealth Armies and the Second World War*. CUP. https://doi.org/10.1017/9781139380881

Gunawan, Y., Anggriawan, M. H. A. R., & Putro, T. A. (2022). Command Responsibility of Autonomous weapons under International Humanitarian Law. *Cogent Social Sciences*, *8*(1), 2139906. https://doi.org/10.1080/23311886.2022.2139906

Hareth, B., & Evans, N. G. (2023). Make Them Rare or Make Them Care. In D. Schoeni & T. Vestner (Eds.), *Ethical Dilemmas in the Global Defense Industry* (pp. 217–236). OUP.

Hassan, F. M., & Osman, N. D. (2019). The Obligation to Prosecute Heads of State Under the Rome Statute of the International Criminal Court (ICC) and Customary International Law: The African And United States' Perspectives. *MJSL*, *7*(1), 33–56. https://doi.org/10.33102/mjsl.v7i1.112

Horowitz, M. C. (2019). When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability. *Journal of Strategic Studies*, *42*(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174

Human Rights Watch. (2020). *Stop Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control*.

Javdani, S., Admoni, H., Srinivasa, S. P. S. S., & Bagnell, J. A. (2018). Shared Autonomy via Hindsight Optimization for Teleoperation and Teaming. *International Journal of Robotics Research*, *37*(7), 717–742. https://doi.org/10.1177/0278364918776060

Kalmanovitz, P. (2022). Can Criminal Organizations Be Non-State Parties to Armed Conflict? In *International Review of the Red Cross* (pp. 1–19). ICRC, CUP. https://doi.org/10.1017/S1816383122000510

Kohama, S. (2019). Territorial Acquisition, Commitment, and Recurrent War. *International Relations of the Asia-Pacific*, *19*(2), 269–295. https://doi.org/10.1093/irap/lcy001

Krishnan, A. (2009). Automating War: The Need for Regulation. *Contemporary Security Policy*, *30*(1), 172–193. https://doi.org/10.1080/13523260902760397

Lawand, K. (2006). Reviewing the legality of new weapons, means and methods of warfare. *International Review of the Red Cross*, *88*, 925–930. https://doi.org/10.1017/S1816383107000884

Lokhorst, G.-J., & Hoven, J. van den. (2012). Responsibility for military robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics*: *The Ethics and Social Implications of Robotics* (pp. 145–156). MIT Press.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the Sky: How People Morally Evaluate Human and Machine

Decisions in a Lethal Strike Dilemma. In M. I. A. Ferreira, J. S. Sequeira, G. S. Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-being* (pp. 111–133). Springer. https://doi.org/10.1007/978-3-030-12524-0_11

McFarland, T., & McCormack, T. (2014). Mind the Gap: Can Developers of Autonomous Weapon Systems be Liable for War Crimes?, *Internatoinal Law Studies*, *90*, 350–362.

Payne, K. (2021). *I, Warbot: The Dawn of Artificially Intelligent Conflict*. OUP.

Pella V. (1950).Towards an International Criminal Court. *American Journal of International Law*, *44*(1), 38–49.

Schabas,William A. (2009). International Crimes. In D. Armstrong (Ed.), *Routledge Handbook of International Law*. Routledge. https://doi.org/10.4324/9780203884621

Totaro, D. L. (2023). Machine or Robot? Thoughts on the Legal Notion of Autonomy in the Context of Self-Driving Vehicles and Intelligent Machines. *European Business Law Review*, *34*(1), 99–114.

Authors information

Fareed Mohd Hassan – PhD, Senior Lecturer, Faculty of Syariah and Law, Universiti Sains Islam Malaysia (USIM) Nilai, Negeri Sembilan, Malaysia
Address: Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
E-mail: fareed@usim.edu.my
ORCID ID: https://orcid.org/0009-0006-7501-0782
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57222041043
Google Scholar ID: https://scholar.google.com/citations?user=2LjkkQ4AAAAJ&hl=en

Noor Dzuhaidah Osman – PhD, Senior Lecturer, Faculty of Syariah and Law, Universiti Sains Islam Malaysia (USIM) Nilai, Negeri Sembilan, Malaysia
Address: Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
E-mail: noordzuhaidah@usim.edu.my
ORCID ID: https://orcid.org/0000-0003-2659-9309
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57226308331
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/AAT-6652-2021
Google Scholar ID: https://scholar.google.com/citations?user=h48_cAIAAAAJ&hl=en

Authors' contributions

Authors equally contribute to the paper.

Conflict of interests

The author declares no conflict of interests.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – February 28, 2023
Date of approval – April 23, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 343.3/.7:341.4:004.8
EDN: https://elibrary.ru/mgreoy
DOI: https://doi.org/10.21202/jdtl.2023.19

Автономное вооружение на основе искусственного интеллекта и индивидуальная уголовная ответственность согласно Римскому статуту

Фарид Мохд Хасан

Исламский научный университет Малайзии
Нилаи, Негри-Сембилан, Малайзия

Нур Джухаида Осман ✉

Исламский научный университет Малайзии
Нилаи, Негри-Сембилан, Малайзия

Ключевые слова

Автономное вооружение,
война,
вооруженный конфликт,
искусственный интеллект,
Международный уголовный суд,
право,
Римский статут,
робототехника,
уголовная ответственность,
цифровые технологии

Аннотация

Цель: международное право обязывает государства преследовать лиц, нарушивших закон в ходе вооруженных конфликтов, чему способствовало создание Международного уголовного суда. Цель данной статьи – рассмотрение ответственности за преступления, совершенные с использованием автономных устройств на основе искусственного интеллекта, согласно положениям Римского статута Международного уголовного суда.

Методы: доктринальный анализ позволил изучить позиции экспертов по вопросу ответственности за преступления, совершенные с использованием автономных устройств на основе искусственного интеллекта, согласно положениям Римского статута Международного уголовного суда.

Результаты: в работе показано, что Международный уголовный суд может отправлять правосудие только в отношении физических лиц, предположительно совершивших преступления в рамках его юрисдикции, но не в отношении автономных вооружений. В статье утверждается, что лица, способствовавшие совершению предполагаемых преступлений, будут, вероятно, нести уголовную ответственность за предоставление средств для совершения предполагаемых преступлений автономными вооружениями на основе искусственного интеллекта согласно статье

✉ Контактное лицо

25(3)(c) Римского статута. Авторы приходят к выводу, что Римский статут дает решение относительно автономного вооружения на основе искусственного интеллекта.

Научная новизна: в статье изучены актуальные вопросы, связанные с ответственностью за преступления, совершенные с использованием автономных устройств на основе искусственного интеллекта, согласно положениям Римского статута Международного уголовного суда.

Практическая значимость: результаты работы могут быть использованы при разработке регулирования автономного вооружения на основе искусственного интеллекта, а также служить основой для будущих исследований в сфере ответственности за использование как автономных вооружений на основе искусственного интеллекта, так и искусственного интеллекта в целом.

Для цитирования

Хасан, Ф. М., Осман, Н. Д. (2023). Автономное вооружение на основе искусственного интеллекта и индивидуальная уголовная ответственность согласно Римскому статуту. *Journal of Digital Technologies and Law*, *1*(2), 464–480. https://doi.org/10.21202/jdtl.2023.9

Список литературы

Ambos, K. (2016). Article 25: Individual Criminal Responsibility. In O. Triffterer, & K. Ambos (Eds.), *The Rome Statute of the International Criminal Court: A Commentary* (3rd ed.). Nomos.

Amoroso, D., & Tamburrini, G. (2021). In Search of the 'Human Element': International Debates on Regulating Autonomous Weapons Systems. *The International Spectator*, *56*(1), 20–38. https://doi.org/10.1080/03932729.2020.1864995

Bantekas, I. (2022). Punishment in Warfare and the Application of Law. In *Principles of Direct and Superior Responsibility in International Humanitarian Law* (pp. 1–37). Manchester UP.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *Military Uses of AI BT.* In *An Introduction to Ethics in Robotics and AI* (pp. 93–99). Springer International Publishing. https://doi.org/10.1007/978-3-030-51110-4_11

Burgess, A. (2017). *The Executive Guide to Artificial Intelligence: How to Identify and Implement Applications for AI in Your Organization*. Springer. https://doi.org/10.1007/978-3-319-63820-1

Ekelhof, M. A. (2017). Complications of a Common Language: Why It Is So Hard to Talk About Autonomous Weapons. *Journal of Conflict and Security Law*, *22*(2), 311–331. https://doi.org/10.1093/jcsl/krw029

Fennell, J. (2019). *Fighting the People's War: The British and Commonwealth Armies and the Second World War*. CUP. https://doi.org/10.1017/9781139380881

Gunawan, Y., Anggriawan, M. H. A. R., & Putro, T. A. (2022). Command Responsibility of Autonomous weapons under International Humanitarian Law. *Cogent Social Sciences*, *8*(1), 2139906. https://doi.org/10.1080/23311886.2022.2139906

Hareth, B., & Evans, N. G. (2023). Make Them Rare or Make Them Care. In D. Schoeni & T. Vestner (Eds.), *Ethical Dilemmas in the Global Defense Industry* (pp. 217–236). OUP.

Hassan, F. M., & Osman, N. D. (2019). The Obligation to Prosecute Heads of State Under the Rome Statute of the International Criminal Court (ICC) and Customary International Law: The African And United States' Perspectives. *MJSL*, *7*(1), 33–56. https://doi.org/10.33102/mjsl.v7i1.112

Horowitz, M. C. (2019). When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability. *Journal of Strategic Studies*, *42*(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174

Human Rights Watch. (2020). *Stop Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control*.

Javdani, S., Admoni, H., Srinivasa, S. P. S. S., & Bagnell, J. A. (2018). Shared Autonomy via Hindsight Optimization for Teleoperation and Teaming. *International Journal of Robotics Research*, *37*(7), 717–742. https://doi.org/10.1177/0278364918776060

Kalmanovitz, P. (2022). Can Criminal Organizations Be Non-State Parties to Armed Conflict? In *International Review of the Red Cross* (pp. 1–19). ICRC, CUP. https://doi.org/10.1017/S1816383122000510

Kohama, S. (2019). Territorial Acquisition, Commitment, and Recurrent War. *International Relations of the Asia-Pacific*, *19*(2), 269–295. https://doi.org/10.1093/irap/lcy001

Krishnan, A. (2009). Automating War: The Need for Regulation. *Contemporary Security Policy*, *30*(1), 172–193. https://doi.org/10.1080/13523260902760397

Lawand, K. (2006). Reviewing the legality of new weapons, means and methods of warfare. *International Review of the Red Cross*, *88*, 925–930. https://doi.org/10.1017/S1816383107000884

Lokhorst, G.-J., & Hoven, J. van den. (2012). Responsibility for military robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), Robot Ethics: *The Ethics and Social Implications of Robotics* (pp. 145–156). MIT Press.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma. In M. I. A. Ferreira, J. S. Sequeira, G. S. Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-being* (pp. 111–133). Springer. https://doi.org/10.1007/978-3-030-12524-0_11

McFarland, T., & McCormack, T. (2014). Mind the Gap: Can Developers of Autonomous Weapon Systems be Liable for War Crimes?, *Internatoinal Law Studies*, *90*, 350–362.

Payne, K. (2021). *I, Warbot: The Dawn of Artificially Intelligent Conflict*. OUP.

Pella V. (1950).Towards an International Criminal Court. *American Journal of International Law*, *44*(1), 38–49.

Schabas,William A. (2009). International Crimes. In D. Armstrong (Ed.), *Routledge Handbook of International Law*. Routledge. https://doi.org/10.4324/9780203884621

Totaro, D. L. (2023). Machine or Robot? Thoughts on the Legal Notion of Autonomy in the Context of Self-Driving Vehicles and Intelligent Machines. *European Business Law Review*, *34*(1), 99–114.

Сведения об авторах

Фарид Мохд Хасан – доктор наук, старший преподаватель факультета шариата и права, Исламский научный университет Малайзии, Нилаи, Негри-Сембилан, Малайзия
Адрес: Бандар Бару Нилаи, 71800 Нилаи, Негри-Сембилан, Малайзия
E-mail: fareed@usim.edu.my
ORCID ID: https://orcid.org/0009-0006-7501-0782
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57222041043
Google Scholar ID: https://scholar.google.com/citations?user=2LjkkQ4AAAAJ&hl=en

Нур Джухаида Осман – доктор наук, старший преподаватель факультета шариата и права, Исламский научный университет Малайзии, Нилаи, Негри-Сембилан, Малайзия
Адрес: Бандар Бару Нилаи, 71800 Нилаи, Негри-Сембилан, Малайзия
E-mail: noordzuhaidah@usim.edu.my
ORCID ID: https://orcid.org/0000-0003-2659-9309
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57226308331
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/AAT-6652-2021
Google Scholar ID: https://scholar.google.com/citations?user=h48_cAIAAAAJ&hl=en

Вклад авторов

Авторы внесли равный вклад в создание статьи.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.87.89 / Международное уголовное право
Специальность ВАК: 5.1.5 / Международно-правовые науки

История статьи

Дата поступления – 28 февраля 2023 г.
Дата одобрения после рецензирования – 23 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.20

Artificial Intelligence Technologies in Criminal Procedural Proving

Mikhail S. Spiridonov

South Ural State University (National Research University)
Chelyabinsk, Russian Federation

Keywords

Artificial intelligence,
criminal case,
criminal procedure,
digital technologies
digitalization,
evidence,
judicial procedure,
law,
neural network,
proving

Abstract

Objective: to summarize and analyze the approaches, established in criminal procedural science, regarding the use of artificial intelligence technologies, to elaborate an author's approach to the prospects of transformation of criminal procedural proving under the influence of artificial intelligence technologies.

Methods: the methodological basis of the research is integrity of general, general scientific and specific legal methods of legal science, including abstract-logical, comparative-legal and prognostic methods.

Results: the main areas of using artificial intelligence technologies in the criminal procedure are defined, such as prophylaxis and detection of crimes, organization of preliminary investigation, criminological support of crime investigation, and assessing evidences at pre-trial and trial stages. The author comes to a conclusion that the rather optimistic approach to this issue, established in the science of criminal procedure, significantly outstrips the actually existing artificial intelligence technologies. The main requirements are identified, which the activity of using artificial intelligence in collecting evidences in a criminal case should satisfy. The author pays attention to the problems of using artificial intelligence technologies in conducting expert assessments, requiring an improved methodology of forensic work. The issue is considered of the prospects of transforming the criminal-procedural proving process under introduction of artificial intelligence technologies. A conclusion is substantiated that the assessment

of evidences with mathematical algorithms, in which preset values of each evidence quality are used, contradict to the principle of free assessment of evidences in the criminal procedure. The author comes to a conclusion that today there are no sufficient grounds for endowing artificial intelligence with legal personality during proving.

Scientific novelty: the work presents an attempt to consider the role of artificial intelligence in the criminal-procedural proving; it specifies the requirements to be met by this technology during evidences collection and analyzes the prospects of transforming the proving process under the introduction of artificial intelligence technologies.

Practical significance: the main provisions and conclusions of the research can be used to improve a mechanism of legal regulation of artificial intelligence technologies in the criminal procedure.

For citation

Spiridonov, M. S. (2023). Artificial Intelligence Technologies in Criminal Procedural Proving. *Journal of Digital Technologies and Law, 1*(2), 481–497. https://doi.org/10.21202/jdtl.2023.20

Contents

Introduction

The theme of artificial intelligence in a criminal procedure seems rather futuristic, as the phenomenon per se can be called purely technical; essentially, it is no more than a mathematical algorithm intended for, to keep it simple, processing the incoming information and forming new information on its basis. However, digital technologies so rapidly enter not only everyday life but also such special spheres as criminal procedure, that science cannot stay aloof and ignore the issues which at first glance seem to have no real practical application.

Researchers justly note that we will not be prepared for new challenges emerging in connection with modern technologies, if we do not discuss them right now (Vesnic-Alujevic et al., 2020; Feijóo & Kwon, 2020; Robles Carrillo, 2020). Hence, it is especially important to study the prospects of artificial intelligence technologies in the criminal procedure, to analyze the problems emerging thereof and search for possible optimal variants of solving them, and to predict the ways of criminal procedure development.

Criminal procedure, like any other branch of legal science, requires an objective and unbiased assessment of how the use of artificial intelligence corresponds to the goals and tasks set before it, what influence the use of these technologies has and will have in the future on the protection of personality against unsubstantiated criminal prosecution. Such assessment should not be excessively and groundlessly optimistic, as in the absence of strong artificial intelligence any statements about robots substituting judges tomorrow cannot be perceived seriously. However, one should not slide to the opposite – to blind negation of the prospects of using new technologies, including in such a complex abstract sphere as the proving process.

One should agree with the researchers who believe that there is a need to analyze the approaches, established in the science of criminal procedure, to assessing the prospects of using artificial intelligence technologies and, based on such analysis, to consider the issue of the role of these digital technologies in the criminal-procedural proving (Silva et al., 2020; Kaur et al., 2023; Wang & Ma, 2022; Kalai et al., 2022; Cascavilla et al., 2021).

1. Mathematical algorithms in the service of criminal procedure

As early as in 1990, Professor Rissland noted that there is a fruitful synergy between law and artificial intelligence; law opens broad opportunities for developing analytical and computational models of artificial intelligence; at the same time, certain characteristics of law make it an especially complicated sphere for artificial intelligence (Rissland, 1990).

Thirty years after these statements we can distinguish several areas in which artificial intelligence technologies are used or are planned to be used in criminal procedure (in a broad sense).

1.1. Prophylaxis and detection of crimes

Scholars mark the potential of the artificial intelligence technologies in the sphere of prophylaxis and detection of crimes. For example, foreign researchers write about the possibility of using artificial intelligence to detect some cybercrimes (Kaur et al., 2023; Kalai et al., 2022; Cascavilla et al., 2021). A work by A. M. Tsirin and E. A. Artemenko describes software based on artificial intelligence technologies, which is intended to analyze the performed monitoring or oversight activities and reveal the true reasons

for damage incurred and the indicators of risk of the potentially dangerous observees (Tsirin & Artemenko, 2023). To struggle against corruption, N. A. Kuzmin proposes to create artificial intelligence based systems of tracking financial transactions, money input and output (Kuzmin, 2021). S. V. Rastoropov presents a review of such software in the sphere of preventing and detecting crime as ShotSpotter, Prepol, CloudWalkTechnology, HART, VAA, which are already successfully used in police activity (Rastoropov, 2020).

1.2. Organization of preliminary investigation

In this area, Yu. A. Tsvetkov, for example, suggests developing "an artificial managerial brain" capable of generating "optimal variants of solutions at all nods a criminal case trajectory" (Tsvetkov, 2021). M. A. Malina considers it possible to apply artificial intelligence technologies to detect and correct various distortions of the form of procedural documents, as well as to identify poor-quality and potentially unreliable information incoming to an investigation officer (Malina, 2021).

1.3. Criminological support of crime investigation

A good example is "Zerkalo" ("Mirror") software described in the article by D. N. Sretentsev and V. R. Volkova; it allows revealing the signs of intraframe editing of videos made by artificial neural networks which synthesize video images of people (deepfake) (Sretentsev & Volkova, 2021). An article by F. Rahman describes a research conducted at Syracuse University, where machine learning was used to classify and identify individual DNA profiles, as well as to analyze large amounts of complex data to reveal patterns some of which can be inaccessible to human analysis (Rahman, 2019). An article by Silva et al. (Silva et al., 2020) researches the possibility of using artificial intelligence to analyze images as evidences in criminal procedure.

1.4. Assessing evidences at pre-trial and trial stages

For example, A. V. Sibilkova assumes that the sufficiency of collected evidences can be estimated on the basis of machine learning of artificial neural networks of the results of criminal investigations of certain categories of cases, i. e. "artificial neural networks must obtain information about what was lacking to convict" (Sibilkova, 2019). As a model for introducing artificial intelligence into the system of justice, A. V. Makutchev sees its functioning "on equal terms, in cooperation with a judge" or substituting a judge with artificial intelligence (Makutchev, 2022). A work by A. A. Sumin and O. V. Khimicheva provides a good example of using artificial intelligence systems for assessing evidences at pre-trial stage in China (Sumin & Khimicheva, 2020). The problems and prospects of using artificial intelligence at that stage of a criminal procedure were also considered by foreign authors (Stoykova, 2023; Yassine et al., 2023; Amariles & Baquero, 2023).

The above brief review shows that the science of criminal procedure have long formed a rather optimistic approach to the issue of using artificial intelligence technologies in various spheres, including in criminal-procedural proving. However, this approach largely outstrips the actually existing artificial intelligence technologies, i. e. it can be stated that researchers mostly express their expectations and prognoses about the possibilities of using new technologies.

Undoubtedly, the use of new technologies in any of the above areas can only be welcomed. However, one should realize that there must be a clear and solid reason to introduce a certain technology into the criminal procedure. This is because a criminal procedure as an indispensible part of a legal procedure does not need any technologies; it is, so to say, self-sufficient. Is any technology needed to initiate a criminal case, collect evidences, submit them to the court where they will be considered and a decision on the merits will be made? Surely not. All this can be done even in the absence of pen and paper, exactly as it was done in the early human history.

Apparently, the areas of using artificial intelligence technologies touch upon all stages of the proving process in a criminal case. Some of them are used to collect evidences, serving in that instance as a tool of obtaining the evidence information in the hands of a subject of proving. Other technologies may be applied in checking and even assessing evidences.

Evidences and proving underlie all procedural decisions in a criminal case, directly influencing the rights and interests of the process participants. Thus, this is one of the most sensitive spheres of the criminal procedure. In the Russian legal tradition, proving is considered to "the core of criminal process" (Sheifer, 2022), "the centerpiece of all procedural activity" (Lupinskaya, 2023). Thus, from the viewpoint of the science of criminal procedure, it is important to define the place of artificial intelligence technologies, including in the future, in proving in a criminal case: should they be perceived exclusively as a tool with the help of which certain data relevant for the criminal case can be obtained, which can be shaped as evidences, or one may speak of the changes in the very nature of proving in a criminal procedure as it transforms from the sphere of human cognition into the sphere of machine cognition, where the role of a human being is reduced to just registering its results in a particular law-enforcement decision.

2. Artificial intelligence in the hands of a subject of proving

Using artificial intelligence technologies to obtain evidential information relevant for the criminal case implies a certain processing of the data contained in the source of evidence. Accordingly, there appears a risk of distorting or transforming this information, which may ultimately influence the reliability of the evidence per se (Stoykova, 2023). The subject of proving and the process participants must have a possibility to estimate how the said processing of information was taking place and make certain of its results.

2.1. Problems of collecting evidences using artificial intelligence technologies

One should agree about the main problems emerging at the stage of collecting evidences using artificial intelligence technologies, which were formulated by Eftychia Bampasika (Bampasika, 2021):

1. Inexplicability – the complexity of understanding the algorithms used by artificial intelligence leads to the impossibility of verifying or challenging such evidence. Indeed, if during evidence collection certain software based on artificial intelligence is used, for example, a facial recognition (identification) system or an image restoration system, then the algorithm underlying this product must be transparent and accessible for studying both by the subject of proving and the process participant.

2. Discrimination and bias – the information on which basis artificial intelligence makes a decision is not always complete and free of bias or distortions. That means that the set of data fed to, say, a neural network for "learning" must be accessible by the participants of the proving process, for them to have an opportunity to reveal any distortion or bias.

3. Lack of responsibility – the functioning of artificial intelligence is in any case based on the human activity which is not sufficiently regulated by law. The activity of creating and developing technologies allowing for collecting of evidences or information, on which basis evidences in a criminal case are subsequently formed, must be not only legislatively regulated but also meet the fundamental principles of a criminal procedure.

The above issues are actually worrying; therefore, any artificial intelligence technology used in a criminal procedure with a view of obtaining and collecting evidences must possess a set of properties which allows obtaining, as a result, evidence admissible from the viewpoint of criminal-procedural law.

2.2. Using artificial intelligence technologies in conducting expert assessments

In this aspect, an important position today belongs to using artificial intelligence technologies when conducting expert assessments. For example, using neural networks, based on machine learning, during expert research must be reflected in the research section of the expert's opinion.

An example of how the use of artificial intelligence technologies can be specified in an expert's opinion is an article by Alessandro Marrone and a group of biologists, which thoroughly describes the methodology of the research determining the bloodstains age by colorimetric analysis, including using five different machine learning approaches (Marrone et al., 2021). The researchers describe each of the machine learning approaches applied and the results obtained. One may easily see how the use of any other artificial intelligence based tool (like a neural network) should be similarly described when conducting a biological

expert assessment. This will allow estimating the content of the expert's opinion from the viewpoint of their reliability and, as a result, provide an opportunity to use these conclusions in proving in a criminal case.

Another example showing the use of artificial intelligence technologies in a criminal procedure as a tool for obtaining and collecting evidences is a deepfake technology. With its development, the accessibility and, accordingly, breadth of its use in criminal sphere will only increase. One may easily imagine using deepfake in banking or insurance fraud and other crimes of such kind. Hence, as early as today we should prepare a scientific basis for relevant expert research, which are, apparently, impossible without applying artificial intelligence tools. If a neural network can generate a fake video, then an expert must have available a no less effective neural network capable of recognizing such video. Hence, a symmetric answer to developing the deepfake technology must be development of a new methodology of portrait and videoscopic expertise taking into account the most vulnerable aspects of this technology, for example, such as the presence or absence of reflection in the eyes of the personages on video. As we have mentioned above, the Russian Ministry of Internal Affairs tries to develop such software in order to counteract the criminal deeds using deepfake. However, besides technical means, one should provide for the relevant methodological support of such activity.

As one can see, collecting evidences with the help of artificial intelligence technologies, provided this activity meets the requirements of openness, controllability, objectivity and is supplied with a relevant liability mechanism, fully complies with the principle of free evaluation of evidence (Article 17 of the Criminal-procedural Code of the Russian Federation[1]). It is worth noting that this is not about appearance of a new type of evidence in the criminal procedure or, as it is sometimes called, "electronic evidence". We fully share the opinion, expressed in science, that the types and forms of evidences stipulated by criminal-procedural law do not need expanding (Golovko, 2019). This is only about regulating the new means of obtaining and collecting evidences, which some researchers denote as "obtaining digital information through machine means" (Aleksandrov, 2018). We believe that the evidences obtained with the help of artificial intelligence technologies will possess the sign of admissibility, hence, can be used in proving in a criminal case.

3. The prospects of transforming the proving process

A rapid development of artificial intelligence technologies and an increasing digitalization of judicial procedures incite some researchers to a conclusion that artificial intelligence may become a subject of legal relations and be endowed with legal personality

[1] Criminal-procedural Code of the Russian Federation of December 18, 2001 No. 174-FZ. *Collection of legislation of the Russian Federation*. December 24, 2001, No. 52 (part I). Article 4921.

(Papysheva, 2022) or even can substitute a judge (Kolokolov, 2020). This position might be exceedingly optimistic and even to some extent futuristic, but it is not unreasonable. As a matter of fact, the problem of a judicial decision approximating the objective truth and the related search for effective means and mechanisms which could exclude or minimize a judicial error occupies the minds of the Russian researchers in criminal-procedural science so much that they are eager to grasp any, even a purely hypothetical, opportunity to solve it. In that instance, the artificial intelligence tools which, as is broadly advertized, "allow solving some humanly impossible tasks", "exclude the influence of a human factor in problem solving", "accelerate the decision-making process", naturally become a rather attractive object to build hypotheses about the directions of a criminal procedure development.

One more reason for researchers to turn to artificial intelligence technologies is unpredictability of judicial decisions, especially as regards trial by jury. According to the Criminal-procedural Code of the Russian Federation, the jurors are not obliged to explain their verdict, which does not allow the parties to assess the relevance of such decision and understand how and around what the judgments of the jury were built. Because of that, jury decisions are often perceived as arbitrary, detached from the proving process.

Besides, none of the process participants, including a professional judge, may influence their decision, the more so participate in their deliberation. Left alone with a question sheet, without special training in proving in a criminal case, but having got instructions from a professional judge as to the necessity to interpret all reasonable doubts in favor of the defendant, the jurors have to turn to their experience and the skills of reflections and logical deductions formed in their lifetime.

As a result, in a case where evidences included videos from a fuel station with an armed assault and testimonies of victims vividly describing the events, the jury may come to a conclusion of the absence of a criminal act and pronounce for the defendants[2]. In another case, where evidences included experts' opinion about the presence of a defendant's DNA on the crime scene and personal belongings of the murdered victims at the defendant's home, the jury may come to a conclusion of the defendant's noninvolvement and also pronounced for them[3].

Undoubtedly, if it were possible to ask the jurors why they came to such decisions, they would probably explain that the evidence presented to them was simply insufficient to conclude that the charges had been proved. This could largely clarify the connection

[2] Case No. 2-3/2020, heard by Stavropol regional court. Available at: https://kraevoy--stv.sudrf.ru/modules.php?name=sud_delo&srv_num=2&name_op=case&case_id=30809862&case_uid=bbd2dd35-a6f1-4f7b-a158-18a71eefa38f&delo_id=1540006

[3] Case No. 2-4/2021, heard by Stavropol regional court. Available at: https://kraevoy--stv.sudrf.ru/modules.php?name=sud_delo&srv_num=2&name_op=doc&number=23537753&delo_id=1540006&new=0&text_number=1

between the verdict and the evidences and would remove the question about the perceived justification of the verdict.

In this situation, when the motifs of the decision made by the subject of proving are concealed for us, the natural question is: do mathematical methods underlying any artificial intelligence technology allow improving the proving process in a criminal case, so that it resulted in a just ruling of the court, maximally approximated to the objective truth? What if every evidence in a criminal case was attributed a certain weight and a neural network was taught to assess these evidences and build certain conclusions based on them?

When answering these questions, on should first of all turn to the essence of the principle of free evaluation of evidence described in Article 17 of the Russian Criminal-procedural Code, which consists, inter alia, in the absence of a preset power of evidences and prohibition of any grading them by quality (Golovko, 2017). When we set precise values to each piece of evidence for the algorithm and suggest it making an evaluative conclusion on that basis, we disregard that the criminal-procedural proving, although being a type cognitive activity, has cardinal differences from other types of cognition. Therefore, formalization of evidences through attributing a certain weight to them, even for the sake of a mathematical algorithm functioning, contradicts the fundamental principle of the criminal procedure. Here one should agree with M. A. Malina in the negative estimation of such an approach to using artificial intelligence technologies in proving in a criminal case (Malina, 2021).

Are there alternative way or one should unequivocally and ultimately reject an idea of transforming the proving process through using artificial intelligence technologies?

Fortunately, there is always an alternative. For example, one of the world leading specialists in evidentiary law, revered Professor Ronald J. Allen, convincingly demonstrated as early as in 2001 that it does not work applying such mathematical methods of estimating probabilities as Bayes theorem to proving in a criminal process and suggested a theory of relative likelihood (Allen, 2001). In a later work, developing this idea, he stated that evidences would be more convincing when they confirm a conclusion about a single hypothesis (for example, this person committed a crime) compared to a competing theory (for example, the crime was committed by someone else), and weaker when they do not exclude probable alternative hypotheses based on alternative suggestions (Allen & Pardo, 2007).

If one looks through the prism of these ideas at the proving process in a jury trial, it becomes obvious that it is just like that: the parties try to convince the jurors that their version of circumstances of the case is more probable. It is on the estimation of probability of evidences that a judge instructs the jury in the opening statement, asking them in the analysis of evidences to interpret in favor of the defendant only reasonable doubts, i. e. those which can be explained, which are based on common sense, not on a biased opinion, suggestions, imagination, sympathy or antipathy, desire to cater for public opinion or meet the expectations of friends, on emotions or fantasies.

Hence, if one considers the possibility of using artificial intelligence technologies at the stage of checking and evaluating evidences, then only through the algorithms based on the theory of relative likelihood. Otherwise we turn on the violation of the principle of free evaluation of evidence.

Can the existing artificial intelligence technologies perform the function of evaluating evidences at the same level as non-professional judges – the jurors? Apparently not. Such a mathematical method has not been developed yet. Hence, it is so far premature to speak of a legal personality of artificial intelligence in the process of criminal-procedural proving.

However, looking into the future with a hope to improve the process of proving, including in the jury trial, the science of criminal procedure should first of all turn to mathematicians. Only elaboration of mathematical methods for analyzing evidences in a criminal case will allow passing from using artificial intelligence technologies exclusively as a tool in the process of obtaining and collecting evidences to forming a new process of proving, in which artificial intelligence will be able to play a cognitive role. How soon it will happen, and whether it will happen at all, time will tell.

Conclusions

The optimistic approach to evaluating the possibilities of using artificial intelligence technologies in a criminal procedure, established in the science of criminal procedure, significantly outstrips the actually existing technologies, i. e. is based on expectations and prognoses. The actual application of the said technologies takes place only in certain spheres of the criminal procedure. Nevertheless, the continuing development of artificial intelligence allows speaking of an inevitable broadening of its spheres of application, including in the process of proving in criminal cases.

The most realistic scenario of using artificial intelligence technologies in the proving process is their use in obtaining and collecting information of evidentiary significance, which may further be structured as evidence in a criminal case. Using artificial intelligence as a tool when collecting evidences must meet such requirements as openness, controllability, objectivity and provision with a relevant liability mechanism. In that instance the evidences obtained with the help of artificial intelligence technologies will possess the sign of admissibility.

Evaluation of evidence with the help of mathematical algorithms in which preset values of quality of each piece of evidence are used, contradicts to the principle of free evaluation of evidence in a criminal procedure. Applying the artificial intelligence technologies based on mathematical algorithms in criminal-procedural proving is only possible under the condition of compliance with the said principle, i. e., for example, on the basis of the theory of relative likelihood of evidences. Hence, it is now premature to say that machine cognition may become the content of criminal-procedural proving.

References

Aleksandrov, A. S. (2018). The problems of the theory of criminal procedural proof, which must be solved in connection with the transition to the digital age. *Judicial authority and criminal process*, *2*, 130–139. (In Russ.).

Allen, Ronald J., & Pardo, M. S. (2007). The Problematic Value of Mathematical Models of Evidence. *Journal of Legal Studies*, *36*, 107–108.

Allen, Ronald J. (2001). Artificial intelligence and the evidentiary process: The challenges of formalism and computation. *Artificial Intelligence and Law*, *9*, 99–114.

Amariles, D. R., & Baquero, P. M. (2023). Promises and limits of law for a human-centric artificial intelligence. *Computer Law & Security Review*, *48*, 105795. https://doi.org/10.1016/j.clsr.2023.105795

Bampasika, E.-V. (2021). Artificial Intelligence as Evidence in Criminal Trial. In *CEUR Workshop Proceedings* (pp. 133–138).

Cascavilla, G., Tamburri, D., & Van Den Heuvel, W.-J. (2021). Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security*, *105*, 102258. https://doi.org/10.1016/j.cose.2021.102258

Feijóo, C., & Kwon, Y. (2020). AI impacts on economy and society: Latest developments, open issues and new policy measures. *Telecommunications Policy*, *44*(6), 101987. https://doi.org/10.1016/j.telpol.2020.101987

Golovko, L. V. (2017). *Course in criminal procedure* (2nd ed., amended). Moscow: Statut. (In Russ.).

Golovko, L. V. (2019). The digitalization in criminal procedure: local optimization or global revolution? *Vestnik ekonomicheskoy bezopasnosti*, *1*, 15–25. (In Russ.).

Kalai, T., Gnanaprakasam, C., Indumathy, M., Khilar, R., & Sathish Kumar, P. J. (2022). Artificial intelligence based optimization for mapping IP addresses to prevent cyber-based attacks. *Measurement: Sensors*, *24*, 100508, https://doi.org/10.1016/j.measen.2022.100508

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Kolokolov, N. A. (2020). Once again on artificial intelligence in justice. *Ugolovnoye sudoproizvodstvo*, *4*, 3–6. (In Russ.).

Kuzmin, N. A. (2021). Prospects for the use of artificial intelligence in combating corruption. *Vestnik Moskovskogo universiteta MVD Rossii*, *3*, 154–156. (In Russ.).

Lupinskaya, P. A. (2023). *Decisions in criminal judicial procedure: theory, legislation, practice:* monograph (3rd ed., stereotyped) Moscow. (In Russ.).

Makutchev, A. V. (2022). Modern Possibilities and Limits of Artificial Intelligence Introduction into the Justice System. *Actual Problems of Russian Law*, *17*(8), 47–58. (In Russ.). https://doi.org/10.17803/1994-1471.2022.141.8.047-058

Malina, M. A. (2021). Digitalization of the Russian criminal procedure: artificial intelligence for an investigator or instead of an investigator. *Rossiiskiy sledovatel*, *2*, 29–32. (In Russ.).

Marrone, A., La Russa, D., Montesanto, A., Lagani, V., La Russa, M. F., & Pellegrino, D. (2021). Short and Long Time Bloodstains Age Determination by Colorimetric Analysis: A Pilot Study. *Molecules*, *26*, 6272. https://doi.org/10.3390/molecules26206272

Papysheva, E. S. (2022). Artificial intelligence as a threat to the principle of assumption of innocence. *Advocate's Practice*, *5*, 2–5. (In Russ.).

Rahman, F. (2019). Introducing Artificial Intelligence in the Criminal Justice System with Special Reference to India, *NCU Law Review*, *2*(1), 49–63.

Rastoropov, S. V. (2020). The use of artificial intelligence for crime prevention and identification (the world experience). *Public International and Private International Law*, *5*, 40–43. (In Russ.). https://doi.org/10.18572/1812-3910-2020-5-40-43

Rissland, Edwina L. (1990). Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning. *Yale Law Journal*, *99*, 1957–1980.

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, *44*(6), 101937, https://doi.org/10.1016/j.telpol.2020.101937

Sheifer, S. A. (2022). *Evidence and proving in criminal cases: problems of theory and legal regulation:* monograph (2nd ed., amend. and compl.). Moscow: Norma. (In Russ.).

Sibilkova, A. V. (2019). Artificial intelligence in the investigator's employ. *Rossiiskiy sledovatel*, *3*, 68–70. (In Russ.).

Silva, I., Marcos do Valle, J., Souza, G., Budke, J., Araújo, D., Carvalho, D., Cacho, N., Sales, H., Lopes, F., & Silva Júnior, R. (2020). Using micro-services and artificial intelligence to analyze images in criminal evidences. *Forensic Science International: Digital Investigation*, *37*, Supplement, 301197. https://doi.org/10.1016/j.fsidi.2021.301197

Sretentsev, D. N., & Volkova, V. R. (2021). Prospects for the introduction of artificial intelligence systems in crime investigation. *Rossiiskiy sledovatel*, *11*, 38–42. (In Russ.).

Stoykova, R. (2023). The right to a fair trial as a conceptual framework for digital evidence rules in criminal investigations. *Computer Law & Security Review*, *49*, 105801. https://doi.org/10.1016/j.clsr.2023.105801

Sumin, A. A., & Khimicheva, O. V. (2020). Artificial intelligence in the criminal procedure of the states of the Asia-Pacific region: a general overview. *Mezhdunarodnoye ugolovnoye pravo i mezhdunarodnaya yustitsiya*, *2*, 18–21. (In Russ.).

Tsirin, A. M., & Artemenko, E. A. (2023). Digital Technologies and Artificial Intelligence as Measures of Preventing Corruption in Control (Supervisory) Activities: Domestic and Foreign Experience. *Journal of Russian Law*, *3*, 126–142. (In Russ.).

Tsvetkov, Yu. A. (2021). Artificial intelligence in management of investigative authorities. *Rossiiskiy sledovatel*, *9*, 29–33. (In Russ.).

Vesnic-Alujevic, L., Nascimento, S., & Pólvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy*, *44*(6), 101961. https://doi.org/10.1016/j.telpol.2020.101961

Wang, H., & Ma, S. (2022). Preventing crimes against public health with artificial intelligence and machine learning capabilities. *Socio-Economic Planning Sciences*, *80*, 101043. https://doi.org/10.1016/j.seps.2021.101043

Yassine, S., Esghir, M., & Ibrihich, O. (2023). Using Artificial Intelligence Tools in the Judicial Domain and the Evaluation of their Impact on the Prediction of Judgments. *Procedia Computer Science*, *220*, 1021–1026. https://doi.org/10.1016/j.procs.2023.03.142

Author information

Mikhail S. Spiridonov – Candidate of Sciences in Jurisprudence, Associate Professor of the Department of Criminal Procedure, Criminology and Forensics, South Ural State University (National Research University)
Address: 76 prospekt Lenina, 454080 Chelyabinsk, Russian Federation
E-mail: spiridonovms@susu.ru
ORCID ID: https://orcid.org/0009-0008-2715-8912
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/AAK-9097-2021
RSCI Author ID: https://elibrary.ru/author_items.asp?authorid=1089837

Conflict of interests

The author declare no conflict of interests.

Financial disclosure

The research was not sponsored.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – April 28, 2023
Date of approval – May 12, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 343.14:004.8
EDN: https://elibrary.ru/acsqxh
DOI: https://doi.org/10.21202/jdtl.2023.20

Технологии искусственного интеллекта в уголовно-процессуальном доказывании

Михаил Сергеевич Спиридонов

Южно-Уральский государственный университет (национальный исследовательский университет)
г. Челябинск, Российская Федерация

Ключевые слова

Доказательство,
доказывание,
искусственный интеллект,
нейронная сеть,
право,
судопроизводство,
уголовное дело,
уголовный процесс,
цифровизация,
цифровые технологии

Аннотация

Цель: обобщение и анализ сложившихся в уголовно-процессуальной науке позиций к применению технологий искусственного интеллекта, выработка авторского подхода к перспективам трансформации уголовно-процессуального доказывания под влиянием технологий искусственного интеллекта.

Методы: методологическую основу исследования составляет единство всеобщего, общенаучных и специально-юридических методов правовой науки, в том числе абстрактно-логического, сравнительно-правового и прогностического.

Результаты: определены основные направления применения технологий искусственного интеллекта в уголовном процессе, такие как профилактика и выявление преступлений, организация предварительного расследования, криминалистическое сопровождение расследования преступлений, оценка доказательств на досудебной и судебной стадиях. Автор приходит к выводу, что сложившийся в науке уголовного процесса достаточно оптимистичный подход по данному вопросу значительно опережает реально существующие в настоящее время технологии искусственного интеллекта. Выявлены основные требования, которым должна отвечать деятельность по применению искусственного интеллекта при сборе доказательств по уголовному делу. Обращено внимание на проблемы применения технологий искусственного интеллекта при проведении судебных экспертиз, которые требуют совершенствования методологии судебно-экспертной работы. Рассмотрен вопрос о перспективах трансформации процесса уголовно-процессуального доказывания в условиях внедрения технологий искусственного интеллекта. Обосновывается вывод, что оценка доказательств с помощью математических алгоритмов, в которых применяются заранее установленные значения качества

каждого доказательства, противоречит принципу свободы оценки доказательств в уголовном процессе. Автор приходит к выводу об отсутствии в настоящее время достаточных оснований для наделения искусственного интеллекта субъектностью в процессе доказывания.

Научная новизна: в работе предпринята попытка рассмотреть место искусственного интеллекта в уголовно-процессуальном доказывании, выявлены требования, которым должно соответствовать применение этой технологии при сборе доказательств, при этом проанализированы перспективы трансформации процесса доказывания с учетом внедрения технологий искусственного интеллекта.

Практическая значимость: основные положения и выводы исследования могут быть использованы для совершенствования механизма правового регулирования технологий искусственного интеллекта в уголовном процессе.

Для цитирования

Спиридонов, М. С. (2023). Технологии искусственного интеллекта в уголовно-процессуальном доказывании. *Journal of Digital Technologies and Law*, *1*(2), 481–497. https://doi.org/10.21202/jdtl.2023.20

Список литературы

Александров, А. С. (2018). Проблемы теории уголовно-процессуального доказывания, которые надо решать в связи с переходом в эпоху цифровых технологий. *Судебная власть и уголовный процесс*, *2*, 130–139. https://elibrary.ru/XWCHZB

Головко, Л. В. (2017). *Курс уголовного процесса* (2-е изд., испр.). Москва: Статут, 2017.

Головко, Л. В. (2019). Цифровизация в уголовном процессе: локальная оптимизация или глобальная революция? *Вестник экономической безопасности*, *1*, 15–25.

Колоколов, Н. А. (2020). Еще раз об искусственном интеллекте в правосудии. *Уголовное судопроизводство*, *4*, 3–6. https://elibrary.ru/OYLHDM

Кузьмин, Н. А. (2021). Перспективы использования искусственного интеллекта в противодействии коррупции. *Вестник Московского университета МВД России*, *3*, 154–156.

Лупинская, П. А. (2023). *Решения в уголовном судопроизводстве: теория, законодательство, практика*: монография (3-е изд., стереотип.). Москва.

Макутчев, А. В. (2022). Современные возможности и пределы внедрения искусственного интеллекта в систему правосудия. *Актуальные проблемы российского права*, *8*, 47–58. https://doi.org/10.17803/1994-1471.2022.141.8.047-058

Малина, М. А. (2021). Цифровизация российского уголовного процесса: искусственный интеллект для следователя или вместо следователя. *Российский следователь*, *2*, 29–32.

Папышева, Е. С. (2022). Искусственный интеллект как угроза принципу презумпции невиновности. *Адвокатская практика, 5*, 2–5.

Расторопов, С. В. (2020). Использование искусственного интеллекта для предупреждения и выявления преступлений (мировой опыт). *Международное публичное и частное право*, *5*, 40–43. https://doi.org/10.18572/1812-3910-2020-5-40-43

Сибилькова, А. В. (2019). Искусственный интеллект на службе у следователя. *Российский следователь*, *9*(3), 68–70. https://elibrary.ru/ZANFTV

Сретенцев, Д. Н., Волкова, Д. Р. (2021). Перспективы внедрения искусственного интеллекта в сферу расследования преступлений. *Российский следователь*, *11*, 38–42. https://elibrary.ru/HLNQDU

Сумин, А. А., Химичева, О. В. (2020). Искусственный интеллект в уголовном процессе государств Азиатско-Тихоокеанского региона: общий обзор. *Международное уголовное право и международная юстиция*, *2*, 18–21. https://elibrary.ru/RGILWR

Цветков, Ю. А. (2021). Искусственный интеллект в управлении следственными органами. Российский следователь, 9, 29–33.

Цирин, А. М., Артеменко, Е. А. (2023). Цифровые технологии и искусственный интеллект как средства профилактики проявлений коррупции в контрольной (надзорной) деятельности: отечественный и зарубежный опыт. *Журнал российского права*, *3*, 126–142.

Шейфер, С. А. (2022). *Доказательства и доказывание по уголовным делам: проблемы теории и правового регулирования*: монография (2-е изд., испр. и доп.). Москва: Норма.

Allen, Ronald J., & Pardo, M. S. (2007). The Problematic Value of Mathematical Models of Evidence. *Journal of Legal Studies*, *36*, 107–108.

Allen, Ronald J. (2001). Artificial intelligence and the evidentiary process: The challenges of formalism and computation. *Artificial Intelligence and Law*, *9*, 99–114.

Amariles, D. R., & Baquero, P. M. (2023). Promises and limits of law for a human-centric artificial intelligence. *Computer Law & Security Review*, *48*, 105795. https://doi.org/10.1016/j.clsr.2023.105795

Bampasika, E.-V. (2021). Artificial Intelligence as Evidence in Criminal Trial. In *CEUR Workshop Proceedings* (pp. 133–138).

Cascavilla, G., Tamburri, D., & Van Den Heuvel, W.-J. (2021). Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security*, *105*, 102258. https://doi.org/10.1016/j.cose.2021.102258

Feijóo, C., & Kwon, Y. (2020). AI impacts on economy and society: Latest developments, open issues and new policy measures. *Telecommunications Policy*, *44*(6), 101987. https://doi.org/10.1016/j.telpol.2020.101987

Kalai, T., Gnanaprakasam, C., Indumathy, M., Khilar, R., & Sathish Kumar, P. J. (2022). Artificial intelligence based optimization for mapping IP addresses to prevent cyber-based attacks. *Measurement: Sensors*, *24*, 100508, https://doi.org/10.1016/j.measen.2022.100508

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*,101804. https://doi.org/10.1016/j.inffus.2023.101804

Marrone, A., La Russa, D., Montesanto, A., Lagani, V., La Russa, M. F., & Pellegrino, D. (2021). Short and Long Time Bloodstains Age Determination by Colorimetric Analysis: A Pilot Study. *Molecules*, *26*, 6272. https://doi.org/10.3390/molecules26206272

Rahman, F. (2019). Introducing Artificial Intelligence in The Criminal Justice System with Special Reference to India, *NCU Law Review*, *2*(1), 49–63.

Rissland, Edwina L. (1990). Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning. *Yale Law Journal*, *99*, 1957–1980.

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, *44*(6), 101937, https://doi.org/10.1016/j.telpol.2020.101937

Silva, I., Marcos do Valle, J., Souza, G., Budke, J., Araújo, D., Carvalho, D., Cacho, N., Sales, H., Lopes, F., & Silva Júnior, R. (2020). Using micro-services and artificial intelligence to analyze images in criminal evidences. *Forensic Science International: Digital Investigation*, *37*, Supplement, 301197. https://doi.org/10.1016/j.fsidi.2021.301197

Stoykova, R. (2023). The right to a fair trial as a conceptual framework for digital evidence rules in criminal investigations. *Computer Law & Security Review*, *49*, 105801. https://doi.org/10.1016/j.clsr.2023.105801

Vesnic-Alujevic, L., Nascimento, S., & Pólvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy*, *44*(6), 101961. https://doi.org/10.1016/j.telpol.2020.101961

Wang, H., & Ma, S. (2022). Preventing crimes against public health with artificial intelligence and machine learning capabilities. *Socio-Economic Planning Sciences*, *80*, 101043. https://doi.org/10.1016/j.seps.2021.101043

Yassine, S., Esghir, M., & Ibrihich, O. (2023). Using Artificial Intelligence Tools in the Judicial Domain and the Evaluation of their Impact on the Prediction of Judgments. *Procedia Computer Science*, *220*, 1021–1026. https://doi.org/10.1016/j.procs.2023.03.142

Сведения об авторе

Спиридонов Михаил Сергеевич – кандидат юридических наук, доцент кафедры уголовного процесса, криминалистики и судебной экспертизы, Южно-Уральский государственный университет (национальный исследовательский университет)
Адрес: 454080, Российская Федерация, г. Челябинск, пр. Ленина, 76
E-mail: spiridonovms@susu.ru
ORCID ID: https://orcid.org/0009-0008-2715-8912
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/AAK-9097-2021
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=1089837

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.79.35 / Доказательства
Специальность ВАК: 5.1.4 / Уголовно-правовые науки

История статьи

Дата поступления – 28 апреля 2023 г.
Дата одобрения после рецензирования – 12 мая 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.21

Recommendations on the Ethical Aspects of Artificial Intelligence, with an Outlook on the World of Work

Zsofia Riczu

University of Miskolc
Miskolc, Hungary

Keywords

Artificial intelligence,
digital technologies,
digitalization,
ethics,
labor law,
labor relations,
labor,
law,
legislation,
principles of law

Abstract

Objective: the spread and wide application of Artificial Intelligence raises ethical questions in addition to data protection measures. That is why the aim of this paper is to examine ethical aspects of Artificial Intelligence and give recommendations for its use in labor law.

Methods: research based on the methods of comparative and empirical analysis. Comparative analysis allowed to examine provisions of the modern labor law in the context of use of Artificial Intelligence. Empirical analysis made it possible to highlight the ethical issues related to Artificial Intelligence in the world of work by examining the disputable cases of the use of Artificial Intelligence in different areas, such as healthcare, education, transport, etc.

Results: the private law aspects of the ethical issues of Artificial Intelligence were examined in the context of ethical and labor law issues that affect the selection process with Artificial Intelligence and the treatment of employees as a set of data from the employers' side. Author outlined the general aspects of ethics and issues of digital ethics. Author described individual international recommendations related to the ethics of Artificial Intelligence.

Scientific novelty:this research focused on the examination of ethical issues of the use of Artificial Intelligence in the specific field of private law – labor law. Authors gave recommendations on ethical aspects of use of Artificial Intelligence in this specific field.

Practical significance: research contributes to the limited literature on the topic. The results of the research could be used in lawmaking process and also as a basis for future research.

For citation

Riczu, Zs. (2023). Recommendations on the Ethical Aspects of Artificial Intelligence, with an Outlook on the World of Work. *Journal of Digital Technologies and Law, 1*(2), 498–519. https://doi.org/10.21202/jdtl.2023.21

Content

Introduction

The rise of Artificial intelligence is indisputable. In many forums, we can find studies that approach the topic from a social perspective. The research clearly shows that artificial intelligence (AI – Artificial Intelligence) has a mixed reception, on the one hand it is suitable for carrying out useful activities in many fields, but opposing opinions already involve conspiracy theories.

The legal regulation of AI is also a significant area of research, primarily with regard to regulatory issues (Cyman et al., 2021; Alikhademi, et al., 2022). Given that it is difficult to keep up with the dynamism of technological development, the legal regulatory background is currently in follow mode: after the appearance of technological achievements, it tries to establish regulations regarding specific issues – on the one hand, this is positive, given that the legislation itself is based on empirical knowledge, and on the other hand, the development of legal regulation requires caution, as it is necessary to take into account issues that have not yet appeared, but which can already be predicted in certain cases. Scientific and technological development, the digital revolution, AI and algorithms are

putting legal regulations to new tests, and the dominance of legal institutions and the rethinking of individual legal categories can be observed (Harmathy, 2019). The development of regulation is greatly aided by community standards, but it is also necessary to carry out regulatory tasks at the national level in accordance with the specificities.

The application and spread of algorithms and artificial intelligence is a very interesting research area. Their application promises speed and accuracy. Recently, the ever-increasing dangers inherent in intelligent robots have become a central topic in many areas of research on artificial intelligence. This point of view, called «alarmist» by László Z. Karvalics, is a logical consequence (and strong ally) of the aging paradigm of «strong AI» and the latest versions of this paradigm. (Karvalics, 2015)

Technology is constantly changing. Just like the previous industrial revolutions, now Industry 4.0 (or even Industry 5.0) can bring new things day by day, hour by hour, minute by minute. The rate of development is increasing exponentially, which law and regulations cannot keep up with. The protection of individuals and their rights is essential both in the digital space and the offline space, since, as Stefán Ibolya put it, the lack or insufficiency of legal regulation can be an obstacle to economic development (Stefán, 2020).

The presence of AI is not only typical in the research world and the scientific field, although it also plays a significant role in supporting the development of science and managing global challenges, from Apple Inc.'s Siri to space exploration to self-driving cars, they have become a part of our daily lives, appearing as convenience devices, given that human capabilities are supplemented with elements of computing. Learning algorithms can perform many tasks, whether it is driving a car or making decisions. These applications make decisions about our lives on many occasions, which is why there is a need for certain guarantees. However, the spread and wide application of AI raises ethical questions in addition to data protection measures. The theme indicated in the title of the thesis is one of the cornerstones of my research.

Recent debates have highlighted the urgent need for work on the social and ethical aspects of algorithmic and data-driven systems that control our lives. Largely focused on machine learning (ML) and artificial intelligence (AI), conversations among scientists, journalists and advocates have begun to address issues of equity, bias, transparency, access, participation and discrimination, often referred to as «AI and ethics». Underlying these discourses are concerns about how to mitigate discrimination and data bias, as well as open questions about whether algorithmic systems can be fair and whether their use will promote an equitable future or, on the contrary, perpetuate or even reinforce existing inequalities. Despite the apparent 'novelty' of these issues, many of the underlying concerns have a long tradition and intellectual pedigree in the field of library and information science (LIS) (Hoffmann et al., 2019).

The pace of technological innovation and the speed of its global adoption, supported by the digital economy, is clearly outpacing the speed of human consciousness. Jerome

Beranger[1] has sought to list and define the main ethical criteria that form the basis of a future frame of reference in order to move towards AI in the service of human intelligence, and to anticipate and foresee the drifts and possible consequences of the development of algorithmic systems (Beranger, 2021).

The size and complexity of the data and algorithms used in artificial intelligence (AI)-based systems pose significant challenges in predicting their ethical, legal and policy implications. Drawing on interviews with stakeholders in AI research, law and policy, Locating the work of artificial intelligence ethics reports that the work of AI ethics is structured by personal values and professional commitments (Fleischmann et al., 2023).

In the first part, I outline the general aspects of ethics, from which I examine the issues of digital ethics. After that, I would like to describe the individual international recommendations related to the ethics of AI. In the present study, however, regarding the ethical issues of AI, I examine its private law (including labor law) aspects in more detail in connection with my research area. Within this topic, I examine ethical and labor law issues that affect the selection process with artificial intelligence and the treatment of employees as a set of data from the employers' side.

1. Ethics – Digital Ethics

The word ethics comes from the Greek word ethos, which covers habit, form of behavior. The subject of ethics is human action and the person unfolding in action. It is often used as a synonym for the term morality, which refers to forms of behavior and activity related to a person's purpose (Turay, 2000). Ethics covers the moral standards that exist subconsciously, the values that lie behind our actions (Müller & Kerényi, 2019). Ethics is the doctrine of moral values, which must be separated from etiquette, which is the science of custom, manners, decorum, i.e. human behavior (Legeza, 2013).

Ethics and applied ethics focus on practical problems and everyday situations. Its area is the lower and higher manifestation of moral phenomena, moral generality – Fobel refers to Kansky's definition and notes that at the XX World Congress of Philosophy the central theme was the problem of applied ethics, within which bioethics, health ethics, environmental protection and business ethics came to the fore sports ethics, as well as issues of technological and legal ethics (Fobel, 2002). Within the scope of technological ethics, digital ethics is the determining factor, the appearance of which can be dated at the same time as the appearance of the Internet and the development of technology. Ethical behavior is just as important on the world wide web, accessible to everyone, as it is outside the digital dimension. Nowadays, anyone can

[1] The scientific expert on the ethical approach of the digital revolution, the cofounder and CEO of ADELIAA and is also an associate researcher in the Inserm 1295 BIOETHICS team at the University of Toulouse.

produce digital content, post comments on a given topic, and express an opinion on certain issues. We express our values during these activities.

Digital ethics, or information ethics in a broader sense, deals with the impact of digital information and communication technologies (ICT) on our society and the environment in general. More narrowly, information ethics (or digital media ethics) examines ethical issues related to the Internet and Internet-based information and communication media, such as mobile phones and navigation services (Capurro, 2022).

The emergence of digital ethics has also generated legal ethical challenges. Digital communication, the emergence of algorithms and the rise of AI also have an impact on people's everyday lives. Based on Dániel Eszteri's point of view, new ethical problems such as automated decision-making without human intervention and the rights of artificial entities may come into focus (Eszteri, 2021). According to Gabriella Németh's view: The appearance of machine entities can redefine man himself, or the values associated with human existence (Németh, 2021). Imre Négyesi examined the ethical issues of AI for military purposes (Négyesi, 2020), Réka Pusztahelyi explained in her work that the ethical principles of AI can be grouped from several points of view, depending on who they target. Accordingly, the field has separated specific, sectoral and comprehensive ethical principles (Pusztahelyi, 2019). I base this study on the latter finding. After all, the ethical principles of AI applied in medicine do not necessarily cover the ethical conceptions of AI applied in the labor market.

AI systems are not fundamentally about injustice and inequity, yet the question of AI systems and their treatment of the unjust world is a moot point. As Knowles has pointed out, one mechanism for dealing with the ethics of AI is the ethics education of AI: it teaches clear moral reasoning, responsible choices and right action to those who build, use and/or submit to AI systems (Knowles, 2021).

The first step to creating ethical AI systems is to explore ethical dilemmas. The community of AI-related researchers and professionals prefers the use of frameworks in AI regulation over ad-hoc standards (Yu et al., 2018). Such regulatory frameworks were created by UNESCO, the European Parliament, and the European Commission, whose recommendations and resolutions I describe in the next chapter.

2. Regulations related to the ethical aspects of AI

The emergence of AI has raised many questions from the regulatory side. These are not only operations of legal conceptualization and categorization, legal responsibility issues related to the ethical challenges of AI form a separate subject area.

The ethical risks of AI are significant. This is supported by Deloitte's 2018 research, according to which 32% of managers rated the ethical risks of AI as so significant that

they stopped their AI initiatives if appropriate[2]. Based on a 2019 survey by the Capgemini Research Institute, nearly half of users have experienced an ethical problem with AI, and 86% of managers stated that they are aware of cases where the use of AI has led to ethical problems (Capgemini Research Institute, 2019) [3]. These researches also prove that it is not enough to define the legal framework of AI, and attention must also be paid to the ethical aspects of AI.

Although the terminology is changing, the essence of AI ethics has slowly emerged. The principles of the OECD aim at the supervision of reliable AI, similar declarations were formulated in the principles of the United Center for Artificial Intelligence (where traceability, reliability and manageability are the most important principles). The EU's high-level expert group focuses on the principles of data protection, data management, transparency, non-discrimination and fairness (Tilesch & Hatamleh, 2021).

2.1. UNESCO – Recommendation on the ethics of artificial intelligence

AI is present in the lives of billions of people, even without their knowledge, transforming society unnoticed. Its application has many advantages, as it helps in finding a job in addition to completing studies. However, in addition to the benefits, these applications also generate risks and challenges. AI has significant social and cultural implications, raising issues of freedom of speech and expression, right to privacy, property rights, discrimination, manipulation and distortion of information. In addition to the former, AI poses challenges related to human cognitive ability and its interaction. Algorithms are able to support the spread of disinformation, and can influence political and ideological attitudes. Deep learning processes can strengthen the institution of bias, which is opposed to the requirement of equal treatment, highlights the asymmetry between individual social strata and groups, increasing the digital divide, thereby also the chances of digital disconnection - states UNESCO's preliminary study on artificial intelligence[4].

In order to mitigate these risks and overcome the challenges, it is necessary to establish both international and national regulatory frameworks based on UNESCO's position. In accordance with the above, in November 2021, UNESCO adopted its recommendation on AI at the General Conference. Regarding its antecedents, at the 40th meeting in November 2019, they voted to develop a global standard-setting tool. Accordingly, the Ad Hoc Expert

---

2 *Deloitte*, 2018. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey-2018.html

3 *Why addressing ethical questions in AI will benefit organization*. https://www.expertbibliotheek.nl/publicaties/data

4 *Recommendation on the Ethics of Artificial Intelligence*. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

Group held multidisciplinary consultations for the comprehensive implementation of the text of the recommendation[5].

UNESCO's recommendation deals with ethical issues related to the fields of AI, in which it approaches AI and the relevant ethical propositions as a normative reflection. It defines ethics as the basis for evaluating AI technologies as a compass. The purpose of the recommendation is not to define the definition of AI, but rather to formulate a recommendation regarding the issues that are of central importance from an ethical point of view[6]. The goal is to create a globally accepted normative instrument, in which values such as human rights, freedoms, dignity, environmental protection, diversity, inclusion, the creation of peaceful and equitable societies and the formulation of basic principles also formulate specific policy recommendations[7], primarily regarding the fact that the member states need to introduce frameworks for ethical impact assessments, during which they carry out risk assessment, supervision measures and create mechanisms for security guarantees[8]. In addition to ethical considerations, it also places great emphasis on data protection, international cooperation and development, and also touches on environmental and gender issues.

2.2. Ethical guidelines of the high-level expert group (HLEG) established by the European Commission

Since its inception, the European Union has faced a number of social and environmental challenges, the range of which has expanded exponentially with the increase in the number of member states. Recognizing the importance of sustainability and climate change, the EU has committed itself to several international initiatives, such as the climate convention[9] or the UN Sustainable Development Goals[10]. Efforts have also been made within the EU along the principles of sustainability, within the framework of which the Commission established the High-Level Expert Group (HLEG) investigating the ethical aspects of AI[11].

In the comprehensive work of the HLEG, the guideline covers all sectors of the application of artificial intelligence and defines not only the development, but also the moral framework of its use (Pusztahelyi, 2019). The recommendations of the expert group contributed to the Commission's initiatives such as Communication on building trust in human-centered

5 *Recommendation on the Ethics of Artificial Intelligence*. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

6 *Ibid*.

7 *Ibid*.

8 *Ibid*.

9 *The Paris Agreement*. https://www.un.org/en/climatechange/paris-agreement

10 *Sustainable Development Goals*. https://sdgs.un.org/goals

11 *High-Level Expert Group (HLEG)*. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

artificial intelligence[12], the White Paper on Artificial Intelligence: A European Approach to Excellence and Trust[13], and the updated Coordinated Plan for Artificial Intelligence[14].

The HLEG already states in the executive summary that three basic conditions must be met in the case of AI, which are the following: legality, ethics, stability. The publication establishes the framework for the implementation of reliable artificial intelligence, which provides guidelines for ethical and stable AI, in terms of legality, as Réka Pusztahelyi draws attention to, it starts from the fact that the application of AI takes place within the framework provided by the standards (Pusztahelyi, 2019).

In the first chapter, starting from fundamental rights, it lays down the foundations of artificial intelligence, and then lays down the purpose of ethics[15], at the same time, it states that a code of ethics for a subfield cannot replace ethical reasoning. Similar to the UNESCO publication, it is based on fundamental rights (values): dignity, freedom, justice, democracy, equality, the rule of law, and non-discrimination appear as ethical and legal entitlements. In addition, it lays down 4 basic principles: respect for human autonomy, prevention of harm, fairness and explainability, which principles go beyond legal requirements, but the fulfillment of these principles causes a serious dilemma for programmers (Pusztahelyi, 2019). In addition to the general principles, it lays down seven requirements, which together must be met to ensure the full application of the ethical principles. These are: human agency and supervision (including fundamental rights, human agency and human supervision), technical stability and security (resistance to attacks and security, accuracy, reliability and reproducibility), data protection and data management (respect for privacy, quality and integrity of data, access to data), transparency (including traceability, explainability), diversity, non-discrimination and fairness (avoidance of unfair bias, accessibility and universal design, and interested parties issues of the participation of parties) social and environmental well-being (emphasizing the line of sustainability and environmental protection) and accountability (emphasizing auditability, minimization of negative effects, compromises and legal remedies)[16]. The publication emphasizes that these requirements are equally important, that their implementation is necessary during the entire cycle of AI application, but at the same time, some of these requirements can also be found in existing legal regulations.

---

12 Communication on building trust in human-centered artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence

13 White Paper on Artificial Intelligence: a European approach to excellence and trust. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

14 Coordinated Plan on Artificial Intelligence 2021 Review, 2021. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

15 Ethics Guidelines for Trustworthy AI, 2021. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

16 *Ibid*.

2.3. European Parliament – Framework for Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies

Following the presentation of UNESCO's recommendation and the guidelines of the HLEG, it is essential to present the framework regulation of the European Parliament (EP). The basic principles of the proposal are to support the use of AI, robotics, and other related technologies and their alignment with ethical principles. The purpose of the proposal is to create a regulation on ethical principles for the development, introduction and use of artificial intelligence, robotics and related technologies[17]. The purpose of the regulation is to establish a comprehensive and durable EU regulatory framework for ethical principles and legal obligations.

Ibán Garcíadel Blanco, the representative responsible for the report, highlighted the advantages and risks of artificial intelligence: the goal is to achieve a more sustainable and just society, but at the same time, a great emphasis must be placed on the protection of privacy and the avoidance of discrimination. It is essential to create a regulation that lays down the basic ethical principles. In addition to building trust, creating security is also necessary to create human-centric AI. Accordingly, the decree is created along principles such as the evaluation of high-risk AI, robotics and similar technologies, ensuring safety, transparency, accountability, guarantees against discrimination, legal remedies and the right to do so, social responsibility, sustainable AI technologies, respecting privacy and applying restrictions on biometric identification, as well as providing appropriate control over the data used and generated by the technologies. The regulation stipulates: in the Union, any artificial intelligence, robotics and related technologies – including the software, algorithms and data used or produced by such technologies – in accordance with EU legislation, human dignity, autonomy and security, as well as other fundamental rights enshrined in the Charter are fully must be developed, implemented and used with due respect[18]. The text of the decree was adopted on October 20, 2020[19], the proposal has been submitted. The legal basis for submitting the proposal is Article 114 of the TFEU on the adoption of measures to ensure the creation and operation of the single internal market. The proposal is one of the central elements of the single digital market strategy. The proposal is based on already existing legal regulations, is proportionate and necessary to achieve its objectives. It takes a risk-based approach

17 *Framework of ethical aspects of artificial intelligence, robotics and related technologies* (2020/2012(INL), 2020). https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1636985&t=d&l=en

18 *Ibid*.

19 *Ibid*.

and imposes a regulatory burden only when AI systems are likely to pose a high risk to fundamental rights or security[20].

It can be seen that the framework regulation of the EP is structured along similar ethical principles as the recommendation of UNESCO and the guidelines of the HLEG.

2.4. Other recommendations and guidelines

The documents detailed above contain general guidelines and recommendations, with a similar logical structure and highlighting the importance of guarantees of almost identical ethical values. In the following, I would like to briefly describe the publications that are not aimed specifically at standard-setting, and that strive to implement sector-specific ethical principles.

The Institute of Electrical and Electronics Engineers (IEEE), an international organization, also issued its AI-related ethical guidelines entitled Ethically Aligned Design (EAD) in 2019. The goal of the IEEE Global Initiative is to provide pragmatic and directional insights and recommendations that serve as a key reference for technologists, educators, and policy makers (Pusztahelyi, 2019). In 2021, the organization introduced the IEEE7000 ethical standard with the aim of enabling the development of ethical and fair AI systems. This standard is already being examined by the European Commission. The standard goes beyond the issue of data protection, transparency and reliability, as they also conducted an analysis of the social consequences of technology, whether technology changes the character of an autonomous person. The standard provides engineers with a clear system design and development framework. It uses various ethical theories to elicit relevant values and then ranks them using corporate or industry value lists. It then derives a new artifact called an “ethical value requirement” (EVR), which is translated into system requirements. The system requirements are derived using a risk assessment (Spiekermann et al., 2022).

The United States Department of Defense also created the Joint Artificial Intelligence Center (JAIC), which launched its mission initiatives in 2019. Although the JAIC formulates guidelines for AI primarily in the military field, I still consider it significant from an ethical point of view, considering that in order to limit the endangerment of innocent civilians, the key priority of AI systems is to comply with the legal provisions, from the first moment the requirements are established until to the last rigorous testing step[21].

---

20 *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

21 *Joint Artificial Intelligence Center*. https://www.defense.gov/News/News-Stories/Article/Article/2418970/joint-artificial-intelligence-center-has-substantially-grown-to-aid-the-warfigh/

Not only organizations established at the international or national level dealt with the ethical issues of AI: IBM also considered it important to incorporate ethics into the design and development process[22], and Adobe also recorded its ethical principles related to AI[23].

The Russian Commission for the Implementation of the AI Ethics Code has identified the protection of human interests and rights as a key priority for the development of AI technologies, which, in addition to addressing liability issues and possible consequences, also stipulates that AI technologies should be used for their intended purpose and integrated only in areas where they benefit people[24].

3. Ethical issues related to AI in the world of work – based on the EPRS paper on AI

AI and related technologies are expected to result in numerous economic and social benefits in all sectors, affecting the financial sector, healthcare, and agriculture. The use of AI is particularly useful for improving forecasts and optimizing certain operations. However, we must be careful that the consequences of AI systems are fundamental rights protected by the Charter of Fundamental Rights, and that AI systems can threaten fundamental rights such as equal treatment, non-discrimination, human dignity, protection of personal data and privacy[25].

AI systems are capable of processing data sets more accurately and faster than humans, but their use carries significant risks in cases where AI makes its decision without human intervention. In the case of systems that avoid human control or operate with minimal human intervention, many ethical questions may arise, so in the absence of a certain level of testing and security guarantees, I believe that these work processes cannot be trusted.

In healthcare, AI can also be used in many cases to perform diagnostic tasks. Machine learning takes place on the basis of samples included in the data set provided to it, however, the data may be distorted due to external influences, and accordingly the result produced by the AI may also be distorted. The ethics of artificial intelligence in radiology: In the presentation on the European and North American multi-social declaration, referring to the IEEE standard, it was emphasized that the priority of the human factor

22 *IBM: everyday ethics for Artificial Intelligence*. https://www.ibm.com/design/ai/ethics/everyday-ethics/

23 *Adobe AI ethics principles*. https://www.adobe.com/about-adobe/aiethics.html

24 *AI Alliance in Russia – AI Ethics Code*. https://ethics.a-ai.ru/

25 *European Parliamentary Research Service: Artificial intelligence act*. (2021). https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

is necessary, that is, that AI should be subordinated to human judgment, supervision and control[26].

Like healthcare, education, transport and energy, finance and banking, employment is also classified by the EP as a high-risk sector from the point of view of AI. AI systems are increasingly used in decision-making processes in the world of work. The consideration of ethical aspects is also significant during the practical use of AI, especially in the case of vulnerable or disadvantaged groups and persons, or in cases where the positions of the parties are unequal, be it the case of business and consumer or employer and employee (relevant from the point of view of this study). In the latter case, it is the actual subordination-superiority relationship that establishes the need for the enforcement of ethical principles.

Accordingly, the Commission presented a proposal for the AI regulatory framework in 2021. The general goal of the proposal is to create the right conditions for the development of reliable AI systems, and to this end, it defines a harmonized legal framework. In addition, it defined objectives that ensure the compliance of AI systems with EU law, legal certainty, and promote the creation and maintenance of the single market. In addition, the new normative framework would establish a technologically neutral definition of AI systems. The proposal analyzes in detail the system of risks related to use. It uses a risk-based approach, based on which it ranks individual AI systems: from an unacceptable rating to systems with high risk or limited risk. The proposal lays down the set of requirements for high-risk systems: the requirement for prior compliance assessment, the service provider's registration in the EU-level database, testing, technical stability, transparency, the requirement for human supervision and cyber security. In addition, it fixes the issue of the coordination of the proposal and the EU standards that are being developed[27].

In Annex III of the proposal, in addition to AI systems suitable for biometric identification and student evaluation in education, the group of high-class AI systems also includes systems related to access to employment and self-employment. Pursuant to this, distributors and service providers of AI systems used to advertise and prescreen job vacancies, recruit people, select employees, and evaluate candidates based on interviews must meet a number of requirements before putting the system into operation, and strict regulation and supervision apply to the entire operation process[28].

---

26 *Ethical Issues of Using AI in radiology*. https://www.neuroct.hu/blog/a-mesterseges-intelligencia-alkalmazasanak-etikai-kerdesei-a-radiologiaban

27 *Harmonising Artificial Intelligence: The Role of Standards in the EU AI Regulation*. https://montrealethics.ai/harmonizing-artificial-intelligence-the-role-of-standards-in-the-eu-ai-regulation/

28 *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

In connection with the employment relationship, we can already encounter work done by AI in the field of selection and recruitment in the first round - often without our knowledge, AI selects our CV from the multitude of applications received. The use of AI for this purpose significantly increases the efficiency of selection from the employer's side. However, in my opinion, an application cannot replace the work of an HR specialist, rather it should be seen as a kind of supplementary and auxiliary activity, I believe that human interaction in the context of a personal interview cannot be reproduced by an AI or even a robot.

During the selection process, AI ensures the conditions of justice, reduces the risk of bias and prejudice - during recruitment, candidates are selected solely on the basis of their professional experience and skills. At the same time, this advantage can generate ethical risks in terms of the organization's receptiveness. An excellent example of this is Amazon Inc.'s HR robot, which was trained based on resumes submitted to the company over a 10-year period. Given that the technical field is characterized by male dominance, based on the CVs submitted by men, the AI has «trained» itself to give preference to applications submitted by men. It automatically rejected resumes that included the word «female», thereby violating gender equality as an ethical principle (Dastin, 2018).

In connection with the above case, the question may arise as to how ethical it is for the employer to treat employees as a set of data? Data sets can typically be modeled, but in the absence of context, the value of the data is reduced or completely meaningless. As an example, Boyd and Crawford cite the graphical modeling of the network of personal relationships mapped on the basis of social media, which may provide data, but does not provide complete and accurate information about the relationships under investigation (Boyd & Crawford, 2012). In the case of selecting employees, we can face a similar ethical dilemma: can a recruiting AI map the candidate's suitability? The received CVs are practically taken out of context, impersonal data sets, from which, although conclusions can be drawn, I believe that we cannot fully rely on them.

A similar dilemma can also arise if AI is used to analyze the workforce to carry out a group downsizing process. In Annex III of the proposal, AI software whose purpose is to promote or terminate work-related contractual relationships and to make related decisions are also classified as high-risk systems. The same classification applies to systems that monitor and evaluate the performance and behavior of persons in a legal relationship. (AI Act Proposal). There is no doubt that AI makes fair decisions based on performance evaluation, but is it ethical to get rid of an employee who performs less well at a given moment, who is raising his children alone, just because the AI ranked him in a lower category during the performance evaluation? Although the usefulness of AI in certain processes is indisputable, I do not support treating employees as a set of data. I believe that when making decisions, we must consider other factors, such as the employee's other opportunities for further training and retraining, the employee's flexibility, and we cannot forget the issue of his/her family background or belonging to other disadvantaged groups. In this regard, when developing the operating regulations

of a member state or a specific organization, in addition to striving for innovation, we must consider the definition of the machine-human relationship and the supervision-responsibility relationships and (in accordance with the EU legislative proposal) it is necessary to ensure the transparency of the application of AI and to ensure the decision the possibility of human override. It can be seen that machines are increasingly involved in ethical decisions that require ethical explanation. Current machine learning algorithms are ethically inscrutable, but not very different from human behaviour. The article by Marius Dorobantu, Yorick Wilks explores the role of rationality and reasoning in traditional ethical thinking and artificial intelligence, emphasising the need for some explanation of actions. He explores Neil Lawrence's embodiment factor as an approach to the differences between human and machine intelligence, linking it to a theological understanding of personhood, and proposes the notion of artificial moral ortheses that can provide ethical explanations for both artificial and human agents as a more promising unifying approach to human and machine ethics (Dorobantu & Wilks, 2019).

Conclusions

In the era of Artificial Intelligence (AI), ethics has taken on a whole new level of importance and debate (Sudhi & Huraimel, 2021). To date, no consensus has been reached regarding the risks of AI. Although it would be a shame to block innovation, there is no doubt that based on some ethical issues, human supervision is necessary for the application of AI. The guidelines, recommendations and draft regulations briefly presented in this study are aimed at this as well, which show that the goal is to create a legal framework in which the technology can be successfully applied in the future.

Innovation is at the heart of every civilisation. The Belmont Report of 1978 outlined three ethical principles: respect for persons, charity and justice, which have formed the basis of human sciences research. However, the Independent Human Research Review Committees and their regulations are struggling to cope with internet research ethics, big data and artificial intelligence research, as evidenced by the 2014 Facebook* Emotional Contagion study and the controversies surrounding the 2016 «AI gaydar» research (Tang, 2020).

As can be seen from the analyzes above, the ethical risk of artificial intelligence capable of autonomous operation poses a significant challenge to regulation. Machine learning has become a popular tool in many criminal justice applications, including sentencing and policing. However, there is also the potential for predictive policing systems to create uneven effects and exacerbate social injustices. Although previous research has shown that machine learning models can effectively handle certain tasks, they are prone to replicating the systemic bias of previous human decision makers. However, little academic research has addressed the importance of fairness in machine learning applications in policing (Alikhademi et al., 2022).

The relevant recommendations provide a framework for the development of regulations; however, it is necessary to harmonize international and national standards and to reorganize the standards systems. Labor regulations can also play a big role in this approach. The binding contracts and resources that set the goal of achieving full employment will be of decisive importance, and the instruments regulating collective redundancies will also play a significant role in this regard (Stefano, 2018). Efforts to regulate AI are visible not only in the European Union: while a risk-based approach to the problem is typical at EU level, more specific guidelines have been proposed in the United States (Sussman, 2021). However, the common feature is that in order to promote research and development and to expand the use of AI technology, it is necessary to create an infrastructure that encourages the cooperation of the actors and ensures operation according to regulatory and, not least, ethical frameworks. The interaction between AI (robots) and humans must be understood, as well as the issues of emotional intelligence - points out József Hajdú (Hajdú, 2020).

States have recognized the importance of AI and have begun to develop a regulatory environment centered on an autonomous AI strategy. On April 25, 2018, Hungary signed the Declaration on Cooperation on Artificial Intelligence together with 25 European states, which recorded the intention of the signatories to cooperate in the field of European AI developments and the support of AI-supported innovation[29].

We are at a turning point in the debate on the ethics of artificial intelligence (AI), because we are witnessing general-purpose AI textual agents, such as GPT-3, that can generate large-scale, highly refined content that appears to be written by humans. On the other side, there is the Natural Language Processing (Krutilla & Kővári, 2022). We can see a lack of discussion in the business community about the ethical issues surrounding the merging of the roles of humans and machines in content production (Illia et al.,2023).

Ethical decision-making is the central issue of our time. In my opinion, autonomous intelligent systems do not consider the human needs appearing on the human side, they make an objective-based decision. Therefore, it is necessary to create a human-centered AI that is in line with the values and ethical principles of society, that is, to put brakes necessary from an ethical point of view in order to protect the institution of fairness and dignity.

Based on the above, it can be said that the application of AI results in a complex change. For this, the creation of regulatory frameworks and ensuring transparency are essential. From this aspect, I wanted to present the individual regulatory solutions in this study, highlighting the question of the ethics of AI applied in labor relations.

"AI will never be ethical. It is a tool, and like any tool, it is used for good and bad. There is no such thing as a good AI, only good and bad humans. We are not smart enough to make AI ethical. We are not smart enough to make AI moral … In the end, I believe that the

[29] *Hungary's Artificial Intelligence Strategy 2020-2030*. https://ai-watch.ec.europa.eu/countries/hungary/hungary-ai-strategy-report_en

only way to avoid an AI arms race is to have no AI at all. This will be the ultimate defense against AI." (Connock, 2023).

* The organization is recognized as extremist, its activity is prohibited in the territory of the Russian Federation.

References

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., & Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, *4*(23), 1–17. https://doi.org/10.1007/s10506-021-09286-4

Beranger, J. (2021). *Societal Responsibility of Artificial Intelligence: Towards an Ethical and Eco-responsible AI*. UK: Wiley-Iste.

Boyd, D., & Crawford, K. (2012). Az adatrengeteg kínos kérdései. *Információs Társadalom*, *12*(2), 7. https://doi.org/10.22503/inftars.xii.2012.2.1

Candriam Academy. (2022). *What is the European Commission's HLEG?*

Capgemini Research Institute. (2019). *Why addressing ethical questions in AI will benefit organization*.

Capurro, R. (2018). Digital Ethics. I*nternational Journal of Applied Research on Information Technology and Computing*, *9*/1, 23–31.

Connock, A. (2023). *Media Management and Artificial Intelligence: Understanding Media Business Models in the Digital Age*. UK: Routledge.

Cyman, D., Gromova, E., & Juchnevicius, E. (2021). Regulation of Artificial Intelligence in BRICS and the European Union, *BRICS Law Journal*, *8*(1), 86–115. https://doi.org/10.21684/2412-2343-2021-8-1-86-115

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Dorobantu, M., & Wilks, Y. (2019). Moral orthoses: a new approach to human and machine ethics. *Zygon Journal of Religion and Science*, *54*(4), 12–23. https://doi.org/10.1111/zygo.12560

Eszteri, D. (2015). A mesterséges intelligencia fejlesztésének és üzemeltetésének egyes felelősségi kérdései. *Infokommunikáció és Jog*, 47–57. ISSN 1786-0776

Fleischmann, S. S., Greenberg, K. R., Verma, N., Cummings, B., Li, L., & Shenefel, C. (2023). Locazing the work of artificial intelligence ethics. *Journal of the Association for Information Science and Technology*, *74*(3), 311–322. https://doi.org/10.1002/asi.24638

Fobel, P. (2002). Alkalmazott filozófia és etika. In S. Karikó, & S. Karikó (Szerk.), *Az alkalmazott filozófia esélyei*. Budapest: Áron Kiadó.

Hajdú, J. (2020). A mesterséges intelligecia hatása a munkaerőpiacra, avagy elveszik-e a robotok az ember munkáját. *Infokommunikáció és Jog*, *7*.

Harmathy, A. (2019). A polgári jog a változó jogrendszerben. In V. Lamm, & A. Sajó, *Studia in honorem Lajos Vékás*. Budapest: HVG-ORAC Lapés Könyvkiadó Kft.

Hoffmann, A. L., Roberts, S. T., Wolf, C. T., & Wood, S. (2019). Beyond fairness, accountability, and transparency in the ethics of algorithms: Contributions and perspectives from LIS. *Proceedings of the Association for Information Science and Technology*, *55*(1), 694–696. https://doi.org/10.1002/pra2.2018.14505501084

Illia, L., Colleoni, E., & Zyglidopoulo, S. (2023). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility*, *32*(1), 201–210. https://doi.org/10.1111/beer.12479

Karvalics, Z. L. (2015). Mesterséges intelligencia – a diskurzusok újratervezésének kora. *Információs Társadalom.*

Knowles, M. A. (2021). Five Motivating Concerns for AI Ethics Instruction. *Proceedings of the Association for Information Science and Technology*, *58*, 472–476. https://doi.org/10.1002/pra2.481

Krutilla, Z., & Kővári, A. (2022). The origin and primary areas of application of natural language processing. In *2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics Automation Computer Science and Robotics* (pp. 293–298). https://doi.org/10.1109/cinti-macro57952.2022.10029432

Legeza, L. (2013). *Mérnöki etika*. Budapest.

Müller, J., & Kerényi, Á. (2019). A biizalom ésetika igénye a digitális korszakban. *Hitelintézeti Szeml*, *18/4*, 8–19.

Négyesi, I. (2020). A mesterséges intelligencia és az etika. *Társadalomtudomány*, *104.*

Németh, G. (2021). Jogászi etikai kihívások a technológiai fejlődés tükrében: az etika és jog innovációjának aktuális kérdései. *Tanulmányok*. http://real.mtak.hu/108838/1/JAP-2020-01_NG.pdf

Pusztahelyi, R. (2019). Bizalmunkra méltó MI – A mesterséges intelligencia fejlesztésének és alkalmazásának erkölcsi-etikai vonatkozásairól. *Publicationes Universitatis Miskolcinensis Sectio Juridica et Politica*, *XXXVII/2*, 99.

Spiekermann, S., Krasnova, H., Hinz, O., Baumann, A., Benlian, A., Grimple, H., & Trenz, M. (2022). Values and Ethics in Information Systems. *Bus Inf Syst Eng*, *64*, 247–264. https://doi.org/10.1007/s12599-021-00734-8

Stefán, I. (2020). A mesterséges intelligencia fogalmának polgári jogi értelmezése. *Pro Futuro*, *1*, 29–39.

Stefano, V. (2018). "Negotiating the algorithm": Automation, artificial intelligence and labour protection. *Employment Working Paper*, 246.

Sudhi, S., & Huraimel, K. (2021). Dealing with Ethics, Privacy, and Security. In *Reimagining Businesses with AI* (pp. 193–206).

Sussman, E. H. (2021). *U. S. Artificial Intelligence Regulaton takes shape.*

Tang, B. (2020). Independent AI Ethics Committees and ESG Corporate Reporting on AI as Emerging Corporate and AI Governance Trends. In S. Chishti, I. Bartoletti, A. Leslie, & S. Millie, *The AI Book: The Artificial Intelligence Handbook for Investors, Entrepreneurs and FinTech Visionaries*. https://doi.org/10.1002/9781119551966.ch48

Tilesch, G., & Hatamleh, O. (2021). *Mesterség és Intelligencia.* Libri Kiadó.

Turay, A. (2000). *Az ember és az erkölcs- Alapvető etika Aquinói Tamás nyomán.* Szeged: Agapé, Ferences Nyomda és Könyvkiadó Kft.

Yu, H., Shen, Zh., Miao, Ch., Leung, C., Lesser, V. R., & Yang, O. (2018). *Building ethics into Artficial Intelligence*. http://arxiv.org/pdf/1812.02953.pdf

Author information

Zsofia Riczu – PhD candidate, Faculty of Law, Agricultural and Labour Law Department, University of Miskolc
Address: Miskolc-Egyetemváros, Miskolc, Hungary
E-mail: jogriczu@uni-miskolc.hu
ORCID ID: https://orcid.org/0000-0002-4024-5833

Conflict of interest

The author declares no conflict of interest.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – April 9, 2023
Date of approval – April 22, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 341.1/8:004.8
EDN: https://elibrary.ru/cpffyw
DOI: https://doi.org/10.21202/jdtl.2023.21

Рекомендации по этическим аспектам искусственного интеллекта в приложении к сфере трудовых отношений

София Рицу

Университет Мишкольца
г. Мишкольц, Венгрия

Ключевые слова

Законодательство,
искусственный интеллект,
право,
принципы права,
труд,
трудовое право,
трудовые отношения,
цифровизация,
цифровые технологии,
этика

Аннотация

Цель: распространение и широкое использование искусственного интеллекта выдвигает на первый план не только проблемы защиты данных, но и этические вопросы. Цель данной статьи – изучение этических аспектов искусственного интеллекта и предложение рекомендаций для его использования в трудовом праве.

Методы: исследование основано на методах сравнительного и эмпирического анализа. Сравнительный анализ позволил изучить положения современного трудового законодательства в контексте искусственного интеллекта. Эмпирический анализ выявил этические проблемы, относящиеся к искусственному интеллекту в сфере труда, путем изучения спорных случаев использования искусственного интеллекта в различных областях, таких как здравоохранение, образование, транспорт и др.

Результаты: частноправовые аспекты этических проблем искусственного интеллекта были изучены в контексте этических и трудовых вопросов права, влияющих на процесс отбора с помощью искусственного интеллекта и на обращение с работниками с точки зрения работодателя. Автор выделяет как общие аспекты этики, так и вопросы цифровой этики. Предложены отдельные международные рекомендации относительно этики искусственного интеллекта.

Научная новизна: исследование посвящено изучению этических аспектов использования искусственного интеллекта в конкретной отрасли частного права – трудовом праве. Автор дает рекомендации относительно этических аспектов использования искусственного интеллекта в данной сфере.

Практическая значимость: исследование восполняет имеющиеся пробелы в научной литературе по указанному вопросу. Результаты работы могут использоваться в процессе законотворчества и служить базой для дальнейших исследований.

Для цитирования

Рицу, С. (2023). Рекомендации по этическим аспектам искусственного интеллекта в приложении к сфере трудовых отношений. *Journal of Digital Technologies and Law*, *1*(2), 498–519. https://doi.org/10.21202/jdtl.2023.21

Список литературы

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., & Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, *4*(23), 1–17. https://doi.org/10.1007/s10506-021-09286-4

Beranger, J. (2021). *Societal Responsibility of Artificial Intelligence: Towards an Ethical and Eco-responsible AI*. UK: Wiley-Iste.

Boyd, D., & Crawford, K. (2012). Az adatrengeteg kínos kérdései. *Információs Társadalom*, *12*(2), 7. https://doi.org/10.22503/inftars.xii.2012.2.1

Candriam Academy. (2022). *What is the European Commission's HLEG?*

Capgemini Research Institute. (2019). *Why addressing ethical questions in AI will benefit organization*.

Capurro, R. (2018). Digital Ethics. I*nternational Journal of Applied Research on Information Technology and Computing*, *9*/1, 23–31.

Connock, A. (2023). *Media Management and Artificial Intelligence: Understanding Media Business Models in the Digital Age*. UK: Routledge.

Cyman, D., Gromova, E., & Juchnevicius, E. (2021). Regulation of Artificial Intelligence in BRICS and the European Union, *BRICS Law Journal*, *8*(1), 86–115. https://doi.org/10.21684/2412-2343-2021-8-1-86-115

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Dorobantu, M., & Wilks, Y. (2019). Moral orthoses: a new approach to human and machine ethics. *Zygon Journal of Religion and Science*, *54*(4), 12–23. https://doi.org/10.1111/zygo.12560

Eszteri, D. (2015). A mesterséges intelligencia fejlesztésének és üzemeltetésének egyes felelősségi kérdései. *Infokommunikáció és Jog*, 47–57. ISSN 1786-0776

Fleischmann, S. S., Greenberg, K. R., Verma, N., Cummings, B., Li, L., & Shenefel, C. (2023). Locazing the work of artificial intelligence ethics. *Journal of the Association for Information Science and Technology*, *74*(3), 311–322. https://doi.org/10.1002/asi.24638

Fobel, P. (2002). Alkalmazott filozófia és etika. In S. Karikó, & S. Karikó (Szerk.), *Az alkalmazott filozófia esélyei*. Budapest: Áron Kiadó.

Hajdú, J. (2020). A mesterséges intelligecia hatása a munkaerőpiacra, avagy elveszik-e a robotok az ember munkáját. *Infokommunikáció és Jog*, *7*.

Harmathy, A. (2019). A polgári jog a változó jogrendszerben. In V. Lamm, & A. Sajó, *Studia in honorem Lajos Vékás*. Budapest: HVG-ORAC Lapés Könyvkiadó Kft.

Hoffmann, A. L., Roberts, S. T., Wolf, C. T., & Wood, S. (2019). Beyond fairness, accountability, and transparency in the ethics of algorithms: Contributions and perspectives from LIS. *Proceedings of the Association for Information Science and Technology*, *55*(1), 694–696. https://doi.org/10.1002/pra2.2018.14505501084

Illia, L., Colleoni, E., & Zyglidopoulo, S. (2023). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility*, *32*(1), 201–210. https://doi.org/10.1111/beer.12479

Karvalics, Z. L. (2015). Mesterséges intelligencia – a diskurzusok újratervezésének kora. *Információs Társadalom*.

Knowles, M. A. (2021). Five Motivating Concerns for AI Ethics Instruction. *Proceedings of the Association for Information Science and Technology*, *58*, 472–476. https://doi.org/10.1002/pra2.481

Krutilla, Z., & Kővári, A. (2022). The origin and primary areas of application of natural language processing. In *2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE*

International Conference on Recent Achievements in Mechatronics Automation Computer Science and Robotics (pp. 293–298). https://doi.org/10.1109/cinti-macro57952.2022.10029432

Legeza, L. (2013). *Mérnöki etika*. Budapest.

Müller, J., & Kerényi, Á. (2019). A biizalom ésetika igénye a digitális korszakban. *Hitelintézeti Szeml*, *18/4*, 8–19.

Négyesi, I. (2020). A mesterséges intelligencia és az etika. *Társadalomtudomány*, *104*.

Németh, G. (2021). Jogászi etikai kihívások a technológiai fejlődés tükrében: az etika és jog innovációjának aktuális kérdései. *Tanulmányok*. http://real.mtak.hu/108838/1/JAP-2020-01_NG.pdf

Pusztahelyi, R. (2019). Bizalmunkra méltó MI – A mesterséges intelligencia fejlesztésének és alkalmazásának erkölcsi-etikai vonatkozásairól. *Publicationes Universitatis Miskolcinensis Sectio Juridica et Politica*, *XXXVII/2*, 99.

Spiekermann, S., Krasnova, H., Hinz, O., Baumann, A., Benlian, A., Grimple, H., & Trenz, M. (2022). Values and Ethics in Information Systems. *Bus Inf Syst Eng*, *64*, 247–264. https://doi.org/10.1007/s12599-021-00734-8

Stefán, I. (2020). A mesterséges intelligencia fogalmának polgári jogi értelmezése. *Pro Futuro*, *1*, 29–39.

Stefano, V. (2018). "Negotiating the algorithm": Automation, artificial intelligence and labour protection. *Employment Working Paper*, 246.

Sudhi, S., & Huraimel, K. (2021). Dealing with Ethics, Privacy, and Security. In *Reimagining Businesses with AI* (pp. 193–206).

Sussman, E. H. (2021). *U. S. Artificial Intelligence Regulaton takes shape.*

Tang, B. (2020). Independent AI Ethics Committees and ESG Corporate Reporting on AI as Emerging Corporate and AI Governance Trends. In S. Chishti, I. Bartoletti, A. Leslie, & S. Millie, *The AI Book: The Artificial Intelligence Handbook for Investors, Entrepreneurs and FinTech Visionaries*. https://doi.org/10.1002/9781119551966.ch48

Tilesch, G., & Hatamleh, O. (2021). *Mesterség és Intelligencia.* Libri Kiadó.

Turay, A. (2000). *Az ember és az erkölcs- Alapvető etika Aquinói Tamás nyomán.* Szeged: Agapé, Ferences Nyomda és Könyvkiadó Kft.

Yu, H., Shen, Zh., Miao, Ch., Leung, C., Lesser, V. R., & Yang, O. (2018). *Building ethics into Artficial Intelligence*. http://arxiv.org/pdf/1812.02953.pdf

Сведения об авторе

София Рицу – аспирант кафедры аграрного и трудового права, Университет Мишкольца
Адрес: Университетский городок, г. Мишкольц, Венгрия
E-mail: jogriczu@uni-miskolc.hu
ORCID ID: https://orcid.org/0000-0002-4024-5833

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.87.91 / Международное право в практике отдельных государств
Специальность ВАК: 5.1.5 / Международно-правовые науки

История статьи

Дата поступления – 9 апреля 2023 г.
Дата одобрения после рецензирования – 22 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.22

Ethical-Legal Models of the Society Interactions with the Artificial Intelligence Technology

Dmitriy V. Bakhteev

Ural State Law University named after V. F. Yakovlev
Ekaterinburg, Russian Federation

Keywords

Artificial Intelligence,
ChatGPT,
digital technologies,
ethics,
law,
machine learning,
model,
regulation,
robot,
society

Abstract

Objective: to explore the modern condition of the artificial intelligence technology in forming prognostic ethical-legal models of the society interactions with the end-to-end technology under study.

Methods: the key research method is modeling. Besides, comparative, abstract-logic and historical methods of scientific cognition were applied.

Results: four ethical-legal models of the society interactions with the artificial intelligence technology were formulated: the tool (based on using an artificial intelligence system by a human), the xenophobia (based on competition between a human and an artificial intelligence system), the empathy (based on empathy and co-adaptation of a human and an artificial intelligence system), and the tolerance (based on mutual exploitation and cooperation between a human and artificial intelligence systems) models. Historical and technical prerequisites for such models formation are presented. Scenarios of the legislator reaction on using this technology are described, such as the need for selective regulation, rejection of regulation, or a full-scale intervention into the technological economy sector. The models are compared by the criteria of implementation conditions, advantages, disadvantages, character of "human – artificial intelligence system" relations, probable legal effects and the need for regulation or rejection of regulation in the sector.

Scientific novelty: the work provides assessment of the existing opinions and approaches, published in the scientific literature and mass media, analyzes the technical solutions and problems occurring in the recent past and present. Theoretical conclusions are confirmed by references to applied

situations of public or legal significance. The work uses interdisciplinary approach, combining legal, ethical and technical constituents, which, in the author's opinion, are criteria for any modern socio-humanitarian researches of the artificial intelligence technologies.

Practical significance: the artificial intelligence phenomenon is associated with the fourth industrial revolution; hence, this digital technology must be researched in a multi-aspectual and interdisciplinary way. The approaches elaborated in the article can be used for further technical developments of intellectual systems, improvements of branch legislation (for example, civil and labor), and for forming and modifying ethical codes in the sphere of development, introduction and use of artificial intelligence systems in various situations.

For citation

Bakhteev, D. V. (2023). Ethical-legal Models of the Society Interactions with the Artificial Intelligence Technology. *Journal of Digital Technologies and Law, 1*(2), 520–539. https://doi.org/10.21202/jdtl.2023.22

Contents

Introduction

Numerous advantages of artificial intelligence systems (including fast learning ability, ability to solve a wide range of tasks, higher that in humans efficiency) together with their increasing penetration into various sphere of life forces a question whether an artificial intelligence system is (or will be) able to perceive itself as an autonomous personality, independent of developers and users, whether artificial intelligence will realize its advantages over people, how it will estimate its position and what it will do if it wants to change it. These questions are at the intersection of the objective areas of ethics, law, and technology; hence, they should be resolved with interdisciplinary research methods (Kazim, 2021). Depending on the answers to these questions and the degree of technology development, the models/scenarios of social reaction described below are probable and, consequently, the rights to artificial intelligence technologies.

1. The tool model

A tool is a product of human activity and a means for manufacturing other objects, including tools. Karl Marx and Martin Heidegger in their works emphasized the differences between a tool and a machine. The former pointed out that cardinally different notions are often mixed up: while a tool demands immediate participation of a human in the labor process, a machine "supersedes the workman, who handles a single tool, by a mechanism operating with a number of similar tools, and set in motion by a single motive power" (Marx, 2001). He also noted that a machine differs from a tool in sufficient autonomy (mostly a resource one, as a machine is directed and controlled by a human anyway). Martin Heidegger, in turn, relying upon Hegel's works, lists such criteria of an object being a machine as autonomy, self-reliance and independence (Heidegger, 1993). For the present research, we may interpret it as follows: at the stage of development and testing, artificial intelligence within the tool concept serves more as a tool, not as a machine, as its functioning is connected with human activity at three levels: during development (a tool and a machine are similar in this aspect), during implementation of the activity previously inherent exclusively to a human, and during the human control over the results of the artificial intelligence activity. This approach is also sometimes called pragmatic (Morley et al., 2021). The "machine character" of the artificial intelligence is, in this case, a derivative characteristic from autonomy, but the studied model denies informational autonomy of the artificial intelligence systems. Accordingly, within this model artificial intelligence is viewed as a tool, including for implementing the needs of the humanity (Watkins & Human, 2023).

The theory of autonomy (informational and resource ones) described in the author's works (Bakhteev, 2021) wholly correlates with the fact that the artificial intelligence systems may be perceived as machines (which, in turn, is proved by the existence of the terms "machine learning" and "machine vision"), and as entities comparable with cognoscitive biological objects. This said, one should bear in mind that, while a tool (in traditional scientific interpretation) is intended for easing human labor, the artificial intelligence may substitute human labor with its activity. At that, it would be incorrect to compare artificial intelligence with machines of the industrial revolution era. Those machines put a lot of people out of work, but they created new jobs at the same time. In case of automation in general and using artificial intelligence systems in particular, we already observe elimination of certain positions, mainly associated with mediation services: consultants, dispatchers, marketers, etc. Standard industrial robots even now, without intellectual modules, have largely optimized assembly line production, which allowed reducing costs and manifoldly increase the product quantity and quality and put out of work a lot of unqualified and low-qualified workers; actually, we observe a new industrial revolution. By 2020, over 3 million industrial robots were used globally (however, by 2023 the growth rate decreased); intellectual system are being integrated into various

spheres of life, which apparently positively influences economy, but incurs certain harm on the society, first of all, by reducing employment. One variant of solving this problem is to legislatively guarantee employment, or to implement retraining programs, supported by the state and large corporations, for those who lost the job.

Development of the artificial intelligence allows it to solve an increasing range of intellectual problems, which creates prerequisites for further cutting jobs in an increasing number of spheres, as well as for elimination of entire professions. "The actual effect of a reduced payroll fund (and the number of jobs. – Author's note) due to introduction of robots is determined by the number of people released and the size of their payroll, as well as the cost of the robots, which, in turn, is determined by the complexity of construction and degree of intellectualization of the robots" (Timofeev, 1978).

Unlike the initial stage of robotization, spreading of the artificial intelligence systems may reduce the employment rate not for vocational jobs only. According to forecasts of experts from Oxford University, in the next 20 years, 47% of jobs in the USA and 77% of jobs in China will be automated. According to one of the leading experts in the sphere of computer facilities M. Vardi, by 2045 about 50% of people will be jobless (Vardi, 2012). One may object that the number of artificial intelligence software developers increases, but the increase of the number of programmers and other persons involved in developing intellectual systems cannot be compared to the decrease of other jobs. Moreover, the desire to create intellectual systems capable of self-replication is also evident, so one may not exclude job cuts in the spheres related to information technologies: for example, ChatGPT in its fourth version can create a simple but compilable code. Thus, in January 2023, Alphabet company announced cutting 12,000 jobs around the world, including due to introduction of various intellectual systems capable of substituting, in particular, marketing specialists, copywriters and illustrators[1]. According to the model developed for analyzing the probability of certain professions disappearing due to the use of intellectual chat-systems, a 100% substitution of humans, at the present stage of the technology development, is possible for mathematicians, taxation specialists, financial analysts, writers, copywriters, web designers, book designers, secretaries of public authorities, and news analysts (Eloundou et al., 2023).

By now, it is the tool model that may be considered the only one with full-fledged implementation: in applied activity, the artificial intelligence systems serve as a tool increasing performance. This, just like during previous technological revolutions, imposes on the state and society the function of preserving jobs and regulating the intensity of the application of the said technology.

[1] Pichar, S. (2023, January 20). *A difficult decision to set us up for the future*. https://blog.google/inside-google/message-ceo/january-update/

2. The xenophobia model

As the discussions on development and use of the artificial intelligence systems activate, the voices of opponents of further research in this sphere become louder. At that, at least a part of them cannot be called obscurants with irrational fear of the technological progress. For example, a world-renowned scientist, popularizer of science S. Hawking said: "…appearance of a full-fledged artificial intelligence may become the end of the human race... Such intelligence will take up the initiative and start improving oneself with an accelerated speed. People's abilities are restricted by too slow an evolution, we cannot contend with the speed of machines and are going to lose". Of a similar opinion is a famous American engineer, IT entrepreneur E. Musk, who stated: "I believe the artificial intelligence will sooner or later kill us all... Facebook*, Google, Amazon, Apple – they all already know a lot about you. The artificial intelligence, which will be created within these corporations, will get an enormous power over people. And concentration of power in one pair of hands always generates great risks". It was also marked that "distribution of functions between an artificial intelligence system and a human must follow the human-centered principle and always leave the opportunity for a human choice. It implies providing human control over working processes within artificial intelligence systems" (Semis-ool, 2019).

These facts determine a need to thoroughly examine the "xenophobia" model of attitude to artificial intelligence. It is worth noting that it develops the tool model along a negative scenario, i.e. both due to the progress in the development and use of the artificial intelligence systems, and to implementation of one or several risks (in the form of instantaneous negative events or long-term crises), as described above.

The term "xenophobia" is formed with two Greek roots: ξένος ("alien") + φόβος ("fear"). Thus, xenophobia is literarily defined as fear, intolerance to something alien, unknown[2].

Researchers are not unanimous about the origins of xenophobia. Some authors mark that it could appear as an adaptation tool during evolution, facilitating survival and transference of genes to offspring. For example, fear of strangers could be, inter alia, based on the observation that aliens could be carriers of new pathogenic microorganism, dangerous for the locals due to the lack of necessary antibodies.

Traditionally, the term "xenophobia" was used to denote fear, distaste for people of other races, nationalities, cultures and religions. However, in our opinion, a research of the process of interaction between humanity and the technological achievements allows using this term to describe a certain type of attitude towards the scientific-technical progress and its fruits – technologies, including, apparently, artificial intelligence.

[2] Ozhegov, S. I., & Shvedova, N. Yu. (2016). *Thesaurus of the Russian language* (p. 300). Moscow: A TEMP.

Finalizing our approach to xenophobia, we should highlight an important aspect that, ultimately, xenophobia is a specific type of fear. According to E. P. Ilyin, fear, as one of many emotions, is "an emotional state reflecting a protective biological reaction of a human or animal experiencing a fake or real danger to their health or wellbeing" (Ilyin, 2016). Further, E. P. Ilyin stated that from the biological point of view fear is, undoubtedly, a useful phenomenon, while for a human as a social creature fear is often an obstacle in achieving the set goals. In this section of our work we research the bases of the potential critical distrust of the society towards the artificial intelligence technology.

The essence of the "xenophobia approach" to artificial intelligence is in viewing it as a real threat to the humanity and its actual position in the world.

A general analysis of the artificial intelligence technology critic makes it possible to identify two main forms of fear (distrust) people fell towards this technology – essential and instrumental.

Essential fear is due to the fact that people fear not the use of the artificial intelligence technology, but artificial intelligence per se as an artificial but quite independent and autonomous intelligence, capable of being, learning, thinking, and perceiving oneself without participation of a human. Emergence of such a "manmade thinking machine", which is capable of thinking not just like a human but better than a human, undermines the human monopoly to cognitive activity which existed during the history of civilization and enabled the humans to take the dominant position among other species. In this respect, artificial intelligence becomes a separate species, which humanity cannot perceive otherwise that a competitive one. At that, it is uncontrollable artificial intelligence that causes fear, i.e., the situation of artificial intelligence gaining self-consciousness as a result of a software break or purposeful actions of a developer. Actually, such situation regarding biological processes should be considered a mutation, but it is doubtful that the processes of technological products development are so much connected with evolutionary mechanisms. That is why the scenario of an "aggressive" artificial intelligence seems extremely unrealistic.

The instrumental fear, in turn, reflects the fear of a human being ousted by the artificial intelligence systems in labor sphere, as was described in the tool model (see above).

After a systematic research of artificial intelligence began in the 1940s – 1950s, i. e. less than one hundred years ago, systems have been created which exceed humans in some types of intellectual activity. Abilities of modern computers still do not allow comprehensive modeling of a human mind or the whole world around, but artificial intelligence can very well cope with abstractions. Games are a good example of abstraction and, what is very important in this case, the results of participating in a game can be accurately assessed. Thus, since as early as the beginning of the 2000s, the world strongest chess players can oppose nothing to a computer; according to G. Kasparov, all professional chess players train by playing against chess computer programs, as a human rival cannot provide

a sufficient depth of search. In 2015, a chess computer program for the first time won a human in a Go game – one of the most complex open information games[3], which had been considered impossible. Artificial neural networks are capable of winning professional players of computer games in cybersports[4], which also had been considered an exclusive prerogative of a human.

The content of xenophobia approach consists in that artificial intelligence may be used by individual persons, organization and states as a means to achieve their malevolent goals.

A typical example of this fear is an uproar emerging soon after the US presidential elections and associated with Cambridge Analytica company. According to some sources, this private company, using the latest methods of information collection and analysis in Facebook* social network, obtained a large array of data, including personal ones, in order to develop a special political advertisement which, according to a number of experts, facilitated electing the present US President. Moreover, this organization is accused of participating in interference into the results of over 200 elections worldwide. A former staff member of Cambridge Analytica Chris Wylie marked: "We used imperfect software of Facebook* to collect millions of user profiles and to build models which allowed us to learn about people and use these data to activate their internal demons"[5].

This case shows that even now the abilities of the artificial intelligence systems are used to collect personal data and manipulate public opinion. In future, artificial intelligence can be used to manipulate large amounts of information, forming the world view for the population of entire states, which creates real threats for democratic institutions, freedom of speech and information dissemination. States may use it for imposing a certain attitude to their populations and the populations of other states; representatives of various corporations – for artificially forming demand to certain goods and services. Finally, representatives of the criminal circles may use it to collect confidential information about citizens and organizations, which may further be sold in the shadow market or used for blackmail or fraud.

Another layer of problems is using artificial intelligence in military activity. A notional combat robot equipped with artificial intelligence, or an army of such robots, is an effective

3 For example, while chess has about 20 moves per turn, the Go game has about 200.

4 Statt, N. (2019, April 13). OpenAI's Dota 2 AI steamrolls world champion e-sports team with back-to-back victories. *The Verge*. https://www.theverge.com/2019/4/13/18309459/openai-five-dota-2-finals-ai-bot-competition-og-e-sports-the-international-champion

5 Chereshnev, E. (March 20, 2018). Defenseless data: how Facebook found itself amidst the greatest controversy ever. *Forbes*. https://www.forbes.ru/tehnologii/358883-bezzashchitnye-dannye-kak-facebook-okazalas-v-centre-samogo-bolshogo-skandala-v

substitute for regular troops. It may execute complex intellectual tasks, act in most unfavorable conditions, requires no rest or sleep, and its destruction does not worry much the public opinion in the country waging the war (at least until that war becomes too expensive from the economical point of view). At that, artificial intelligence is able to solve the previously human tasks with an un-human, machine rationality. If wrongly designed, the artificial intelligence system will have no problem in using illegal means of fighting a war, killing civilians, etc.

Another aspect of xenophobic approach is related to phenomena described in the section about the tool approach. The Hollywood guild of script writers, together with thousands of artists and illustrators around the world consider the results of the artificial intelligence systems to be a priori plagiarism, as they do not represent creativity as such, but just a mixture of already revealed meanings. It should be noted, however, that a significant part of human creative works is made along the same lines.

Thus, the xenophobia approach to assessing artificial intelligence cannot be called definitely ungrounded. There are good reasons to fear an uncontrollable development of artificial intelligence technologies and their further integration into various aspects of human life. This approach, implying either strict control over research in this sphere or, in a more radical form, their complete rejection, is apparently not free from shortcomings. These include: hindering the scientific-technical progress, impossibility to optimize people's practical activity by using the artificial intelligence systems, hap between scientific achievements and their integration into practice, hence, lowering the authority computer science in the public's eyes. An alternative to digital technologies, whose flagman is artificial intelligence, is usually said to be biotechnologies, thus, the rejection of artificial intelligence development, disillusionment in this technology may lead to the development of medicine, physiology, genetics, etc. Nevertheless, we believe that one should deny the advantages of this approach, which implies a more weighted estimation of the technology and its application; elaboration of the tools for forecasting and accessing the risks of further study and use of artificial intelligence, which can be with corrections used in other fields of knowledge; stimulation of development of other sciences related to the development of a human and human potential; providing a new impulse for comprehending a human and their place in the world.

3. The empathy model

According to this model, the society, positively perceiving household and other social intellectual robots and software assistants, favorably accepts the idea of the technology dissemination and does not exclude the possibility to endow the artificial intelligence systems with legal subject properties (in a limited sense).

This model is based on an advanced and broadened sense of humanism and human responsibility not only for themselves but to those around. Intellectual and autonomous systems, according to this model, cease being considered tools or competitors to a human but are viewed as companions, but in a limited sense, as pet companions. It is the ethical and legal norms regulating attitude to animals that form the basis of this model implementation. Actually, this model is transitional between the tool and the tolerance ones and cannot be viewed as something long-term.

Let us consider the examples confirming that this model is being partially implemented in the society.

The following experiment was described by K. Darling and S. Hauert. Six groups of eight people each were given toys shaped as dinosaurs, a size of a small cat. The participants were offered to interact with them. Then each group was ordered to "strangle", "break he head" or otherwise "kill" the toys, which was toughly opposed: the participants not just refused to "kill" their "dinosaurs" but also tried to defend them from other people and experienced serious discomfort seeing how a dinosaur "died"[6]. At that, only one of 48 toys was "killed".

Another example is switching off of the servers maintaining Jibo robot toys, which many users perceived as a death of their companion and reacted very emotionally.

People often reach in a similar way to the problems and death of literature, movie, or game personages, although they exist only virtually.

There is an opinion in psychology that people like communicating with chat bots (like ChatGPT) to discuss their psychological problem, as an intellectual system is rarely capable of reprimand which is characteristic to humans[7]. Assumingly, this phenomenon will be even more frequent in the future.

The described situations, although being particular cases and not reflecting the common public attitude to intellectual systems, demonstrate that in certain cases a person or groups of people may treat apparently inanimate (and, admittedly, even not intellectual) objects as pets. It is also notable that empathy directly depends on the appearance (exterior?) of a cyberphysical system and the vocabulary used. For example, the degree of empathy and trust towards a cyberphysical system of anthropomorphic phenotype may also depend on the "facial" features. M. B. Mathur and D. B. Reichling found that the reliability of a robot varies depending on its face similarity to a human's, does not increase linearly with a human image, but falls when an agent is very realistic but is not completely similar to a human (Mathur & Reichling, 2016). This phenomenon, initially described in 1978 by a Japanese

6 See: Darling, K., & Hauert, S. (2013, March 8). Giving rights to robots. *Robohub.org*. RobotsPodcast No. 125. https://robohub.org/robots-giving-rights-to-robots/

7 An expert: People use neural network as a psychologist due to a fear of reprimand on the part of a real person. (23 March, 2023). *"Moskva" Agency of city news*. https://www.mskagency.ru/materials/3286743

scientist M. Mori, is called “uncanny valley”: the most unusual anthropomorphous robots suddenly appeared to seem unpleasant due to elusive inconsistencies in appearance and behavior, which caused discomfort and fear (Mori, 2012). Thus, there is a probability that, with the robotics development, the empathy model may shift towards not the tolerance but xenophobia model. The psychological risks were reflected in the Code of ethical standards in robotics and AI, developed by the British Standards Institution: a user of an intellectual system must not feel uncomfortable; they must not experience anxiety or stress (Winfield, 2019).

Specification of this model also requires disclosing the phenomenon of mutual training by a human and a machine. Interacting with intellectual systems, a human transforms, but the changes touch upon not only the sphere of technological skills but also physiology and moral-ethical sphere. For example, a research by a group of Swiss scientists yielded an experimentally confirmed result that “the repeated movements along the smooth surface of a sensor screen change sensor reactions and, therefore, the brain’s ideas on the consequences of touching” (Balerna & Ghosh, 2018): when fingers touch a surface with the intensity similar to that when managing a smart phone, the brain of a modern human expects the “image” before their eyes to change.

Other features significant for considering this model include worsened memory and attention concentration due to the possibility to quickly find the necessary information via a smart phone or a voice assistant, and improved visual skills allowing a more rapid and better comprehensive perception of complex visual objects. Generally, one should not assume that integration of intellectual systems into the society results in degradation of the latter. The Flynn effect demonstrates that an average intellectual level of each new generation increases, i.e., the current average intellectual coefficient corresponds to a higher intellectual coefficient of the previous generation (Flynn, 2009). This said, modern research shows that with the spreading of digital technologies the Flynn effect decreased or even disappeared (Teasdale, 2005). However, this research cannot be considered reliable, as it was performed on the intelligence of army draftees, that is, the conclusions may be explained by social reasons, not the actual decrease of the intellectual level. Assumingly, intellect has not decreased but reshaped (Bukatov, 2018); for example, a student today remembers less than their predecessors in the 20th century but possesses a much larger range of techniques for searching and analyzing information. Accordingly, we observe a graduate substitution of substantive knowledge for skill for working with information. “One should take into account the biological co-adaptation and co-evolution of the human sense organs, the broadening of a range of our perceptions, which is ensured by technical advances” (Ogurtsov, 2006). At the same time, one should not exclude the factors of attention obtusion, reduction of perceived responsibility and professionalism of decision-makers “counseled” by an artificial intelligence system. Thus, there appears a situation of “shifting responsibility” for a mistake or illegal action onto an artificial intelligence system.

Besides physiological and intellectual aspect, the empathy model comprises probable changes in the emotional sphere. For example, M. Scheutz points out: "Social robots establish emotional contact with people and make the latter deeply trust them, which, in turn, may be used to manipulate people in previously impossible ways. For example, a company may use unique relations of a robot with its owner for the robot to persuade the owner to purchase the products which the company wants to promote. Consider the human relations, where, under normal circumstances, social emotional mechanisms like empathy and guilt will prevent escalation of such scenarios" (Scheutz, 2009).

As is known, many states stipulate criminal liability for cruelty towards animals, say nothing of taking its life without due grounds. For example, the European Convention for the Protection of Pet Animals points at inadmissibility of unnecessary pain and sufferings[8]. Continuing the comparison of the artificial intelligence systems with pets, will not a significant reprogramming of such a system incur pain on it, similar to how a cosmetic surgery intervention does, which is prohibited by the said Convention?

Accordingly, within the frameworks of the model under study this approach is transferred onto an artificial intelligence system and with strong reservations acts in a part of the society. However, its full-fledged implementation requires, at least, a conditional and socially accepted answer to the question whether an artificial intelligence system can feel pain and sufferings.

4. The tolerance model

The steady development of scientific-technical progress and perseverance of interest towards improving the said technologies may lead to the above-mentioned situation – emergence of a "strong" artificial intelligence or, at least, broad spreading of intellectual assistant systems. In the former case, technical restrictions will be leveled; artificial intelligence systems will obtain sufficient autonomy, the only framework limitation of which may be legal norms and the technical restrictions derived from them. Such conditional equity (or, at least, attributability) of the humanity and the artificial intelligence may lead to both positive and negative phenomena.

Within this model, an artificial intelligence serves as a "partner" of the humanity, provides the function of a restricting mechanism, obviates conflict escalation, and implements general and specific prevention of law breaches. The negative scenarios described above remain unfulfilled, as the wellbeing of both the humanity and the

[8] *European Convention for the Protection of Pet Animals* (1987, November 13). Council of Europe: official website. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168007a67d

artificial intelligence are interdependent, and both entities provide their own stability and development (for the humanity – first of all, social, for the artificial intelligence systems – technical) through cooperation, not competition. "Involvement into market relations requires mutual account of interests and rights… It is hard to recall any other, except mutual use, principle, with which equality and justice were established in human relationships so naturally and spontaneously" (Apresyan, 1995). Relationships between a society and artificial intelligence systems can hardly be called human in full sense, but it is rather expedient to expect mutual direct and reverse usefulness from them. For example, the National Standard "Requirements to safety for industrial robots" uses the following wording: "collaborative robot": a robot designed for immediate interaction with a human within a certain collaborative workspace; "collaborative workspace": a working space within a protected area where a robot and a human may perform works together during production"[9]. Apparently, a robot in these definitions is not a tool but a subject of interaction, collaboration, joint work, which looks precipitate so far, nevertheless.

Thus, success of artificial intelligence systems in scientific and creative activity may lead to growing incomes which will be directed to further development of an artificial intelligence system. Projects based on artificial intelligence become not just payable but super-profitable. However, one should remember that any economic growth is limited both in volumes and in time.

Development of the technology may result in a qualitative change of life: artificial intelligence (for example, as the author of a work of art or an invention) may act for itself at court, compete with representatives of various professions, which, in turn, forms motivation and stimulates the society and human development.

The tolerance model obviously remains the only one possible when creating a "strong" artificial intelligence, but it can also be implemented in the nearest future even in absence of the latter: with the growing number of situations of effective riskless practical use of artificial intelligence systems, trust in them at corporate and state levels will increase. For example, a Canadian insurance company Kanetrix.ca uses such systems for a client to choose the insurance product to purchase. Given that the artificial neural networks are used for that, it would be strange to except transparency and solvability, but these characteristics

[9] *GOST R 60.1.2.2-2016/ISO 10218-2:2011 Robots and robotic devices. Requirements to safety for industrial robots. Part 2: Robotic systems and their integration*. (2016). Moscow: Standartinform.

were not required: in this case the artificial intelligence system won convincingly compared to a human[10].

Such conditions may become true only in case the issue of liability limits of artificial intelligence systems is completely resolved and criteria for the presence of consciousness and will in decision-making are defined, without which artificial intelligence systems cannot be endowed with the features of a subject of law.

The drawbacks of this model lie in the sphere of both public and private law: the society cannot recognize artificial intelligence systems to be a subject equal to a human without answering the questions about the very essence and criteria of imposing liability for its actions upon such a system, i. e. referring it to an object or subject of law. For example, V. A. Laptev believes that the consequences of actions and decisions of an artificial intelligence may be considered as a force majeure circumstance, i. e. the one excluding the very question of liability, or a compulsory insurance against third party risks for the developer of an artificial intelligence system must by introduced (Laptev, 2017).

If artificial intelligence systems (or cyberphysical systems) achieve a certain level of cognitive abilities, i. e. if they obtain an apparent moral significance, such as intellect or sensitivity, then they probably will aspire to recognition of their moral status and must have rights, that is, a certain part of privileges, claims, authorities, or immunities (Gunkel, 2018). This is only possible under a significant qualitative technological progress. Apparently, the description of this model contains too many words “if” and “probably”. It reflects the degree of diffidence about the possibility for the artificial intelligence technology to develop up to such limits, but one cannot exclude such possibility. Stemming from the rate of technology development, experts forecast that, under the worst scenario, a full-fledged artificial intelligence comparable to a human will be designed by the end of the 21st century.

Conclusion

In a summarized form, the correlation of the described models is shown in Table.

It should be noted also that these models do not reflect the sequence of social relations’ development in the context of machine learning; they may be implemented simultaneously in different regions, economic sectors, law branches, etc.

[10] McWaters, R. J. et al. *Navigating Uncharted Waters. A roadmap to responsible innovation with AI in financial services.* World Economic Forum. http://www3.weforum.org/docs/WEF_Navigating_Uncharted_Waters_Report.pdf

Correlation of the society and law interaction models with the artificial intelligence technology

Criterion for comparison	The tool model	The xenophobia model	The empathy model	The tolerance model
Condition of implementation	Implemented today	If significant crises occur	Development of empathy attitudes in the society, progress in robotics and intellectual assistant systems	Achievement of technological singularity, emergence of a "strong" (general) artificial intelligence
Advantages	Low level of social and legal risks when using artificial intelligence systems while providing industrial and intellectual progress, possibility to preserve the current approaches to regulating technologies	Development of medicine, physiology, genetics	Development of public morals and humanism (in a broad sense)	Development of both technologies and law and other socio-humanitarian fields of knowledge
Disadvantages	Stagnation in socio-humanitarian fields of knowledge, rejection of the concept of a "strong" (general) artificial intelligence, anthropocentricism	Hindering of the scientific-technical progress in the sphere of digital and computer technologies	Decrease of rationality in favor of emotionality, transformations in the emotional sphere, mass problems with memory and attention	Reduced requirements to the artificial intelligence systems efficiency. Possibility to use an artificial intelligence system, possessing a legal personality, as a "proxy"; a need to review a large number of legal and other social norms
Character of "human – artificial intelligence" relationships	Exploitation	Competition	Empathy, co-adaptation	Mutual exploitation, cooperation
Legal consequences	A human (operator or developer) is liable for negative decisions and actions of an artificial intelligence system	Introduction of permissive regulation of the sector	Increased legal protection of intellectual systems without making them a legal subject	Making artificial intelligence systems a legal subject

Summarizing the above, it is worth highlighting that it is necessary to further research the artificial intelligence systems potential, including their properties during integration and spreading in the society, which is inevitable under these processes. These models reflect the facets of reality, both the existing and the potential one. Modeling of such situations should be done differentially, and the described ethical-legal models may contribute to that.

* The organization is recognized as extremist, its activity is prohibited in the territory of the Russian Federation.

References

Apresyan, R. G. (1995). Normative models of moral rationality. In *Morals and rationality* (pp. 94–118). Moscow: Institut filosofii RAN. (In Russ.).

Bakhteev, D. V. (2021). *Artificial intelligence: ethical-legal approach*. Moscow: Prospekt. (In Russ.).

Balerna, M., & Ghosh, A. (2018). The details of past actions on a smartphone touchscreen are reflected by intrinsic sensorimotor dynamics. *Digital Med*, *1*, Article 4. https://doi.org/10.1038/s41746-017-0011-3

Bukatov, V. M. (2018). Clip changes in the perception, understanding and thinking of modern schoolchildren – negative neoplasm of postindustrial way or long-awaited resuscitation of the psychic nature? *Actual Problems of Psychological Knowledge*, *4*(49), 5–19. (In Russ.).

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023, March 17). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*.

Flynn, J. R. (2009). *What Is Intelligence: Beyond the Flynn Effect*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511605253

Gunkel, D. J. (2018). *Robot rights*. Cambridge, MA: MIT Press.

Heidegger, M. (1993). The Question Concerning Technology. In *Time and being: articles and speeches*. Moscow: Respublika. (In Russ.).

Ilyin, E. P. (2016). *Emotions and feelings*. (2d ed.). Saint Petersburg: Piter. (In Russ.).

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns*, *3*(9). https://doi.org/10.1016/j.patter.2021.100314

Laptev, V. A. (2017). Responsibility of the "future": legal essence and evidence evaluation issue. *Civil Law*, *3*, 32–35. (In Russ.).

Marx, K. (2001). *Capital* (Vol. 1). Moscow: AST. (In Russ.).

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition, 146,* 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

Mori, M. (2012). The uncanny valley. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/mra.2012.2192811

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., & Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI Ethics. *Minds and Machines*, *31*, https://doi.org/10.1007/s11023-021-09563-w

Ogurtsov, A. P. (2006). Opportunities and difficulties in modeling intelligence. In D. I. Dubrovskii, & V. A. Lektorskii (Eds.), *Artificial intelligence: interdisciplinary approach* (pp. 32–48). Moscow: IIntELL. (In Russ.).

Scheutz, M. (2009). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. *Workshop on Roboethics at ICRA*.

Semis-ool, I. S. (2019). "Trustworthy" artificial intelligence. In D. V. Bakhteev (Ed.), *Technologies of the 21st century in jurisprudence: works of the All-Russia scientific-practical conference (Yekaterinburg, May 24, 2019)* (pp. 145–149). Yekaterinburg: Uralskiy gosudarstvenniy yuridicheskiy universitet. (In Russ.).

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn Effect in reverse. *Personality and Individual Differences*, *39*(4), 837–843. https://doi.org/10.1016/j.paid.2005.01.029

Timofeev, A. V. (1978). *Robots and artificial intelligence*. Moscow: Glavnaya redaktsiya fiziko-matematicheskoy literatury izdatelstva "Nauka". (In Russ.).

Vardi, M. (2012). Artificial Intelligence: Past and Future. *Communications of the ACM*, *55*, 5. https://doi.org/10.1145/2063176.2063177

Watkins, R., & Human, S. (2023). Needs-aware artificial intelligence: AI that 'serves [human] needs'. *AI Ethics*, *3*. https://doi.org/10.1007/s43681-022-00181-5

Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics*, *2*, 46–48. https://doi.org/10.1038/s41928-019-0213-6

Author information

Dmitriy V. Bakhteev – Doctor of Law, Associate Professor, Department of Criminology, Ural State Law University named after V. F. Yakovlev, Head of CrimLib.info project group
Address: 21 Komsomolskaya Str., 620137, Ekaterinburg, Russian Federation
E-mail: ae@crimlib.info
ORCID ID: https://orcid.org/0000-0002-0869-601X
ScopusAuthorID:https://www.scopus.com/authid/detail.uri?authorId=57208909117
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/ABA-1494-2020
Google Scholar ID: https://scholar.google.ru/citations?user=h0zOOdcAAAAJ
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=762765

Conflict of interests

The author declare no conflict of interests.

Financial disclosure

The research was not sponsored.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – March 25, 2023
Date of approval – April 24, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 340.143:004.8
EDN: https://elibrary.ru/ekeudk
DOI: https://doi.org/10.21202/jdtl.2023.22

Этико-правовые модели взаимоотношений общества с технологией искусственного интеллекта

Дмитрий Валерьевич Бахтеев

Уральский государственный юридический университет имени В. Ф. Яковлева
г. Екатеринбург, Российская Федерация

Ключевые слова

ChatGPT,
искусственный интеллект,
машинное обучение,
модель,
общество,
право,
регулирование,
робот,
цифровые технологии,
этика

Аннотация

Цель: исследование современного состояния технологии искусственного интеллекта в формировании прогностических этико-правовых моделей взаимоотношений общества с рассматриваемой сквозной цифровой технологией.

Методы: основным методом исследования является моделирование. Помимо него, в работе использованы сравнительный, абстрактно-логический и исторический методы научного познания.

Результаты: сформулированы четыре этико-правовые модели взаимоотношений общества с технологией искусственного интеллекта: инструментальная (на основе использования человеком системы искусственного интеллекта), ксенофобная (на основе конкуренции человека и системы искусственного интеллекта), эмпатическая (на основе сочувствия и соадаптации человека и систем искусственного интеллекта), толерантная (на основе взаимоиспользования и сотрудничества между человеком и системами искусственного интеллекта). Приведены исторические и технические предпосылки формирования таких моделей. Описаны сценарии реакций законодателя на ситуации использования этой технологии, такие как необходимость точечного регулирования, отказа от регулирования либо же полномасштабного вмешательства в технологическую отрасль экономики. Произведено сравнение моделей по критериям условий реализации, достоинства, недостатков, характера отношений «человек – система искусственного интеллекта», возможных правовых последствий и необходимости регулирования отрасли либо отказа от такового.

Научная новизна: в работе приведена оценка существующих в научной литературе, публицистике мнений и подходов, проанализированы технические решения и проблемы, возникшие в недавнем прошлом и настоящем. Теоретические выводы подтверждаются ссылками на прикладные ситуации, имеющие общественную или правовую значимость. В работе использован междисциплинарный подход, объединяющий правовую, этическую и техническую составляющие, которые, по мнению автора, являются критериальными для любых современных социогуманитарных исследований технологий искусственного интеллекта.

Практическая значимость: феномен искусственного интеллекта связывают с четвертой промышленной революцией, соответственно, эта цифровая технология должна быть изучена многоаспектно и междисциплинарно. Выработанные в научной статье подходы могут быть использованы при дальнейших технических разработках интеллектуальных систем, совершенствования отраслевого законодательства (например, гражданского и трудового), а также при формировании и модификации этических кодексов в сфере разработки, внедрения и использования систем искусственного интеллекта в различных ситуациях.

Для цитирования

Бахтеев, Д. В. (2023). Этико-правовые модели взаимоотношений общества с технологией искусственного интеллекта. *Journal of Digital Technologies and Law*, *1*(2), 520–539. https://doi.org/10.21202/jdtl.2023.22

Список литературы

Апресян, Р. Г. (1995). Нормативные модели моральной рациональности. В кн. *Мораль и рациональность* (с. 94–118). Москва: Институт философии РАН.

Бахтеев, Д. В. (2021). *Искусственный интеллект: этико-правовой подход*: монография. Москва: Проспект.

Букатов, В. М. (2018). Клиповые изменения в восприятии, понимании и мышлении современных школьников – досадное новообразование «постиндустриального уклада» или долгожданная реанимация психического естества? *Актуальные проблемы психологического знания*, *4*(49), 5–19.

Ильин, Е. П. (2016). *Эмоции и чувства*. (2-е изд.). Санкт-Петербург: Питер.

Лаптев, В. А. (2017). Ответственность «будущего»: правовое существо и вопрос оценки доказательств. *Гражданское право*, *3*, 32–35.

Маркс, К. (2001). *Капитал* (Т. 1). Москва: АСТ.

Огурцов, А. П. (2006). Возможности и трудности в моделировании интеллекта. В кн. Д. И. Дубровский, В. А. Лекторский (ред.) *Искусственный интеллект: междисциплинарный подход* (с. 32–48). Москва: ИИнтеЛЛ.

Семис-оол, И. С. (2019). «Заслуживающий доверия» искусственный интеллект. В сб. Д. В. Бахтеев, *Технологии XXI века в юриспруденции: мат-лы Всерос. науч.-практ. конф. (Екатеринбург, 24 мая 2019 года)* (с. 145–149). Екатеринбург: Уральский государственный юридический университет.

Тимофеев, А. В. (1978). *Роботы и искусственный интеллект*. Москва: Главная редакция физико-математической литературы издательства «Наука».

Хайдеггер, М. (1993). Вопрос о технике. В сб. *Время и бытие: статьи и выступления*. Москва: Республика.

Balerna, M., & Ghosh, A. (2018). The details of past actions on a smartphone touchscreen are reflected by intrinsic sensorimotor dynamics. *Digital Med*, *1*, Article 4. https://doi.org/10.1038/s41746-017-0011-3

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023, Marth 17). GPTs are GPTs: An *Early Look at the Labor Market Impact Potential of Large Language Models*.

Flynn, J. R. (2009). *What Is Intelligence: Beyond the Flynn Effect*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511605253

Gunkel, D. J. (2018). *Robot rights*. Cambridge, MA: MIT Press.
Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns*, *3*(9). https://doi.org/10.1016/j.patter.2021.100314
Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition, 146,* 22–32. https://doi.org/10.1016/j.cognition.2015.09.008
Mori, M. (2012). The uncanny valley. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/mra.2012.2192811
Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., & Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI Ethics. *Minds and Machines*, *31*, https://doi.org/10.1007/s11023-021-09563-w
Scheutz, M. (2009). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. *Workshop on Roboethics at ICRA*.
Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: *The Flynn Effect in reverse. Personality and Individual Differences*, *39*(4), 837–843. https://doi.org/10.1016/j.paid.2005.01.029
Vardi, M. (2012). Artificial Intelligence: Past and Future. *Communications of the ACM*, *55*, 5. https://doi.org/10.1145/2063176.2063177
Watkins, R., & Human, S. (2023). Needs-aware artificial intelligence: AI that 'serves [human] needs. *AI Ethics*, *3*. https://doi.org/10.1007/s43681-022-00181-5
Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics*, *2*, 46–48. https://doi.org/10.1038/s41928-019-0213-6

Сведения об авторе

Бахтеев Дмитрий Валерьевич – доктор юридических наук, доцент, доцент кафедры криминалистики, Уральский государственный юридический университет имени В. Ф. Яковлева, руководитель группы проектов CrimLib.info
Адрес: 620137, Российская Федерация, г. Екатеринбург, ул. Комсомольская, 21
E-mail: ae@crimlib.info
ORCID ID: https://orcid.org/0000-0002-0869-601X
ScopusAuthorID:https://www.scopus.com/authid/detail.uri?authorId=57208909117
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/ABA-1494-2020
Google Scholar ID: https://scholar.google.ru/citations?user=h0zOOdcAAAAJ
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=762765

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.07.49 / Планирование и прогнозирование в праве
Специальность ВАК: 5.1.1 / Теоретико-исторические правовые науки

История статьи

Дата поступления – 25 марта 2023 г.
Дата одобрения после рецензирования – 24 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.23

Artificial Intelligence as an Auxiliary Tool for Limiting Religious Freedom in China

Natalya I. Shumakova ✉

South Ural State University (National Research University)
Chelyabinsk, Russian Federation

Elena V. Titova

South Ural State University (National Research University)
Chelyabinsk, Russian Federation

Keywords

Artificial intelligence,
China,
digital technologies,
extremism,
freedom of worship,
human rights,
law,
neuron network,
regulation,
religion

Abstract

Objective: based on studying the statistics of crimes, national legislation and norms of international law, to give a legal assessment to restrictions of the right to worship implemented with the use of artificial intelligence technologies in China.

Methods: the methodological basis of the research is the set of methods of scientific cognition, including specific sociological (analysis of statistical data and other documents), formal-legal (examining legal categories and definitions), formal-logical (analysis and synthesis), general scientific (induction, deduction), and other methods.

Results: the work researches prerequisites for using artificial intelligence technologies in China to control public relations arising during religious activity both in the digital space and beyond; analyzes the legal framework of the measures implemented; gives a legal assessment to restrictions of the religious freedom using artificial intelligence technologies; forecasts the further development of Chinese legislation and foreign policy associated with religious freedom. Additionally, the work analyzes materials of human rights organizations aimed at hindering the Chinese policy of "sinicisation" and "de-extremification" of ethnic and religious minorities, including with the help of control and propaganda using modern digital technologies.

✉ Corresponding author

Scientific novelty: the work researches the attempt of China to regulate the challenges related to religious activity, arising during rapid digitalization of the society and state, which the Republic faces being a developing, multinational and polyconfessional country. The established restrictions of religious freedom using artificial intelligence technologies are considered along with the relevant criminal statistics. The legal assessment of using artificial intelligence as a tool for restricting the right to worship is given from the standpoint of international law, as well as with the account of Chinese national legislation.

Practical significance: the research results can be used to elaborate a consistent legal framework for using artificial intelligence technologies to counteract extremism.

For citation

Shumakova, N. I., Titova, E. V. (2023). Artificial Intelligence as an Auxiliary Tool for Limiting Religious Freedom in China. *Journal of Digital Technologies and Law, 1*(2), 540–563. https://doi.org/10.21202/jdtl.2023.23

Contents

Introduction

The process of digitalization of social relations transfers into a new plane the implementation of fundamental human rights, in particular, the right to religious freedom. K. A. Bingaman, considering this process to be irreversible, highlights that it blur the line between the real and the virtual in the aspects of administrating religious services and rituals (Bingaman, 2023). A similar concept is maintained by E. Marique and Y. Marique, who describe digital space as “contemporary public area” uniting virtual actions with real consequences, hence, requiring special regulation (Marique & Marique, 2020). C. Ashraf also speaks of the possibility to single out digital religion as an independent institution (Ashraf, 2020). This theory was confirmed by his colleagues from Catalonia, who called the platforms of digital religious

activity "new public spaces facilitating the creation of a social identity of young people" (Bosch et al., 2017).

Emergence and development of online organizations rendering religious services, as well as social relations occurring during online communications on religious topics, create additional challenges, often leading to limiting freedom of worship. In particular, speaking of the phenomenon of online extremism, K. V. Bhatia, turning to the results by L. Dawson and D. Cowan (Bhatia, 2021), marks that an unlimited access to creating, distributing and consuming online information leads to a crisis of power in the digital environment, as well as to a crisis of authenticity, associated with the content of the disseminated materials.

OSCE also points out the growth of online dissemination of extremist materials[1], while the UNO General Assembly[2] calls for paying attention to the statistics of incidents related to suing digital platforms for inciting hatred and for supporting and financing extremist groups.

The necessity of protection against threats, associated with the growing digitalization of all spheres of the life of society and state, is being discussed by the Chinese scholars, too. For example, X. Wei in the research work "A critical evaluation of China's legal responses to cyberterrorism" calls cyberterrorism a special type of a threat to stability and security of states and characterizes the current Chinese legislation, as well as measures for its implementation, as insufficient for ensuring elimination of this threat (Wei, 2022). Such a categorical view on the activity in digital space is implicitly confirmed by the research of the disseminated content, which leads to "mass polarization" of the society's attitude towards terrorism and extremism (Guan & Liu, 2019).

A. Sabic-El-Rayess also puts social networks among main sources of disseminating the information which facilitates radicalization of the population and, as a consequence, acts of violence (Sabic-El-Rayess, 2012). Referring to notable psychological changes, associated with the process of communication in the digital environment and leading to radicalization, a number of researchers, as a measure of prevention and counteraction, suggest using machine learning not only to detect certain types of messages, but also to research longitudinal open data of social networks in order to identify individual changes in publications and certain types of messages (Smith et al., 2020).

L. F S. Meneses, also mentioning negative psychological changes, calls the uncontrolled dissemination of unverified information in the digital environment "a crisis of truth" and focuses on the destructive influence of this phenomenon on the individual's ability for critical

---

1 OSCE PA vs. COVID-19. Stage 1. Reflections, policy contributions and recommendations presented by OSCE PA President George Tsereteli. (2020). *OSCE PA Reports*. https://www.oscepa.org/en/documents/president/reports-22/

2 UNO General Assembly. (2020, June 18). Racial discrimination and emerging digital technologies: a human rights analysis : report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance. *Report of the Special Procedure of the Human Rights Council*. https://digitallibrary.un.org/record/3879751?ln=ru

thinking (Meneses, 2021). Researches of a focus group in Hong Kong during 2019 protests also prove that demonstration of certain images and videos can significantly influence critical thinking (Chan, 2019).

Stemming from the analysis of past research, it becomes apparent that the impossibility of using traditional regulation mechanisms in the Internet is the immediate reason for using artificial intelligence (further – AI) as a tool of state control in the digital space. The possibility of using its algorithms for studying religious activity and religion per se was pointed out by R. Reed, who marked that collecting, sorting and storing information would inevitably produce the impression of total surveillance and suggested thinking about the role of religion in the society accepting certain ideas (Reed, 2021).

At the same time, modern digital technologies may be viewed as auxiliary tools for struggling against crimes aimed at inciting hatred on ethnic and religious grounds, violating territorial integrity and overthrowing state power and other deeds beyond the digital space.

The need for the law enforcement bodies to use AI was pointed out by several researchers (Kao & Sapp, 2022; Fontes et al., 2022), mentioning the side effects of AI abuse, the main being destabilization of the society.

The objective of this work is to provide a legal evaluation of using artificial intelligence and other digital technologies to control religious activity in the People's Republic of China (further – PRC). To achieve this objective, during the research we identified the feature of implementing the right to freedom of worship in PRC, analyzed the statistics of crimes associated with extremism and religious intolerance, studied the materials of a hearing conducted by the US Commission on International Religious Freedom (USCIRF). Finally, an attempt is made to forecast further development of Chinese policy and legislation in relation to religious freedom.

1. Guarantees and limitations of religious freedom

Article 36 of the PRC Constitution stipulates that "citizens of the People's Republic of China shall enjoy freedom of religious belief"[3] and prohibits coercion of citizens "to believe in or not to believe" or discrimination based on religion. At the same time, Article 36 contains an important provision: the state is only obliged to "protect normal religious activities". However, there is no clear definition of the term "normal religious activities" in the Chinese Constitution, it is just outlined. Stemming from the text, it is understood as the religious activity which does not disrupt public order, impair the health of citizens, contain signs of social activity or interfere with the state's education system. The Article

[3] Constitution of the People's Republic of China of December 4, 1982 (with amendments and additions in the edition of March 11, 2018). *The National People's Congress of the People's Republic of China*. https://www.npc.gov.cn/englishnpc/constitution2019/201911/1f65146fb6104dd3a2793875d19b5b29.shtml

under consideration also indicates that religious groups and relations associated with their activity shall not be subject to control by foreign forces.

Generally, China has a wide range of laws and other legal acts aimed at ensuring, protecting and regulating public relations associated with implementation of freedom of religion, including the Criminal Code of PRC (1997)[4], Hong Kong Basic Law (1997)[5], Macao Basic Law (1999)[6], Labor Law (1985)[7], Law on compulsory education (1986)[8], Law on regional national autonomies of PRC (1984)[9], Law on meetings and demonstrations in PRC (1989)[10], Law on advertising[11], etc.

Almost any normative legal act in PRC guarantees unacceptability of religious discrimination and inalienability of religious freedom. At the same time, having studied Chinese legislation, one may conclude that it is aimed at unacceptability of religion penetrating social institutions; moreover, it becomes obvious that China strives to turn religion into a tool of state power. As early as in 2005, after the first edition of Religious Affairs Regulations came into force[12], this was pointed out by B. Leung (Leung, 2005).

Later research just confirm the attempts of the Chinese government to "sinicise", that is, to adapt religion to the system of socialist values (Lavička, 2021), with the ultimate goal of detail regulation of the religious activity, suppression of unsanctioned religious groups and reduction of religious influence (Lin, 2018). To achieve these goals, all religious relations, in any way touching upon the interests of the state and society, are subject to state regulation.

---

4 中华人民共和国主席令 of March 14, 1997. 新华网 (year of publication not indicated). https://www.npc.gov.cn/zgrdw/npc/lfzt/rlys/2008-08/21/content_1882895.htm

5 Basic Law of April 4, 1990. *The Government of the Hong Kong Special Administrative Region, Hong Kong Basic Law Drafting Committee* (year of publication not indicated). https://www.basiclaw.gov.hk/en/basiclaw/chapter1.html

6 The Basic Law of the Macao SAR of March 31, 1993 (2014). *Macao government web*. https://www.zlb.gov.cn/2014-06/26/c_126677086_4.htm

7 中华人民共和国劳动法 of July 5, 1994. 中国政府门户网站，全国人大法规库 (year of publication not indicated). https://www.gov.cn/banshi/2005-05/25/content_905.htm

8 中华人民共和国义务教育法 中华人民共和国主席令 38 of April 4, 1986 (2006). *China Educational and Research Network*. https://www.edu.cn/edu/zheng_ce_gs_gui/jiao_yu_fa_lv/200603/t20060303_165119.shtml

9 中华人民共和国民族区域自治法 of May 5, 1984, No. GJXFJ-0000-2014-00084 (with amendments and additions in the edition of February 28, 2001) (2005). 中国政府门户网站. https://www.gov.cn/ziliao/flfg/2005-09/12/content_31168.htm

10 中华人民共和国集会游行示威法 of October 31, 1989, No. GJXFJ-0000-2014-00084 (2014, May 12). 国家信访局门户网站. https://www.gjxfj.gov.cn/gjxfj/fgwj/flfg/webinfo/2014/05/1601761496620028.htm

11 中华人民共和国广告法 of October 27, 1994 (with amendments and additions in the edition of April 24, 2015) (2015). 中国人大网【字体：大 中 小】打印. https://www.gov.cn/guoqing/2021-10/29/content_5647620.htm

12 Religious Affairs Regulations of July 7, 2004, No. 426. *State Council of the People's Republic of China*. https://www.refworld.org/pdfid/474150382.pdf

As a guarantee to ensure freedom of worship, these norms are stipulated in Article 6 of Religious Affairs Regulations[13] (2017). Besides, on March 1, 2022, "Measures to administrate rendering religious services in the Internet" came into force in China[14] – a headline-making subordinate legislation, adopted by the National Religious Affairs Administration in collaboration with Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security and Ministry of State Security of PRC. It became a logical continuation of the Law on cybersecurity of Chinese citizens[15] (2017), Religious Affairs Regulations[16] (2017), as well as "Measures to administrate rendering information services in the Internet"[17] (2000). Apparently, its adoption was a reaction towards emerging services of virtual performance of religious rituals, an outburst of which occurred during the coronavirus pandemic. For example, there appeared applications for virtual censing, chats for reading mantras and prayers and other means of manifesting religious activity in the digital environment.

The official press release to "Measures to administrate rendering religious services in the Internet"[18], published in the Zhejiang Nationalities and Religious Affairs Committee website, speaks of the need to standardize online service with religious topics. In the same press release, the National Religious Affairs Administration emphasizes that the digital space of China cannot be regarded as "a special zone for religious activity" or an "enclave of religious or public opinion".

The enacted legislation[19] sets the standards of providing religious services and distribution of any religious information. For example, in compliance with Article 6, only organizations registered in the territory of China may perform this activity, while any participation of foreigners in it is prohibited. Provisions of this Article are expanded to any information associated with religion, including visual information distributed through

13 宗教事务条例中华人民共和国国务院令686 of August 26, 2017, No. 000014349/2017-00167. (2017, September 7). 中华人民共和国国务院令. https://www.gov.cn/zhengce/content/2017-09/07/content_5223282.htm

14 互联网宗教信息服务管理办法 of December 3, 2021, No. 17. (2021). 国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部、中华人民共和国国家安全部令. https://www.gov.cn/gongbao/content/2022/content_5678093.htm

15 中华人民共和国网络安全法 of November 7, 2016 (2016). 新华社北京. https://www.gov.cn/xinwen/2016-11/07/content_5129723.htm

16 宗教事务条例 中华人民共和国国务院令 686 of August 26, 2017, No. 000014349/2017-00167. (2017, September 7). 中华人民共和国国务院令. https://www.gov.cn/zhengce/content/2017-09/07/content_5223282.htm

17 互联网信息服务管理办法 of September 25, 2000, No. 292 (with amendments and additions in the edition of January 8, 2011). 日中华人民共和国国务院令第. https://www.gov.cn/zhengce/2020-12/26/content_5574367.htm

18 国家宗教事务局相关负责人就《互联网宗教信息服务管理办法》答记者问 of February 28, 2022, No. 002482103/2022-00007. (2022, February 28). 浙江省民宗委. https://mzw.zj.gov.cn/art/2022/2/28/art_1229468422_2394644.html

19 互联网宗教信息服务管理办法 of December 3, 2021, No. 17. (2021). 国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部、中华人民共和国国家安全部令. https://www.gov.cn/gongbao/content/2022/content_5678093.htm

websites, applications, forums, blogs, microblogs, messengers, online broadcasts, etc. Besides, one should remember that only five religions are officially recognized in China[20]: Buddhism, Daoism, Islam, Protestantism, and Catholicism. That means that representatives of other confessions, previously engaged in missionary activity and rendering religious services in the territory of China online, are now deprived of such opportunity.

International reaction to developing and applying such measures was not long in coming. On September 13, 2022, the US Commission on International Religious Freedom (USCIRF) conducted hearings[21], during which it accused China of hindering the implementation of the right to freedom of worship at three levels: public, private and digital.

As followed from the hearing testimonies, the Chinese government uses modern digital technologies:

– to trace the location of believers;
– to trace any user activity both in the digital space and in real life;
– to recognize speech;
– to recognize faces;
– to collect information about religious groups in the Internet.

As follows from the materials of the hearing, introduction of "Measures to administrate rendering religious services in the Internet"[22] may aggravate religious persecution in the country, according to the party of charge. PRC was also accused of selecting the course towards strengthening people's unity and "sinicisation" of religions which is actually aimed at elimination of traditional culture and values. As for using modern technologies, including AI technologies as an auxiliary tool for implementing the said legislation, it was characterized as manifestation of "digital authoritarianism". Besides, the party of charge demanded applying sanctions and measures to China, aimed at infringement of its sovereignty, including the USA interference into the country's internal policy.

Earlier, similar accusations were set forth to China by the UN Human Rights Office[23] in the Assessment of human rights concerns in the Xinjiang Uyghur Autonomous Region.

---

20 中国宗教概况. 宗教局, 中国政府门户网站 (year of publication not indicated). https://www.gov.cn/test/2005-06/22/content_8406.htm

21 Hearing Before the Congressional-Executive Commission on China. September 13, 2022. (2023). *Control of Religion in China through Digital Authoritarianism. Washington*: U.S. Government publishing office. https://www.govinfo.gov/content/pkg/CHRG-117jhrg48647/pdf/CHRG-117jhrg48647.pdf

22 互联网宗教信息服务管理办法 of December 3, 2021, No. 17. (2021). 国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部、中华人民共和国国家安全部令. https://www.gov.cn/gongbao/content/2022/content_5678093.htm

23 OHCHR Assessment of human rights concerns in the Xinjiang Uyghur Autonomous Region, People's Republic of China of August 31, 2022. (2022). *UN Human Rights Office*. https://www.ohchr.org/sites/default/files/documents/countries/2022-08-31/22-08-31-final-assesment.pdf

2. Legal framework for establishing control over religious activity

The common basis for using AI technologies for ensuring digital safety and struggling against cybercrime, including with distribution of extremist religious materials, are the following Articles of the Law on cybersecurity of PRC citizens[24]:

– Article 21, according to which communications service providers are obliged to classify and store for half a year registers of users' network activity;

– Article 46, prohibiting creation of websites and communication groups to use them for illegal purposes or for distributing illegal information;

– Article 47, obliging communications service providers to strengthen management of information published by the users and to immediately stop data transfer in case illegal activity is detected, then to take measures to process and eliminate the data which violate the country's legislation. The preserved proofs must be handed over t the authorities;

– Article 51, stipulating the establishment of a state system of digital security monitoring. As follows from the Article, state departments of cybersecurity and informatization must execute general coordination of other relevant department with a view of improving the efficiency of collecting, analyzing and presenting the information on cybersecurity for early prevention of risks of digital security violation;

– Article 58, which allows applying temporary measures to network communications in certain regions, including measures restricting them to protect national security and public order, as well as to react to large accidents undermining public security;

– Article 68, stipulating liability of communications service providers for non-application of response measures in case of detecting distribution of illegal information.

China intends to apply these norms also to regulating generative AI. For example, Article 12 of the draft "Measures for managing the generative AI services"[25] equals suppliers of such services to content producers and obliges them to make sure that the generated texts, photo- and audiovisual materials are not in any way discriminating. In the same draft, Article 19 stipulates liability of the users of generative AI services for creating illegal materials.

3. Using artificial intelligence to implement control over religious activity

The way the provisions of the "Measures to administrate rendering religious services in the Internet"[26] are planned to be implemented is of no less interest. The above-mentioned

---

24 中华人民共和国网络安全法 of November 7, 2016. (2016). 新华社北京. https://www.gov.cn/xinwen/2016-11/07/content_5129723.htm

25 生成式人工智能服务管理办法（征求意见稿） of April 11, 2023 (2023). 国家互联网信息办公室. https://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm

26 互联网宗教信息服务管理办法 of December 3, 2021, No. 17. (2021). 国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部、中华人民共和国国家安全部令. https://www.gov.cn/gongbao/content/2022/content_5678093.htm

press release states that this will be done cooperatively by the National Religious Affairs Administration, Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security and Ministry of State Security, hence, the companies affiliated with them, such as:

– iFLYTEK Co., Ltd.[27] – a Chine state company applying AI to study and recognize speech, including rare languages and dialects (including Tibetan and Uighur);

– CloudWalk Technology Co., Ltd.[28] – as follows from its official website, this company, engaged in developing technologies for recognizing a human face, body and voice, also performs profiling and studying behavior patterns;

– China Electronics Technology Group Corporation[29] – a state company engaged, inter alia, in developing software and innovations in the sphere of providing security and implementation of AI;

– Qihoo 360[30] – a provider of free-of-charge internet- and mobile security, owning 360 Total Security, 360 Mobile Security, 360 Security and other security products providing national security using AI technologies. The company is also engaged in training specialists in this sphere;

– Huawei Technologies[31] – one of the world largest producers of telecommunication equipment. As follows from publicly available advertisement materials placed in its official website, their devices may help identify individuals by voice, manage ideological re-education of convicts, trace political activists, use facial recognition technologies to trace consumers, etc. The company also develops software to scan faces in order to transfer information to law enforcement.

Besides, one should not forget that AI technologies for religious control are also used outside the digital space. In particular, according to the 2019 data of the US Commission on International Religious Freedom[32], by 2035 Hangzhou Hikvision Digital Technology Co., Ltd. plans to install cameras in 967 mosques, which will allow them not only to trace all visitors through facial recognition system, but also control that the sermons adhere to the letter

27 *About iFLYTEK Create A Better World With A.I.* https://global.iflytek.com

28 *CloudWalk Technology Co., Ltd*. https://www.cloudwalk.com

29 *China Electronics Technology Group Corporation*. https://www.cetcei.com/enzgdzkj/about_us/introduction29/index.html

30 Qihoo 360. https://www.360.cn/

31 *Nowhere to hide: Building safe cities with technology enablers and AI*. https://www.huawei.com/en/huaweitech/publication/winwin/AI/nowhere-to-hide

32 Religious Freedom in China's High-Tech Surveillance State. (2019, September). *USCIRF Country Update: China*. https://www.uscirf.gov/publications

of the law. According to Article 9 of "Provision of Xinjiang Uygur Autonomous Region on de-extremification"[33] (2017), AI will recognize the following signs of extremism:

– wearing a burqa, closing the face, or trying to make others wear it, as well as wearing other extremist symbols (clause 7);

– distributing religious fanatic ideas through wearing a large beard and/or using certain names (clause 8);

– conclusion or dissolution of marriage according to religious traditions without state registration.

Clause 5 of the law under consideration deserves special attention, as it states as a manifestation of extremism refusal to use television or radio, which implies the presence of automated monitoring of such activity.

4. Statistics of crime as the basis for establishing control over religious activity

Control over religious activity in problem regions, especially in Xinjiang Uygur Autonomous Region (further – XUAR), as follows from the White Paper of the Chinese government "The Fight Against Terrorism and Extremism and Human Rights Protection in Xinjiang"[34] (2019), is a part of the state program for struggling against terrorism and extremism, aimed at protection of representatives of other confessions and ethnic groups and protection of human rights in general.

The same document presents the following statistics of crimes associated with separatism and religious extremism, marking that it is not complete:

a) terrorist attacks and threat of public safety:

– 05.02.1992 – exploding two passenger buses in Urumqi (3 killed, 23 wounded);

– 25.02.1997 – exploding three passenger buses in Urumqi (9 killed, 68 wounded);

– 23.05.1998 – terrorists placed over 40 self-priming devices in crowded locations of Urumqi, which led to 15 arsons;

– 07.03.2008 – an attempt of explosion on board of CZ6901 flight from Urumqi to Beijing;

– 30.06.2011 – in Kashgar two terrorists took possession of a truck killing the driver, then purposefully slammed into a crowd. Then they attacked pedestrians with cold weapons (8 killed, 27 wounded);

– 31.06.2011 – terrorists with cold weapons attacked pedestrians in Kashgar (6 killed, 16 wounded);

– 29.07.2012 – an attempt to capture an aircraft at GS7554 flight from Hotan to Urumqi;

---

33 新疆维吾尔自治区去极端化条例 of March 27, 2017, No. mzj/2020-00007 (with amendments and additions in the edition of May 9, 2019) (2019). 新疆人大网 更新时间. https://www.xjpcsc.gov.cn/article/225/lfgz.html

34 新疆維吾爾自治區打擊恐怖主義和極端主義和保護人權 of March 2019 (2021). 中华人民共和国国务院新闻办公室. https://geneva.china-mission.gov.cn/chn/ztjs/ajljdiowqjknew/baipishu/202110/t20211014_9587970.htm

– 28.10.2013 – terrorist attack in the centre of Beijing: terrorists drove a jeep with 31 barrels of petrol towards tourists and duty policemen, then burnt it (2 killed, 40 wounded);

– 01.03.2014 – terrorists with cold weapons attacked people at Kunming railway station (31 killed, 141 wounded);

– 30.04.2014 – a terrorist attacked passengers at Urumqi railway station, another simultaneously enacted an explosive (3 killed, 79 wounded);

– 22.05.2014 – five terrorists on cars slammed into a crowd in Urumqi, then enacted an explosive (39 killed, 94 wounded);

– 18.09.2015 – terrorists attacked a coal mine in Aksu region (16 killed, 18 wounded);

b) killings of religious leaders:

– 24.08.1993 – two terrorists attempted to kill imam Mullah Abulizi in Kashgar;

– 22.03.1996 – two people in masks shot Vice President of the Islamic Association of Xinhe County and assistant imam Akemusidike Aji;

– 12.05.1996 – terrorists attempted to kill Aronghan Aji, vice president of the China Islamic Association and president of Xinjiang Islamic Association, and hatip of Id Kah Mosque in Kashgar;

– 06.10.1997 – terrorists shot Senior Mullah Younusi Sidike, member of the China Islamic Association, president of Aksu Islamic Association and imam of the Great Mosque of Baicheng County;

– 27.01.1998 – terrorists shot Abulizi Aji, imam of the Great Mosque of Baicheng County;

– 30.10.2014 – Senior Mullah Juma Tayier, vice president of Xinjiang Islamic Association and imam of the Id Kah Mosque, was brutally killed by three terrorists on his way home after morning Fajr prayer;

c) attacks on state bodies and public riots:

– 05.04.1990 – a group of terrorists of over 200 people attacked a government building of Baren Township, Akto County, Kizilsu Kirgiz Autonomous Prefecture, kidnapping 10 people and killing 6 armed police officers;

– 27.08.1996 – a government residence was attacked in Jianggelesi Township of Yecheng County in Kashgar prefecture, a deputy township head, a policeman on duty and three security men were killed;

– 05–08.02.1997 – during riots in Yining, 7 people died and 198 were wounded;

– 24.10.1999 – armed terrorists attacked a police station in Saili Township, Zepu County, in Kashgar prefecture. A public security guard and a criminal suspect in custody were shot dead; a policeman and a public security guard were injured;

– 05.07.2009 – a riot in Urumqi: thousands of terrorists attacked civilians, government organs, public security and police officers, residential houses, stores and public transportation facilities; 197 people died and over 1700 were wounded;

– 04.08.2008 – terrorists drove a stolen dump truck into armed frontier police at drill on Seman Road, Kashgar City, and threw homemade grenades; 16 people died and 16 were wounded;

– 23.04.2013 – community workers found terrorists making explosives at a private home in Selibuya Town, Bachu County, Kashgar Prefecture, and were killed on the spot. Then terrorists attacked local government staff and police, as a result of which 15 people died and 2 were badly wounded;

– 26.06.2013 – terrorists launched attacks at the police station, patrol squadron, seat of local government and construction sites of Lukeqin Township – 24 people died and 25 were wounded;

– 28.07.2014 – terrorists with knives and axes attacked the government building and police station of Ailixihu Town, Shache County, Kashgar Prefecture. Some then moved on to Huangdi Town where they attacked civilians. 37 people died and 13 were wounded;

– 21.09.2014 – the police station and farmer's market of Yangxia Town, the police station of Tierekebazha Town, and a store at the Luntai county seat, Bayingol Mongolian Autonomous Prefecture were attacked. Ten people died and 54 were wounded;

– 28.12.2016 – terrorists drove into the courtyard of Moyu County government, Hotan Prefecture, detonated a homemade explosive device, and attacked government staff; 2 people died and 3 were wounded.

The statistics cited in the said White Paper of the Government[35] cannot be confirmed or complemented from other official sources, as the access to information is either archived or restricted for foreigners. In general, there are doubts concerning the accuracy of the statistical data in open access. For example, according to the data of Bureau of Counterterrorism[36], in China there were no terrorist attacks or extremist acts associated with violence since 2016. This contradicts to earlier information about an explosion of a handmade device and death of five people near Pishan County of XUAR in 2017.[37] A report of People's prosecutor's office of XUAR as of 2017[38] does not highlight this episode in the overall statistics of crimes.

According to a report of Human Rights Watch[39], referring to a presumably official statistics from 2017 to 2022 provided by PRC, by February 2022 the overall number of condemnatory judgments in extremist cases was 540,826 in this region. The same report once again emphasizes that China regularly blanks out the complete data; inter alia, access to judicial sentences is completely forbidden, thus, it is impossible to learn for which crimes associated with extremism these sentences were passed.

---

35 新疆維吾爾自治區打擊恐怖主義和極端主義和保護人權 of March 2019 (2021). 中华人民共和国国务院新闻办公室. https://geneva.china-mission.gov.cn/chn/ztjs/ajljdiowqjknew/baipishu/202110/t20211014_9587970.htm

36 Bureau of Counterterrorism. (2020). *Country Reports on Terrorism 2020: China (Hong Kong and Macau)*. https://www.state.gov/reports/country-reports-on-terrorism-2020/china

37 Bureau of Counterterrorism. (2020). *Country Reports on Terrorism 2017: Country Reports on Terrorism*. https://www.state.gov/reports/country-reports-on-terrorism-2017/

38 自治区人民检察院工作报告 of January 23, 2018. (2018). 新疆维吾尔自治区人民检察院. https://archive.ph/5MLLE

39 China: Xinjiang Official Figures Reveal Higher Prisoner Count of September 14, 2022. (2022). *Human Rights Watch*. https://www.hrw.org/news/2022/09/14/china-xinjiang-official-figures-reveal-higher-prisoner-count

A rather impressive statistics was given by X. Wei in the above-mentioned work (Wei, 2022), according to which terrorism in China in the 21st century shifted to the digital space and its main manifestations in the country are such auxiliary acts, leading to real terrorist attacks and violations of public order, as using the Internet for disseminating illegal information, incitement to crimes and extremist-related propaganda (64 %), and possessing audiovisual materials and other items which can be used for committing terrorist attacks (33.9 %).

5. Using artificial intelligence and digital technologies to control religious activity in problem regions

Speaking of XUAR, one should mind that the basis for establishing digital control over implementation of the right to religious freedom is the local legal act "Provision of Xinjiang Uygur Autonomous Region on de-extremification"[40] (2017). In particular, clause 13 of Article 9 of this act prohibits "publishing, dissemination, downloading, storing of audio- and video materials of extremist content", as well as access to such materials, while Article 26 endows telecommunication departments and operators with the right to monitor systems and apply technical means to exchanging voice messengers, talking on mobile and telephone devices and using other telecommunication tools. In case of revealing information of extremist content, an operator is obliged to interrupt transmission, delete all relevant materials, register proofs, and report about the incident. At the same time, Article 3 provides a definition of religious extremism – it is "dissemination of radical religious ideology through statements and actions, denial of normal production and living and impeding them" and "suggestions and actions using distortion of religious doctrines or other means to ignite hatred or discrimination and propagate violence". "Provision of Xinjiang Uygur Autonomous Region on de-extremification"[41] is a part of the Strike Hard Campaign against Violent Terrorism, developed to fight extremism and terrorism in XUAR[42].

Assumingly, AI technologies will also be used for digital control in Tibet. Although public riots in this region are traditionally not considered terrorist activity, given its historical interrelations with China, the latter apprehends a new growth of separatist sentiments in it. The main premises for such fears are the protests which were organized by Buddhist monks in March 2008 in Lhasa and turned into mass unrest, during which,

40 新疆维吾尔自治区去极端化条例 of March 27, 2017, No. mzj/2020-00007 (with amendments and additions in the edition of May 9, 2019) (2019). 新疆人大网 更新时间. https://www.xjpcsc.gov.cn/article/225/lfgz.html

41 *Ibid.*

42 公安部开展严厉打击暴力恐怖活动专项行动 (2014, May 25). 中央政府门户网站//新华社. https://www.gov.cn/xinwen/2014-05/25/content_2686705.htm

inter alia, arson of the main city mosque was committed (Van Wie Davis, 2009)[43]. Today, it is rather difficult to get access to official statistics, but Van Wie Davis states in her research of this issue, referring to the data provided by the local government, that 16 people were killed during protests, four of which were hotel employees and shopkeepers, intentionally burnt alive.

Also in March 2008, protesters with homemade explosion devices attacked government offices, police stations, hospitals, schools, banks and markets under the slogan "Freedom to Tibet!" in the south-west part of Sichuan province. Van Wie Davis, referring to the police data, states that attacks were organized by Buddhist monks and aimed, inter alia, at Muslim residents, forcing police to cordon off the Muslim quarter of Lhasa on March 15, 2008[44].

The Chinese government has taken a number of preventive measures to de-radicalize Tibet, among which of utmost interest for our research is the enacted "Provision on creating an exemplary zone of ethnic unity and progress in the Tibet Autonomous Region"[45] (2020). The Provision is primarily aimed at the above-mentioned "sinicisation" of religion and integration of the idea of ethnic unity into it, as follows from Article 19 of this legal act. The idea of ethnic unity and progress must be also propagated by online agencies, as stipulated by Article 27. Sanctions are mentioned in Article 46 – punishment is stipulated for undermining the idea of national unity and progress by disseminating rumors, producing information, or by taking a categorical opposing stand. Article 24 is worth special attention, as it demands that Tibetans must promote the above ideas in families, in order to bring up exemplary citizens, as follows from Article 34.

Conclusions

The research results showed that China strives to create a legal framework for state regulation of the public relations arising during implementation of the right to worship, in order to preserve people's unity, public order and the country's territorial integrity. At the same time, the Chinese legislator strives to timely answer such challenges of today as dissemination of extremist materials and radicalization of the population in the digital space, in particular, by creating and distributing illegal content using generative AI. As for implementation of laws and other legal acts aimed at controlling and restricting religious freedom, including the so called sinicisation and de-extremification

43 Official data being inaccessible, statistics is given by Van Wie Davis, E. Tibetan separatism in China. (2009). *Korean Journal of Defense Analysis*, *21*(2), 155–170. https://doi.org/10.1080/10163270902872135

44 *Ibid.*

45 西藏自治区民族团结进步模范区创建条例 第四章 宣传教育 of January 11, 2020, No. mzj/2020-00007. (2020). 地区民宗局. https://www.al.gov.cn/info/1258/28851.htm

of the population of certain regions, they are ensured, inter alia, by using artificial intelligence technologies. The listed measures are conditioned by actual threats to the territorial integrity and stability of PRC and security of its citizens, which follows from the analysis of statistics of crimes committed on ethnic and religious grounds, associated with mass disturbances and terrorist attacks with the ultimate goal of certain regions' secession from the state.

According to Article 36 of the Constitution of the People's Republic of China[46], the state obliges to protect only a "normal religious activity", i. e. such form of religious freedom implementation that does not violate public order, nor threaten life and health of citizens or incites to subvert the political system, etc. This Article is the starting point for the whole Chinese policy and, accordingly, legislation regarding religion. Thus, the laws and other legal norms adopted based on Article 36 inherently comply with the Constitution.

At the first glance, using AI for recognition of textual and audiovisual messages also does not contradict the Basic Law, including, as one may assume from Article 34 "Provision on creating an exemplary zone of ethnic unity and progress in the Tibet Autonomous Region"[47], personal conversations between family members, as Article 40 of the PRC Constitution stipulates an abridgment of the privacy of correspondence in cases necessary for providing national security. Nevertheless, the same Article explicitly states that such activity may be implemented exclusively by criminal investigations agencies, public security bodies or prosecutor's agencies, but on no account by organizations or physical persons. It is against this norm to involve third-party organizations or specialists in implementation of the "Measures to administrate rendering religious services in the Internet", which, as follows from the official press release to this legal act, is being planned by the National Religious Affairs Administration in collaboration with Cyberspace Administration of China.

Moreover, preventive measures against dissemination of extremist materials, implying intervention into correspondence and other types of interaction through telecommunications, stipulated by the "Measures to administrate rendering religious services in the Internet"[48],

---

46 Constitution of the People's Republic of China of December 4, 1982, with amendments and additions in the edition of March 11, 2018. *The National People's Congress of the People's Republic of China*. https://www.npc.gov.cn/englishnpc/constitution2019/201911/1f65146fb6104dd3a2793875d19b5b29.shtml

47 西藏自治区民族团结进步模范区创建条例 第四章 宣传教育 of January 11, 2020, No. mzj/2020-00007. (2020). 地区民宗局. https://www.al.gov.cn/info/1258/28851.htm

48 互联网宗教信息服务管理办法 of December 3, 2021, No. 17. (2021). 国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部、中华人民共和国国家安全部令. https://www.gov.cn/gongbao/content/2022/content_5678093.htm

"Measures to administrate rendering information services in the Internet"[49] and the Law on cybersecurity[50], contradict international law, in particular, Article 12 of Universal Declaration of Human Rights[51], which does not allow intervention into private life and secrecy of correspondence, as well as Article 19, according to which everyone has the right to the freedom of thought and to disseminating one's ideas.

It is worth noting that the said legal acts, as well as "Provision on creating an exemplary zone of ethnic unity and progress in the Tibet Autonomous Region"[52] and "Provision of Xinjiang Uygur Autonomous Region on de-extremification"[53] together with Article 36 of the PRC Constitution contradict Article 18 of Universal Declaration of Human Rights[54], which guarantees freedom of thought, freedom of conscience and worship to everyone. Hence, attempts to implement them in the digital space also contradict international law. Such attempts include using generative AI to distribute state ideology and other types of propaganda with the aim to influence citizen's convictions.

The conducted research also showed that control over religious activity in China has different intensity depending on the region and implies special increased measures in ethnic autonomies, such as XUAR and Tibet. This makes apparent that "Provision on creating an exemplary zone of ethnic unity and progress in the Tibet Autonomous Region"[55] and "Provision of Xinjiang Uygur Autonomous Region on de-extremification"[56] contradict the norms of Article 33 of the PRC Constitution, according to which all citizens of China are equal before the law. Consequently, all measures for their implementation, including using "smart" cameras in certain regions, which help to collect and sort out the information about the appearance, behavior patterns of parishioners, etc., as well as the content speeches during religious rituals and services, contradict the basic law of the state and the norms of international law, in particular, the provisions of International Convention on the Elimination

---

49 互联网信息服务管理办法 of September 25, 2000, No. 292 (with amendments and additions in the edition of January 8, 2011). 日中华人民共和国国务院令第. https://www.gov.cn/zhengce/2020-12/26/content_5574367.htm

50 中华人民共和国网络安全法ofNovember7,2016.(2016).新华社北京.https://www.gov.cn/xinwen/2016-11/07/content_5129723.htm

51 Universal Declaration of Human Rights of December 10, 1948. *KonsultantPlyus*. https://www.consultant.ru/document/cons_doc_LAW_120805/

52 西藏自治区民族团结进步模范区创建条例 第四章 宣传教育 of January 11, 2020, No. mzj/2020-00007. (2020). 地区民宗局. https://www.al.gov.cn/info/1258/28851.htm

53 新疆维吾尔自治区去极端化条例 of March 27, 2017, No. mzj/2020-00007 (with amendments and additions in the edition of May 9, 2019). (2019). 新疆人大网 更新时间. https://www.xjpcsc.gov.cn/article/225/lfgz.html

54 Universal Declaration of Human Rights of December 10, 1948. *KonsultantPlyus*. https://www.consultant.ru/document/cons_doc_LAW_120805/

55 西藏自治区民族团结进步模范区创建条例 第四章 宣传教育 of January 11, 2020, No. mzj/2020-00007. (2020). 地区民宗局. https://www.al.gov.cn/info/1258/28851.htm

56 新疆维吾尔自治区去极端化条例 of March 27, 2017, No. mzj/2020-00007 (with amendments and additions in the edition of May 9, 2019) (2019). 新疆人大网 更新时间. https://www.xjpcsc.gov.cn/article/225/lfgz.html

of all Forms of Racial Discrimination[57]. The norms stipulated by this Convention, ratified by China, do not correlate with the policy of people's unity, implying forced "sinicisation" of ethnic minorities and their beliefs.

The features of measures helping China to homogenize its population were mentioned in the work by S. Jiménez-Tovar and M. Lavička, which studied the government introducing certain ideas using audiovisual and other means, influencing the formation of national identity (Jiménez-Tovar & Lavička, 2020).

In another work M. Lavička (Lavička, 2021) proves that stricter rules act in XUAR, which can be interpreted as an attempt to eradicate religion in this territory, including by segregating the younger generation from religious traditions and customs. However, one should realize that this process had started long before socialist values began to be propagated in China, originating in Confucianism (Ma, 2006). Nevertheless, following this course in the 21st century directly contradicts Article 4 of the said Convention, aimed against the ideas of ethnic supremacy. However, it would have been wrong to say that its provisions are not implemented by the Chinese government at least partially. For example, it follows from the text of the White Paper "The Fight Against Terrorism and Extremism and Human Rights Protection in Xinjiang"[58] that China executes its duties of counteracting the dissemination of the ideas of ethnic and religious supremacy by Uighur Muslims, but counteraction to their dissemination using modern digital technologies to propagate similar ideas cannot be considered an adequate response measure, corresponding to the requirements of the said Convention and norms of international law.

Stemming from the research results, one may assert that China has a well-developed legal framework for using AI technologies as a tool for regulating religious activity, and its use for these purposes by state authorities and other organizations corresponds to the norms of Chinese legislation. At the same time, the Chinese legislation per se contains a number of collisions and gaps, which may hinder the proper implementation of the norms aimed both at guaranteeing and limiting the citizens' rights and freedoms. Moreover, a number of Chinese national laws and legal act, in compliance with which AI is supposed to be used, contradict international conventions ratified by the PRC.

Today, there is no reason to believe that China will change its policy of extending state regulation onto the digital space with which its citizens interact, and will stop using AI as a control tool both within and beyond. Consequently, the national legislation of the People's Republic of China will develop in the direction of establishing state control not only over public relations, but also over AI itself and its learning (for example, the draft "Measures

[57] International Convention on the Elimination of all Forms of Racial Discrimination of March 7, 1966. Information-legal portal *GARANT.RU*. https://base.garant.ru/2540327

[58] The Fight Against Terrorism and Extremism and Human Rights Protection in Xinjiang of June 2, 2016 (2016). *The State Council Information Office of the People's Republic of China*. https://english.www.gov.cn/archive/white_paper/2016/06/02/content_281475363031504.htm

on managing the services of generative intelligence"[59] obliges to re-train generative AI in case it creates illegal materials or materials which can be used to commit illegal deeds). The conducted analysis also showed that a Chinese legislator is trying to promptly respond to gaps and collisions in law; therefore, it can be assumed that the above-mentioned internal contradictions in the national legislation will soon be eliminated.

Given the current geopolitical situation, one may assume that, under further increase of pressure on China from Western countries, including the application of measures to influence the "Global EU Sanctions Regime for Human Rights Violations", China will no longer be able to justify the existing restrictions of Article 29 of Universal Declaration of Human Rights[60], allowing them "in order to ensure proper recognition and respect for the rights and freedoms of others and to meet the just requirements of morality and public order". Chinese Academicians have already pointed out the fact that the Collective West ignores threats to China's territorial integrity and social stability under the auspices of human rights protection (Ji, 2014). Under such course of events, China either may autonomously exit from international conventions and organizations, membership in which hinders development along the chosen political course and undermines state sovereignty, or will be excluded from them – such a precedent has already formed when the Russian Federation was excluded from the Council of Europe.

References

Ashraf, C. (2020). Exploring the impacts of artificial intelligence on freedom of religion or belief online. *The International Journal of Human Rights*, *26*(5), 764. https://doi.org/10.1080/13642987.2021.1968376

Bhatia, K. V. (2021). Religious Subjectivities and Digital Collectivities on Social Networking Sites in India. *Studies in Indian Politics*, *9*(1), 22. https://doi.org/10.1177/2321023021999141

Bingaman, K. A. (2023). Religion in the Digital Age: An Irreversible Process. *Religions*, *14*(1), https://doi.org/10.3390/rel14010108

Bosch, M. D. et al. (2017). Typing my Religion. Digital use of religious webs and apps by adolescents and youth for religious and interreligious dialogue. *Church, Communication and Culture*, *2*(2), 122–135. https://doi.org/10.1080/23753234.2017.1347800

Chan, C. (2019). Using digital storytelling to facilitate critical thinking disposition in youth civic engagement: A randomized control trial. *Children and Youth Services Review*, *107*. https://doi.org/10.1016/j.childyouth.2019.104522

Fontes, C, et al. (2022). AI-powered public surveillance systems: why we (might) need them and how we want them. *Technology in Society*, *71*. https://doi.org/10.1016/j.techsoc.2022.102137

Guan, T., & Liu, T. (2019). Globalized fears, localized securities: 'Terrorism' in political polarization in a one-party state. *Communist and Post-Communist Studies*, *52*(4), 343–344. https://doi.org/10.1016/j.postcomstud.2019.10.008

Ji, F. Y. (2014). Talking Past Each Other: Chinese and Western Discourses on Ethnic Conflict. *Procedia-Social and Behavioral Sciences*, *155*(6), 434–441. https://doi.org/10.1016/j.sbspro.2014.10.318

Jiménez-Tovar, S., & Lavička, M. (2020). Folklorized politics: how Chinese soft power works inCentral Asia. *Asian Ethnicity*, *21*(2), 244–268. https://doi.org/10.1080/14631369.2019.1610355

[59] 生成式人工智能服务管理办法（征求意见稿） of April 11, 2023 (2023). 国家互联网信息办公室. https://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm

[60] Universal Declaration of Human Rights of December 10, 1948. *KonsultantPlyus*. https://www.consultant.ru/document/cons_doc_LAW_120805

Kao, Y., & Sapp, S. G. (2022). The effect of cultural values and institutional trust on public perceptions of government use of network surveillance. *Technology in Society*, *70*. https://doi.org/10.1016/j.techsoc.2022.102047

Lavička, M. (2021). Changes in Chinese legal narratives about religious affairs in Xinjiang. *Asian Ethnicity*, *22*(1), 61–76. https://doi.org/10.1080/14631369.2020.1793100

Leung, B. (2005). China's Religious Freedom Policy: The Art of Managing Religious Activity. *The China Quarterly*, *184*, 894. https://doi.org/10.1017/s030574100500055x

Lin, W. (2018). Religion as an object of state power: The People's Republic of China and its domestic religious geopolitics after 1978. *Political Geography*, *67*, 1–11. https://doi.org/10.1016/j.polgeo.2018.09.003

Ma, R. (2006). Ethnic Relations in Contemporary China: Cultural Tradition and Ethnic Policies Since 1949. *Policy and Society*, *25*(1), 85–90. https://doi.org/10.1016/s1449-4035(06)70128-x

Marique, E., & Marique, Y. (2020). Sanctions on digital platforms: Balancing proportionality in a modern public square. *Computer Law & Security Review*, *36*. https://doi.org/10.1016/j.clsr.2019.105372

Meneses, L. F. S. (2021). Thinking critically through controversial issues on digital media: Dispositions and key criteria for content evaluation. *Thinking Skills and Creativity*, *42*. https://doi.org/10.1016/j.tsc.2021.100927

Reed, R. (2021). A.I. in Religion, A.I. for Religion, A.I. and Religion: Towards a Theory of Religious Studies and Artificial Intelligence. *Religions*, *12*(6). https://doi.org/10.3390/rel12060401

Sabic-El-Rayess, A. (2012). How do people radicalize?, *International Journal of Educational Development*, *87*. https://doi.org/10.1016/j.ijedudev.2021.102499

Smith, L. G. E. et al. (2020). Detecting psychological change through mobilizing interactions and changes in extremist linguistic style. *Computers in Human Behavior*, *108*. https://doi.org/10.1016/j.chb.2020.106298

Van Wie Davis, E. (2009). Tibetan separatism in China. *Korean Journal of Defense Analysis*, *21*(2), 155–170. https://doi.org/10.1080/10163270902872135

Wei, X. (2022). A critical evaluation of China's legal responses to cyberterrorism. *Computer Law & Security Review*, *47*. https://doi.org/10.1016/j.clsr.2022.105768

Authors information

Natalya I. Shumakova – post-graduate student of the Department of Constitutional and Administrative Law, South Ural State University (National Research University)
Address: 76 prospekt Lenina, 454080 Chelyabinsk, Russian Federation
E-mail: territoryoflaw@mail.ru
ORCID ID: https://orcid.org/0009-0004-6063-0650

Elena V. Titova – Doctor of Law, Associate Professor, Head of the Department of Constitutional and Administrative Law, Director of Law Institute, South Ural State University (National Research University)
Address: 76 prospekt Lenina, 454080 Chelyabinsk, Russian Federation
E-mail: territoryoflaw@mail.ru
ORCID ID: https://orcid.org/0000-0001-9453-3550
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57201640405
Google Scholar ID: https://scholar.google.ru/citations?user=Pqj6OiQAAAAJ
RSCI Author ID: https://www.elibrary.ru/author_items.asp?authorid=451302

Authors' contributions

Elena V. Titova performed overall guidance and setting the research tasks, search and selection of scientific literature, critical evaluation of the research results interpretation.
Natalya I. Shumakova performed analysis of the Chinese national legislation, studied criminal statistics, interpreted the research results and prepared the manuscript.

Conflict of interests

The authors declare no conflict of interests.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – April 17, 2023
Date of approval – April 24, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 342.731:004.8
EDN: https://elibrary.ru/fetbvu
DOI: https://doi.org/10.21202/jdtl.2023.23

Искусственный интеллект как вспомогательный инструмент ограничения религиозной свободы в Китае

Наталья Игоревна Шумакова ✉

Южно-Уральский государственный университет (национальный исследовательский университет)
г. Челябинск, Российская Федерация

Елена Викторовна Титова

Южно-Уральский государственный университет (национальный исследовательский университет)
г. Челябинск, Российская Федерация

Ключевые слова

Искусственный интеллект,
Китай,
нейронная сеть,
права человека,
право,
регулирование,
религия,
свобода вероисповедания,
цифровые технологии,
экстремизм

Аннотация

Цель: на основании изучения статистики преступлений, национального законодательства и норм международного права дать правовую оценку ограничениям права на свободу вероисповедания с применением технологий искусственного интеллекта в Китае.

Методы: методологическую основу исследования составляет совокупность методов научного познания, в том числе конкретно-социологический (анализ статистических данных и иных документов), формально-юридический (изучение правовых категорий и дефиниций), формально-логические методы (анализ и синтез), общенаучные методы (индукция, дедукция) и др.

Результаты: в работе изучены предпосылки использования в Китае технологий искусственного интеллекта для контроля за общественными отношениями, возникающими в процессе религиозной активности как в цифровом пространстве, так и за его пределами; проанализирована правовая база применения указанных мер; дана правовая оценка ограничения религиозной свободы с использованием технологий искусственного интеллекта; сделан прогноз дальнейшего развития китайских законодательства и внешней политики, связанных с религиозной свободой. Дополнительно в работе проанализированы материалы правозащитных организаций, направленных на сдерживание

✉ Контактное лицо

политики КНР по «китаизации» и «деэкстремификации» этнических и религиозных меньшинств, в том числе при помощи контроля и пропаганды с использованием современных цифровых технологий.

Научная новизна: в работе исследована попытка Китая урегулировать связанные с религиозной активностью вызовы, возникающие в процессе стремительной цифровизации общества и государства, с которыми республика сталкивается как развивающаяся, многонациональная и поликофессиональная страна. Установленные ограничения религиозной свободы с применением технологий искусственного интеллекта рассмотрены в комплексе с релевантной статистикой преступлений. Правовая оценка применения искусственного интеллекта как инструмента ограничения свободы вероисповедания дана не только с точки зрения международного права, но и с учетом норм китайского национального законодательства.

Практическая значимость: результаты проведенного исследования могут быть использованы для разработки непротиворечивой нормативной правовой базы для использования технологий искусственного интеллекта в целях противоборства экстремизму.

Для цитирования

Шумакова, Н. И., Титова, Е. В. (2023). Искусственный интеллект как вспомогательный инструмент ограничения религиозной свободы в Китае. *Journal of Digital Technologies and Law*, *1*(2), 540–563. https://doi.org/10.21202/jdtl.2023.23

Список литературы

Ashraf, C. (2020). Exploring the impacts of artificial intelligence on freedom of religion or belief online. *The International Journal of Human Rights*, *26*(5), 764. https://doi.org/10.1080/13642987.2021.1968376

Bhatia, K. V. (2021). Religious Subjectivities and Digital Collectivities on Social Networking Sites in India. *Studies in Indian Politics*, *9*(1), 22. https://doi.org/10.1177/2321023021999141

Bingaman, K. A. (2023). Religion in the Digital Age: An Irreversible Process. *Religions*, *14*(1), https://doi.org/10.3390/rel14010108

Bosch, M. D. et al. (2017). Typing my Religion. Digital use of religious webs and apps by adolescents and youth for religious and interreligious dialogue. *Church, Communication and Culture*, *2*(2), 122–135. https://doi.org/10.1080/23753234.2017.1347800

Chan, C. (2019). Using digital storytelling to facilitate critical thinking disposition in youth civic engagement: A randomized control trial. *Children and Youth Services Review*, *107*. https://doi.org/10.1016/j.childyouth.2019.104522

Fontes, C, et al. (2022). AI-powered public surveillance systems: why we (might) need them and how we want them. *Technology in Society*, *71*. https://doi.org/10.1016/j.techsoc.2022.102137

Guan, T., & Liu, T. (2019). Globalized fears, localized securities: 'Terrorism' in political polarization in a one-party state. *Communist and Post-Communist Studies*, *52*(4), 343–344. https://doi.org/10.1016/j.postcomstud.2019.10.008

Ji, F. Y. (2014). Talking Past Each Other: Chinese and Western Discourses on Ethnic Conflict. *Procedia-Social and Behavioral Sciences*, *155*(6), 434–441. https://doi.org/10.1016/j.sbspro.2014.10.318

Jiménez-Tovar, S., & Lavička, M. (2020). Folklorized politics: how Chinese soft power works inCentral Asia. *Asian Ethnicity*, *21*(2), 244–268. https://doi.org/10.1080/14631369.2019.1610355

Kao, Y., & Sapp, S. G. (2022). The effect of cultural values and institutional trust on public perceptions of government use of network surveillance. *Technology in Society*, *70*. https://doi.org/10.1016/j.techsoc.2022.102047

Lavička, M. (2021). Changes in Chinese legal narratives about religious affairs in Xinjiang. *Asian Ethnicity*, *22*(1), 61–76. https://doi.org/10.1080/14631369.2020.1793100

Leung, B. (2005). China's Religious Freedom Policy: The Art of Managing Religious Activity. *The China Quarterly*, *184*, 894. https://doi.org/10.1017/s030574100500055x

Lin, W. (2018). Religion as an object of state power: The People's Republic of China and its domestic religious geopolitics after 1978. *Political Geography*, *67*, 1–11. https://doi.org/10.1016/j.polgeo.2018.09.003

Ma, R. (2006). Ethnic Relations in Contemporary China: Cultural Tradition and Ethnic Policies Since 1949. *Policy and Society*, *25*(1), 85–90. https://doi.org/10.1016/s1449-4035(06)70128-x

Marique, E., & Marique, Y. (2020). Sanctions on digital platforms: Balancing proportionality in a modern public square. *Computer Law & Security Review*, *36*. https://doi.org/10.1016/j.clsr.2019.105372

Meneses, L. F. S. (2021). Thinking critically through controversial issues on digital media: Dispositions and key criteria for content evaluation. *Thinking Skills and Creativity*, *42*. https://doi.org/10.1016/j.tsc.2021.100927

Reed, R. (2021). A.I. in Religion, A.I. for Religion, A.I. and Religion: Towards a Theory of Religious Studies and Artificial Intelligence. *Religions*, *12*(6). https://doi.org/10.3390/rel12060401

Sabic-El-Rayess, A. (2012). How do people radicalize?, *International Journal of Educational Development*, *87*. https://doi.org/10.1016/j.ijedudev.2021.102499

Smith, L. G. E. et al. (2020). Detecting psychological change through mobilizing interactions and changes in extremist linguistic style. *Computers in Human Behavior*, *108*. https://doi.org/10.1016/j.chb.2020.106298

Van Wie Davis, E. (2009). Tibetan separatism in China. *Korean Journal of Defense Analysis*, *21*(2), 155–170. https://doi.org/10.1080/10163270902872135

Wei, X. (2022). A critical evaluation of China's legal responses to cyberterrorism. *Computer Law & Security Review*, *47*. https://doi.org/10.1016/j.clsr.2022.105768

Сведения об авторах

Шумакова Наталья Игоревна – аспирант кафедры конституционного и административного права, Южно-Уральский государственный университет (национальный исследовательский университет)
Адрес: 454080, Российская Федерация, г. Челябинск, пр. Ленина, 76
E-mail: territoryoflaw@mail.ru
ORCID ID: https://orcid.org/0009-0004-6063-0650

Титова Елена Викторовна – доктор юридических наук, доцент, заведующий кафедрой конституционного и административного права, директор юридического института, Южно-Уральский государственный университет (национальный исследовательский университет)
Адрес: 454080, Российская Федерация, г. Челябинск, пр. Ленина, 76
E-mail: territoryoflaw@mail.ru
ORCID ID: https://orcid.org/0000-0001-9453-3550
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57201640405
Google Scholar ID: https://scholar.google.ru/citations?user=Pqj6OiQAAAAJ
РИНЦ Author ID: https://www.elibrary.ru/author_items.asp?authorid=451302

Вклад авторов

Е. В. Титова осуществляла общее руководство и постановку задач исследования, поиск и подбор научной литературы, критическую оценку интерпретации результатов исследования.
Н. И. Шумакова осуществляла анализ национального законодательства Китая, изучала криминальную статистику, выполняла интерпретацию результатов исследования и подготовку рукописи.

Конфликт интересов

Авторы сообщают об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.15.59 / Права и свободы человека и гражданина
Специальность ВАК: 5.1.2 / Публично-правовые (государственно-правовые) науки

История статьи

Дата поступления – 3 марта 2023 г.
Дата одобрения после рецензирования – 4 мая 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.24

The Possibility and Necessity of the Human-Centered AI in Legal Theory and Practice

Andrey V. Rezaev ✉

Saint Petersburg State University
Saint Petersburg, Russian Federation

Natalia D. Tregubova

Saint Petersburg State University
Saint Petersburg, Russian Federation

Keywords

Algorithm,
artificial intelligence,
artificial sociality,
digital economy,
digital technologies,
human,
human-centered artificial intelligence,
law,
regulation,
sociology

Abstract

Objective: the paper aims to define the problems juridical theory and practice face with the progress of AI technologies in everyday life and correlate these problems with the human-centered approach to exploring artificial intelligence (Human-Centered AI).

Methods: the research critically analyzes the relevant literature from various disciplines: jurisprudence, sociology, philosophy, and computer sciences.

Results: the article articulates the prospects and problems the legal system confronts with the advancement of digital technologies in general and the tools of AI specifically. The identified problems are correlated with the provisions of the human-centered approach to AI. The authors acknowledge the necessity for AI inventors, as well as the owners of companies participating in the race to develop artificial intelligence technologies, to place humans, not machines, into the focus of attention as a primary value. In particular, special effort should be directed towards collecting and analyzing high-quality data for the organization of artificial intelligence tools development, taking into account that nowadays, the tools of AI are as practical as the data on which they are trained are effective.

✉ Corresponding author

The English translation of the original text has been provided by the Editorial Office of the Journal of Digital Technologies and Law.

The authors formulate three principles of human-centered AI for the legal sphere: 1) a human as a necessary link in the chain of making and executing legal decisions; 2) the need to regulate artificial intelligence at the international law level; 3) formulating "a taboo" for introducing the artificial intelligence technologies.

Scientific novelty: the article manifests one of the first attempts in the Russian-language scientific literature to outline the prospects of developing human-centered AI methodology in jurisprudence. Based on an analysis of special literature, the authors formulate three principles of including artificial intelligence into juridical theory and practice according to the assumptions of a human-centered approach to AI.

Practical significance: the principles and arguments the article advances can be helpful in the legal regulation of artificial intelligence technologies and their harmonious inclusion into legal practices.

For citation

Rezaev, A. V., Tregubova, N. D. (2023). The Possibility and Necessity of the Human-Centered AI in Legal Theory and Practice. *Journal of Digital Technologies and Law, 1*(2), 564–580. https://doi.org/10.21202/jdtl.2023.24

Contents

Introduction

In 1948, Norbert Wiener, a founder of cybernetics, wrote: "we are already in a position to construct artificial machines of almost any degree of elaborateness of performance. Long before Nagasaki and the public awareness of the atomic bomb, it had occurred to me that we were here in the presence of another social potentiality of unheard-of importance for good and evil" (Wiener, 1983).

Today, "artificial machines" are already solving (or will be able to solve soon) multiple problems humanity faces. However, these machines undoubtedly, created new problems,

too[1]. What Wiener called "artificial machines" is now, in this or that form, a part of the life of society, and we can hardly imagine our life without artificial intelligence technologies. Therefore, it is not a surprise that in recent years there has been a lot of information «noise» around artificial intelligence and its potential to radically change the world we live and work in.

The objective of our considerations here is to show that as artificial intelligence technologies are being developed and introduced into our daily life, the necessity is proportionally increasing for the software developers, designers, and owners of the companies participating in the race to introduce new AI tools, to have humans and their needs but not machines and their efficiency, as the primary value and goal of advancement. It is not the goals of one person or company, not the technologies or machines, but a human and the human attitude that serves as the measure of morals and humanness. Realizing that good-hearted calls for humanness may sound abstract in the logic of technologically oriented development, we would like to discuss more specifically how the need to work with artificial intelligence within the approach called Human-Centered AI (HCAI).

The problem of the artificial intelligence orientation towards the good of humans is acute in all spheres of life but especially sensitive in some of them. These include education, medicine, and jurisprudence, where the price of a mistake – of a human or an algorithm – is the highest. Juridical solutions regulate human life and their relations with others and sometimes refer to existential issues – life, death, and justice.

In this article, we consider the problems the juridical theory and practice face with the advancement of AI in everyday life and how these problems correlate with the human-centered approach to AI. We define artificial intelligence as "an ensemble of rational, logical, and formalized instrumental rules developed and coded by human beings that organize the processes and activities to emulate rational/intellectual structures and fabricate and reproduce goal-oriented practices as well as the mechanisms for constructing further coding and decision making." (Rezaev & Tregubova, 2019).

Today, one of the factors determining the development of artificial intelligence technologies is online culture – "an ensemble of communication networks, devices, algorithms, formal and informal rules of interaction, patterns of behavior, cultural symbols, which allow and structure people's activity in the internet and similar networks, providing remote access to creating, exchanging and obtaining information" (Rezaev & Tregubova, 2019).

[1] To confirm this, one may cite a recent statement by Sam Altman, a founder of OpenAI company which developed a famous ChatGPT chatbot: "I think where we are right now is not where we want to be. The way this should work is that there are extremely wide bounds of what these systems can do, that are decided by, not Microsoft or OpenAI, but society, governments, something like that, some version of that, people, direct democracy. <…> It's very new technology. We don't know how to handle it." *Bing's Revenge and Google's AI FacePlant*. https://www.nytimes.com/2023/02/10/podcasts/bings-revenge-and-googles-ai-face-plant.html

The Internet provides vast data on which artificial intelligence algorithms are trained and the "platform" for these algorithms to act.

As a result of the simultaneous development of computation capacities of artificial intelligence online culture, artificial intelligence is increasingly involved in everyday life and human relationships. "Artificial sociality" appears: artificial intelligence becomes an active mediator and participant in social interactions (Rezaev & Tregubova, 2019).

From its inception, the AI project had an a-disciplinary character. The artificial intelligence developers strived to reproduce human intelligence, hence, boldly borrowed the necessary provisions from mathematics, psychology, cybernetics, etc. (Russell & Norvig, 2007). However, while developing the machines reproducing the functioning of the human mind required turning to the achievements from various fields of knowledge, this is all the more true to understand how these machines enter the everyday life of society and are built into social relations. In other words, researching the problems of artificial sociality has an interdisciplinary and potentially – "a-disciplinary" character. That is why, in this article, we rely on both the philosophical and sociological analysis of the problems of AI, and the results of research in jurisprudence and law.

Further reflections are organized as follows. First, we will pay attention to several vital aspects associated with the introduction of digital technologies into legal practices. Then we will consider the problems and prospects of the rapid penetration of AI into the everyday life of society, which changes the characteristics of juridical work and the structure of legal systems. Finally, we will turn to the human-centered approach to AI and its consequences for the legal sphere.

1. Digital technologies and law

Summarizing the influence of digital technologies on the legal system, one should emphasize the following.

First, digital technologies simplified access to legal information via online databases, legal search systems, and other online resources. This fact, accordingly, created opportunities for nearly every Internet user (that is, almost 90% of the Russian population)[2] to perform online research to obtain legal information. The Internet revolutionized search in all spheres of human life (Utekhin, 2019) and legal information is no exception.

Second, digital technologies improved communication between lawyers, clients, and other actors in the legal system. For example, videoconferencing allows lawyers to communicate with their clients distantly. Thus, access to legal services for residents of remote districts and regions is improved.

---

[2] *Dmitriy Chernyshenko: Russia has about 130 million Internet users today – almost 90% of the population.* http://government.ru/news/46639/

Third, digital technologies made it possible to submit and store legal documents electronically. In other words, organizing juridical practices with digital technologies significantly decreases the need for paper documents and simplifies information search and exchange (Rusakova, 2020; Stepanov et al., 2021).

Fourth, digital technologies have led to automating many legal processes, such as routine checks of documents and compiling contracts (including the so-called smart contracts (Efimova et al., 2020)). Accordingly, the demand for routine manual labor of lawyers and their assistants has significantly decreased.

Fifth, using digital technologies in legal practices gave rise to new branches of law, such as cyberlaw/law in cyberspace (Mazhorina, 2020), intellectual property law, and data protection law (Voinikanis, 2020).

Thus, digital technologies have already significantly influenced the development of law, making legal services and practices more accessible, efficient, and effective. At the same time, practicing lawyers, special literature, mass media, and everyday legal service consumers almost unanimously emphasize that digital technologies generate new problems for the legal system development. These are, first of all, the issues of confidentiality (Talapina, 2022) and accessibility (Panchenko, 2012) of legal databases and the problem of critical assessment of the information obtained from the Internet (Greger, 2017).

The current stage of digital technologies development in online culture suggests paying attention to how artificial intelligence transforms and shapes further development of legal practices. What are the advantages and disadvantages of using AI technologies in routine legal practice?

2. Artificial intelligence in legal practice and theory: *pro et contra*

Using artificial intelligence technologies in routine legal practice has both advantages and disadvantages. The main benefits are the following:

– Effective organization of the lawyers' practices. The AI instruments automate and accelerate the performance of such tasks as document review, preliminary juridical examination of literature sources, and analysis of contracts (Talapina, 2021).

– Artificial intelligence may perform specific tasks more accurately than people, for example, find regularities in data or check documents for factual mistakes, and grammatical or stylistic inconsistencies (Andreev et al., 2020).

– The use of AI technologies, reducing the need for manual work, saves the funds of juridical companies and their clients.

– Artificial intelligence technologies provide lawyers with more complete, comprehensive, and detailed information allowing them to make more grounded decisions.

The main disadvantages of using artificial intelligence are the following:

– The common disadvantage of using artificial intelligence for all professions is that some professions disappear while others will appear and come to the fore (Lee, 2019). A broad use of AI technologies in legal practice is still a potential, but very soon, it will inevitably lead to a review of jobs nomenclature within the juridical system structure; this will especially touch upon paralegals and other auxiliary staff (Lessig, 2019).

– artificial intelligence systems are, to a certain extent, translators of bias and prejudices characteristic of their creators (Gorokhova, 2021). The artificial intelligence algorithms may be biased or erroneous due to, at least, two circumstances: a) if they are based on and were developed with biased or erroneous data arrays; b) if they are misused. Hence, the introduction of AI technologies implies searching for ways to ensure just and bias-free artificial intelligence systems.

– artificial intelligence technologies, like any other technologies, bear safety risks. AI technologies cannot guarantee complete cybersecurity (O'Neil, 2018). Artificial intelligence may minimize but not eliminate data leakage or hacking. Accordingly, confidentiality – the cornerstone of legal practices – is threatened when using artificial intelligence technologies.

Thus, using AI technologies in everyday legal practices provides multiple advantages, but they should be weighed with potential risks and drawbacks. Lawyers must not only realize the capabilities of AI, but also thoroughly review their use and see their limitations and potential risks.

Besides the problems with using the algorithms and machines which are already manifested in everyday life, one should also keep in mind the actual problems generated by the ubiquitous penetration of artificial intelligence into legal practices:

– confidentiality problems. The effective performance of artificial intelligence systems often requires access to large amounts of personal data, which causes concerns about confidentiality and data protection. Regulators and legislators must find a balance between privacy protection and promoting innovations in the sphere of artificial intelligence (Gorokhova, 2021).

– legal liability for the actions performed by artificial intelligence. As artificial intelligence systems become more autonomous and make decisions without human interference, questions arise about who is responsible for their actions (Vavilin, 2021; Baturin & Polubinskaya, 2022). For example, if an AI driverless car caused an accident, should

a developer, a user, or the artificial intelligence system per se be liable (Rudenko, 2020)? Who will bear responsibility if something goes wrong when AI instruments are used? Who will be responsible for accidents or mistakes caused by an artificial intelligence system: a programmer, an owner (of what?) or the AI designer? These already are the juridical questions of today.

– the issues related to intellectual property rights to the products created by artificial intelligence technologies (Lee et al., 2021). For example, who will be deemed an inventor or an artist if an artificial intelligence system creates a work of art or invents a new technology?

– a critical element is using artificial intelligence tools (for example, ChatGPT) for juridical interpretation of documents and applying legal norms, especially regarding the complex and nuanced character of juridical substantiation of a certain decision. There are grounds to fear that artificial intelligence will not be able to comprehensively grasp human considerations and judgments necessary for effective juridical decision-making (Tsvetkov, 2021).

– lack of communication and real-life human contacts. This is a significant disadvantage for legal practices, which may, by default, touch upon existential matters of life and death, restriction of freedom. Judges note: justice is impossible without a holistic view of the situation, including moral and emotional aspects, which is inaccessible for AI (Bykov & Narskaya, 2022).

Noteworthy, the very question of whether AI technologies should be regulated and how is the object of discussion (Etzioni & Etzioni, 2017). The frameworks are just starting to be elaborated in this sphere, the "pioneers" often being the legislators of the European Union (Hickman & Petrin, 2021; Fink & Finck, 2022; Ulnicane, 2022). In Russia, legal regulation of artificial intelligence technologies is also being developed. In 2019, the National strategy of artificial intelligence development up to 2030[3] was adopted, specifying the basic definitions and general principles of using AI technologies.

Thus, the development of artificial sociality poses both a practical and conceptual problem for jurisprudence. The practical/functional problem is how artificial intelligence technologies will change legal practices, while the conceptual problem refers to their legal regulation.

We believe that a set of problems which have already emerged and which are bound to emerge in the future in legal practice can be solved more effectively with the approach called a human-centered AI.

[3] On the development of artificial intelligence in the Russian Federation: Executive Order of the President of the Russian Federation. http://static.kremlin.ru/media/events/files/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf

3. Human-centered artificial intelligence

The approach called a Human-Centered AI in the scientific literature (Ford et al., 2015; Shneiderman, 2021)[4] implies, first of all, understanding the straightforward fact that people and machines are not the same[5]. There is no need to aim at making an artificial intelligence tool similar to a human being. On the contrary, success will probably be achieved in the opposite direction when a human stays a human with their intellect, consciousness, subconsciousness, and emotional and spiritual world. At the same time, machines and algorithms will be developed by a human and, during "self-training," follow their own logic of development, different from that of a human.

Unfortunately, this circumstance is being neglected, just like the human-centered approach to AI in general. Most technological leaders in the USA and other countries continue spending a lot on developing software that can do just what people can do. Developers very well realize that they can earn easy money by selling their products to corporations having no other orientations in their development except those set by the logic of the market and profit (Zuboff, 2022). Everyone is focused on using artificial intelligence to reduce costs for the working force while caring little about the essence of the social progress and development of a moral human being and a just society.

Human-centered AI requires immediate attention to collecting and analyzing high-quality data to organize the development of artificial intelligence tools. The artificial intelligence algorithms are as effective as the data on which they are trained are effective. In contrast, partial or incomplete data may lead to not only unjust/false results but to ones opposite to the initial goal. Collecting information for self-learning models must be diverse and representative; the data must reflect the real world we live in and the people we work with, regardless of their social and class differences.

Elsewhere, we have already emphasized the fact that, under the current stage of capitalism development, under the extreme orientation towards financial indicators, profit, and functional efficiency, it is practically impossible to solve these problems (Rezaev, 2021)[6]. Nevertheless, it would be wrong for the tactics of social sciences development not to consider them at all and not to attempt to propose variants of their solution.

---

4 Notably, in 2019 a Human-Centered AI Institute was established at Stanford University (USA) – the largest research center in this area.

5 This statement has been repeatedly made in philosophy and social sciences. See (Dreyfus, 1978; Wolfe, 1993; Esposito, 2017).

6 For example, Elon Musk (who sponsored OpenAI company which developed ChatGPT) said with obvious regret that he donated money (US$ 1 billion) to create an open platform aimed at free open access, while now ChatGPT is an opposite model – close and fully aimed at profits. However, Elon Musk executes no control over OpenAI or ChatGPT at the moment. *Elon Musk at the 2023 World Government Summit in Dubai*. https://www.youtube.com/watch?v=jmNrlNgXx_U&ab_channel=ElonAlerts

The market has never been and cannot be (even under artificial sociality) the touchstone of beauty, goodness, and truth. Strategically, social knowledge substantiated the impossibility of a harmonious, moral, and just world without exploitation of human by human and without social and cultural inequality within the framework of a capitalist economy[7]. However, what the problems for society are and what the variants of the social development trajectories are under the still uncontrolled spreading of AI tools – these topics are just starting to be considered belatedly.

Characterizing the features of artificial intelligence development, one should remember that AI technologies are not neutral. Humans create them, and algorithms reproduce their creators' values, biases, and prejudices. Thus, AI designers and producers must adhere to ethical and human-oriented approaches. It means, among other things, accounting for various viewpoints and opinions in the design process, providing transparency and accountability, and paying priority attention to the human personality and wellbeing of society in general, not that of individual subjects or technological systems.

The key point is the understanding that AI instruments are already a powerful means of solving some of the most burning problems facing society, but they are not a panacea. While defining and formulating the directions of social development, one should not rely exclusively on artificial intelligence when solving social, economic, cultural, and political problems. Even under "artificial sociality," people must remain within the reality of human experience and admit that social progress requires more than just technological solutions.

[7] An example is "The Wealth of Nations" by Adam Smith. Although Smith is often called the first theoretician of political economy and an advocate of capitalism, his works are critical regarding many aspects of a capitalist economy. For example, he postulates that a rush for profits may lead to a lack of concern about the well-being of workers and society as a whole and that to provide a just and equal society a certain form of state intervention is necessary. In his work "The Great Transformation" Karl Polanyi states that capitalism is a historically recent phenomenon that generated absolutely negative consequences for social development, including turning labor into a fictitious commodity, destroying a traditional way of life, and rising nationalistic and fascist movements. Thorstein Veblen in his book "The Theory of the Leisure Class" showed that capitalism creates a conspicuous consumption and wastefulness culture when people are praised for their ability to consume and demonstrate their wealth, not for their contribution to society. A contemporary Canadian researcher Naomi Klein asserts that capitalism is often imposed on society by violence and coercion and is often used by the wealthy elite to maintain their political and economic power (Klein, 2007). See for more details (Harvey, 2014).

Conclusion

This essay began with a citation from Norbert Wiener, the founder of cybernetics. We want to conclude it with the judgments presented by one of the founders of artificial intelligence research – Joseph Weizenbaum. Weizenbaum argued that the use of computers should be banned or at least restricted in two cases (Weizenbaum, 1982). The first case is connected with attempts to replace a human with a machine in the areas related to interpersonal relationships, love, and understanding. The second case is using computers in a situation where it can lead to irreversible consequences. In our opinion, Weizenbaum correctly formulated the basic principles of human-centered artificial intelligence, which relate to the general spread of AI technologies and their use in the theory and practice of jurisprudence, in particular.

In conclusion, we will formulate three principles of including AI into the legal theory and practice according to the methodological principles of human-centered approach (HCAI).

First. A human being must always remain within the chain of making/executing legal decisions. Legal scientists have persistently formulated this thesis. Artificial intelligence technologies may take on many tasks in legal practice, but it is a human being that must control, check, conceptualize, and weigh the actions and decisions of the artificial intelligence.

Second. Today, we need to elaborate the laws determining a rational and understandable modus vivendi for the activity of artificial intelligence in social systems aimed at a human being, not at profit and market. This is almost impossible within one state, especially a capitalist one. That is why the world faces the need to create international law for AI evolvement in society. Like any rule, the law may be violated – by mistake or out of malice. But violation of the law does not repeal the law itself; it just reveals the malicious persons who distorted the law.

Third. The progress of AI in everyday life of people poses the need for prohibitions, including juridical ones, a taboo for using artificial intelligence in certain spheres of human life (Rezaev, 2021). These are, first of all, spheres associated with existential issues. For example, an important issue is whether one should use artificial intelligence to determine if a person is lying (Oravec, 2022), or whether artificial intelligence may serve as an autonomous weapon (International Committee, 2020). Defining such spheres at international, national, and local levels, formulating legal prohibitions, and creating law-enforcement mechanisms is one of the priority tasks for Human-Centered AI.

References

Andreev, V. K., Laptev, V. A., & Chucha S. Yu. (2020). Artificial intelligence in the system of electronic justice by consideration of corporate disputes. *Vestnik of Saint Petersburg University. Law*, *11*(1), 19–34. (In Russ.). https://doi.org/10.21638/spbu14.2020.102

Baturin, Yu. M., & Polubinskaya, S. V. (2022). Artificial intelligence: legal status or legal regime? *Gosudarstvoiparvo*, *10*, 141. (In Russ.). https://doi.org/10.31857/s102694520022606-7

Bykov, A. V., & Narskaya, A. I. (2022). Law, Morality, and Machine Learning: Judges' Perspective on the Essence of Justice and the Prospects of Its Robotization. *Monitoring of Public Opinion: Economic and Social Changes Journal (Public Opinion Monitoring)*, *5*, 278–298. (In Russ.). https://doi.org/10.14515/monitoring.2022.5.2137

Dreyfus, H. (1978). *What computers can't do: A critique of artificial reason.* Moscow: Progress. (In Russ.).

Efimova, L., Mikheeva, I., & Chub, D. (2020). Comparative Analysis of Doctrinal Concepts of Legal Regulating Smart Contracts in Russia and Foreign States. *Journal of the Higher School of Economics*, *4*, 78–105. (In Russ.).

Esposito, E. (2017). Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift für Soziologie*, *46*(4), 249–265. https://doi.org/10.1515/zfsoz-2017-1014

Etzioni, A., & Etzioni, O. (2017). Should Artificial Intelligence Be Regulated? *Issues in Science and Technology*, *33*(4), 32–36.

Fink, M., & Finck, M. (2022). Reasoned A(I) administration: explanation requirements in EU law and the automation of public administration. *European Law Review*, *47*(3), 376–392.

Ford, K. M., Hayes, P. J., Glymour, C., & Allen, J. (2015). Cognitive Orthoses: Toward Human-Centered AI. *AI Magazine*, *36*(4), 5–8. https://doi.org/10.1609/aimag.v36i4.2629

Gorokhova, S. S. (2021). Artificial intelligence: an instrument ensuring cybersecurity of the financial sphere or a cyber threat to banks? *Banking Law*, *1*, 35–46. (In Russ.). https://doi.org/10.18572/1812-3945-2021-1-35-46

Greger, R. (2017). Judge as an Internet Surfer. Identification of the Circumstances of the Case on the Internet. *Herald of Civil Procedure,* 7(4), 161–173. (In Russ.). https://doi.org/10.24031/2226-0781-2017-7-4-161-173

Harvey, D. (2014). *Seventeen Contradictions and the End of Capitalism*. Oxford: Oxford University Press.

Hickman, E., & Petrin, M. (2021). Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective. *European Business Organization Law Review*, *22*, 593–625. https://doi.org/10.1007/s40804-021-00224-0

International Committee of the Red Cross (2020). Artificial intelligence and machine learning in armed conflict: A human-centred approach. *International Review of the Red Cross*, *102*(913), 463–479. https://doi.org/10.1007/s40804-021-00224-0

Klein, N. (2007). *The Shock Doctrine: The Rise of Disaster Capitalism*. New York: Henry Holt.

Lee, J.-A., Hilty, R. M, & Liu, K.-C. (Eds.). (2021). *Artificial Intelligence and Intellectual Property*. Oxford: Oxford University Press.

Lee, K.-F. (2019). *AI Superpowers: China, Silicon Valley and the new world order*. Moscow: Mann, Ivanov i Ferber. (In Russ.).

Lessig, L. (2019). Artificial intelligence is going to oust a wide circle of lawyers. *Zakon*, *5*, 8–30. (In Russ.).

Mazhorina, M. (2020). Cyberplace and Methodology of International Private Law. *Journal of the Higher School of Economics*, *2*, 230–253. (In Russ.).

O'Neil, C. (2018). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Moscow: AST. (In Russ.).

Oravec, J. A. (2022). The emergence of "truth machines"? Artificial intelligence approaches to lie detection. *Ethics and Information Technology*, *24*, 6. https://doi.org/10.1007/s10676-022-09621-6

Panchenko, V. Yu. (2012). Information availability of legal assistance: ideal and real state. *Agrarnoye i zemelnoye parvo*, *11*(95), 95–102. (In Russ.).

Rezaev, A. V. (2021). Twelve Theses on Artificial Intelligence and Artificial Sociality. *Monitoring of Public Opinion: Economic and Social Changes*, *1*, 20–30. https://doi.org/10.14515/monitoring.2021.1.1894

Rezaev, A. V., Tregubova, N. D. (2019). Artificial intelligence, On-line Culture, Artificial Sociality: Definition of the Terms. Monitoring of Public Opinion: *Economic and Social Changes*, *6*, 35–47. https://doi.org/10.14515/monitoring.2019.6.03

Rudenko, N. (2020). Sociotechnical barriers to developing autonomous vehicles in Russia. In L. Zemnukhova, K. Glazkov, O. Logunova, A. Maksimova, D. Sivkov, & N. Rudenko, *The Adventures of Technologies: Digitalization Barriers in Russia* (17–70). Moscow – Saint Petersburg: FNISTS RAN. (In Russ.). https://doi.org/10.31119/978-5-89697-339-3

Rusakova, E. (2020). The integration of modern digital technologies to the legal proceedings of People's Republic of China and Singapore, *Gosudarstvo i parvo*, *9*, 102. (In Russ.). https://doi.org/10.31857/s102694520011323-6

Russell, S., & Norvig, P. (2007). *Artificial Intelligence: A Modern Approach* (2d ed.). Moscow: Vilyams. (InRuss.).

Shneiderman, B. (2021). Human-centered AI. *Issues in Science and Technology*, 37(2), 56–61.

Stepanov, O., Pechegin, D., & Diakonova, M. (2021). Towards the Issue of Digitalization of Judicial Activities. *Journal of the Higher School of Economics*, *5*, 4–23. (In Russ.). https://doi.org/10.17323/2072-8166.2021.5.4.23

Talapina, E. V. (2021). Artificial intelligence and legal expertise in public administration. *Vestnik of Saint Petersburg University. Law*, *12*(4), 865–881. (In Russ.). https://doi.org/10.21638/spbu14.2021.404

Talapina, E. V. (2022). The right to informational self-determination: on the edge of public and private. *Law. Journal of the Higher School of Economics*, *15*(5), 24–43. (In Russ.).

Tsvetkov, Yu. A. (2021). Artificial Intelligence in Justice. *Zakon*, *4*, 91–107. (In Russ.).

Ulnicane, I. (2022). Artificial Intelligence in the European Union: policy, ethics and regulation. In T. Hoerber, I. Cabras, G. Weber (Eds.). *Routledge Handbook of European Integrations* (pp. 254–269). London: Routledge. https://doi.org/10.4324/9780429262081-19

Utekhin, I. (2019). Search and Interfaces for Search. *Laboratorium: Russian Review of Social Research*, *11*(1), 152–165. (In Russ.). https://doi.org/10.25285/2078-1938-2019-11-1-152-165

Vavilin, E. V. (2021). Artificial intelligence as a participant in civil relations: the transformation of law. *Vestnik Tomskogo gosudarstvennogo universiteta. Pravo*, *42*, 135–146. (In Russ.). https://doi.org/10.17223/22253513/42/11

Voinikanis, E. A. (2020). Regulation of big data and intellectual property right: common approaches, problems and prospects of development. *Zakon*, *7*, 135–156. (In Russ.).

Weizenbaum, J. (1982). *Computer Power and Human Reason: From Judgment to Calculation*. Moscow: Radio i svyaz. (In Russ.).

Wiener, N. (1983). *Cybernetics: Or Control and Communication in the Animal and the Machine* (2d ed.). Moscow: Nauka; Glavnaya redaktsiya izdanii dlya zarubezhnykh stran. (In Russ.).

Wolfe, A. (1993). *The Human Difference: Animals, Computers, and the Necessity of Social Science*. Berkeley: University of California Press.

Zuboff, Sh. (2022). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Moscow: Izd-vo Instituta Gaidara. (In Russ.).

Authors information

Andrey V. Rezaev – Doctor of Philosophical Sciences, Professor, Head of the International Research Laboratory TANDEM at the Faculty of Sociology, St Petersburg State University
Address: 1/3 Smolnogo Str., 191124 Saint Petersburg, Russia
E-mail: rezaev@hotmail.com
ORCID ID: https://orcid.org/0000-0002-3918-835X
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=13004674100
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/K-3472-2013
Google Scholar ID: https://scholar.google.ru/citations?user=Uzv39ccAAAAJ
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=648768

Natalia D. Tregubova – PhD (Sociology), Associate Professor of the Department of Comparative Sociology, Saint Petersburg State University
Address: 1/3 Smolnogo Str., 191124 Saint Petersburg, Russia
E-mail: n.tregubova@spbu.ru
ORCID ID: https://orcid.org/0000-0003-3259-5566
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=56645016900
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/K-3487-2013
Google Scholar ID: https://scholar.google.com/citations?user=8dhGr3gAAAAJ&hl
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=832705

Authors' contributions

A. V. Rezaev and N. D. Tregubova contributed equally to the formulation of the article's key provisions and preparation of the manuscript for publication.

Conflict of interests

The authors declare no conflict of interests.

Financial disclosure

The study was supported by RFBR and MOST, the research project No. 21-511-52002.

Thematic rubrics:

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – March 3, 2023
Date of approval – May 4, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023

Научная статья
УДК 340.143:004.8
EDN: https://elibrary.ru/sadrzw
DOI: https://doi.org/10.21202/jdtl.2023.24

Возможность и необходимость человеко-ориентированного искусственного интеллекта в юридической теории и практике

Андрей Владимирович Резаев ✉

Санкт-Петербургский государственный университет
г. Санкт-Петербург, Российская Федерация

Наталья Дамировна Трегубова

Санкт-Петербургский государственный университет
г. Санкт-Петербург, Российская Федерация

Ключевые слова

Алгоритм,
искусственная социальность,
искусственный интеллект,
право,
регулирование,
социология,
цифровая экономика,
цифровые технологии,
человек,
человеко-ориентированный искусственный интеллект

Аннотация

Цель: определение проблем, которые ставит распространение технологий искусственного интеллекта перед юридической теорией и практикой, и соотнесение этих проблем с человеко-ориентированным подходом к искусственному интеллекту (Human-Centered AI).

Методы: исследование основано на критическом анализе релевантной литературы из разных дисциплинарных областей: юриспруденции, социологии, философии, компьютерных наук.

Результаты: в статье сформулированы основные перспективы и проблемы, которые ставит перед правовой системой развитие цифровых технологий в целом и алгоритмов искусственного интеллекта в частности. Выделенные проблемы соотнесены с положениями человеко-ориентированного подхода к искусственному интеллекту. Авторы утверждают необходимость того, чтобы разработчики программ искусственного интеллекта, владельцы компаний, участвующих в гонке по внедрению технологий искусственного интеллекта, сосредоточивались на том, чтобы люди и человек как базовая ценность общества находились в центре внимания. В частности, специальные усилия следует направить на сбор и анализ высококачественных данных для организации разработок инструментов искусственного интеллекта, поскольку сегодня алгоритмы искусственного интеллекта эффективны

✉ Контактное лицо

настолько, насколько эффективны данные, на которых они обучаются. Авторы формулируют три принципа человеко-ориентированного искусственного интеллекта для правовой сферы: 1) человек как необходимое звено в цепочке принятия и исполнения правовых решений; 2) необходимость регулирования искусственного интеллекта на уровне международного права; 3) формулировка «табу» для внедрения технологий искусственного интеллекта.

Научная новизна: статья представляет собой первую в русскоязычной научной литературе попытку обозначить перспективы развития области человеко-ориентированного искусственного интеллекта в юриспруденции. На основании анализа специальной литературы авторы формулируют три принципа включения искусственного интеллекта в юридическую теорию и практику с точки зрения человеко-ориентированного подхода к искусственному интеллекту.

Практическая значимость: принципы, сформулированные в статье, будут полезны как для правового регулирования технологий искусственного интеллекта, так и для гармоничного их включения в юридические практики.

Для цитирования

Резаев, А. В., Трегубова, Н. Д. (2023). Возможность и необходимость человеко-ориентированного искусственного интеллекта в юридической теории и практике. *Journal of Digital Technologies and Law*, *1*(2), 564–580. https://doi.org/10.21202/jdtl.2023.24

Список литературы

Андреев, В. К., Лаптев, В. А., *Чуча*, С. Ю. (2020). Искусственный интеллект в системе электронного правосудия при рассмотрении корпоративных споров. *Вестник Санкт-Петербургского университета. Право*, *1*, 19–34. https://doi.org/10.21638/spbu14.2020.102

Батурин, Ю. М., *Полубинская*, С. В. (2022). Искусственный интеллект: правовой статус или правовой режим? *Государство и право*, *10*, 141–154. https://doi.org/10.31857/s102694520022606-7

Быков, А. В., *Нарская*, А. И. (2022). Закон, мораль и машинное обучение: взгляд судей на сущность и перспективы роботизации правосудия. *Мониторинг общественного мнения: экономические и социальные перемены*, *5*, 278–298. https://doi.org/10.14515/monitoring.2022.5.2137

Вавилин, Е. В. (2021). Искусственный интеллект как участник гражданских отношений: трансформация права. *Вестник Томского государственного университета. Право*, *42*, 135–146. https://doi.org/10.17223/22253513/42/11

Вейценбаум, Дж. (1982). *Возможности вычислительных машин и человеческий разум. От суждений к вычислениям*. Москва: Радио и связь.

Винер, Н. (1983). *Кибернетика, или управление и связь в животном и машине* (2-е изд.). Москва: Наука; Главная редакция изданий для зарубежных стран.

Войниканис, Е. А. (2020). Регулирование больших данных и право интеллектуальной собственности: общие подходы, проблемы и перспективы развития. *Закон*, *7*, 135–156.

Горохова, С. С. (2021). Искусственный интеллект: инструмент обеспечения кибербезопасности финансовой сферы или киберугроза для банков. *Банковское право*, *1*, 35–46. https://doi.org/10.18572/1812-3945-2021-1-35-46

Грегер, Р. (2017). Судья как интернет-серфер. Выяснение обстоятельств дела в Интернете. *Вестник гражданского процесса*, *7*(4), 161–173. https://doi.org/10.24031/2226-0781-2017-7-4-161-173

Дрейфус, Х. (1978). *Чего не могут вычислительные машины: Критика искусственного разума*. Москва: Прогресс.

Ефимова, Л. Г., Михеева, И. В., Чуб, Д. В. (2020). Сравнительный анализ доктринальных концепций правового регулирования смарт-контрактов в России и в зарубежных странах. *Право. Журнал Высшей школы экономики*, *4*, 78–105. https://doi.org/10.17323/2072-8166.2020.4.78.105

Зубофф, Ш. (2022). *Эпоха надзорного капитализма. Битва за человеческое будущее на новых рубежах власти*. Москва: Изд-во Института Гайдара.

Лессиг, Л. (2019). Искусственный интеллект вытеснит широкий пласт юристов. *Закон*, 5, 8–30.

Ли, К.-Ф. (2019). *Сверхдержавы искусственного интеллекта. Китай, Кремниевая долина и новый мировой порядок*. Москва: Манн, Иванов и Фербер.

Мажорина, М. В. (2020). Киберпространство и методология международного частного права. *Право. Журнал Высшей школы экономики*, *2*, 230–253. https://doi.org/10.17323/2072-8166.2020.2.230.253

О'Нил, К. (2018). *Убийственные большие данные. Как математика превратилась в оружие массового поражения*. Москва: АСТ.

Панченко, В. Ю. (2012). Информационная доступность юридической помощи: идеальная модель и реальное состояние. *Аграрное и земельное право*, *11*(95), 95–102.

Рассел, С., Норвиг, П. (2007). *Искусственный интеллект: современный подход* (2-е изд.). Москва: Вильямс.

Резаев, А. В., Трегубова, Н. Д. (2019). «Искусственный интеллект», «онлайн-культура», «искусственная социальность»: определение понятий. *Мониторинг общественного мнения: Экономические и социальные перемены*, *6*, 35–47. https://doi.org/10.14515/monitoring.2019.6.03

Руденко, Н. И. (2020). Социотехнические барьеры разработки беспилотных автомобилей в России. В кн. Л. В. Земнухова и др. *Приключения технологий: барьеры цифровизации в России* (с. 17–70). Москва – Санкт-Петербург: ФНИСЦ РАН. https://doi.org/10.31119/978-5-89697-339-3

Русакова, Е. П. (2020). Интегрирование современных цифровых технологий в судопроизводство Китайской Народной Республики и Сингапура. *Государство и право*, *9*, 102–109. https://doi.org/10.31857/s102694520011323-6

Степанов, О. А., Печегин, Д. А., Дьяконова, М. О. (2021). К вопросу о цифровизации судебной деятельности. *Право. Журнал Высшей школы экономики*, *5*, 4–23. https://doi.org/10.17323/2072-8166.2021.5.4.23

Талапина, Э. В. (2021). Искусственный интеллект и правовые экспертизы в государственном управлении. *Вестник Санкт-Петербургского университета. Право*, *4*, 865–881. https://doi.org/10.21638/spbu14.2021.404

Талапина, Э. В. (2022). Право на информационное самоопределение: на грани публичного и частного. *Право. Журнал Высшей школы экономики*, *15*(5), 24–43.

Утехин, И. (2019). Поиск и его интерфейсы. *Laboratorium: журнал социальных исследований*, *11*(1), 152–165. https://doi.org/10.25285/2078-1938-2019-11-1-152-165

Цветков, *Ю. А*. (2021). Искусственный интеллект в правосудии. *Закон*, *4*, 91–107.

Esposito, E. (2017). Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift für Soziologie*, *46*(4), 249–265. https://doi.org/10.1515/zfsoz-2017-1014

Etzioni, A., & Etzioni, O. (2017). Should Artificial Intelligence Be Regulated? *Issues in Science and Technology*, *33*(4), 32–36.

Fink, M., & Finck, M. (2022). Reasoned A(I)administration: explanation requirements in EU law and the automation of public administration. *European Law Review*, *47*(3), 376–392.

Ford, K. M., Hayes, P. J., Glymour, C., & Allen, J. (2015). Cognitive Orthoses: Toward Human-Centered AI. *AI Magazine*, *36*(4), 5–8. https://doi.org/10.1609/aimag.v36i4.2629

Harvey, D. (2014). *Seventeen Contradictions and the End of Capitalism*. Oxford: Oxford University Press.

Hickman, E., & Petrin, M. (2021). Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective. *European Business Organization Law Review*, *22*, 593–625. https://doi.org/10.1007/s40804-021-00224-0

International Committee of the Red Cross (2020). Artificial intelligence and machine learning in armed conflict: A human-centred approach. *International Review of the Red Cross*, *102*(913), 463–479. https://doi.org/10.1007/s40804-021-00224-0

Klein, N. (2007). *The Shock Doctrine: The Rise of Disaster Capitalism*. New York: Henry Holt.

Lee, J.-A., Hilty, R. M, & Liu, K.-C. (Eds.). (2021). *Artificial Intelligence and Intellectual Property*. Oxford: Oxford University Press.

Oravec, J. A. (2022). Oravec, J. A. (2022). The emergence of "truth machines"? Artificial intelligence approaches to lie detection. E*thics and Information Technology*, *24*, 6. https://doi.org/10.1007/s10676-022-09621-6

Rezaev, A. V. (2021). Twelve Theses on Artificial Intelligence and Artificial Sociality. *Monitoring of Public Opinion: Economic and Social Changes*, *1*, 20–30. https://doi.org/10.14515/monitoring.2021.1.1894

Shneiderman, B. (2021). Human-centered AI. *Issues in Science and Technology*, *37*(2), 56–61.

Ulnicane, I. (2022). Artificial Intelligence in the European Union: policy, ethics and regulation. In T. Hoerber, I. Cabras, G. Weber (Eds.). *Routledge Handbook of European Integrations* (pp. 254–269). London: Routledge. https://doi.org/10.4324/9780429262081-19

Wolfe, A. (1993). *The Human Difference: Animals, Computers, and the Necessity of Social Science*. Berkeley: University of California Press.

Сведения об авторах

Резаев Андрей Владимирович – доктор философских наук, профессор, руководитель Международной исследовательской лаборатории ТАНДЕМ, Санкт-Петербургский государственный университет
Адрес: 191124, Российская Федерация, г. Санкт-Петербург, ул. Смольного, 1/3
E-mail: rezaev@hotmail.com
ORCID ID: https://orcid.org/0000-0002-3918-835X
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=13004674100
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/K-3472-2013
Google Scholar ID: https://scholar.google.ru/citations?user=Uzv39ccAAAAJ
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=648768

Трегубова Наталья Дамировна – кандидат социологических наук, доцент кафедры сравнительной социологии, Санкт-Петербургский государственный университет
Адрес: 191124, Российская Федерация, г. Санкт-Петербург, ул. Смольного, 1/3
E-mail: n.tregubova@spbu.ru
ORCID ID: https://orcid.org/0000-0003-3259-5566
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=56645016900
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/K-3487-2013
Google Scholar ID: https://scholar.google.com/citations?user=8dhGr3gAAAAJ&hl
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=832705

Вклад авторов

А. В. Резаев и Н. Д. Трегубова внесли равный вклад как в формулировку ключевых положений статьи, так и в подготовку рукописи к публикации.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Финансирование

Исследование выполнено при финансовой поддержке РФФИ и Министерства по науке и технологиям Тайваня в рамках научного проекта № 21-511-52002.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.07.49 / Планирование и прогнозирование в праве
Специальность ВАК: 5.1.1 / Теоретико-исторические правовые науки

История статьи

Дата поступления – 3 марта 2023 г.
Дата одобрения после рецензирования – 4 мая 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.