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.