Legal Issues in the Digital Age. 2025. Vol. 6, no. 2.
Вопросы права в цифровую эпоху. 2025. Том 6. № 2.

E-Government

Research article
JEL: K23, K 38
UDK: 342.5, 342.7
DOI:10.17323/2713-2749.2025.2.161.182

Transparency in Public Administration in the Digital Age: Legal, Institutional, and Technical Mechanisms

Pavel P. Kabytov[1], Nikita A. Nazarov[2]

[1, 2] Institute of Legislation and Comparative Law under the Government of the Russian Federation, 34 Bolshaya Cheryomushkinskaya Str., Moscow 117218, Russia,

[1] kapavel.v@yandex.ru, https://orcid.org/0000-0001-8656-5317

[2] naznikitaal@gmail.com, https://orcid.org/0000-0002-3997-0886

Abstract

The article contains a comprehensive analysis of the very relevant topic of ensuring transparency and explainability of public administration bodies in the context of an ever-increasing introduction of automated decision-making systems and artificial intelligence systems in their operations. Authors focus on legal, organisational and technical mechanisms designed to implement the principles of transparency and explainability, as well as on challenges to their operation. The purpose is to describe the existing and proposed approaches in a comprehensive and systematic manner, identify the key risks caused by the non-transparency of automated decision-making systems, and to evaluate critically the potential that various tools can have to minimise such risks. The methodological basis of the study is general scientific methods (analysis, synthesis, system approach), and private-scientific methods of legal science, including legalistic and comparative legal analysis. The work explores the conceptual foundations of the principle of transparency of public administration in the conditions of technology transformation. In particular, the issue of the "black box" that undermines trust in state institutions and creates obstacles to juridical protection, is explored. It analyses preventive (*ex ante*) legal mechanisms, such as mandatory disclosure of the use of automated decision-making systems, the order and

© Kabytov P.P., Nazarov. N.A., 2025
This work is licensed under a Creative Commons Attribution-Sharealike 4.0 International License
https://creativecommons.org/licenses/by-sa/4.0/legalcode.en

logic of their operation, information on the data used, and the introduction of pre-audit, certification and human rights impact assessment procedures. Legal mechanisms for *ex post* follow-up are reviewed, including the evolving concept of the "right to explanation" of a particular decision, the use of counterfactual explanations, and ensuring that users have access to the data that gave rise to a particular automated decision. The authors pay particular attention to the inextricable link between legal requirements, and institutional and technical solutions. The main conclusions are that none of the mechanisms under review are universally applicable. The necessary effect may only be reached through their comprehensive application, adaptation to the specific context and level of risk, and close integration of legal norms with technical standards and practical tools. The study highlights the need to further improve laws aimed at detailing the responsibilities of developers and operators of the automated decision-making system, and to foster a culture of transparency and responsibility to maintain public administration accountability in the interests of society and every citizen.

Keywords

transparency; explainability; automated decision-making; artificial intelligence; legal regulation; *ex ante* mechanisms; *ex post* mechanisms; right to explanation; black box; protection of citizens' rights.

Acknowledgements: The research was carried out with the Russian Science Foundation grant No. 23-78-01254, https://rscf.ru/project/23-78-01254/.

For citation: Kabytov P.P., Nazarov N.A. (2025) Transparency in Public Administration in the Digital Age: Legal, Institutional and Mechanisms. *Legal Issues in the Digital Age,* vol. 6, no. 2, pp. 161–182. DOI:10.17323/ 2713-2749.2025.2.161.182

Introduction

Introduction of automated decision-making systems and artificial intelligence (AI) systems into the operations of public administration bodies marks a new era in the development of public administration, which can be loosely described as the "automation of public administration." Its main purpose is to increase efficiency, optimise resources and enhance the quality of government services that may be provided automatically, i.e. without direct human involvement. In this case, citizens interact directly with the technology envelope of public administration. Hence, this creates a range of challenges, and maintaining transparency and explainability of the decisions taken holds a special place among them.

Historically, the principle of transparency (openness) of public authority activities evolved as a fundamental guarantee that the authorities would be accountable to society, citizens' rights would be protected, and

the basis for trust between the state and its citizens would be laid. As decisions affecting the rights and legitimate interests of individuals are increasingly made or drafted without the direct participation of a human person (public servant), the so-called "black box" problem arises that consists in the opacity of the internal decision-making logic and the prerequisites for making a certain final decision.

Thus, the lack of understanding how and on what grounds the automated decision-making system has come to a particular conclusion undermines trust in state institutions, creates obstacles to juridical protection and is capable to lead to systemic violations of legal guarantees, and human and civil rights. Therefore, our article aims to provide a comprehensive analysis of and offer a system for existing and proposed legal mechanisms aimed at ensuring transparency and explainability of automated decision-making systems and AI systems in public administration. It explores the conceptual foundations of the transparency principle in the context of new technology realities, identifies the key risks associated with the opacity of algorithm systems, and critically assesses the potential and limitations of various legal instruments (both preventive ones, *ex ante*, and subsequent control ones, *ex post*) in addressing the issue under review. A special emphasis is placed on the need to integrate legal, organisational and technical approaches in order to establish an effective system of safeguards.

1. Conceptual Foundations of and Challenges to Opacity in the Context of Automation of Public Administration

1.1. Automated Decision-Making and Artificial Intelligence in the Public Sphere: Essence and Key Parameters

In the past years, public administration has been actively exploring the potential of automated decision-making, i.e. the procedure of making decisions where information technologies are used either to facilitate the formation of judgements by decision-makers, or to replace them, partially or completely. In this situation, it should not be of critical importance which particular technology (whether a simple rules-based system or a neural network) has influenced the outcome. Undoubtedly, the specificity of technology must be taken into account in creating a regulatory requirements framework for the development, implementation and operation of such systems. At the same time, the very fact that

the process of making a decision that affects the rights and freedoms of a person is automated plays the decisive role in determining the item subject to regulation. The existing law-enforcement practice confirms this: even automated decision-making systems that use software code and that, according to some classifications, do not belong to AI systems (e.g., self-learning systems) in a strict sense can influence the lives of citizens and the activities of organisations in very serious and sometimes critical ways[1]. In view of the above, one should positively assess the approaches of such systems of justice where the "automated decision-making process" as such is the special subject of regulatory influence, regardless of the complexity of the underlying system. It enables a broader and more technology-neutral legal regulation and thus covers the risks associated with automation[2].

Automated decision-making systems can be classified on various grounds:

by their application sphere: law enforcement, legislative, judicial activities;

by the level of their automation: partially automated (a human operator supports the decision-making process), delegated (the system initiates and makes the decision but hands over to a human operator in case of a problem), and fully automated decision-making;

by their legal significance: decisions that have direct legal consequences; intra-organisational decisions; decisions that have other significant effects.

by the technologies used: systems based on rigidly defined rules, systems based on statistical methods, AI-based systems (machine learning, deep learning, etc.).

---

[1] See: Automating Society Report 2020. Available at: URL: https://automatingsociety.algorithmwatch.org (accessed: 07.05.2025); Automating Society 2019. Available at: URL: https://algorithmwatch.org/en/automating-society-2019/ (accessed: 07.05.2025)

[2] See: Directive on Automated Decision-Making. 2019. URL: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 (accessed: 11.12.2023); Gesetz über die Möglichkeit des Einsatzes von datengetriebenen Informationstechnologien bei öffentlich-rechtlicher Verwaltungstätigkeit (IT-Einsatz-Gesetz ITEG) Vom 16. März 2022. Available at:мURL: https://www.gesetze-rechtsprechung.sh.juris.de/bssh/document/jlr-ITEGSHpP1 (accessed: 10.12.2023); Article 28(1) Förvaltningslag (2017:900); Articles 41 и 42 Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público; Article 35a Verwaltungsverfahrensgesetz (VwVfG); Articles L311-3-P311-3-1-2 Code des relations entre le public et l'administration

At the same time, by introducing automated decision-making, the state seeks to improve the efficiency of public administration, minimise "human factor" errors, cut costs, and reduce corruption risks. However, these advantages come with major challenges including threats to human rights, difficulty to ensure human control, the problems of diffusing responsibility and, of particular importance for our study, the fundamental issue of making such systems transparent and explainable.

1.2. The Principle of Transparency in Public Administration: Theoretical and Legal Dimension

The principle of transparency (openness) of the activities of public administration is the cornerstone of a modern state governed by the rule of law. Historically, the idea of the openness of power has come a long way from the first legislative acts (for example, the Swedish Law 'On Freedom of the Press' of 1776) to its global recognition and enshrinement in international documents and national legal systems, including the Russian Federation Constitution (Part 2, Article 24).

However, to characterise the phenomenon in question, modern Russian legal doctrine and legislation use terms that are different, although close in meaning: 'transparency', 'openness', 'transparency', 'glasnost', 'publicity', 'publicity' [Silkin V.V., 2021: 20–31]. Such diversity, as noted in the literature, "results in a certain conventionality in the use of this or that term, the blurring of the concepts in question" [Pogodina I.V., 2023: 29–31]. This may make it difficult to develop a unified approach to their enshrinement in law and to their enforcement in the specific context of automated decision-making systems.

Despite the nuances in terminology, the essence of the principle lies in a mode of functioning of public authorities, which ensures that information on their activities is accessible to the society, creates conditions for public control and participation of citizens in the management of state affairs, and promotes the development of mutual trust between the state and society.

The transparency principle in Russian law includes the following key elements:

information openness: the obligation of state and local self-government bodies to actively publish information about their activities (e.g., on official websites, and in the media) and provide this information upon requests from citizens and organisations. Federal Law No. 8-FZ

"On Access to Information on the Activities of State Bodies and Local Self-Government Bodies" of 09.02.2009 describes in detail the possible ways of ensuring access to information. These include its publication in the mass media (Art. 12), placement on the Internet (Art. 13, 14), placement in the premises occupied by the authorities (Art. 16), provision of information upon request (Art. 18), and others. Federal Law No. 149-FZ of 27.07.2006 "On Information, Information Technologies and Information Protection" also enshrines the openness of information on the activities of government bodies and free access to such information as one of the principles of legal regulation (Art. 3);

comprehensibility and accessibility of information: information should be provided in a form that ensures that it can be perceived and understood by a wide range of people, and not specialists only. As the Concept of Openness of Federal Executive Bodies (approved by the order of the Government of the Russian Federation of 30.01.2014 No. 93-r) notes, the "comprehensibility" of information is important;

civil society involvement and public control: transparency creates prerequisites for a constructive dialogue between the authorities and society, for citizen participation in the process of developing and making decisions. Federal Law No. 212-FZ of 21.07. 2014 " Fundamentals of Public Control in the Russian Federation" explicitly states that one of the tasks of public control is "to increase the level of trust of citizens in the activities of the state, as well as to ensure close cooperation between the state and civil society institutions" (part 2, Article 2). For example, the Rules for Disclosure by Federal Executive Authorities of Information on the Preparation of Draft Regulatory Legal Acts and the Results of their Discussion (approved by Resolution of the Russian Federation Government No. 851 of 25.08.2012) are aimed at implementing the principle of transparency. These Rules provide for compulsory posting of draft regulatory legal acts on the portal <regulation.gov.ru>.

accountability and responsibility of the authorities: the openness of the activities of the authorities allows the public to assess their effectiveness, identify violations, and hold officials accountable for their actions. As academician O.E. Kutafin rightly emphasised, in the modern period "state power responsible to the people and the law" is one of the main criteria for the establishment of constitutionalism [Kutafin O.E., 2008: 18].

developing and maintaining trust between the state and society: as enshrined in Article 75.1 of the Russian Constitution, "conditions shall be created in the Russian Federation for sustainable economic growth of

the country and improvement of the welfare of citizens, for mutual trust between the state and society." Trust, in turn, serves as the basis of social institutions, "uniting people, guaranteeing them security, the success of collective endeavours and allowing them to direct their combined energies for the common good" [Narutto S.V., Nikitina A.V., 2022: 13–18].

Thus, transparency is not just a desirable attribute, but a fundamental legal principle of public authorities' activity in a modern state governed by the rule of law. It has deep roots and has been enshrined in international acts and national laws including the Russian Constitution. The contents of this principle is quite diverse: it includes information openness, clarity and accessibility; society involvement; accountability and responsibility of authorities; and society's confidence in the government.

Implementation of the transparency principle helps enhance mutual trust between the state and society, improve public administration efficiency, prevent corruption, and protect citizens' rights. Still, to achieve real transparency it would be necessary not only to pass laws and regulations, but also to develop the corresponding culture in government bodies, and for civil society to take a pro-active stance. It is important to balance openness with the need to protect legal interests.

In this context the article offers a comprehensive analysis of the very relevant topic of ensuring public administration bodies' transparency and explainability in the context of an ever-increasing implementation of automated decision-making systems and artificial intelligence systems in their operations.

1.3. From Legislative towards Scientific Understanding of the Prerequisites for Maintaining the Transparency of Automated and Artificial Intelligence for Public Administration

Scholars in the sphere of legal science emphasize that in addition to enshrining transparency as a basic principle of public administration there are other prerequisites to enshrine the requirement of transparency and explainability of automated decision-making systems:

Trust is a significant aspect of automated decision-making, and explainability and transparency are necessary to increase and fortify this trust [Fine Licht de K., Fine Licht de J., 2020: 917]. Algorithm explainability is more important than algorithm transparency both for the ordi-

nary citizen and for the person making decisions [Grimmelikhuijsen S., 2023: 242] because explainability allows to reveal the cause-and-effect relationship between the input data, the logic of the system operation, and the automated decision made, thus contributing to understanding its validity.

In addition, sociological surveys compare citizens' trust in the case of decision-making with or without human involvement. E.g., one of them noted that when an AI system solved a "technical" job scheduling task, there was no difference in ranking, but for tasks requiring "human judgement," namely making a hiring decision, algorithms were perceived as less trustworthy [Lee M.K., 2018: 1-16]. Another study shows that citizens have less trust in automated decisions that "lack transparency." However, there is no transparency in the decision-making process even for the decision makers themselves [Schiff D.S., Schiff K.J., Pierson P., 2022: 653–573].

Explanation and transparency contribute to the creation of a safer and more reliable product, and enable collecting evidence for accountability [Sokol K., Flach P.A., 2019: 1–4]. This is especially important in the field of diagnosis and treatment, because in the absence of such requirements, the fundamental principles of medical ethics are jeopardised, which may negatively affect the safety of the individual and society.

Transparency encourages the human user to participate in the decision-making process, and explanations allow to correct and find technical errors in the automated decision-making system [Srinivasu P.N. et al., 2022: 1–20].

Explainability and transparency are necessary conditions of accountability for both the decision-maker and the operator of the automated decision-making system. Transparency is an informational aspect of accountability and as such is a prerequisite for accountability. And the individual right to information or clarification is only one of the elements of a broader structure of regulation and supervision [Wischmeyer T., Rademacher T., 2020: 75–101].

Lack of algorithm transparency can hide discrimination, create room for manipulation, or make people blindly trust algorithm-based decision-making [Drunen M. Z., Helberger N., Bastian M., 2019: 220–235]. Price discrimination can be identified in addition to gender discrimination, which creates inequality among different segments of the population [Veale M., Edwards L., 2018: 401–402].

Transparency allows to remove information asymmetry between all actors. As a result of the use of automated systems, an information asymmetry may develop, first of all between a state agency (the system operator) and a citizen (the subject of the decision), where the advantage of one person arises precisely owing to information about the other person (including information against the other person). Information asymmetry can be used both to the advantage and to the disadvantage.

Explainability and transparency ensure the decision-making procedure is legitimate [Fine Licht de K., Fine Licht de J., 2020: 918–926]. Moreover, an automated decision, made in a way that is explainable and procedurally fair, helps to ensure that the decision is legitimate and that the decision-making body has credibility among citizens.

Explanation and transparency may be helpful to the applicant by helping to understand which inputs had the strongest influence on the decision made [Verma S., Boonsanong V. et al., 2022: 2]. In addition, these requirements allow an applicant to challenge a decision, for example, if their race was critical in determining the outcome. This may also be useful for organisations when testing their algorithms for systematic biases.

In some cases, explanation and transparency provide the applicant with feedback on the basis of which they can take action to get the desired outcome in the future.

Explanation helps to adhere to laws related to machine decisions, including Regulation No 2016/679 of the European Parliament and of the Council of the European Union "On the protection of natural persons with regard to the processing of personal data and on the free circulation of such data and repealing Directive 95/46/EC (General Data Protection Regulation)" (hereinafter GDPR).

At the same time, there are also opposing views arguing the requirements of explainability and transparency are unnecessary, especially in the context of public administration. The arguments proposed are that the pace of technology development, multiple transparency concepts, uncertainty about where transparency is required, how best to approach communication with different stakeholders, and how to build transparency measures into meaningful and organisationally realistic accountability measures all pose challenges to implementing these requirements, despite seemingly general agreement this is important [Felzmann H., Fosch-Villaronga E., Lutz C. et al., 2020: 3355]. These chal-

lenges may also include the risks of disclosure of algorithm developers' trade secrets, the possibility of system manipulation by knowledgeable actors ("gaming" the system), and the significant costs of developing and implementing truly effective explainability mechanisms for complex AI systems. Furthermore, there is a concern that excessive transparency requirements may slow down the adoption of innovative technologies in public administration.

2. Legal Mechanisms to Maintain Transparency and Explainability of Automated Decision-making and AI Systems

2.1. Mechanism Categories: *ex ante* Approach and *ex post* Approach

Contemporary and proposed mechanisms aimed at maintaining transparency and explainability of automated decision-making systems and AI systems in public administration may be categorised on various grounds. Legal doctrine and related fields of knowledge offer various grounds for categorising such mechanisms[3].

Categorisation by the goal of transparency and explainability. Under this approach, items that fall under the requirement of transparency and explainability, can be grouped by the two main aspects:

Transparency and explainability of the decision-making process (algorithm) implies disclosure of information about the system itself, its architecture, logic of functioning and data used (e.g., what factors the system takes into account and how when making decisions);

Transparency and explainability of the outcome (a particular decision): focuses on providing information that justifies a specific decision made by the system in relation to a particular actor or situation (e.g., why a particular decision was made in this case and what data of the actor influenced it).

Categorisation by the timing of the explanation and the nature of the transparency. This approach differentiates mechanisms depending on

[3] See: Explaining decisions made with AI. 2022. Available at: URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/ (accessed: 07 April 2025)

the stage in the life cycle of automated decision making and AI systems in which they are implemented, and subdivides transparency into:

ex ante mechanisms. These mechanisms are implemented before automated decisions are made, and independently of a particular decision. Their purpose is to prevent risks, ensure the predictability of the system's operation, and inform the public and stakeholders about the principles of its operation and potential consequences;

ex post mechanisms. These mechanisms are applied after the automated decision has been made, especially if it affects the rights and legitimate interests of the subjects. Their purpose is to ensure accountability, enable effective appeal, correct errors and analyse the performance of the system for future improvement.

Categorisation by the levels and types of transparency. The following interrelated levels and types of transparency can be identified depending on the item of information disclosure:

data transparency: disclosure of information about the data used. This aspect is critical because the quality and characteristics of the data directly affect the functioning and performance of the system and the AI.

algorithm transparency: disclosure of information about the algorithm itself. The purpose is to maintain understanding of how the system processes information and arrives at conclusions. In some cases, this may involve disclosure of the source code or model of the AI, although this carries risks to intellectual property, various secrets, and information security (e.g., identifying system vulnerabilities);

results transparency: the ability of a system or its associated mechanisms to explain in ways understandable to a human why a particular decision was made and how certain inputs led to a particular conclusion.

These approaches to classification emphasise the multidimensionality of the concepts of "explainability" and "transparency" in relation to the automated decision-making system. At the same time, different types of transparency and explainability mechanisms are not mutually exclusive. On the contrary, they should complement each other, forming a comprehensive system at all stages of the system's life cycle in public administration.

Our analysis of foreign academic literature, laws and law enforcement practices allows us to identify a number of basic legal, institutional and technical mechanisms aimed at ensuring the transparency and explain-

ability of the system. We believe in the beginning it would be expedient to group them according to one of the key classifications, namely the timing of the explanation (*ex ante* and *ex post*):

2.2. Mechanisms of Preventive Control (*ex ante*)

Ex ante mechanisms create conditions for inherent predictability, controllability and legitimacy of automated decision making.

Disclosure of the use of an automated decision-making system. Obligation to inform actors that a decision has been made using the above system. This is a basic requirement related to the right to information and necessary for the realisation of other rights (request for information; call for human intervention: right to appeal a decision made using an automated decision-making system);

Disclosure of the order or logic of decision-making (under personal data laws). Personal data law (e.g., the general requirements for informing the person contained in Federal Law No. 152-FZ "On Personal Data" of 27.07.2006), contains rules requiring operators to explain how automated decisions are made or to provide "meaningful information about the logic involved", although the level of detail isn't as significant as in Articles 13-15 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27.04.2016 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) in the EU. However, this mechanism is limited in scope only to decisions based solely on automated processing of personal data with legal consequences. Other limitations include aspects of protection of IP and trade secrets, and the difficulty of explaining the logic of complicated models to non-specialists. Moreover, even if such a right does exist, its implementation may be hampered by the lack of clear criteria for the "meaningfulness" of information about the logic and about the limits on the disclosure of such information so as not to infringe the rights of developers. Another open question is the efficiency of such disclosure for complex self-learning AI systems because their logic is not always deterministic and is able to evolve over time;

Disclosure of information about the data used for the development and operation of the automated decision-making system. Provision of information about the sources, types, and characteristics of data on which the system has been trained and operates. This allows the poten-

tial impact of the system to be assessed, the impact of the system to be investigated, and biases to be identified. This also includes disclosure of data in the form of open data sets (with due observation of confidentiality), which facilitates public scrutiny and encourages innovation;

Disclosure of the programme code and (or) AI model. Providing access to the source code or detailed description of the model. This mechanism allows for the most in-depth public scrutiny. On the other hand, it faces serious constraints related to the protection of intellectual property and trade secrets. International practice offers various examples in this respect.

Pre-audit, certification, and impact assessment. Independent checks of automated decision-making systems prior to implementation for their compliance with the law, ethical standards, and to identify risks. These may range from government oversight mechanisms to voluntary certification or internal audit systems. Such internal audit may assess the suitability of the system for its stated purposes, the quality and representativeness of the data used for training, the existence of discrimination prevention mechanisms, the reliability and security of the system, and the adequacy of measures to ensure transparency and explainability. Another promising field is developing standardised methodologies for conducting such assessments, including criteria for assessing data and algorithm biases, and accrediting independent auditors with relevant competencies in both legal and technical areas.

2.3. Legal Mechanisms of Subsequent Control (*ex post*)

The aim of this mechanisms is to maintain basis of a decision already taken is understood and may be challenged.

"Right to an explanation" of an individual decision. An evolving concept involving the legislated ability of a person affected by an automated decision to receive comprehensible explanations of the system's role in a particular decision and its underlying determinants. An example of enshrining such a right is Article 86 of Regulation 2024/1689 of the European Parliament and of the Council of 13.06.2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). This right implies not only stating that an automated decision-making system was

used, but providing the user with personalise information about the factors that influenced a particular decision and, if possible, the logic that guided the system. However, implementation of this right directly depends on the technical ability to provide such an explanation in a form understandable to the human user, in particular in the case of complex AI systems;

Counterfactual explanations. Providing information about what changes in the inputs or conditions could have led to a different (e.g., desired) outcome. This approach helps in understanding the logic of the system and its sensitivity to various factors, and offers practical pieces of advice to the user. Counterfacts answer the question "what if" and can reveal hidden biases. On the other hand, in implementing this approach, one is faced with the multiple possible explanations problem ("the Rashomon effect") and the difficulty to take into account all the relevant factors. In addition, generating counterfactual explanations can be computationally expensive and resource intensive. It is also important to bear in mind that while such explanations can be useful for understanding the system's sensitivity to changes in the inputs, they sometimes fail to show the real cause for the decision made; instead of it they show how a different outcome could have been achieved. That said, they have a significant potential in enhancing the understanding and extending the field of the user's opportunities to interact with the system;

Disclosure of the data that served as the basis for a particular automated decision. Ensuring that user has access to specific data that the automated decision-making systems used to make a decision about him or her. This allows to check the data for correctness and completeness, identify irrelevant or discriminatory factors, exercise the right to correct the data; furthermore, it is the basis for a reasoned challenge to the decision.

3. The Role of Organizational and Technical Solutions in the Legal Support of Transparency

On the other hand, the purpose of legal mechanisms largely depends on the existence of adequate organisational and technical tools for implementing them. The rules of law that enshrine principles and duties require adequate technical tools for putting them into practice. Without proper technology solutions, many legal requirements, such as the right to explanation or the obligation to disclose the system's logic, may

remain declarations. That enhances the role of technological methods that can either make the systems inherently more understandable or provide tools for *ex post factum* analysis of how they work.

3.1. *Ex ante* Organisational and Technical Approaches: Interpretable Models and "Transparency by Design"

The key strategy is to create and use systems designed with interpretability or explainability features. That includes:

artificial intelligence models that are interpretable and explainable "by default." Initially interpretable or explainable models thus directly promote the implementation of *ex ante* legal mechanisms. For instance, the use of such models facilitates due diligence audit and certification, because their logic is more open to analysis. Besides, it facilitates disclosure of information about their decision-making procedure or logic and about the data used to develop the system. Legislative codification of requirements or recommendations to use such models, especially for automated high-risk decision-making in public administration, could be an important step towards building transparent automated decision-making systems. That may be implemented via standards, guidelines for developers and state clients, and also via assessment criteria used in the procurement of artificial intelligence systems for public needs;

forming publicly accessible registers of the automatic decision-making systems used in public administration, indicating their purpose, applications (specific state functions or services), type of the data used (including the availability and sources of personal data), degree of automation (decision-making support system or fully or partially automated decision), developer and operator information, information on conformity assessment or audit passed (where applicable), and contact information for requesting explanations or appealing against decisions. Such registers should be easily accessible to citizens and be updated regularly. Keeping them could be entrusted to a special authority or integrated into existing State service and open data portals;

delegation of specific powers to an existing or newly established public authority to supervise automated decision-making systems in the public sector. Such powers might include: keeping the above-mentioned register, development of transparency and explainability guidelines and standards, holding scheduled and extraordinary checks (audits), issuing orders to correct any irregularities, and initiating studies to assess the

risks and the automated decision-making systems' impact on individuals' rights. Given the specifics of operating within the public administration system, it is important to make sure that such an authority is independent, impartial, and possesses the required expertise and technical resources. A potential mechanism that could strengthen confidence in the findings is the adoption of procedures that keep the audit findings unchanged and truthful using e.g. distributed registry technology or other cryptographic methods to record the findings, and in some cases expressly defined by law, for records on formal aspects of the audit, notarisation;

Adherence to Privacy/Transparency by Design approaches. As noted by L. Edwards and M. Veale, the newly passed GDPR introduces a number of new provisions that attempt to create an environment in which less "toxic" automated systems will be built in future. These ideas come out of the long evolution of Privacy by Design engineering as a way to build privacy-aware or privacy-friendly systems, generally in a voluntary rather than mandated way. [Edwards L., Veale M., 2018: 46–54]. While, historically, Privacy by Design focused on privacy, its principles (proactivity, integration in design, and focus on the user) are also applicable to the pursuit of transparency and explainability in a broader sense as they lay a basis for Transparency by Design.

Besides, the above concept should extend into a principle of heightened requirements to models for high-stakes decisions. Thus, one study states that the legislator should call for greater efforts to ensure the safety of, and confidence in, machine learning models that support high-stake and highly significant decisions [Rudin C., 2019: 206–215]. This principle leads developers and customers to choose or create more reliable and, potentially, more transparent models at the *ex ante* stage for critical automated decision-making systems in public governance.

3.2. *Ex post* Organisational and Technical Approaches: Explainable Artificial Intelligence Tools for Decision Analysis

The analysis of decisions already taken by systems (especially, by "black boxes") employs methods of an explainable AI system (Explainable AI, XAI):

Explainable artificial intelligence (XAI) methods. A set of techniques that help generate explanations for individual decisions that suit the specific case and the user's level of understanding (e.g., LIME — Local

Interpretable Model-agnostic Explanations; SHAP — SHapley Additive exPlanations);

Interactive visualisation and What If analysis tools. These enable both users and experts to examine the model's behaviour and understand how different input data will affect the result, which is closely related to counterfactual explanations;

Intelligent decision assistance. Automated decision-making was shown to have many benefits for both business and society, but that comes at a cost. It has long been known highly automated decision-making may have various drawbacks such as biased decisions and loss of professional skills by employees. Authors have analysed those two disadvantages to develop a new decision support system, namely Intelligent Decision Assistance [Schemmer M., Kühl N. et al. 2021: 1–10]. That system complements the human decision-making process with explainable AI, while offering no concrete recommendations. Such an approach may be used *ex post*, so that the human reviewer can understand AI contribution to the decision taken and assess it for relevance, which is important for human supervision and challenge mechanisms;

Establishing a procedure for challenging automated decisions. Development of an administrative and judicial procedure for appealing against decisions that were taken using automated decision-making systems, including the definition of the standard of proof and burden of proof distribution. Human control must remain in place and permit revision of an automated decision. Thus, whatever the automation level may be, there should remain an opportunity to appeal to a human and have the decision revised. The procedure should also take into account the specifics of automated decision-making systems, e.g. permit requesting the system's technical logs (subject to any limitations on access to legally protected secrets) and engaging artificial intelligence experts to analyse whether the system is functioning correctly.

Those tools form an institutional and technical basis for exercising the right to explanation and can be used for system audit by individuals and supervisory authorities.

3.3. Integration of Organisational and Technical Solutions into Legal Regulation: the Need and Prospects

Automated decision-making and artificial intelligence systems cannot efficiently be made transparent and explainable unless the legal rules

are closely integrated with the development, implementation and use of the relevant technical standards, tools, and methods. That is because legal regulation should not just proclaim duties and principles, but also create efficient mechanisms for putting them into practice by stimulating technological development and channelling it towards the observance of human rights and good governance.

Firstly, assessment of how the above legal mechanisms are codified in the light of the state policy is one of the key modalities. The authors of the current paper believe that such mechanisms should accompany every stage of an automatic decision-making system's lifecycle. As an additional reference point, we can consider developing criteria and clear recommendations for the developers of those systems that could help create reliable systems with an emphasis on the protection of the state's core values and the rights of individuals.

That may be achieved particularly by establishing:

minimum requirements on the interpretability of artificial intelligence models depending on the degree of risk and the significance of the decisions taken (e.g. mandatory use of verifiable and explainable models for high-risk systems);

formats and protocols for giving explanations that make them understandable to various categories of users (laypersons, officials and/or experts);

standards and requirements on data quality that guarantee the reliability of that fundamental element of artificial intelligence by providing accurate, up-to-date, representative and complete data that will underlie an automated decision;

requirements on logging the automated decision-making system's actions, which is critical for conducting audit, investigating incidents and providing evidence in a decision challenge process. The said logs must contain information about the input data, key information processing stages, and the resultant decision indicating time.

Secondly, law should encourage and regulate the use of specific technical tools and techniques that enhance transparency. This includes:

supporting the development and implementation of explainable artificial intelligence (XAI) tools such as LIME, SHAP or analogues, adapted for use in state information systems. The state could either commission such developments or facilitate their advent into the market;

creation and support of platforms for testing and verifying an automated decision-making system for compliance with transparency and non-discrimination requirements and with other ethical and legal rules. Such sandboxes could be used by developers as well as supervisory authorities;

development of methods for assessing the automatic decision-making systems' effect on human rights (Human Rights Impact Assessment), to include technical aspects of system analysis and the assessment of potential social consequences of their adoption.

Thirdly, legal conditions should be created to support efficient use of technically generated explanations and data in legal procedures. Law should establish requirements on the quality, completeness and understandability of technically generated explanations so that individuals can use them to protect their rights, and courts and administrative authorities can use them to assess decisions for lawfulness. The legal status and evidential force of information obtained from an automated decision-making system (such as logs and explanations) should be defined. This should include development and codification of procedures for requesting, receiving and challenging such explanations that will guarantee prompt provision of understandable information and easy access to the procedure itself.

Fourthly, it is important to develop interdisciplinary co-operation. The transparency of an automated decision-making system can only be successfully and efficiently achieved through deep integration of legal, organisational and technical solutions. Close co-operation and interaction among lawyers, AI developers, researchers, ethicists, and members of civil society is thus required. Here we should assume the very implementation of automated public administration is impossible without a better "digital literacy" and understanding of the work of AI by two groups: on the one hand, by public officials, judges and other law enforcers. On the other hand, by citizens who are both recipients of such decisions and the principal actors, and are thus expected to know and understand their own rights and duties, including procedure for challenging an automated decision.

Consequently, amid rapid evolution of the governance paradigm, any well-developed legal order aiming to protect human rights and interests as the supreme value should include, as justified and necessary actions, active studies of the world's best practices (including the approaches embedded in the EU AI Act) and encouraging domestic research and

practical developments in the field of explainable and trusted artificial intelligence. Such an interdisciplinary and international approach will support the adoption of advanced technology subject to the basic principles of a state governed by the rule of law, where transparency is central to the government's accountability to society.

Conclusion

One of the key objectives of current law and order is to maintain the transparency and explainability of automated decision-making and use of artificial intelligence systems in public administration.

Analysis has shown that, despite active development of law and doctrine in the field, the existing legal mechanisms — both preventive (*ex ante*) ones and those providing for posterior (*ex post*) control — are fraught with certain limitations and cannot always and fully protect citizens' rights and keep the authorities accountable amid algorithm-based governance.

None of the mechanisms discussed is a universal solution; efficiency can only be achieved through their comprehensive application and adaptation to the specific context around the use of automatic decision-making and artificial intelligence systems, with due regard to the risk level associated with the decisions in question, their social significance and the technical complexity of the systems being used. Most importantly, legal requirements must be closely integrated with the development and implementation of relevant organisational and technical solutions that can ensure real, not declarative, transparency and explainability.

Development of legislation and jurisprudence should aim to specify the obligations of developers and operators of automated decision-making systems, establish clear-cut criteria for assessing the adequacy of the explanations returned, and to strike the optimal balance between the needs for openness, protection of intellectual property and trade secrets, and information security.

A deep-rooted and conscious culture of transparency and responsibility should become an important feature, both in public authorities and among developers and operators of artificial intelligence systems. Only in this way can we ensure that the adoption of advanced information technology really fosters safer and more equitable, efficient and accountable governance that meets both society's and every citizen's interests.

References

1. De Fine Licht K., De Fine Licht J. (2020) Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations are Key When Trying to Produce Perceived Legitimacy. *AI & Society*, no. 35, pp. 917–926. doi: https://doi. org/10.1007/s00146-020-00960-w

2. Drunen M.Z., Helberger N., Bastian M. (2019) Know Your Algorithm: What Media Organizations Need to Explain to their Users about News Personalization. *International Data Privacy Law, v*ol. 9, no. 4, pp. 220–235. doi: https://doi.org/10.1093/idpl/ipz011.

3. Edwards L., Veale M. (2018) Enslaving the Algorithm: From a Right to an Explanation to a Right to Better Decisions? *IEEE Security & Privacy*. no 3, pp. 46–54. doi: https://doi.org/10.1109/MSP.2018.2701152.

4. Felzmann H., Fosch-Villaronga E., Lutz C. et al. (2020) Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics,* no. 6, pp. 3333–3361. doi: https://doi.org/10.1007/s11948-020-00276-4.

5. Grimmelikhuijsen S. (2023) Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making. *Public Administration Review,* no. 2, pp. 241–262. doi: https://doi.org/10.1111/puar.13483.

6. Kutafin O.E. (2008) The Russian Constitutionalism. Textbook. Moscow*:* Norma, 544 p. (in Russ.)

7. Lee M.K. (2018) Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society,* vol. 5, no. 1, pp. 1–16. doi: https://doi.org/10.1177/2053951718756684.

8. Narutto S.V., Nikitina A.V. (2022) Constitutional Principle of Trust in Modern Russian Society. *Konstitucionnoe i municipalnoe pravo*= Constitutional and Municipal Law, no. 7, pp. 13–18 (in Russ.)

9. Pogodina I.V. (2023) Forming Culture of Transparency with Help of ICTs. *Gosudarstvennaya vlast i mestnoe samoupravlenie*=State Power and Local Self-Government, no. 11, pp. 29–31(in Russ.)

10. Rudin C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence,* no. 5, pp. 206–215.

11. Schemmer M., Kühl N. et al. (2021) Intelligent Decision Assistance versus Automated Decision-Making: Enhancing Knowledge Work through Explainable Artificial Intelligence, pp. 1–10. doi: https://doi.org/10.48550/ARXIV.2109.13827.

12. Schiff D.S., Schiff K.J., Pierson P. (2022) Assessing Public Value Failure in Government Adoption of Artificial Intelligence. *Public Administration,* vol. 100, no. 3, pp. 653–673. doi: https://doi.org/10.1111/padm.12742.

13. Silkin V.V. (2021) Transparency of the Executive Power in the Digital Age. *Rossijskij yuridicheskij zhurnal*=Russian Law Journal, no. 4, pp. 20–31 (in Russ.)

14. Sokol K., Flach P.A. (2019) Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety. *Safe AI AAAI*, pp. 1–4.

15. Srinivasu P.N., Sandhya N. et al. (2020) From Black Box to Explainable AI in Healthcare: Existing Tools and Case Studies. doi: https://doi.org/10.1155/2022/8167821.

16. Veale M., Edwards L. (2018) Clarity, Surprises, and Further Questions in the Article 29 of Working Party Draft G=Guidance on Automated Decision-Making and Profiling. *Computer Law & Security Review*, no. 2, pp. 398–404. doi: https://doi.org/10.1016/j.clsr.2017.12.002.

17. Verma S., Boonsanong V. et al. (2022) Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. Counterfactual Explanations and Algorithmic Recourses for Machine Learning. arXiv:2010.10596 [cs, stat]. arXiv, pp. 1–23.

18. 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, v*ol. 7, no. 2, pp. 76–99. doi: https://doi.org/10.1093/idpl/ipx005.

19. Wischmeyer T., Rademacher T. (2020) Artificial Intelligence and Transparency: Opening the Black Box. Regulating Artificial Intelligence. Cham: Springer International Publishing, pp. 75–101.

Information about the authors:

P. P. Kabytov – Candidate of Sciences (Law), Leading Researcher.

N. A. Nazarov – Junior Researcher.

Contribution of the authors:

P.P. Kabytov–Introduction, Conclusion; N.A. Nazarov–Chapters 1,2,3; Introduction, Conclusion.

The article was submitted to editorial office 06.03.2025; approved after reviewing 21.04.2025; accepted for publication 12.05.2025.