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

Research article
JEL: K00
UDK: 340
DOI:10.17323/2713-2749.2025.2.69.86

Trust in Artificial Intelligence: Regulatory Challenges and Prospects

Svetlana S. Vashurina

National Research University Higher School of Economics, 20 Myasnitskaya Str., Moscow 101000, Russia,

svashurina@hse.ru, ORCID: 0000-0002-4892-1971.

Abstract

The last few years have witnessed a rapid penetration of artificial intelligence (AI) into different walks of life including medicine, judicial system, public governance and other important activities. Despite multiple benefits of these technologies, their widespread dissemination raises serious concerns as to whether they are trustworthy. The article provides an analysis of the key factors behind public mistrust in AI while discussing ways to build confidence. To understand the reasons of mistrust, the author invokes the historical context, social study findings as well as judicial practices. A special focus is made on the security of AI use, AI visibility to users and on decision-making responsibility. The author also discusses the current regulatory models in this area including the development of universally applicable legal framework, regulatory sandboxes and self-regulation mechanisms for the sector, with multidisciplinary collaboration and adaptation of the effective legal system to become a key factor of this process. Only this approach will producer a balanced development and use of AI systems in the interest of all stakeholders, from their vendors to end users. For a more exhaustive coverage of this subject, the following general methods are proposed: analysis, synthesis and systematization; special legal (comparative legal and historic legal) research methods. In analyzing the available data, the author argues for a comprehensive approach to make AI trustworthy. The following hypothesis is proposed based on the study's findings. Trust in AI is a cornerstone of efficient regulation of AI development and use in various areas. The author is convinced that, with AI made transparent, safe and reliable one, provided with human oversight through adequate regulation, the government will maintain purposeful collaboration between man and technologies thus setting the stage for AI use in critical infrastructures affecting life, health and basic rights and interests of individuals.

Keywords

artificial intelligence; trust in AI systems; system transparency; system visibility; system security; system reliability; regulatory model; regulatory sandbox; self-regulation for AI system development.

For citation: Vashurina S.S. (2025) Trust in Artificial Intelligence: Regulatory Challenges and Prospects. *Legal Issues in the Digital Age,* vol. 6, no. 2, pp. 69–86. DOI:10.17323/ 2713-2749.2025.2.69.86

Background

According to statistics, the Russian public perceives AI mostly in a neutral positive light, a fact confirmed, in particular, by the popular belief that AI would never get out of human control[1]. A survey by Pegasystems showed that only 24% of all those polled in North America, Europe, Near East and Africa, and the Asian Pacific region believed in AI getting out of human control while almost 40% did not agree that AI could handle customer service better than man[2]. Thus, trust in AI cannot be judged as high. However, one has to agree that confidence in AI systems is a key factor of further technological revolution [Leshkevich T.G., 2023: 36]. AI applications can have sizeable impact on people, up to legally binding implications [Vinogradov V.A., 2023: 164]. Obviously, the general criticism of the algorithms based on machine learning comes from their dependance on data quality. Once the source data is biased, the software will generate biased results [O'Neil C., 2016: 87].

Ubiquitous introduction of AI systems raises a critical regulatory issue, that of human trust in AI. In this context, one has to agree with professor Vinogradov that AI systems should be visible and comprehensible to users [Vinogradov V.A., 2023: 157–166]. In this study, author attempts to formulate problem of trust in technologies and its impact on legal regulation of AI. The study primarily purports to discuss what causes mistrust in AI and how to overcome it.

Making AI trustworthy is a prerequisite of regulatory regime that will make AI more intelligible and transparent to users and reduce the risks

[1] Available at: URL: https://ai.gov.ru/knowledgebase/etika-i-bezopasnost-ii/202_ncrii/?ysclid=lvt627n4mj432190293 (accessed: 23.04.2024). Trust in AI: URL: https://wciom.ru/analytical-reviews/analiticheskii-obzor/doverie-k-ii (accessed: 25.04.2025)

[2] Available at: URL: https://www.pega.com/ai-survey (accessed: 23.04.2024)

of violation of human rights. Thus, the challenge is twofold: firstly, to identify what causes public mistrust in AI and, secondly, to discuss regulatory models adopted worldwide and, based on the available regulatory experience, propose ways to offset the causes of mistrust. It is worth noting this study is multidisciplinary with a focus on a comprehensive issue, thus requiring not only to invoke purely legal arguments and assertions but also to apply social study findings and those from related fields of knowledge. In particular, the article refers to examples from history to illustrate socioeconomic implications of high levels of mistrust and human concerns raised by the emergence of new technologies, as well as causes of mistrust and ways to overcome it.

The article provides an analysis of different aspects of social relationships to be regulated amidst complications brought by AI, in particular, those dealing with AI development and introduction, ethical aspects of designing, using and ensuring oversight of AI, human trust in AI, as well as adapting legal regulation of social relationships to the emergence of new technologies.

1. Mistrust in the Emerging Technologies and its Causes

Discussion about human trust in technologies requires a focus on psychological and sociological studies since it is human attitude to innovations that largely foreshadows provisions to regulate a certain area of social relations. While regulation cannot (nor should) anticipate the development of socioeconomic relations, legal provisions, in responding to social conflicts that have taken place, can become an relevant way to address them.

In psychological studies, trust is defined as "emotional attitude, optimistic perception of a thing" [Jones K., 1996: 5], or "psychological attitude consisting of the emotional, cognitive and behavioral components" [Kupreichenko A.B., 2008: 571]. Trust is a critical element of social collaboration expressed in various forms such as trust in government, public agencies, laws. Interestingly, S. Stepkin views trust as relying, among other things, on a balance of individual rights and duties, a reasonably commensurable balance of private and public interests, stability and predictability, openness of government agencies, independence and impartiality of judicial authorities, reliability and consistency of official information [Stepkin S.P., 2023: 32]. It is important to invoke A.N. Kokotov' view whereby the relations built on trust or mistrust define the essence of law and its meaningful functional and formal manifestations

[Kokotov A.N., 2020: 42]. Psychological attitude to a phenomenon will be thus reflected in a legal content.

Indeed, trust in technologies critically depends, in our view, not only on human response to innovations but also on what these technologies are capable of. In discussing this question, it is necessary to identify at least three aspects that clearly illustrate the problem of human trust in technologies:

Changes to the nature of work from AI used in production;

AI safety and reliability;

AI visibility and transparency.

Analysis of the key challenges related to mistrust in technologies will allow to make practical proposals for better regulation of this area.

1.1. Changes to the Nature of Work from AI Used in Production

As a result of the 19th century industrial revolution, machines stepped out as a partial replacement of human functional duties and physical capabilities, with less qualified workers put in charge of automated processes largely to control the equipment. This trend led to gradual ousting of the skilled factory workforce from economic relations associated with production of goods. The introduction of novel and improved capital goods was caused by a desire to make manufacturing better, faster and cheaper. Despite the clearly positive changes for society from automated equipment in different production sectors, these new technologies met with fierce opposition[3]. With a transition from manual to machine work, automation changed the nature of work, only to impact socioeconomic relations.

Mankind is now approaching the fourth industrial revolution caused by AI and big data systems. It is fair to say that current technologies can be a substitute for not just physical but also intellectual human capabilities, being able to process large quantities of data within minimum time, propose graphical or text solutions, create works of art. However, AI use in many areas is not regulated and can potentially become a key issue leading to human rights violations.

[3] In Lancashire automatic equipment ousted manual work in cotton spinning, only to cause violent riots in 1768 and 1779. Available at: URL: https://historyofinformation.com/detail.php?id=443 (accessed: 20.11.2024). In 1866, Belgian workers on strike demolished a glass factory following the introduction of glass melting furnaces. See: G. Deneckere. 1900 België op het breukvlak van twee eeuwen. Tielt, 2006, pp. 70–71.

Bloomberg Intelligence is expecting a 30-fold growth of the generative AI market up to USD 1.3 trillion by 2032[4] as generative AI-enabled solutions will constantly transform industrial operations over the next decades[5]. As of late 2024, generative AI had a major impact on the existing labor market with considerable competitive pressures on different walks of life. Thus, according to a study of the freelance market in Russia, the generative AI — in particular, the rise in popularity of ChatGPT — hit the text processing segment of translators, copyrighters and editors[6]. Meanwhile, the International Labor Organization (ILO) believes that AI will help create more jobs despite that a majority of current occupations will be fully or partially automated[7].

However, the ongoing automation of jobs and partial or full replacement of man in production processes does not always accord well with law, only to cause a negative response by trades. Thus, the United States have become a focal point of strike action, with the Writers Guild of America protesting against the Producers' Alliance for Cinema and Television practices of using AI to write and rerecord any material, and using screen writers' output for machine learning[8]. Meanwhile, the WGA also made proposals to regulate AI use across the industry in the first ever attempt to prohibit using AI as a substitute for workers. The Screen Actors Guild held a no less important strike in the U.S. against video game publishers over a concern that generative AI could be trained to reproduce voice, only to push actors out of work[9].

---

[4] ChatGPT to Fuel $1.3 Trillion AI Market by 2032, New Report Says. Available at: URL: https://www.bloomberg.com/news/articles/2023-06-01/chatgpt-to-fuel-1-3-trillion-ai-market-by-2032-bi-report-says (accessed: 20.11.2024)

[5] Labor market 30 years after: neural networks as the core tool. Available at: URL: https://trends.rbc.ru/trends/education/64ee043f9a79472565f6efde?from=copy (accessed: 20.11.2024)

[6] Labor market impact of artificial intelligence. Availablde at: URL: https://www.tadviser.ru/index.php/%D1%F2%E0%F2%FC%FF:%C2%EB%E8%FF%ED%E8%E5_%E8%F1%EA%F3%F1%F1%F2%E2%E5%ED%ED%EE%E3%EE_%E8%ED%F2%E5%EB%EB%E5%EA%F2%E0_%ED%E0_%F0%FB%ED%EE%EA_%F2%F0%F3%E4%E0 (accessed: 20.11.2024)

[7] Available at: URL: https://rg.ru/2023/08/29/chisto-avtomaticheski.html (accessed: 20.11.2024)

[8] AI can't replace humans yet — but if the WGA writers don't win, it might not matter. Available at: URL: https://www.polygon.com/23742770/ai-writers-strike-chat-gpt-explained (accessed: 20.11.2024)

[9] Video game actors to go on strike over AI // URL: https://www.gamefile.news/p/video-game-actors-strike-sag-aftra, see also: Actors say Hollywood studios want their AI replicas — for free, forever. Available at: URL: https://www.theverge.

However, that it is not only strike action but also trials that dramatically exemplify the rejection of new technologies. In many instances, content providers accused one or more companies of stealing intellectual assets to train large language models[10]. The matter of dispute unambiguously points out that society represented by professional communities is still fearful of losing jobs or incomes. Obviously, those involved in creative occupations, routine work and text processing (translators, copywriters, editors) are all at risk. However, the changing nature of work will generate new jobs required to service AI (like cyber-security specialists, prompt engineers, AI system trainers etc.).

1.2. Security and Reliability of Technologies

A critical aspect of trust in technologies is their security and reliability from a human perspective. The emergence of new technological solutions impacting social relations gives rise to relevant provisions to make technologies trustworthy. With AI systems gradually penetrating all human activities across the board — from leisure to contacts with public authorities — the success and efficiency of their use in areas critical for individual life and rights depend on a high level of security and reliability. In a number of such areas, AI is already around[11].

How should AI safety and reliability be manifested? First of all, AI systems should be resistant to external exposure as a key aspect of cybersecurity. The issue of AI security and reliability is largely related to the stable operation of the system itself, predictability of its behavior and possibility to maintain human oversight. No secure and reliable use of AI in critical infrastructures is possible unless there is an assurance that the system is under control of its owner and/or developer and is able to resist outside attacks and to operate correctly in an uncertain environment. Above all, AI security and reliability criteria come from technical

---

com/2023/7/13/23794224/sag-aftra-actors-strike-ai-image-rights (accessed: 20.11.2024)

[10] Available at: URL: https://www.fastcompany.com/91179905/openai-anthropic-and-meta-tracking-the-lawsuits-filed-against-the-major-ai-companies (accessed: 20.11.2024)

[11] In particular, to analyze medical images in health care; personalize web searches and recommendations; improve road traffic and accessibility of public transport; make public governance more efficient and less costly; provide for maximum comfort in the delivery of public services; ensure face recognition in fighting crime; facilitate and automate routine processes at court, for instance, in predictive administration of justice, etc.

documents and standards regulating the development, introduction and use of this technology but strategic AI regulations should envisage, in our opinion, mandatory drafting and, possibly, harmonization of security and reliability criteria depending on where AI is to be used.

Notably, legal regulation of technologies should meet individual interests, in particular, via the requirements of security and reliability, while, on the other hand, avoid arresting or retarding technological development. The experience of legal regulation of technologies in the 19th century Britain vividly demonstrates provisions meant for safe use of technological achievements can put obstacles to industrial development[12], as evidenced by the automotive sector. This example demonstrates the legislator's strife to enhance other parties' trust in self-propelled vehicles via mandatory traffic hazard warning but the chosen mechanism proved to be inefficient, only to result in provisions that significantly obstructed the sector's development.

1.3. AI Visibility and Transparency to Users

The issue of making AI systems trustworthy is also hinged on AI visibility to users and possibility of authentication and verification of information that AI can generate and disseminate.

As noted above, AI is increasingly harnessed to serve daily needs prompting the widespread use of many technologies. In this regard, it has to be admitted that "the simplicity of using and creating basic products, the emergence of applications for a wide range of users have resulted in the risk of misuse and threats of illicit behavior enabled by technology" [Vinogradov V.A., Kuznetsova D.V., 2024: 218].

Deepfake, a technology harnessed not only to create entertaining content but also to achieve critical business objectives (in cinema, advertising etc.) exemplifies the problem of AI visibility. Meanwhile, this

[12] Under the British Locomotive Act (also known as the Red Flag Act) passed in the second half of the 19th century (1865), the speed of horseless vehicles was limited to 2 miles/hour in urban and 4 miles/hour in rural areas (1 mile/hour=1.61 km/hour). Under the Act, each vehicle was to have three drivers — two in the vehicle and one walking in front with a red flag to warn others of a self-propelled vehicle on the road. Such way of regulating the emerging technologies was clearly contrary to the interests of sectoral development. (see: The Locomotives Act 1865 (Victoriae Reginae 28&29, p. 83 — legislation.gov.uk. Available at: URL: https://www.legislation.gov.uk/ukpga/Vict/28-29/83/pdfs/ukpga_18650083_en.pdf (accessed: 23.04.2024); The Red Flag Act. Available at: URL: https://law-school.open.ac.uk/blog/red-flag-act; Available at: URL: https://Red_Flag_Act_Locomotive_1865_Cars_Speed_Limits_Man_Running_Carrying_A.htm (accessed: 23.04.2024)

technology can be used both for good and evil purposes since it assumes employing AI to manipulate audio, photo and video materials to make them look like original images, videos or sound tracks. In illicitly using a deep fake, the wrongdoer attempts to produce and disseminate AI-generated information that is false and misleading, an equivalent of intentional deception and breach of trust. As a result, this technology is used to commit a crime for personal gain.

However, it is not only deep fake technologies that can lead to a breach of trust and misinformation. With AI capability for self-learning and data generation giving rise to chat bot technologies, a popular AI-enabled chat bot generated false allegation of sexual harassment against a George Washington University professor involving a female student[13]. The chatbot generated on its own a Washington Post article with false information about the crime and would produce upon request a quotation from this article as if it were real. Following this story, Jonathan Turley, US lawyer and legal analyst, called for cautious use of AI stressing the threat of misinformation that this technology can disseminate.

Meanwhile, algorithms are used not only in routine situations but also in human contacts with public authorities, with examples of mistrust also found in the area of justice. Notably, relief in court is inalienable human right to be observed, guaranteed and enforced by the government, so that decision-making algorithms are to be regulated and made visible and comprehensible to trial parties. Because a court decision has an enormous impact on individual rights, especially in criminal proceedings, there should be a mechanism to make sure that algorithmic decision-making is never unfair or inaccurate[14].

AI COMPAS, a system used in the United States for administration of justice, is often subject to criticism. In an important precedent in 2013 involving a certain Mr. Loomis detained in the State of Wisconsin, software (AI COMPAS) was used for risk assessment. The defense argued that this software was used in violation of the right to due process since the accused could not challenge either the evidence for or the accuracy of the text behind the system's decision. Notably, in delivering the sentence, the judge took into account the person's prior criminal history as well as the assessment produced by AI COMPAS.

[13] Available at: URL: https://www.foxnews.com/media/chatgpt-falsely-accuses-jonathan-turley-sexual-harassment-concocts-fake-wapo-story-support-allegation (accessed: 20.08.2024)

[14] Available at: URL: https://towardsdatascience.com/bias-in-the-ai-court-decision-making-spot-it-before-you-fight-it-52acf8903b11 (accessed: 24.04.2024)

AI COMPAS is based on a patented algorithm that takes in account some of the answers to a questionnaire. The algorithm is proprietary and not disclosable to an indefinite range of persons. Under this criminal case, AI COMPAS has identified the accused as being subject to a high risk of relapse, with Loomis convicted to six years in prison. Responding to an appeal, the Supreme Court of the State of Wisconsin has ruled that algorithmic risk assessment used by the first instance court in delivering the sentence did not violate the accused person's right to due process despite this assessment was not disclosed either to the court or the accused[15]. As follows from the above example, the judge has delivered the sentence with reliance on algorithmic decision of a software which was neither transparent nor intelligible to the trial participants. Thus, the guilty sentence relied on a decision generated by the machine has analyzed input data through mathematical calculation.

Thus, whether judicial decisions are fair and correct depends exclusively on the quality of data the developer uploaded to the software. Once introduced not only to the judicial system, but also that of public governance, decision-making algorithms predicting human behavior will eventually result in a technocratic and bureaucratic governance and declining percentage of human decisions [Janssen M., Kuk G., 2016: 371–377], with final decision-making guided by conclusions of an automatic system with minimum human control, only to aggravate the problem of algorithmic responsibility. In this context, building trust in AI will become crucial since a dramatic decline of trust in AI and related algorithmic systems may lead to a still graver crisis of trust in social institutions such as government, businesses and community organizations [Jian J.-Y., Bisantz A.M., Drury C.G., 2000: 53–55].

Algorithmic complexity is a major cause of non-transparency. Widespread use of AI in critically important areas of social life is only feasible if an algorithm as a possible substitute for human decision-making is able to make a decision at least as fair and justified as human person would. In this regard, it is argued that AI systems could be trustworthy if they are legitimate, ethical and reliable[16]. In this context, it is crucial to understand that the issues of legitimacy and ethics cannot be addressed

[15] Loomis v. Wisconsin, 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S.Ct. 2290 2017. Available at: URL: https://harvardlawreview.org/print/vol-130/state-v-loomis/ (accessed: 24.04.2024)

[16] Ethics guidelines for trustworthy AI // European Commission. Available at: URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed: 24.04.2024)

without involving the regulators that adopt regulations governing AI development and use.

2. Regulatory Models for AI

Regulation has a particular task of making AI visible to users and of creating conditions for more trustworthy AI [Vinogradov V.A. et al., 2023: 157–166]. As for the need to create a legal framework regulating AI development, implementation and use, it is worth noting that it is important for the legislator to identify a balanced approach to regulation of technologies. In social relations complicated by the use of technologies in a digital environment, regulatory challenges come from the fact regulation must not hold back the technological change. Notably, these value-based reference points are contradictory [Barfield W., Pagallo U., 2018: 53]. Thus, one needs to strike a balance between regulation based on constitutional principles and an enabling technological environment. The study of international regulatory experience with regard to AI suggests that neither legal system has so far drafted and adopted a comprehensive instrument addressing all challenges in this area. Meanwhile, several models are worth considering to deal with this task:

Adoption of an overarching regulation;

Adoption of sandbox regulations applicable to AI and other technologies;

Self-regulation of the sector.

Notably, this study, while not considering all regulations approved and came in force in different jurisdictions, is focused only at those that vividly demonstrate regulatory models and purport to enhance trust in technologies.

2.1. Overarching Regulation (Exemplified by the European Union)

In March 2024 the European Union has passed Artificial Intelligence Act (AI Act) to establish an overarching legal framework for AI use. It h`s took force on 1 August 2024 with provisions to be applied gradually over the next 6 to 36 months. The Act's declared purposes were: better functioning of the internal market and promoting the uptake of human-centric and trustworthy AI[17]. It is important AI Act implements a risk-

[17] Art. 1 AI ACT. Available at: URL: https: //eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021 (accessed: 24.04.2024)

oriented approach based primarily on guarantees of human rights and trustworthy AI systems.

The starting point of regulatory approach to AI in the European Union was the White Paper on AI[18] identifying not only risks from the use of AI systems, but also a priority task of making them trustworthy. Thus, the risk-oriented approach enshrined in the Act identifies four risk categories that AI systems could be attributed to. The method of regulation varies considerably depending on the said categories. For instance, AI systems posing a clear threat to security, livelihoods and human rights are to be prohibited.

AI systems classified as high risk will be subject to tougher requirements. Thus, high-risk AI systems include critical infrastructures likely to put at risk the life and health of individuals; educational and vocational trainings determining job access or admission; administration of justice and democratic processes; critical private and public services etc. Notably, the key requirements applicable to high-risk AI systems are visibility and transparency to users[19], human oversight[20] and also high quality of databases for AI learning[21] that would allow to minimize risks for users and generate non-discriminatory outcomes. Moreover, users of limited risk systems subject to only specific transparency requirements[22] will be advised that they deal with an AI system with an option of either continue or reject further use. Obviously, AI system transparency is crucial for the regulator for establishing legal regulation in this area.

In addition, the AI Act provides for a multi-level governance system and support for innovations in the AI sector. On the one hand, this system is expected to ensure efficient oversight over the development, deployment and use of AI across sectors while, on the other hand, to support R&D and law enforcement practices of member states at the national level, with public agencies such as the European Commission on AI, European AI Office, Advisory Forum and Panel of Independent

[18] Available at: URL: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (accessed: 24.04.2024)

[19] Art. 13 of AI ACT. Available at: URL: https: //eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021 (accessed: 24.04.2024)

[20] Ibid. Art. 14.

[21] Ibid. Art. 10.

[22] Ibid. Art. 50.

Experts to be set up. Moreover, the Act obliges EU member states to establish AI regulatory sandboxes at the national level[23]. It is worth noting that the national-level regulation in the form of sandboxes offers a considerable potential as it allows to assess the effectiveness of provisions that regulate social relations in this domain.

2.2. Regulatory Sandboxes

While many countries have opted for regulations establishing regulatory sandboxes, it is necessary to identify the benefits of this regime as a whole before discussing how it is used across countries. The institution of regulatory sandboxes is renowned in jurisdictions and advised by international organizations such as the OECD and the International Telecommunication Union [Efremov A.A., 2019: 21-23]. Essentially, a regulatory sandbox allows regulators to establish a special legal regime for innovative businesses in sectors such as IT, finance[24], transportation[25], health[26], public and municipal services.

The institution of regulatory sandboxes allows to drop certain requirements that often hold back the technological development. Using regulatory sandboxes, the companies involved in the development of innovative products and services can test them on practice without running the risk of non-compliance. As regards AI, such sandboxes are used in Germany[27], Russia[28], Canada[29] and other countries.

---

[23] Ibid. Art. 57.

[24] See:CSA Regulatory Sandbox. Available at: URL: https://www.securities-administrators.ca/industry_resources.aspx?id=1588 (accessed: 24.04.2024)

[25] Available at: URL: https://www. timesofisrael.com/new-legislation-paves-path-for-trial-of-driverless-autonomous-taxis-in-israel/; 2020 Autonomous Vehicles Readiness Index. Available at: URL: https://assets.kpmg/content/dam/kpmg/es/pdf/2020/07/2020_KPMG_Autonomous_Vehicles_Readiness_Index.pdf; Self-driving vehicles. Available at: URL: https://www.government.nl/topics/mobility-public-transport-and-road-safety/ self-driving-vehicles (accessed: 24.04.2024)

[26] Health and Biosciences: Targeted Regulatory Review—Regulatory Roadmap. Available at: URL: https://www.canada.ca/en/health-canada/corporate/about-health-canada/legislation-guidelines/acts-regulations/targetedregulatory-reviews/health-biosciences-sector-regulatory-review/roadmap.html (accessed: 15.07.2022)

[27] Making space for innovation vehicles. Available at: URL: https://www.bmwi.de/Redaktion/EN/Publikationen/Digitale-Welt/handbook-regulatory-sandboxes.html (accessed: 24.04.2024)

[28] Federal Law No. 258-FZ “On the Experimental Legal Regimes for Digital Innovations in Russia” of 31.07.2020 (as amended). Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_358738/ (accessed: 24.04.2024)

[29] CSA Regulatory Sandbox...

Once established, regulatory sandboxes allow to have "smart" regulation of information technologies to account for the needs of IT system vendors and users, once the experiment's outcomes are validated. Among the benefits of regulatory sandboxes are lower information asymmetry and regulatory costs, higher capital commitments of companies involved in the experiment, and better understanding of technological innovations by control and supervisory bodies.

For the study of sandbox law provisions for more trustworthy AI, one needs to refer to Federal Law No. 258-FZ "On Experimental Legal Regimes for Digital Innovations in Russia" of 31 July 2020. In particular, as was noted above, the issues of security and regulation of responsibility are crucial here. Thus, FZ No. 258 was amended in 2024 to include a procedure for processing claims on harm to life, health or assets of natural or legal persons from solutions developed by AI under the experimental legal regime. This procedure provides for setting up a special commission to clarify circumstances of the harm being caused. It is worth noting its members may represent not only the regulator but also other stakeholders: sandbox participants, business community, expert community etc. Another equally positive aspect is the principle of open deliberations of such commissions[30] that, in our view, also serves to enhance trust both in AI systems themselves and their vendors.

Meanwhile, there is an issue of assessing the extent of sandbox success. In this regard, one has to accept A.A. Efremov's approach that "a successful experiment may be both the one that yielded positive outcomes and the one whose outcomes cannot be deemed successful for further large-scale application" [Efremov A.A., 2022: 21]. In our view, the main advantage of sandboxes is that the regulator can, via a legal experiment, identify the best approach to effective regulation that strikes a balance between the law-protected values and the imperatives of technological change.

2.3. Sector Self-Regulation

The regulatory models for R&D in artificial intelligence, discussed above, are to be established by public regulators. Meanwhile, they will be less effective where AI vendors are not interested in elaborating shared approaches and principles of AI development, deployment and use at the level of self-regulation. Importantly, it is self-regulation instruments

[30] Para 5, Article 18.1, Federal Law No. 258-FZ // SPS Consultant Plus.

that laid down the early principles and defined business values of major IT companies in this market in what came to be called "codes of good conduct".

The first code of this kind was developed in the United States by Google, a renowned IT giant, in 2018[31]and contained important principles of AI development including data security and privacy.

In Russia and China, self-regulation of the AI sector is also widespread, with large membership associations such as Russia's AI Alliance bringing together IT market leaders (like Sber, Yandex, VK, Uralkhim or Rusagro), and in China — web companies (like Baidu or Tencet), telecom (Huawei) or financial companies (Ping An).

In the People's Republic of China, the focus on self-regulation is made at the level of strategic documents approved by the authorities, with the Next Generation AI Plan stressing the importance of self-regulation at the corporate level, and the White Paper on AI Governance considering AI companies as key entities for future regulation of the sector[32]. Moreover, the interim measures to regulate generative AI services taken in summer of 2023 encouraged collaboration between businesses, universities, research institutions and public agencies in the AI sector, as well as participation of Chinese representatives in the development of international rules for generative AI[33]. For lack of regulation over a long time, several entities (mostly Internet companies) set up in-house AI governance systems and collaborated with other businesses to design a framework for self-regulation and promote the guiding principles for the sector.

Self-regulation sector-by-sector is based on the fundamental principle of bona fide conduct by the parties to legal relationships. Ethical standards for more severe requirements to the development, introduction and use of AI systems are crucial for a balanced regulation of technologies. Undoubtedly, trust between the government and society is not possible without *bona fide* conduct on both sides in the widest sense[34].

In Russia, the parties to the AI Alliance have endorsed in 2021 an AI Code of Ethics as a starting point for self-regulation in developing,

[31] AI at Google: our principles //Available at: URL: https://blog.google/technology/ai/ai-principles/ (accessed: 09.11.2024)

[32] Global Atlas of AI Regulation / Ed. by A.V. Neznamov. Moscow, 2023.

[33] Available at: URL: https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm (accessed: 24.04.2024).

[34] Ethics and Law: Correlation and Mechanisms of Reciprocal Impact / Ed. by V.A. Vinogradov. Moscow, 2023.

introducing and using AI at all stages of its lifecycle not regulated by law and/or technical standards[35]. In guiding the development of technologies in Russia, this document is also expected to build confidence in AI on the part of users, society and government. With 363 companies endorsing the AI Code of Ethics as official signatories[36], these come not only from Russia, but also other countries like Nigeria, Zambia, Cyprus, Senegal, Uganda, Kenya, Uzbekistan, Cuba, etc.

In 2024, the parties to the AI Alliance signed a declaration on responsible development and use of generative AI (Declaration) to establish ethical principles and recommendations for responsible treatment of AI not only for vendors but also users of neural network services[37]. The Declaration builds on advisory provisions of the AI Code of Ethics since "the parties have agreed on the principles of security and transparency, ethical treatment of sensitive issues, measures to prevent abuse and misinformation, as well as on promoting user awareness of the capabilities of new technologies"[38]. To achieve the purposes of the Code, a national Commission for Implementation of the AI Code of Ethics was set up as a body in charge of the implementation of its provisions and related performance monitoring of AI actors; collaboration and exchange of the best practices of AI ethics; drafting proposals on AI development priorities related to ethical aspects. Apparently, such practices can make codes of ethics very efficient as a method of the so-called soft regulation. A controlling authority in place will engage more parties into self-regulation and help develop standardized approaches in this area. Notably, the rules of self-regulation also strive to make AI transparent and intelligible to users which is indicative of a general trend shared both by regulators and businesses themselves.

Conclusions

To sum up the findings of this study, it appears necessary to formulate the following points.

---

[35] Available at: URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/%D0%9A%D0%BE%D0%B4%D0%B5%D0%BA%D1%81_%D1%8D%D1%82%D0%B8%D0%BA%D0%B8_20_10_1.pdf (accessed: 24.04.2024)

[36] The signatories of the AI Code of Ethics. Available at: URL: https://ethics.a-ai.ru/ (accessed: 24.04.2024)

[37] Available at: URL: https://ai.gov.ru/mediacenter/uchastniki-alyansa-v-sfere-ii-podpisali-deklaratsiyu-ob-otvetstvennoy-razrabotke-i-ispolzovanii-gene/ (accessed: 24.04.2024)

[38] Available at: URL: https://tass.ru/ekonomika/20221995 (accessed: 24.04.2024)

Firstly, making AI trustworthy both through regulation and self-regulation (by companies which develop and introduce AI) is a priority task in a number of jurisdictions. In drafting regulatory provisions for transparency and intelligibility of AI actions, it is necessary to strike the right balance between individual rights and liberties associated with AI use and the interests of the sector since excessive administrative procedures behind complicated bureaucratic processes can become a major obstacle to technological development.

Secondly, the problem of trustworthy AI largely depends on its security and reliability as well as on its visibility and transparency to users. As demonstrated by international regulatory experience, the issues of AI visibility and transparency to human users could be addressed, in particular, by mandatory marking AI systems and advising users accordingly. The AI visibility challenge is pending for all legal systems as this criteria will enhance trust in AI systems and AI-enabled decision-making. Trustworthy AI will allow to overcome the digital divide caused not so much by technologically ill-equipped territories as by psychological perception of AI systems by different categories of individuals. Moreover, the experience of the People's Republic of China to step up the liability for AI-enabled misinformation appears useful and promising[39]. As was noted above, massive use of technologies has resulted in illicit ways to harness them. Introducing criminal liability for using AI to deceive or mislead the parties to legal relationships will enhance the society's trust in technologies.

The issue of AI security and reliability is largely related to the system's sustainable, predictable operation and a possibility to maintain human oversight. However, this issue depends, in particular, on security of person and law-protected data behind AI training. The legislator must provide for mechanisms to protect this data. Thus, strike action by creative trades to protest against making copyrighted material or biometric data of celebrities (such as voice) available for AI training is a clear demonstration of the professional communities' rejection of such training practices. It would appear that only the legislator is well-placed to settle the arising controversies.

Thirdly, this discussion of different regulatory models suggests that effective regulation of AI development and introduction in various walks of life requires as a crucial and promising aspect both comprehensive

[39] China seeks to root out fake news and deep fakes with new online content rules //Available at: URL: https://www.reuters.com/article/us-china-technology/china-seeks-to-root-out-fake-news-and-deepfakes-with-new-online-content-rules-idUSKBN1Y30VU/ (accessed: 20.11.2024)

regulation by competent public authorities and self-regulation by the key market players. In this regard, it is worth noting regulatory sandboxes appear to be a shrewd way to proceed since this specific arrangement will facilitate an experiment based on the envisaged purposes, objectives and key indicators of success or failure. Let author of article to believe such sandboxes will allow the regulator to strike a necessary balance for effective regulation of this sector.

Thus, trustworthy AI is currently crucial and trendsetting for further progress in regulating AI development and use. Addressing this challenge will contribute to efficient introduction of these systems into critical spheres of social life including justice, electoral process and other democratic procedures, health, public security, transport accessibility. Apparently, using regulation to build trust in AI is the main vector for legal systems both domestically and internationally wherever one aspires to become a global AI leader.

References

1. Barfield W., Pagallo U. (2018) Research Handbook on the Law of Artificial Intelligence. Northampton: Edward Elgar, 736 p.

2. Efremov A.A. (2019) Experimental Legal Regimes for Digital Innovations: International Experience and Domestic Prospects. *Informatcionnoe pravo*=Information Law, no. 3, pp. 21–23 (in Russ.)

3. Efremov A.A. (2022) Experiments in Public Administration: Aspects of Delivery and Efficiency. *Gosudarstvennaya sluzhba*=Public Service, vol. 24, no. 1, pp. 19–28 (in Russ.)

4. Ethics and Law. Relationship and Mechanisms of Mutual Influence (2023). Monograph. Ed. by V.A. Vinogradov. Moscow: Prospekt, 272 p. (in Russ.)

5. Global Atlas of AI Regulation (2023) Ed. by A.V. Neznamov. Moscow: no publisher, 308 p. (in Russ.)

6. Janssen M., Kuk G. (2016) Challenges and Limits of Big Data Algorithms in Technocratic Governance. *Government Information Quarterly*, vol. 33, pp. 371–377.

7. Jian J.-Y., Bisantz A. M., Drury C. G. (2000) Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71.

8. Jones K. (1996) Trust as Affective Attitude. *Ethics*, vol. 107, no. 1, pp. 4–25.

9. Kokotov A.N. (2020) Trust. Mistrust. Law. Moscow: Norma, 192 p. (in Russ.)

10. Kupreychenko A.B. (2008) *Psychology of Trust and Mistrust.* Moscow: Kogito-Center, 739 p. (in Russ.)

11. Leshkevich T.G. (2023) The Paradox of Trust in AI and its Justification. *Filosofiya nauki* i *tekhniki*=Philosophy of Science and Technology, vol. 28, no. 1, pp. 37–47 (in Russ.)

12. O'Neil C. (2016) Weapons of Math Destruction: how Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers, 209 p.

13. Stepkin S.P. (2023) The Concept of "Trust" in Modern Constitutional Law: Emergence, Prospects of Development and Assessment. *Aktualnye problemy rossiyskogo prava*=Urgent Issues of Russian Law, vol. 18, no. 10, pp. 30–44 (in Russ.)

14. Vinogradov V.A. (2023) Legal Aspects of Development of AI Systems. *Zakon*=Pravo, no. 12, pp. 157–166 (in Russ.)

15. Vinogradov V.A., Kuznetsova D.V. (2024) Deep Fake Technology: International Regulatory Experience. *Pravo*. *Zhurnal Vysshey shkoly ekonomiki*=Law. Journal of the Higher School of Economics, vol. 17, no. 2, pp. 215–240 (in Russ.)

Information about the author:

S.S. Vashurina — Postgraduate Student, Lecturer.

The article was submitted to editorial office 19.03.2025; approved after reviewing 24.04.2025; accepted for publication 14.05.2025.