DOI: 10.17323/2587-814X.2022.2.62.73

Trusted artificial intelligence: Strengthening digital protection

Sergey M. Avdoshin
E-mail: savdoshin@hse.ru

Elena Yu. Pesotskaya
E-mail: epesotskaya@hse.ru

HSE University
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

This article is devoted to aspects associated with the up-coming need for mass implementation of neural networks in the modern society. On the one hand, the latter will fully expand the capabilities of state institutions and society delegated to perform numerous tasks with higher efficiency. However, a significant threat to democratic institutions obliges society to set out the concept of reliable artificial intelligence (AI). The authors explore a new concept of a trusted AI necessary for the scientific and international community to counter improper future digital penetration. Explaining to what extent digital transformation is mandatory, the authors emphasize the numerous dangers associated with the applications of artificial intelligence. The purpose of the article is to study the potential hazards of neural networks' abuse by the authorities and the resistance to them with reliance on the trusted AI. Studying various aspects of digital transformation and the use of artificial intelligence technologies, the authors formalize the dangers associated with the emergence and propose an approach to the use of digital protection technologies that can be trusted.

Keywords: neural networks, digital protection, trust, artificial intelligence, society

Citation: Avdoshin S.M., Pesotskaya E.Yu. (2022) Trusted artificial intelligence: Strengthening digital protection. *Business Informatics*, vol. 16, no. 2, pp. 62–73. DOI: 10.17323/2587-814X.2022.2.62.73

Introduction

Undoubtedly, these days the general trend is around the rapid introduction of modern digital technologies into multiple processes within society, where artificial intelligence (AI) is playing a central role [1, 5]. Numerous reforms are noted within both mechanisms already in place and those just emerging to bring humanity to a new stage in its development. Spheres that are actively implementing digital technologies in order to modernize processes and accelerate economic development are no exception.

McCarthy coined the concept of artificial intelligence (AI) in 1955 [6, 7]. AI is a system's ability to interpret data, to learn from such data, and to use gained data to achieve specific goals and tasks through flexible adaptation [6].

We take note that the discussion on the definition of AI has not yet led to a clear result satisfying all stakeholders of AI technologies. Since 2016, the artificial intelligence industry has broken out with the support of cloud computing and big data. Today cloud based artificial neural networks and deep learning form the basis of most applications we know under the label of AI.

The Artificial Neural Network is to some extent modeled on the structure of the biological brain. With these networks, various problems can be solved in a computer-based way. It consists of an abstracted model of interconnected neurons, whose special arrangement and linking can be used to solve computer-based application problems in various fields such as statistics, technology or economics [8].

AI's analytical and cognitive tools allow technology owners to analyze significant amounts of data, meaning they can immediately detect and effectively respond to changes in the agenda. Relying on complex mathematical algorithms, it is possible to increase the level of transparency, optimize internal processes of interdepartmental interaction and stimulate innovative activities, ultimately establishing a higher level of trust [9].

The consequences of the mass introduction of AI are expected to be beneficial for society as a whole. This means that the issue of creating and subsequent implementation of a centralized digital ecosystem aimed at improving the interconnection and interaction of all stakeholders (government, business, associations and individuals) is on the agenda for many companies and organizations [9–11]. The main role in this transformation will be assigned to AI.

The field for AI introduction is truly vast. However, nowadays, there is a noticeable discord, undermining whether the widespread use of AI is timely [12]. On top of issues with privacy, hacker attacks, technological singularity, etc., already widely scrutinized, there is concrete evidence of an equally vital danger [5, 13–15]. The discussion here lies around the accumulation of digital power in the hands of a narrow group of people. It can definitely be argued that a set of modern algorithms opens up almost limitless possibilities. The detailed scenario reflects the concept of the emergence of digital dictatorship in the modern world and the need for digital protection to regulate the unfair use of technology.

Thus, the concept of trusted AI has been proposed as a countermeasure to the unethical use of neural networks [6, 12]. The concept's framework is supposed to find the golden mean between the progressiveness of AI application strategies and the protection of ethical and moral aspects of human life. However, the existing legislative base is negligible and international consensus is absent. Therefore, the delegation of any tasks to neural networks seems to be of high risk now. Not only would the traditional model of public administration be challenged, but also core human values might be threatened [6, 7, 12]. Moreover, understanding the immense gap between the principles of international law and reality, where these principles are constantly violated, we have to realize the insufficiency of only defining abstract principles. Specific counteractions to prevent the use of neural networks against society should be determined. A comprehensive regulatory system is needed to encourage technological progress and define concrete steps to combat rights violations.

Thus, the purpose of this article is to study the potential hazards of neural networks' abuse by technology owners and ways to resist them based on the concept of trust. To avoid problems that can harm a person by distorting, stealing or leaking data, it is necessary to make sure that the results of AI work can be trusted. In this paper, potential problems related to issues of trust, confidentiality and reliability are investigated. The concept of "trusted artificial intelligence" is considered, as well as the phenomenon of digital protection. Section 1 examines in detail the nature of violations of citizens' rights from the point of view of four different spheres of life: political, social, cultural and economic. We point out the irreversibility of the process of implementing neural networks in the context of digital transformation and describe the idea of trusted AI. Section 2 is devoted to discussing possible response strategies and proposals for establishing digital protection. Our study concludes with a summary of key findings.

1. Prospects and methods of artificial intelligence

1.1. New possibilities of artificial intelligence

Scientific and technological progress cannot be stopped. Under the pressure of the rapid development of the IT sector of the world economy, countries are forced to meet the ever-changing demands of business or risk facing a real digital abyss in management [9]. Artificial intelligence has become a part of our lives as smart systems are used in many areas, from client analytics and search engines to voice assistants and medical research. In the medical field, systems that recognize pathologies from video recordings of endoscopic examinations are being developed; in transport – autopilots and traffic management systems; in finance – systems that identify customers or identify suspicious transactions that may indicate tax evasion or money laundering. It is safe to say that further global economic development and progress directly depend on how effectively various industries learn to use artificial intelligence. However, with the development of technology, the problem of trusted artificial intelligence has become more acute, as any problem can have serious consequences. Users want to be sure that the model has a high degree of accuracy and that its results are fair and easily interpreted. For example, incorrectly calibrated sensors of a car equipped with autopilot can cause an accident. Errors in the control of the infrastructure that AI relies upon can lead to a leak of patients' personal data, incorrect medical diagnosis or identity theft. Regarding business and industry, a low level of AI software can lead to delays in transportation, damaging the supply chain.

In today's agenda, special attention is paid to the social aspect of AI's mass introduction. Neural networks' computational and analytical capabilities, far superior to human performance, open up new horizons for public institutions. In addition, AI does not have a limited reserve of endurance and is always available. Consequently, significant amelioration is expected in traffic systems, healthcare, maintaining public order and public services personalization, including education [15]. Impressive progress has already been achieved in providing public services to citizens and legal entities. For now, the scope of AI's implementation is quite limited. The main categories are processing of requests (social payments, migration, citizens' questions, etc.), filling out and searching for documents, translating texts and drafting documents [4].

Significant successes are predicted for neural networks in the field of economic management. Plans are devoted mainly to the improvement of resource allocation and logistics efficiency. It is necessary to restructure and optimize supply channels, warehouse systems, and recycling [15]. Neural networks will be crucial for the promising concept of the "smart city," with control of CCTV cameras, electricity grids, water supply, transport systems, etc. delegated to them [7, 16].

Note that significant changes have occurred under the influence of the COVID-19 pandemic, in particular, in the field of teaching and learning. Academic institutions are moving to digital technologies to provide their students with more resources. Thanks to technology, students now have more opportunities to learn and improve skills at their own pace and on an individual trajectory, now having the opportunity to pass control stages using online tests. Online proctoring services are gaining more and more popularity, in which the subject's face is identified and analyzed to predict his emotions. In addition, aspects such as a phone, a book, or the presence of another person are detected. This combination of models creates an intelligent rule-based inference system that can determine whether there has been any cheating during an exam or test.

Here, the question about the correctness of the system and the adequacy of the assessment of behavior may arise. Any failures and abuses are fraught with negative consequences in terms of academic integrity, discrediting the idea of both offline and online learning. Among the key risks, it is worth highlighting a violation of confidentiality, compromised availability or compromised accounts, leaks of personal data, or distortion of results.

Big data is necessary for the successful development of machine learning models. The quantity, quality and availability of big data affect the efficiency and accuracy of the models being trained. Therefore, many companies are interested in continuous data collection about their consumers. Many systems collect information that is not subject to disclosure: videos and photos from video cameras, speech recordings and financial transactions. Unreasonable use of this information, errors in the model or data theft, can cause threats to the security of individuals or even enterprises and government organizations.

1.2. Potential problems and threats of digital penetration

Data leakage for artificial intelligence is especially dangerous due to the fact that big data usually carries a lot of confidential information from which you can get information about the object that was attacked. Simultaneously, data leakage can occur at any stage of development: training or using a ready-made model.

Violation of confidentiality is another important detail since it is personal information that acts as a catalyst for any digital transformation, becoming the basis for learning models. The process of developing many neural networks is practically inseparable from relying on the collected data, including speech, Internet activity, images, financial flows, medical indicators, etc. Thus, the issue of access to big data turns out to be one of the most significant problems associated with the integration of neural networks into society [13]. Any third party will be aware of the potential risks of using information collected by technology owners regarding user information, their right to privacy and the protection of their personal data [17, 18]. This aspect is widely discussed within society, as there is a direct threat to human health and life behind it. The relevance of this problem has been raised by prominent scientists and statesmen many times. Also, related reports and studies have been repeatedly presented in international discussions (IEEE, EU Committees, OECD, etc.) [6]. The core idea is that AI has only those “ethical values” that have been defined by the developer.

By publicly guaranteeing transparency, full audit and objectivity through the introduction of neural networks, technology owners in the era of digitalization are able to perform any manipulation. This possibility arises directly from the lack of understanding between society users of digital technologies of the structure and the principles of algorithms. Complex AI models perform colossal calculations which cannot be fully understood even by the creators [5, 7]. Thus, for most users, the process of neural networks will be opaque [15]. Scientists refer to it as a “black box” problem. Taking advantage of this phenomenon, unscrupulous developers can use neural networks at their discretion.

Another problem with training data is its low availability. Often, small amounts of data belong to different persons who have no reason to trust each other or the developer, and it is impossible to compile one dataset of sufficient size. If a dataset of the required volume exists, it may still be unavailable if the data contained in it is confidential. Even if suitable data can be found, it is necessary to ensure that they reflect the real state of affairs. In particular, they should not contain hidden biases, as, for example, happened to some facial recognition systems: due to the imbalance of data, they, for example, coped much better with the recognition of light-skinned men than dark-skinned women.

The data that a ready-made AI system works with may also be of poor quality: they may come from unreliable sources or contain information with a high degree of uncertainty. In addition, the databases that the system interacts with may be at risk if the system itself is hacked. For example, a biometric data verification system may be subject to several types of attacks in order to force it to accept an attacker as the owner. Noise is added to the processed data so that the already trained model identifies the object in the photo in the wrong class. Such attacks can be used in computer vision, for example, forcing the model to incorrectly identify road signs.

The bias of the model can have an extremely negative impact on the results of using digital technologies. Among the possible causes of bias are the uneven distribution of data in the training sample, algorithmically embedded preferences and a biased attitude toward individual groups of individuals. Even the classic spectrum of potential crisis cases is huge, starting with ageism and sexism when hiring, and ending with racism when identifying potential criminals [7].

Moreover, it is unacceptable to exclude situations where the bias of the model is a meaningful policy of its creators dictated by their interests, which is even more dangerous. Without proper control, developers can gain serious power over society. Simultaneously, it is difficult to overestimate the degree of destructiveness of the consequences of discrimination: systematic violations of the rights of certain social groups will lead to the definition of AI as an inhumane mechanism.

To avoid problems that can harm a person by distorting, stealing, or leaking data, it is necessary to ensure that the results of AI work can be trusted. Thus, there is a need for the concept of "trusted artificial intelligence": an AI system in respect of which the user can be sure that it is capable of performing the tasks qualitatively [19].

2. Results and discussion: Implementation of trusted artificial intelligence

It is worth admitting that the introduction of artificial intelligence seems to be a very profitable process. However, it is trust that is the key factor in the use of AI, since the rejection of this technology by the masses will exclude all potential benefits [5, 8, 13]. Consequently, the establishment of rational, trusting relationships, excluding excessive trust or its absence, will make it possible to achieve benefits for all of society. The essence of this idea is reflected in the concept of trusted AI.

Trust in AI at the physical level means confidence in the correct operation of all its physical components, such as sensors, and in the quality of the data received by the system. Trust in the infrastructure surrounding AI means confidence in the security of the data with which AI interacts, and in control over access to the system itself. Trust at the application level means confidence in the correct operation of the software.

If an AI system is considered trusted, that is, trust in it is manifested at all three levels, then such a system can be allowed to solve problems with a hugely positive outcome, since the user can be confident in the results of its work.

The trust issue is quite complex. It directly depends on various features of the human psyche [13, 20, 21]. Nevertheless, researchers emphasize that trust is the desire and willingness of an individual to depend on the other party's actions to extract some benefit, despite the potential risks from being in a vulnerable position. This phenomenon is based on the coincidence of the moral values of the parties, which allows a person to predict the further actions of the party subject to trust, and confidence in its sufficient competence [20]. Additionally, the cumulative nature of trust makes development of relationships dependent on the first experience. Therefore, strategies for building a trusted AI should be developed before its mass introduction in strict accordance with ethics and robustness.

Next, we will present several key approaches that appear to be the most important for implementing trusted artificial intelligence for each stakeholder.

A. International community

The first stage is to develop a legislative framework regulating the work of AI in each industry. Already various international committees and organizations are busy drafting general recommendations and guidelines for action [21]. While some attempts to control the technological agenda were made earlier [22], the OECD document "Principles on Artificial Intelligence" is considered the reference point in the discussion around ethics and competence of AI [23, 24]. Having underscored the initial outlines and core principles, the authors provided the basis for more comprehensive documents in the future [23, 25, 26]. Uniting the most vital criteria of trusted AI, they must encourage leading countries in this field to start elaborating their policies [20].

Key principles here include but are not limited to constant human oversight, resilience, accountability, privacy and transparency. Future conventions might also have notions about non-discrimination and fairness, perseverance of social and environmental well-being, attention to mitigating circumstances and the introduction of the so-called "right to explanation" [23, 25–27]. It is essential to underscore that while government agencies will be obliged not only to constantly adapt the legal sphere but also to expand it, international conventions must stand for the announced principles. This will let the authorities develop a balanced system of norms, delineating the areas of regulation between soft and hard laws if these principles are followed [7]. Moreover, this process has to be original. Issues of AI control need to be solved based on the existing legislative structure and not contrary to it. Authorities will give ethical reasons to trust AI if bias is successfully avoided. However, despite the cornerstone importance of future standards, they only point to the importance of compliance with the law in matters of safety and quality, while maintaining abstract rules.

The concept of trusted AI is illustrated in *Fig. 1* and includes many aspects.

B. Scientific community

First of all, it is worth focusing on the importance of scientists and developers to protect neural networks. Otherwise, the basic requirements of trusted intelligence would be undermined, and the situation would potentially shift towards citizens' digital dependence from third parties. However, the evolution of defense methods is almost synchronous with a similar process for AI attacks, which can be illustrated by the following list of the most potential ones:

1. Privacy breaches are one of the most likely issues among the expected ones. Leaks of personal data can occur at almost any point in the neural network's operation, from training to outputting results. Individuals' details are extremely lucrative to certain third parties in many matters. The focus of defense is on privacy-enhancing technologies (PET). The most notable of them is OPAL (the Open Algorithms project). It gives algorithms remote-controlled access to information instead of sending anonymized data. Consequently, only the aggregated result will be returned to the model developers. Owing to the full operations' recording, the entire learning phase is audited [28, 29].

2. Another scenario is data poisoning. The idea here is to inject false information in the training sample, either devaluing the results of the entire system or pushing the neural network towards making wrong decisions beneficial to a third party. Such technologies can both cause significant damage to the reputation of an individual and affect the behavior of the masses (for example, during an election) [20]. No specific solution has yet been proposed here. However, despite all its threatening potential, the risks here are still minimal since these technologies are at an early

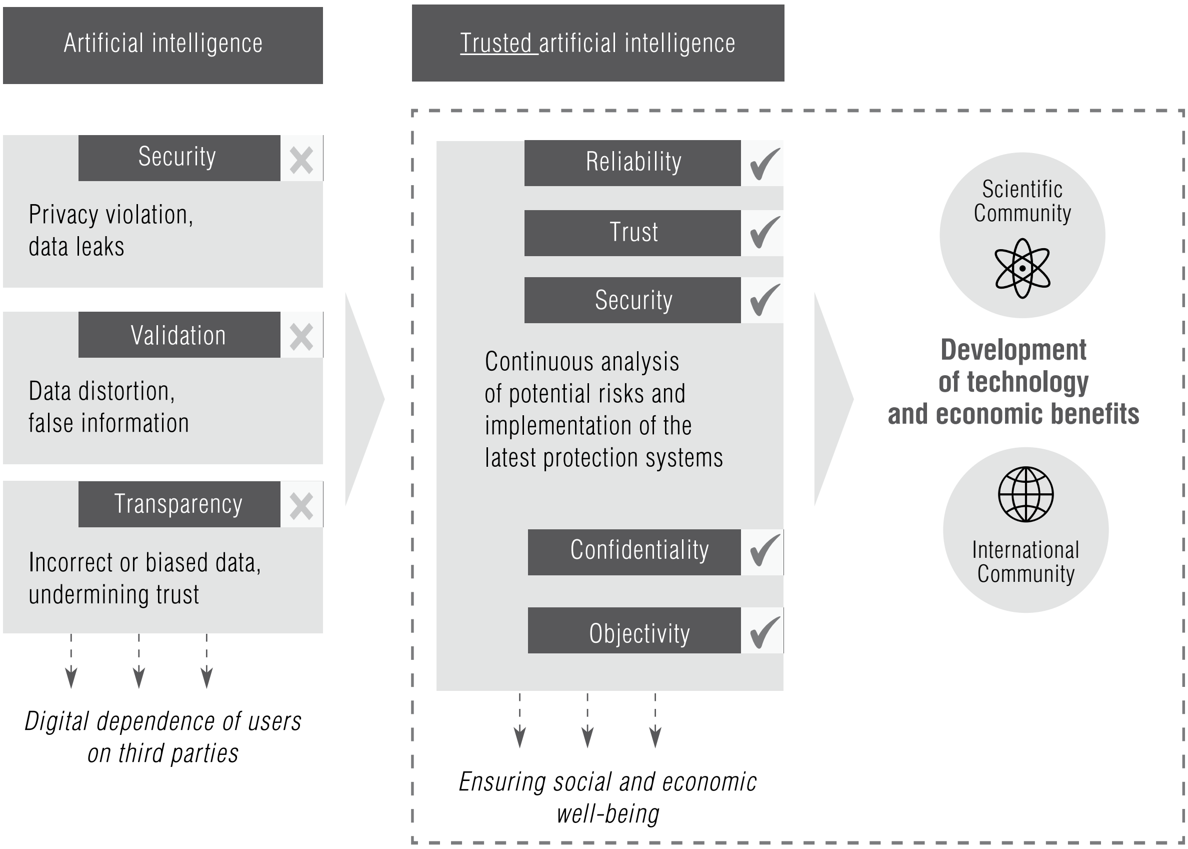
Artificial intelligence
Security
Privacy violation, data leaks
Validation
Data distortion, false information
Transparency
Incorrect or biased data, undermining trust
Digital dependence of users on third parties
Trusted artificial intelligence
Reliability
Trust
Security
Continuous analysis of potential risks and implementation of the latest protection systems
Confidentiality
Objectivity
Ensuring social and economic well-being
Scientific Community
Development of technology and economic benefits
International Community

Fig. 1. The concept of using trusted AI.

development stage which does not imply any substantial threats. There are already several models in the literature for combating data poisoning, almost all of them focused on human control and protection.

3. The so-called evasion attack also causes serious concerns. Unlike data poisoning, which is used during the training phase, this threat appears at the stage of applying AI. It is possible to get a radically new network response by imperceptibly modifying the input values. Even the slightest transformations can lead to sufficient consequences [7]. There are several mechanisms for countering it, but the most widely described is adversarial training, which implies that developers include intentionally incorrect data during training, bolstering the model to ignore potential noise in the future [17, 30].

4. The intellectual value of an already trained neural network is obvious, especially when using big data collected by the government. Model extraction can potentially mean large leaks of personal data. Such a violation of confidentiality can lead to unpredictably catastrophic outcomes (for example, medical records) [7]. Since model inversion mainly exists in scientific articles and abstract models, methods of countering it are still theoretical. The private aggregation of teaching ensembles (PATE) is among the most recognized ones. The concept's idea is to separate data into several sets, each training a separate neural network. Then these independent models, called

"teachers," are combined to train the neural network named "student" by voting, not giving the latter access to the original data [18, 31].

5. Several metrics allow detection of model biases with sufficient efficiency, including equal opportunity, disparate impact, difference in means and normalized mutual information. Developers might discover the imperfections of neural networks and make appropriate alterations using these methods. In turn, the revealed bias can be mitigated by either pre-processing, in-processing or post-processing algorithms [32].

Lastly, an important aspect of the united scientific community's policy is to present specific technologies that would allow the asserted principles to be implemented more effectively. Even now, the gap between the algorithms employed and the abstract principles is obvious. Therefore, in-depth research is constantly taking place, exploring possible scenarios and tools.

Several approaches have already been presented to counter the above issues, one of which encourages us to scrutinize blockchain, a continuous chain of blocks connected in reverse order through hash sums. Each block, in addition to its hash and the hash of the previous block, contains some information. Thus, the blockchain is called a distributed public registry providing data storage and transmission with high robustness and almost zero chance of interference. Smart contracts (the program code ensuring the fulfillment of all the established rules) in conjunction with AI will guarantee the reliability of the final results. Therefore, combining blockchain technology with AI will create a decentralized system that maintains information of any value and provides it for neural network training. As a result, not only is data security guaranteed, but ethical concerns are also addressed owing to a comprehensive history of operations that rules out external interference or pre-programmed bias [24, 33]

Conclusion

The world community has to prepare itself for the up-coming challenges on the inevitable road to a digital society. The latter, indeed, is becoming a crucial stage in improving both governments' operations and human life. The ubiquitous integration of digital technologies and the creation of decentralized ecosystems can open new horizons, eliminating a number of current issues. The key role of neural networks in this process forces us to pay special attention to each potential menace. The establishment and monitoring of trusted AI principles will therefore enable both the scientific community and international organizations to create and regularly update mechanisms to counter risks. Moreover, humanity as a whole will have a chance not to miss out on all the political experience accumulated over countless centuries.

However, the presented technological strategies are nothing more than a theoretical speculation on the topic of future processes. That is why it is vital to constantly adapt strategies to changing realities, to expand the legislative framework and to look for new solutions. At the same time, a strong collective commitment to maintaining democratic institutions is imperative.

Nevertheless, AI is already present in our lives. Although the extent of its adoption is relatively modest, its prospects are truly breathtaking. If we meet the upcoming challenges with dignity, the gains listed above will elevate humanity to a completely new level of existence. ■

References

1. Vogt T., Winter P., Nessler B., Doms T. (2021) *Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications*. Vienna: TÜV Austria Holding AG.

2. Kuleshov A., Ignatiev A., Abramova A., Marshalko G. (2020) Addressing AI ethics through codification. *2020 International Conference Engineering Technologies and Computer Science (EnT)*, pp. 24–30. https://doi.org/10.1109/EnT48576.2020.00011

3. Harrison T., Luna-Reyes L. (2021) Cultivating trustworthy artificial intelligence in digital government. *Social Science Computer Review*, vol. 40, no. 2, pp. 494–511. https://doi.org/10.1177/0894439320980122

4. OECD (2014) *Recommendation of the council on digital government strategies*. Paris: OECD Publishing. Available at: https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf (accessed 22 September 2021).

5. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. (2018) *Adversarial attacks and defences: A survey*. Working paper arXiv: 1810.00069. https://doi.org/10.48550/arXiv.1810.00069

6. Haenlein M., Kaplan A. (2019) A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, vol. 61, no. 4, pp. 5–14. https://doi.org/10.1177/0008125619864925

7. Wei J. (2018) Research progress and application of computer artificial intelligence technology. *MATEC Web of Conferences*, vol. 176, Article Number 01043. https://doi.org/10.1051/matecconf/201817601043

8. Mijwil M., Esen A., Alsaadi A. (2019) *Overview of neural networks*. Available at: https://www.researchgate.net/publication/332655457_Overview_of_Neural_Networks (accessed 22 September 2021).

9. Sibai F.N. (2020) AI crimes: A classification. *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pp. 1–8. https://doi.org/10.1109/CyberSecurity49315.2020.9138891

10. Jastroch N. (2020) Trusted artificial intelligence: On the use of private data. In: *Product Lifecycle Management Enabling Smart X. PLM 2020* (eds. F. Nyffenegger, J. Ríos, L. Rivest, A. Bouras). IFIP Advances in Information and Communication Technology, vol. 594. https://doi.org/10.1007/978-3-030-62807-9_52

11. Nemitz P. (2018) Constitutional democracy and technology in the age of artificial intelligence. *Phil. Trans. R. Soc. A.*, vol. 376. http://doi.org/10.1098/rsta.2018.0089

12. Misra S.K., Das S., Gupta S., Sharma S.K. (2020) Public Policy and Regulatory Challenges of Artificial Intelligence (AI). In: *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation. TDIT 2020* (eds. S.K. Sharma, Y.K. Dwivedi, B. Metri, N.P. Rana). IFIP Advances in Information and Communication Technology, vol. 617. https://doi.org/10.1007/978-3-030-64849-7_10

13. Pavlutenkova M. (2019) Electronic government vs digital government in context of digital transformation. *Monitoring of Public Opinion: Economic and Social Changes Journal*, no. 5, pp. 120–135 (in Russian). https://doi.org/10.14515/monitoring.2019.5.07

14. Kochetkov A.P., Vasilenko I.A., Volodenkov S.V., Gadzhiev K.S., Kovalenko V.I., Soloviev A.I., Kirsanova E.G. (2021) Political Project for Russia: Prospects for implementation in the context of challenges and risks of digitalization of society. *Vlast' (The Authority)*, vol. 29, no. 1, pp. 317–331 (in Russian). https://doi.org/10.31171/vlast.v29i1.7963

15. Williams M., Valayer C. (2018) *Digital government benchmark – study on digital government transformation*. European Union. Available at: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/report-digital-government-benchmark-study-digital-government-transformation (accessed 22 September 2021).

16. Carter D. (2020) Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review*, vol. 37, no. 2. pp. 60–68. https://doi.org/10.1177/0266382120923962

17. Kamat G. (2014) *Algorithms for private data analysis. Lecture 14 – Private ML and stats: Modern ML.* Available at: http://www.gautamkamath.com/CS860notes/lec14.pdf (accessed 22 September 2021).

18. Hinnefeld J., Cooman P., Mammo N., Deese R. (2018) *Evaluating fairness metrics in the presence of dataset bias*. Working paper arXiv: 1809.09245. https://doi.org/10.48550/arXiv.1809.09245

19. National Standard of the Russian Federation (2021) *Artificial intelligence systems. Methods for ensuring trust. General. GOST R 59276-2020* (in Russian).

20. Tinholt D., Carrara W., Linden N. (2017) *Unleashing the potential of Artificial Intelligence in the Public Sector*. Capgemini. Available at: https://www.capgemini.com/consulting/wp-content/uploads/sites/30/2017/10/ai-in-public-sector.pdf (accessed 22 September 2021).

21. Sharma G.D., Yadav A., Chopra R. (2020) Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures*, vol. 2, Article ID 100004. https://doi.org/10.1016/j.sftr.2019.100004

22. McCormick T.R., Min D. (2020) *Principles of Bioethics*. University of Washington. Available at: https://depts.washington.edu/bhdept/ethics-medicine/bioethics-topics/articles/principles-bioethics (accessed 22 September 2021).

23. Lindgren I., Veenstra A.F. (2018) Digital government transformation: a case illustrating public e-service development as part of public sector transformation. Proceedings of the *19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, The Netherlands*, pp. 1–6.

24. IEEE (2021) *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent System*s, *Version 2.* The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (accessed 22 September 2021).

25. Bradul N.V., Lebezova E.M. (2020) Conceptualization of Smart Government: A scientometric approach. *Upravlenets (The Manager)*, vol. 11, no. 3, pp. 33–45. https://doi.org/10.29141/2218-5003-2020-11-3-3 (in Russian).

26. Thiebes S., Lins S., Sunyaev A. (2021) Trustworthy artificial intelligence. *Electron Markets*, vol. 31, pp. 447–464. https://doi.org/10.1007/s12525-020-00441-4

27. Falco G., Viswanathan A., Caldera C., Shrobe H. (2018) A master attack methodology for an AI-based automated attack planner for smart cities. *IEEE Access*, vol. 6, pp. 48360-48373. https://doi.org/ 10.1109/ACCESS.2018.2867556.

28. Bellamy R. et al. (2018) *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.* Working paper arXiv: 1810.01943. https://doi.org/10.48550/arXiv.1810.01943

29. Sarpatwar K. et al. (2019) Towards enabling trusted artificial intelligence via Blockchain. In: Calo, S., Bertino, E., Verma, D. (eds) *Policy-Based Autonomic Data Governance. Lecture Notes in Computer Science*, vol. 11550, pp. 137–153. Springer, Cham. https://doi.org/10.1007/978-3-030-17277-0_8

30. Montjoye Y.D., Farzanehfar A., Hendrickx J., Rocher L. (2017) Solving artificial intelligence's privacy problem. *Field Actions Science Reports*, Special Issue 17, pp. 80–83. Available at: https://journals.openedition.org/factsreports/pdf/4494 (accessed 22 September 2021).

31. DataCollaboratives.org (2021) *Open Algorithms (OPAL) Project*. Available at: https://datacollaboratives.org/cases/open-algorithms-opal-project.html (accessed 22 September 2021).

32. Salah K., Rehman M.H.U., Nizamuddin N., Al-Fuqaha A. (2019) Blockchain for AI: Review and open research challenges. *IEEE Access*, 2019, vol. 7, pp. 10127–10149. https://doi.org/10.1109/ACCESS.2018.2890507

33. Baker-Brunnbauer J. (2021) Management perspective of ethics in artificial intelligence. *AI and Ethics*, vol. 1, pp. 173–181. https://doi.org/10.1007/s43681-020-00022-3

About the authors

Sergey M. Avdoshin

Cand. Sci. (Tech.);

Professor, School of Computer Engineering, HSE Tikhonov Moscow Institute of Electronics and Mathematics (MIEM HSE), National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: savdoshin@hse.ru

ORCID: 0000-0001-8473-8077

Elena Yu. Pesotskaya

Cand. Sci. (Econ.);

Associate Professor, School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: epesotskaya@hse.ru

ORCID: 0000-0003-2129-4645