***E. Korzhovska**, BA student*
***O. Dienichieva**, PhD in Ped., As. Prof.*
Zhytomyr Polytechnic State University

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN CRITICAL SYSTEMS

The current stage of digital transformation is characterized by the widespread use of intelligent systems in areas where algorithmic errors may lead to serious consequences. This applies to fields such as healthcare, financial analytics, autonomous transportation, and defense technologies. In these domains, the use of complex models whose operation is difficult to explain (the so-called "black boxes") may create significant risks. The lack of transparency in the internal decision-making mechanisms of deep neural networks hinders their full deployment, as users cannot verify how models produce their predictions**.** For this reason, the concept of Explainable Artificial Intelligence (XAI) is transforming from a supplementary technical tool into a fundamental requirement for the architecture of modern information systems.

The purpose of this study is to conduct a comprehensive analysis of the role of explainability as a means of ensuring comprehensibility, verifiability, and trust in automated decision-making within critical infrastructure. According to systematic reviews of the current academic literature, the implementation of XAI methods helps reduce the gap between the high predictive power of complex models and the need for human oversight. In medicine, for instance, explainability enables specialists not only to obtain diagnostic results but also to identify features or regions in medical images that influence the AI system's output**.** This not only increases the level of trust among medical professionals but also establishes a basis for the distribution of legal responsibility within the "developer – algorithm – user" relationship [1, p. 18].

In the financial sector, the application of explainable models has become a response to strict regulatory requirements aimed at preventing algorithmic discrimination. The use of approaches such as model-agnostic explanation methods, including LIME and SHAP, makes it possible to detect hidden correlations and biases in data, which is critical for ensuring fairness [2, p. 52]. Moreover, in the field of national security and defense, the ability to verify every step of an AI system is an indispensable condition for its integration into decision-support systems, since algorithmic transparency is directly related to the predictability of system behavior under extreme conditions [3].

It is important to emphasize that XAI should not be considered merely a technical extension for data visualization. Rather, it represents a complex mechanism for establishing a new form of cognitive interaction between humans and machines, where explanations serve as a bridge between human reasoning and machine-generated decisions, facilitating intellectual alignment. The problem of "accuracy versus interpretability" is gradually being addressed through the development of hybrid architectures that combine the mathematical power of deep neural networks with the logical transparency of expert systems.

In conclusion, explainable artificial intelligence can be considered a key direction in the development of critical systems over the coming decade. Ensuring the transparency of algorithmic decisions not only minimizes technical risks but also

contributes to the formation of ethical principles for the safe coexistence of human intelligence and autonomous technologies. Future research should focus on the development of standardized criteria for evaluating the quality of explanations, taking into account the specifics of particular domains and the level of expertise of system operators.

REFERENCES

1. Ethics Guidelines for Trustworthy AI / European Commission High-Level Expert Group on AI. Brussels, 2019. 39 p.
2. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024. 144 p.
3. ISO/IEC 23894:2023 Information technology – Artificial intelligence – Risk management. International Organization for Standardization, 2023.
4. OECD Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments, 2019 (updated 2023).
5. Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges / Ahmed F., Naz N. S., Khan S., Rehman A. U., Ismael W. M., Khan M. A. BMC Medical Imaging, 2026. P.2-29.
6. Explainable Artificial Intelligence for Biomedical and Healthcare Applications / Khamparia A., Gupta D. (Eds.). CRC Press, 2024. 302 p.
7. Explainable AI in the military domain / Wood N. G. Ethics and Information Technology, 2024. P.1-13.