ETHICAL CHALLENGES IN THE DEVELOPMENT OF AI: A GUIDE FOR TECHNOLOGISTS

Oleksii Kiselyov [1]

Received: 2025-10-26
Accepted: 2025-12-25
DOI: https://doi.org/10.5281/zenodo.18305009

Abstract. Analyzing the ethical issues faced by artificial intelligence (AI) systems, the study covers algorithmic bias, privacy violations, and lack of explanation in judgments. An ethical framework for AI development is proposed, consisting of explainability methods (SHAP, LIME), federated learning, data documentation, and diverse teams. Evaluation through frequent audits and the ability to adapt to the cultural environment makes the process fair and transparent, and therefore, trustworthy for AI systems in the financial, medical, and judicial sectors. It is proven that the implementation of a comprehensive ethical approach to the development and use of AI systems helps minimize the risks of discrimination and increase the level of social responsibility of technological solutions. The need to create interdisciplinary regulatory standards that will ensure a balance between innovation and respect for human rights is outlined. The results of the study can be used to formulate practical recommendations for the ethical integration of artificial intelligence into various areas of social activity. The importance of constant monitoring and updating of ethical protocols in accordance with the dynamic development of artificial intelligence technologies is emphasized. Prospects for further research lie in the development of universal methods for assessing the ethics of AI systems and their practical testing in real operating conditions.

Keywords: algorithmic fairness, personal data processing, artificial intelligence audit.

[1] Founder of the Company "Blind Tech", Ukraine
Master`s Degree in Economy, Alfred Nobel University, Ukraine
E-mail: alekseykiselov93@gmail.com
https://orcid.org/0009-0005-5256-5251

Introduction

With radical changes in people's attitudes toward technology, the 21st century is being transformed by artificial intelligence (AI) in finance, healthcare, education, transportation services, and justice, among other areas. Whether it's a recommendation system on social media, credit scoring, disease diagnosis, or behavior prediction, AI is making its mark in every area of life. However, innovation comes with some really important ethical questions that pose serious threats to society, and these questions have legal and social implications. Algorithmic systems can directly or indirectly reproduce historical biases, such as prejudice against certain demographic groups due to biased data, infringe on privacy, or serve as a "black box" whose decisions are difficult to explain. For example, credit risk scoring systems may discriminate against minority groups if the training data contains records of past financial inequality. Traditional approaches to software development, which are guided by ideas of technical efficiency, which may include speed and accuracy, do not take ethical issues into account, and in the case of AI systems that affect human lives, this can no longer be tolerated. The challenge for technologists is to create systems that are not only productive but also fair, transparent, and accountable, and this requires holistic thinking about bias, privacy, and accountability.

Analysis of recent studies. Current research focuses on key ethical aspects of AI: bias, privacy, and accountability, proposing new methods for creating responsible systems. C. Bura, S. Kamatala, and P. K. Myakala [1, pp. 2–3] emphasize the complexity of ethical challenges in data science, noting that algorithms can discriminate due to biased data, for example, in insurance risk assessment systems, where certain groups receive unfairly high rates. M. R. K. Kakarala and S. K. Rongali [2, pp. 549-550] systematize the problems of ethical AI, including algorithmic bias, transparency, accountability, and regulatory compliance, emphasizing that these aspects are interrelated: transparency helps to identify bias but requires additional resources for documentation. H. K. Vuyyuru [3, pp. 1416-1417] proposes a framework for the responsible development of generative AI that creates text, images, or audio, emphasizing the need to integrate ethical principles at an early stage to avoid problems with misinformation or copyright infringement, for example, in synthetic media that imitates real people. S. A. Atmaja [4, pp. 620-621] explores fairness in algorithmic decision-making, noting that it depends on cultural context: an algorithm that is fair in Europe may be biased in Asia due to differences in social norms. S. Muvva [5, pp. 1-2] develops a technical framework for ethical AI, proposing techniques such as differential privacy, which adds noise to data to protect privacy, which is critical for large data pipelines in banking systems. L. Barbudo Carrasco [6, pp. 1-2] emphasizes the role of metadata in mitigating bias, proposing detailed documentation of data sources to identify biases, for example, in medical datasets where older people are underrepresented. A. Chowdhary [7, pp. 1747-1748] proposes a "privacy-first" architecture, including federated learning, which allows models to be trained on local devices without transferring data, which is key for sensitive areas such as cancer diagnosis. A. K. Sharma, D. Kyosev, and J. Kunkel [8, pp. 1-2] have developed a standardized framework for evaluating ethical AI, which includes metrics for transparency and accountability, such as the SHAP and LIME explainability methods. R. Grover [9, pp. 401-402] analyzes ethical challenges in AI and robotics, noting that AI's interaction with physical equipment, such as in driverless cars, creates dilemmas such as choosing priorities in emergency situations. L. K. BD [10, pp. 1-2] explores the implementation of ethical AI in the IT industry, proposing regular audits of algorithms to ensure fairness in corporate systems, such as automated HR platforms. N. Petrukha, I. Demydonok, and O. Hubanov [11, pp. 4-5] consider the ethical aspects of technology in the context of social challenges, such as the implementation of AI in infrastructure restoration, emphasizing the adaptation of ethical principles to local conditions.

Results

In 2025, ethical issues related to the development of artificial intelligence (AI) will remain among the most pressing, as these technologies will be widely used in healthcare, finance, justice, transportation, and marketing. These issues include bias in data sets, data privacy, and, ultimately, the lack of accountability in decision-making systems. Adding to these problems is the emergence of generative AI, which creates its own ethical issues related to misinformation, legal protection of identified copyrighted objects, and impact on the socio-economic context. Responsible AI must be developed by actively addressing bias issues and making it transparent and accountable in order to balance technical efficiency and social responsibility. Real-world examples from 2025 demonstrate the high level of these challenges and insist on the inclusion of ethical principles in AI development to adapt it to societal demands and regulatory requirements.

Algorithmic bias remains one of the most pressing issues in AI development, as systems can unfairly treat certain groups by reproducing historical patterns of discrimination or reinforcing stereotypes through unrepresentative data. In 2025, a credit risk assessment system developed by a major European bank was suspended after regulators discovered its tendency to assign lower ratings to women based on historical data showing that they were less likely to obtain loans [1, pp. 4-5]. This incident prompted an investigation into the system's compliance with fairness standards, highlighting the need for careful analysis of training data to avoid systematic discrimination. L. Barbudo Carrasco notes that metadata documenting data sources, collection methods, and characteristics are key to identifying biases [6, pp. 1-16]. For example, in medical systems for diagnosing heart disease in 2025, the underrepresentation of older people in datasets led to reduced prediction accuracy for this demographic group, requiring the use of stratified sampling and synthetic data to fill the gaps [6, pp. 1-16]. Another example from 2025 concerns a content recommendation system in Asia that reinforced cultural stereotypes due to local data characteristics, causing public discontent and forcing developers to revise algorithms to ensure fairness [2, pp. 549-554]. To mitigate such problems, developers are engaging diverse interdisciplinary teams that include ethicists and sociologists to identify hidden biases early in development, and are applying data rebalancing techniques to ensure representativeness.

Data privacy remains a critical issue in 2025 due to the need for AI systems to access large amounts of personal data, which creates risks of information leaks and privacy violations. In 2025, there was a high-profile case in the US where a misconfigured AI medical system database led to the leak of information about millions of patients, raising questions about data security in sensitive areas [7, pp. 1747-1755]. Amber Chowdhary emphasizes the importance of the principles of data minimization, pseudonymization, and federated learning, which allows models to be trained on local devices without transferring data to a central server [7, pp. 1747-1755]. For example, in European hospital cancer diagnosis systems in 2025, patient data is processed locally, which reduces the risk of information leaks and ensures compliance with regulatory requirements [7, pp. 1747-1755]. Global regulations, such as the General Data Protection Regulation (GDPR) in Europe and the Act on the Protection of Personal Information in Japan (APPI), continue to shape AI development practices, forcing companies to implement differential privacy, which adds noise to data to make it impossible to identify individuals [7, pp. 1747-1755]. In 2025, personalized recommendation systems on social networks actively use differential privacy to protect user data while maintaining the quality of recommendations [7, pp. 1747-1755]. In addition, in 2025, the Executive Order on Removing Barriers to Leadership in AI in the United States eased some regulatory restrictions, but at the same time emphasized the need for transparent data protection practices to prevent abuse [1, pp. 4-5].

Accountability in decision-making systems is a key aspect of ethical AI, as opaque decisions make it difficult to determine responsibility and undermine trust in the systems. In 2025, a recidivism prediction system in the US justice system led to lawsuits over racial discrimination because its decisions remained opaque to users and regulators [4, pp. 620-627]. This case highlighted the need for explainability methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which allow the logic behind model decisions to be revealed [8, pp. 1-8]. For example, in credit scoring systems in 2025, SHAP and LIME methods are used to explain which factors, such as income or credit history, influenced the decision, facilitating auditing and increasing trust [8, pp. 1-8]. In the field of robotics, in 2025, an unmanned vehicle in Europe faced an ethical dilemma when the system was unable to explain its choice of priorities in an emergency situation, leading to public debate about the need for transparency and accountability [4, pp. 620-627]. A. K. Sharma and co-authors propose an evaluation framework that includes traceability, documenting all stages of development, from data collection to model deployment, to ensure accountability [8, pp. 1-8]. In 2025, companies such as Microsoft and Telefonica, through the UNESCO AI Ethics Council, developed an ethical impact assessment tool that helps document decisions and increase trust in systems [1, pp. 4-5]. Transparency is achieved through detailed documentation of data sources, algorithms, their limitations, and risks, which facilitates external control and promotes compliance with global ethical standards [5, pp. 1-8]. Regular monitoring of systems for bias and fairness is necessary because models can become biased due to changes in data, as was the case with content recommendation systems in Asia in 2025, which reinforced cultural stereotypes, requiring immediate audit and correction [2, pp. 549-554].

Generative AI in 2025 has created new ethical challenges, particularly related to misinformation and copyright protection. For example, Disney and Universal's lawsuit against Midjourney in 2025 for alleged copyright infringement in image generation raised questions about the legal limits of AI use in creative industries [1, pp. 4-5]. In addition, in 2025, experts discovered a vulnerability in ChatGPT that allowed censorship to be bypassed in 20 minutes, leading to the generation of harmful content and demonstrating the need to strengthen model security [1, pp. 4-5]. These cases underscore the importance of implementing tools such as Nvidia's Llama Guard, which ensure security and ethical compliance in generative models [3, pp. 1416-1423]. To create ethical AI, developers must integrate principles of fairness, transparency, and accountability from the outset, conducting assessments at all stages of development [3, pp. 1416-1423]. Diverse teams that include people of different ages, genders, races, and professional backgrounds are better at identifying bias, as demonstrated in hiring systems in 2025, where algorithms rejected candidates with non-standard resumes due to a lack of diversity in the data [9, pp. 401-405].

Table 1 illustrates the differences between traditional and ethical approaches to AI development in 2025, highlighting a shift in priorities toward fairness, transparency, and accountability. It serves as a practical guide for developers, demonstrating how to integrate ethical principles to create systems that meet societal expectations and regulatory requirements in the context of today's challenges.

Table 1 demonstrates the need for a paradigm shift in AI development in 2025, where ethical principles become an integral part of the process. The traditional approach, focused solely on technical performance, does not take into account social and ethical implications, which can lead to bias, opacity, and loss of trust. An ethical approach integrates fairness, transparency, and accountability, which is in line with global standards such as the UNESCO Recommendation on AI Ethics [1, pp. 4-5]. For example, in 2025, companies that ignore comprehensive bias testing face regulatory fines and public criticism, as in the case of the credit risk assessment system [1, pp. 4-5].

Table 1.

Comparison of traditional and ethical approaches to AI development

Aspect	Traditional approach	Ethical approach
Development focus	Maximizing accuracy, speed, and efficiency	Balance of productivity, fairness, transparency, and ethics
Testing	Verification of technical parameters such as accuracy and speed	Comprehensive testing, including assessment of bias, fairness, and ethical risks
Data	Collecting the maximum amount of data to improve performance	Data minimization, pseudonymization, anonymization, use of federated learning
Team	Homogeneous technical teams with a focus on engineering skills	Diverse interdisciplinary teams, including technologists, ethicists, sociologists, and lawyers
Documentation	Technical specifications of algorithms and systems	Comprehensive documentation including data sources, algorithms, limitations, and ethical assessments
Monitoring	Verification of performance and technical metrics	Continuous monitoring of bias, fairness, transparency, and social impact
Explainability	Non-transparent algorithms, "black box"	Explainable models, accessible for audit, using SHAP and LIME methods

Figure 1 illustrates the key stages of ethical AI system development in 2025, emphasizing a continuous process that includes six components: planning, data collection, algorithm development, testing, deployment, and monitoring. The model emphasizes the integration of ethical principles at each stage to ensure fairness, transparency, and accountability in the context of today's technological and regulatory challenges.

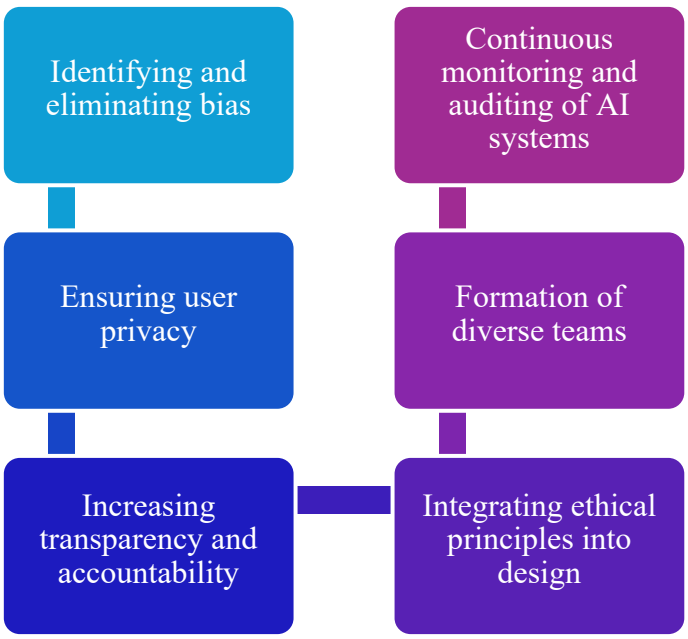

Figure 1. Key stages of ethical AI development

This model emphasizes once again that developing ethical artificial intelligence in 2025 is an ongoing process that requires interdisciplinary collaboration. Best practices suggest that, at the planning stage, developers set ethical goals and assess potential risks, including bias or privacy violations. Data collection rules and principles focus on representativeness and data protection. There is an explanatory approach to algorithm development, bias testing, and a way to deploy and monitor ongoing oversight. This approach allows for the creation of systems that are standardised in terms of both technology and ethics, thereby gaining public trust. The case of lawsuits against Midjourney for copyright infringement or the ChatGPT exploit in 2025 proves why such an approach is necessary [1, pp. 4-5].

Conclusions

Developing ethical AI is a complex task that requires a systematic approach to addressing issues of bias, privacy, and accountability, integrating these principles at all stages of system development. Algorithmic bias arising from historical data, insufficient representativeness, or validation mechanisms can be reduced through careful documentation of data sources and regular audits to identify biases, such as in credit scoring systems that unfairly lower ratings for women. Data privacy, shaped by global regulatory requirements such as GDPR, requires the implementation of data minimization, pseudonymization, and federated learning principles that protect users without sacrificing performance, for example, in medical cancer diagnosis systems where data remains local. Accountability is achieved through explainability of decisions, traceability of processes, and clear definition of responsibility, enabling the creation of transparent systems, such as in the justice system, where recidivism prediction algorithms must be understandable to judges and citizens. Technologists play a key role in forming diverse teams that are better at identifying biases, such as in hiring systems, where diversity helps avoid gender or racial stereotypes. Transparency through detailed documentation facilitates auditing and increases trust, as in the case of recommendation systems that require a clear explanation of the logic behind decisions. Interdisciplinary collaboration with ethicists, sociologists, and lawyers ensures that AI is adapted to cultural and social contexts, such as in infrastructure recovery systems, where local characteristics influence ethical requirements. The practical value of this research lies in providing clear recommendations to technologists: from integrating ethical design to regularly monitoring systems for bias and fairness. Future research should focus on developing automated tools for detecting bias, improving decision explanation methods such as SHAP and LIME, and creating innovative approaches to privacy protection so that AI meets the diverse needs of the global community.

References

1. Bura, C., Kamatala, S., & Myakala, P. K. (2025). Ethical challenges in data science: Navigating the complex landscape of responsibility and fairness. International Journal of Current Science Research and Review, 8(3). https://doi.org/10.47191/ijcsrr/v8-i3-09
2. Kakarala, M. R. K., & Rongali, S. K. (2025). Existing challenges in ethical AI: Addressing algorithmic bias, transparency, accountability and regulatory compliance. World Journal of Advanced Research and Reviews, 25(3), 549–554. https://doi.org/10.30574/wjarr.2025.25.3.0554
3. Vuyyuru, H. K. (2025). Navigating ethical dilemmas in generative AI development: A framework for responsible innovation. International Journal of Science and Research Archive, 14(1), 1416–1423. https://doi.org/10.30574/ijsra.2025.14.1.0166

4. Atmaja, S. A. (2025). Ethical considerations in algorithmic decision-making: Towards fair and transparent AI systems. Riwayat: Educational Journal of History and Humanities, 8(1), 620–627. https://doi.org/10.24815/jr.v8i1.44112
5. Muvva, S. (2025). Ethical AI and responsible data engineering: A framework for bias mitigation and privacy preservation in large-scale data pipelines. International Journal of Scientific Research in Engineering and Management, 9(1), 1–8. https://doi.org/10.55041/ijsrem10633
6. Barbudo Carrasco, L. (2025). Mitigating bias and advocating for data sovereignty: The role of metadata and paradata in ethical AI-driven information systems. Journal of Library Metadata, 1–16. https://doi.org/10.1080/19386389.2025.2515766
7. Chowdhary, A. (2025). Implementing privacy-first architecture: A technical guide to ethical data pipelines and AI systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(1), 1747–1755. https://doi.org/10.32628/cseit251112153
8. Sharma, A. K., Kyosev, D., & Kunkel, J. (2025). Ethical AI: Towards defining a collective evaluation framework (Version 1). arXiv. https://doi.org/10.48550/arXiv.2506.00233
9. Grover, R. (2025). Ethical challenges in AI and robotics: Balancing innovation and responsibility. International Research Journal on Advanced Engineering and Management, 3(3), 401–405. https://doi.org/10.47392/irjaem.2025.0064
10. K, B. Dr. L. (2025). Ethical AI in the IT industry: Addressing bias, transparency, and accountability in algorithmic decision-making. International Journal of Scientific Research in Engineering and Management, 9(3), 1–9. https://doi.org/10.55041/ijsrem42652
11. Petrukha, N., Demydonok, I., & Hubanov, O. (2024). Ethical aspects of bioeconomy in post-war reconstruction projects in Ukraine. Economics, Finance and Management Review, (4), 4–17. https://doi.org/10.36690/2674-5208-2024-4-4-17
12. Kiselyov, O. V. (2025). Overview of fundamental machine learning algorithms and their applications. SWorldJournal, 32(1), 91–104. https://doi.org/10.30888/2663-5712.2025-32-01-029