INTERNATIONAL JOURNAL OF STATISTICAL ANALYSIS

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

AI ETHICS: FAIRNESS, TRANSPARENCY AND ACCOUNTABILITY CHALLENGES

Jomonqulova Fazilat Esirgapovna

Associate Professor, Department of Information Technologies, Samarkand Institute of Economics and Service.

Alisher Sattorov Joshqin ugli

student, group BH-526, Samarkand Institute of Economics and Service.

Abstract

This article presents a theoretical analysis of the three core principles of AI ethics: fairness, transparency and accountability. It examines racial, gender and social forms of algorithmic bias and its sources at the data, modeling and deployment stages. The article also discusses the “black box” problem, the interpretability of decisions (Explainable AI) and their practical and legal limits. In addition, it justifies the “Ethical by design” principle, which builds ethical requirements into a technology’s life cycle from the very beginning, and its use in education and other high-risk areas. The main conclusion is that AI ethics is not a restriction added after the technology is built, but a basic precondition for a system’s reliability, legality and social acceptability.

Keywords

AI ethics, fairness, algorithmic bias, transparency, accountability, “black box”, Explainable AI (XAI), Ethical by design

Introduction. Artificial intelligence (AI) systems today take part in decisions in areas that directly affect people’s lives, such as lending, recruitment, medical diagnosis, criminal risk assessment and the analysis of student performance in education. The social impact of this technology raises not only technical questions but also normative ones: Who does the algorithm treat fairly? To whom, and how, is its decision explained? Who is responsible when an error occurs? For this reason, AI ethics should not be seen as a “moral decoration” added to technology from outside, but as a condition for the system’s reliability and for its acceptance by society.

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

At the international level, this field is increasingly taking a legal form. The 2021 UNESCO Recommendation on the Ethics of Artificial Intelligence, adopted by its member states, sets out human rights, fairness, transparency and human oversight as core values. The European Union's 2024 AI Act classifies systems by risk level, places applications in areas such as education, recruitment and law enforcement in the high-risk category, and requires transparency, human oversight and documentation for them. An analysis by Jobin and colleagues shows that, among more than one hundred AI ethics documents produced by different countries and organizations, there is considerable convergence on the principles of transparency, justice, non-maleficence, responsibility and privacy, but disagreement remains over how to interpret and implement them. The aim of this article is to explain the content of these principles, to analyze the mechanisms through which they are violated in practice, and to justify ways of ensuring them at the design stage.

The topic is especially relevant for Uzbekistan. The Strategy for the Development of Artificial Intelligence Technologies until 2030, approved by Presidential Resolution No. PQ-358 of 14 October 2024, names compliance with ethics and data security as an important direction alongside the introduction of the technology. The national recommendations on ethics in the use of artificial intelligence also set out principles such as legality, human oversight, fairness and non-discrimination, openness and transparency, and responsibility and accountability. The issue of AI and morality is also being discussed in Uzbek academic literature. The task, therefore, is not limited to expanding the use of AI: it is also necessary to make sure that its use is fair, understandable and subject to accountability.

Fairness, transparency and accountability: a triad of principles. These three principles are interconnected, but they cannot be reduced to one another. Fairness means that a system's decisions do not differ without justification between groups and individuals and, where appropriate, that equality and equal opportunity are ensured. Transparency means making clear and open to stakeholders the purpose for which the system works, the data it relies on and the logic it follows. Accountability

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

means that specific parties are responsible for the consequences of the system's impact and that a person who has been harmed can object and demand a remedy. They depend on each other in the following way: without transparency, unfairness cannot be detected; if unfairness is detected but no responsible party is identified, it cannot be corrected. Conversely, without an accountability mechanism, transparency ends with simply providing information. Floridi and colleagues describe AI ethics as a framework of principles of beneficence, non-maleficence, autonomy, justice and explicability, in which explicability serves as an "enabling" principle that makes it possible to check the other principles in practice. In this sense, the three principles form a whole: fairness is the goal, transparency is the means of verification, and accountability is the guarantee of correction.

The problem of algorithmic bias: racial, gender and social inequality. Algorithmic bias means a systematic and unjustified deviation of an AI system's results against certain groups or individuals. According to a common misconception, an algorithm is neutral because it is mathematical. In fact, a model is a product of the data it learns from, the objective it is given and the conditions in which it is used; therefore, it can repeat and even strengthen existing inequalities in society. In the systematic review by Mehrabi and colleagues, sources of bias are divided into those in the data, in the algorithm and in the interaction with users. Their main forms are presented below.

The first is historical bias. If data reflect past decisions, the model learns the inequality in those decisions as a "norm". For example, if a recruitment system is trained on résumés of people hired in the past, who were mostly of one gender, it may learn this historical choice as a "sign of success" and rate candidates of the other gender lower. In such a case, the root of the problem is not a technical defect but the social history contained in the data.

The second is representation and sampling bias. If the training data do not cover some groups sufficiently, the model works inaccurately for them. In the "Gender Shades" study by Buolamwini and Gebru, the error rate of commercial facial

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

recognition systems reached about 35 percent for darker-skinned women, but was below one percent for lighter-skinned men. Although overall accuracy was high, the large difference between groups was hidden. This is also a risk for systems that work in relatively low-resource languages such as Uzbek: if the data do not sufficiently reflect language, region or social group, the system may work poorly for some users.

The third is measurement and proxy bias. When a system uses a substitute indicator instead of a quantity that cannot be measured directly, this indicator may be wrong for some groups. Obermeyer and colleagues found that a widely used healthcare algorithm in the United States took healthcare costs as a proxy instead of the severity of a patient's illness. Because Black patients spent less on medical care, at the same risk score they were actually sicker, yet they were referred to additional programs less often [10]. The problem is not that the algorithm knew race, but that socio-economic conditions entered the model unnoticed through costs.

The fourth is deployment and institutional bias. Even a technically sound model gives unfair results when it is used in the wrong environment. A ProPublica analysis of COMPAS, a system used in the United States to assess the risk of criminal reoffending, showed that Black defendants were more often wrongly labeled "high risk". In 2020 in the United Kingdom, when exams could not be held because of the pandemic, an algorithm used to standardize grades lowered many students' grades compared with their teachers' predictions, and it affected pupils from historically low-performing schools the most. After public protests, it was withdrawn. This example shows how dangerous it is in education when an individual's achievement is replaced by a school's past results.

There is another theoretical difficulty in eliminating bias: there is no single definition of fairness. As theoretical studies have shown, when the base rates differ between groups, criteria such as calibration and equal error rates cannot be fully satisfied at the same time. Therefore, choosing a "fair model" is not only a technical decision but a value-based one: which error has the heavier social cost must be decided in advance, through open discussion. Discriminatory effects often arise not

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

from intent but from design decisions that look “neutral”.

The “black box” problem and the interpretability of decisions (Explainable AI). The term “black box” refers to systems whose inputs and outputs are known but whose internal logic is not understandable to humans. Deep neural networks have millions or even billions of parameters, and how they reach a particular decision cannot be shown as a simple chain of logic. At least three sources of opacity can be distinguished: secrecy linked to commercial or state secrets, users’ lack of technical literacy, and the complexity of the algorithm itself. This distinction is important because each requires a different solution: the first requires legal openness, the second requires education, and the third requires technical explanation methods. Explainable AI (XAI) emerged as a response to this problem and has two main directions. The first is models that are understandable from the start (inherently interpretable), such as decision trees, linear models and rule-based models. The second is explanation methods applied after the fact to a complex model: LIME shows the factors that influenced a single decision through local approximation; SHAP estimates the contribution of each feature based on the Shapley value from game theory; and counterfactual explanations answer the question “What would have had to change for the decision to be different?” The counterfactual approach is especially useful: even without opening the model’s internal structure, it makes clear to a person why a decision was made and how it can be changed.

However, a critical approach to interpretability is necessary. Rudin argues that in high-stakes areas (healthcare, courts) it is better to build an interpretable model from the start than to try to explain a black box afterwards, and that the widespread idea of an “inevitable conflict between accuracy and interpretability” is unfounded in many cases. Post-hoc explanations may not fully reflect the model’s real logic; that is, they may give an explanation that is “plausible but wrong”. It also matters whom the explanation is for: a technical explanation suitable for a developer will not be clear to a student, a patient or a defendant. Therefore, the quality of an explanation is measured not only by technical correctness but also by how useful it is for its

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

audience and whether it allows the decision to be reasonably challenged. The legal and regulatory direction is also clear: the national ethical recommendations in Uzbek name "understandability", "openness and transparency" and "constant human oversight" as separate principles, and the EU regulation makes transparency and human oversight mandatory for high-risk systems. So, for systems that influence important decisions, the answer "the system said so" is not enough.

Accountability mechanisms. Unlike transparency, accountability answers the question "Who is responsible, and how is the problem fixed?" In practice, the biggest risk is the "accountability gap", that is, when the distribution of responsibility among the developer, the deployer and the user remains unclear. The UNESCO Recommendation also lists responsibility and accountability, that is, ensuring audit and human oversight throughout the system's life cycle, as a separate requirement, and the national recommendations also include "responsibility and accountability" as an independent item. In practice, standard documentation for datasets and models (for example, datasheets and model cards), independent audits, incident registers, and the right of an affected person to appeal and to have a decision reviewed are very important. Most importantly, final responsibility for important decisions must never be placed on the algorithm; it must remain with a specific person or organization.

The "Ethical by design" principle. Adding ethical requirements as a "fix" after a system is ready is usually ineffective and expensive. The Ethical by design approach changes this logic and proposes including ethical requirements among the technical requirements from the earliest stage of a project. The approach is inspired by the concept of "Privacy by Design", whose main idea is to build protection into the architecture itself as the default rather than adding it later. In practice, this principle applies to the whole life cycle of a system; the table below summarizes its main stages.

Table 1. Ethical by design: ethical questions and tools at each stage of the life cycle

Stage	Key ethical question	Tools used

https://creativecommons.org/licenses/by/4.0/

INTERNATIONAL JOURNAL OF STATISTICAL ANALYSIS
ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

1. Defining requirements	Who is affected? Whose rights and interests are at risk?	Stakeholder analysis, ethical impact assessment
2. Data collection	Whom does the data cover, and whom does it leave out?	Datasheets for Datasets, representativeness check
3. Modeling	Are the objective function and proxy variables fair?	Fairness constraints, choosing an interpretable model
4. Testing and validation	How does the error rate differ between groups?	Evaluation by group, Model Cards
5. Deployment	Is there human oversight and a way to object?	Human-in-the-loop, appeal procedure
6. Monitoring	Has the system's behavior changed over time?	Independent audit, incident register, re-evaluation

As the table shows, ethical oversight is not a one-time check but a continuous process. One important aspect is the risk-based approach: the more serious a system's impact is (for example, assessing a student's "academic risk" or distributing scholarships), the stricter the oversight should be. The EU regulation also classifies systems by risk level in this way and sets requirements in proportion to risk. Another important condition is the participation of stakeholders, that is, the people who will be affected by the system: an educational algorithm created without the voice of teachers, students and parents may not be socially accepted, even if it is technically perfect.

Significance for Uzbekistan's education system. Because AI decisions in education (admission, scholarships, assessment, identifying "risk groups") directly affect a person's future, ethical requirements must be especially strict. Analytical modules in digital platforms such as HEMIS can be useful for monitoring student performance, but they should remain advisory, and the final decision should be made by a person on the basis of transparent criteria. In addition, national legislation on the protection of personal data requires that the processing of student data follow the

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

principles of purpose limitation and data minimization. Increasing the representation of Uzbek-language data across regions and social groups is also an important condition for reducing the risk of bias.

Conclusions and recommendations

The analysis shows that the AI ethics principles of fairness, transparency and accountability complement each other, and neglecting one of them reduces the effectiveness of the others. Algorithmic bias should be seen not as a technical defect but as an echo of social inequalities in the data, the objective function and the deployment environment; and choosing a balance between different definitions of fairness is a value-based decision, not a technical one. Interpretable methods only partly ease the “black box” problem, but in high-risk areas it is necessary to give priority to inherently interpretable models and to check how useful explanations really are. The Ethical by design approach makes ethics an integral part of the system’s life cycle and allows these problems to be solved in the cheapest and most effective way.

Scientific, theoretical and practical recommendations:

1. For AI systems used in high-risk areas (education, recruitment, finance, healthcare), a mandatory ethical impact assessment and risk classification procedure should be introduced before deployment.
2. A requirement for standard documentation of datasets and models (datasheets, model cards) should be established; evaluation should be carried out by group, not only by overall accuracy.
3. For important decisions (grades, scholarships, admission, disciplinary measures), final responsibility should remain with a human; an affected person should be guaranteed the right to demand an explanation of the decision and to file an appeal.
4. In high-risk applications, priority should be given, wherever possible, to inherently interpretable models; when complex models are used, explanation methods

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

(for example, counterfactual explanations) should be presented in a form that users can understand.

5. An independent audit and monitoring mechanism should be created, and a register of detected errors and incidents should be kept.

6. The creation of datasets in Uzbek that are representative across regions should be supported; training programs for teachers and specialists on ethics, AI literacy and data protection should be expanded.

It is clear that the strength of a technology is measured by its reliability. Only fair, understandable and accountable artificial intelligence will earn society's trust and bring long-term benefits.

References:

1. Law of the Republic of Uzbekistan "On Personal Data", 2 July 2019, No. ZRU-547 (in Uzbek). Website: lex.uz.

2. UNESCO. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO, 2021. Website: unesdoc.unesco.org.

3. Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act). Official Journal of the European Union, 2024. Website: eur-lex.europa.eu.

4. Jobin A., Ienca M., Vayena E. The global landscape of AI ethics guidelines // Nature Machine Intelligence. 2019. Website: nature.com.

5. Resolution of the President of the Republic of Uzbekistan of 14 October 2024, No. PQ-358, "On approving the Strategy for the development of artificial intelligence technologies until 2030" (in Uzbek). Website: lex.uz.

6. Recommendations on ethics in the use of artificial intelligence technologies (in Uzbek). Website: ai.gov.uz.

7. Primov B. A. Artificial intelligence and morality in the information society (in Uzbek). 2025. Pp. 312–317. Website: oriens.uz.

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 30.10.2026
PUBLISHED: 31.10.2026

8. Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A. A Survey on Bias and Fairness in Machine Learning // ACM Computing Surveys. 2021. Vol. 54, No. 6. Website: dl.acm.org.

9. Buolamwini J., Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Geader Classification // Proceedings of Machine Learning Research. 2018. Website: proceedings.mlr.press.

10. Obermeyer Z., Powers B., Vogeli C., Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations // Science. 2019. Website: science.org.

11. Angwin J., Larson J., Mattu S., Kirchner L. Machine Bias // ProPublica. 2016, 23 May. Website: propublica.org.

12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // Nature Machine Intelligence. 2019. Website: nature.com.

https://creativecommons.org/licenses/by/4.0/