STM JOURNALS

International Journal of
Information Security Engineering

ISSN: 3049-1150
Volume 4, Issue 1, 2026
January–June
DOI (Journal): 10.37591/IJISE

https://journals.stmjournals.com/ijise

Review IJISE

Ethical and Responsible AI: A Comprehensive Review of Principles, Methods, and Tools

Divyansh Rajat[1,*], Akshita[2], Jaspreet Kaur[3]

Abstract

Quick development of artificial intelligence (AI) has revolutionized a number of industries, including healthcare, banking, and government, by providing creative answers to challenging issues. However, there are serious ethical issues with growing integration of AI into crucial decision-making processes, including prejudice, a lack of transparency, abuses of data privacy, and accountability gaps. A systematic strategy that incorporates technical solutions, legal frameworks, and ethical standards is needed to address these issues. With an emphasis on fundamental concepts like equity, responsibility, transparency, privacy, and inclusion, this study offers a thorough analysis of ethical AI. It examines the strategies and resources created to guarantee the responsible application of AI, such as explainability tactics, algorithms that improve fairness, and privacy-preserving measures like federated learning and differential privacy. The study also looks into federated learning frameworks that support ethical compliance as well as AI governance solutions like AI Fairness 360, SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and others. Notwithstanding developments, problems, including moral trade-offs, scalability problems, and inconsistent regulations, still exist. In order to create strong, moral AI systems, this study emphasizes the value of interdisciplinary cooperation between academics, decision makers, and business executives. Improving AI interpretability, tackling socio-technical biases, and promoting international collaboration on AI ethics are some future research avenues. In order to ensure justice, security, and accountability in AI-driven applications, stakeholders seeking to align AI technologies with ethical norms may find this paper to be a useful resource.

Keywords: AI Governance, ethical AI, explainability, fairness, privacy-preserving AI, responsible AI

INTRODUCTION

***Author for Correspondence**
Divyansh Rajat
E-mail: divyanshrajat1999@gmail.com

[1]Student, Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India
[2]Student, Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India
[3]Assistant Professor, Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India

Received Date: April 11, 2025
Accepted Date: June 18, 2025
Published Date: March 27, 2026

Citation: Divyansh Rajat, Akshita, Jaspreet Kaur. Ethical and Responsible AI: A Comprehensive Review of Principles, Methods, and Tools. International Journal of Information Security Engineering. 2026; 4(1):23–34p.

By automating decision-making procedures and streamlining operations, artificial intelligence (AI) has become a game-changing technology that is impacting industries like healthcare, banking, education, and governance [1]. Even though AI has many advantages, there are serious ethical issues associated with its growing use in high-stakes applications, such as hazards to privacy, accountability, transparency, and fairness. By ensuring that AI systems function in a manner that is consistent with legal requirements, human values, and social norms, ethical and responsible AI seeks to address these problems [2].

Algorithmic bias, which occurs when AI models trained on biased datasets perpetuate prevailing societal imbalances, is one of the most urgent issues in AI ethics [3]. Research has shown that when identifying members of marginalized groups, facial

recognition systems such as those employed by law enforcement have greater error rates [4]. Underrepresented minorities have been disproportionately impacted by biased AI systems in lending and hiring [5]. Researchers have created fairness-enhancing algorithms, such as adversarial debiasing and re-weighting approaches, to lessen these biases and encourage equitable decision-making [6].

It is challenging to understand the decision-making processes of many AI systems, particularly deep learning models, because they function as "black boxes" [7]. In crucial industries, such as healthcare and banking, where explainability is crucial for both user confidence and regulatory compliance, this lack of openness erodes trust [8]. To help stakeholders comprehend and validate algorithmic outputs, techniques such as Shapley Additive Explanations (SHAP) and LIME (Local Interpretable Model-Agnostic Explanations) have been developed to offer insights into AI model behavior [9].

Large volumes of data are frequently required for AI applications, which raises questions regarding data security and user privacy. Techniques such as federated learning and differential privacy have become popular as ways to reduce data exposure while preserving AI performance because of increased regulatory scrutiny [10]. While federated learning allows AI models to be trained on decentralized data sources without centralizing sensitive information, differential privacy guarantees that individual user data in datasets stays anonymous [11]. Maintaining ethical AI methods requires adherence to legislative frameworks such as the California Consumer Privacy Act (CCPA) and General Data Protection Regulation (GDPR) [12].

Another significant ethical dilemma is determining who is responsible for decisions made by AI. It is also unclear who should be held accountable when AI systems generate inaccurate or detrimental results, such as developers, data sources, or end-users [13]. To create accountability, a number of organizations and regulatory agencies have put out AI governance frameworks such as the OECD AI Principles, IEEE's Ethically Aligned Design, and the European Union's AI Act [14]. These recommendations place strong emphasis on the necessity of risk assessment procedures, human supervision, and unambiguous ethical standards while implementing AI.

Even with increased understanding, putting ethical AI into practice remains difficult. Many businesses find it difficult to strike a compromise between performance efficiency and privacy, transparency, and fairness [15]. Furthermore, consistent evaluation measures to evaluate AI's interpretability, fairness, and adherence to ethical standards are lacking [16]. Establishing global AI ethics norms is made more difficult by cultural and geopolitical variations, underscoring the necessity for international cooperation [17].

This study offers a thorough analysis of ethical AI concepts, practices, and resources, while looking at the most recent developments in explainability, bias reduction, privacy-preserving AI, and regulatory compliance. Several ethical AI frameworks have been compared, emphasizing their advantages, disadvantages, and usefulness. To promote responsible AI growth, this paper also addresses new trends, persistent issues, and potential research avenues. This highlights the necessity of multidisciplinary cooperation between academics and decision makers.

LITERATURE REVIEW

Ethical Principles in AI

Adherence to fundamental principles, such as equity, accountability, transparency, privacy, inclusivity, and safety, is necessary for the ethical development and application of AI. Guidelines for governing AI governance have been produced by international organizations such as the European Union (EU), IEEE, and the Organization for Economic Co-operation and Development (OECD) [1, 2]. Ensuring fairness is essential to prevent discriminatory outcomes, as AI models often inherit biases from historical data, leading to unfair treatment of marginalized groups [3]. Studies have shown that facial recognition systems exhibit higher error rates for darker-skinned individuals and women, while AI-driven hiring tools have favored male candidates owing to historical employment patterns [4, 5].

Although pre-processing, in-processing, and post-processing techniques are among the strategies for mitigating bias, tools such as Microsoft's FairLearn and IBM's AI Fairness 360 aid in detecting and mitigating bias [6, 7]; attaining total fairness is still difficult because of competing definitions of fairness, such as equalized odds and demographic parity [8].

Although explainability and transparency are essential for fostering trust in AI, deep learning models frequently operate as "black boxes," making it challenging to understand how they make decisions [9]. Concerns have been raised by a lack of transparency in industries, including healthcare and finance [10]. AI decision-making is clarified by methods like SHapley Additive exPlanations(SHAP) and LIME [11, 12]. By presenting alternate inputs that would produce different results, counterfactual explanations further enhance comprehension [13]. However, increasing the explainability can reduce the model accuracy, creating a trade-off that must be managed [14].

Privacy is a major concern because AI systems rely on large datasets that contain sensitive user information. Guidelines for appropriate data management are established by laws such as the CCPA and GDPR [15, 16]. Techniques that improve privacy include homomorphic encryption, which enables computations on encrypted data without decryption [18]; federated learning, which trains models on decentralized data [19]; and differential privacy, which adds statistical noise to datasets [17]. Notwithstanding these developments, privacy-preserving AI techniques frequently result in computational inefficiencies that affect model performance [20].

Because liability frameworks for AI-driven decisions are still unclear, accountability is still a major concern, especially in fields such as driverless cars and AI-powered medical diagnostics [21]. Algorithmic Impact Assessments (AIA), which assess risks and biases prior to deployment, AI audits, which independently evaluate adherence to ethical standards, and human-in-the-loop (HITL) systems, which necessitate human oversight in AI decision-making, are some examples of proposed governance models [22–24]. Regulatory bodies, including the EU AI Act, OECD AI Principles, and IEEE Ethically Aligned Design, emphasize the need for clear AI governance and risk assessment frameworks [25].

AI systems need to be developed to serve a variety of demographics. Models that are unable to generalize across various demographic groups owing to a lack of variety in training datasets provide biased findings [26]. Medical disparities could be exacerbated, for instance, if healthcare AI models trained on Western populations perform poorly in non-Western regions [27]. Ensuring diverse datasets, integrating cultural sensitivity into AI development, and including stakeholders from different communities in decision-making are all necessary to address this issue [28, 29].

Safety and security are essential for preventing unintended harm from AI systems. Adversarial attacks, in which malicious inputs trick AI models, present serious security risks [30]. Small changes in images can lead to deep learning models misclassifying objects, which can lead to vulnerabilities in applications such as facial recognition and autonomous driving [31]. To improve AI security, adversarial defense mechanisms, ethical hacking, and red teaming to test AI vulnerabilities and create strong AI security standards [32–34] are necessary.

Assuring fairness, transparency, accountability, inclusivity, and security will allow AI to benefit society while minimizing the risks associated with its widespread adoption. Ethical AI is not just a regulatory requirement but a fundamental necessity for sustainable and responsible technological advancement. By incorporating these ethical principles, AI can be developed and deployed responsibly to balance innovation with societal well-being (Table 1).

Methods for Ensuring Ethical AI

A combination of technical interventions, governance frameworks, and regulatory compliance is required to ensure ethical AI. Unbalanced datasets, flawed model architectures, and systemic discrimination embedded in training data are the sources of bias in AI, and researchers have developed

a number of strategies to mitigate bias, such as re-weighting and resampling, pre-processing techniques to balance dataset representation, and data augmentation to expand underrepresented groups to improve fairness [1, 2]. Post-processing techniques, such as equalized odds adjustments and calibration methods, further align results across demographic groups [5, 6], whereas toolkits such as Google's What-If Tool and IBM's AI Fairness 360 support bias mitigation and fairness evaluation [7]. In-processing techniques, such as adversarial debiasing, elimination of discriminatory patterns, and fairness regularization, incorporate fairness constraints into optimization functions [3, 4].

Explainability is essential because AI models, particularly deep learning systems, frequently operate as "black boxes." To simulate complicated AI judgments, feature attribution techniques such as SHAP and LIME assign relevance scores to the input variables and create local surrogate models [8, 9]. To improve interpretability, counterfactual explanations offer substitute inputs that can influence AI predictions [10]. By utilizing transparent models such as decision trees and rule-based algorithms rather than intricate neural networks, several AI systems place a higher prioritize intrinsic interpretability [11]. However, attaining explainability frequently means sacrificing forecast accuracy, resulting in a trade-off that needs to be handled properly [12].

Another crucial component of responsible AI is privacy, especially in light of laws such as the CCPA and GDPR, which enforce stringent data protection guidelines [13]. Differential privacy is a method of improving privacy that prevents individual user identification while maintaining statistical utility by adding statistical noise to the datasets [14]. Federated learning reduces privacy threats by enabling AI models to be trained over dispersed edge devices without the need for central data storage [15]. While secure multiparty computing enables AI collaboration without disclosing sensitive information, homomorphic encryption allows computations on encrypted data without decryption [16, 17]. These privacy-preserving techniques have been incorporated into AI applications using tools such as PySyft from OpenMined and TensorFlow Privacy [18].

For AI to be deployed ethically, regulatory compliance and governance are essential. User permission, data protection, and explainability in AI-driven choices are required under the GDPR [19]. The EU Artificial Intelligence Act establishes fairness and transparency requirements for high-risk AI applications and categorizes AI systems according to risk levels [20]. IEEE's Ethically Aligned Design describes best practices for governance and risk mitigation, whereas the OECD AI Principles place an emphasis on responsibility, inclusion, and human rights [21, 22]. To maintain compliance, organizations use algorithmic transparency reports, third-party AI audits, and Algorithmic Impact Assessments (AIA), which assess AI risks and record decision-making procedures [23–25]. It is still difficult to strike a balance between innovation and regulation because too many regulations can impede the development of AI [26].

Another major concern is AI security, as risks such as data poisoning, adversarial attacks, and model inversion jeopardize the integrity of decisions [27]. Adversarial training improves the model's robustness by subjecting AI systems to adversarial attacks during training [28]. While AI red teaming mimics actual assaults to find security flaws, feature engineering techniques assist in lessening AI's vulnerability to manipulated inputs [29, 30]. By guaranteeing data integrity and offering auditability via decentralized ledgers, blockchain technology improves AI security [31]. Best practices for safeguarding AI applications are established by standards such as ISO/IEC 27001, especially in high-stakes industries, such as healthcare and finance [32].

The development of ethical AI requires ongoing improvements in security, privacy, governance, transparency, and fairness. AI can be used responsibly to advance society while lowering risks by combining explainability strategies, privacy-preserving mechanisms, regulatory frameworks, security measures, and bias reduction techniques.

Tools for Ethical AI Implementation

Important components of ethical AI include security, privacy, transparency, and fairness, all of which call for the employment of certain frameworks and tools. Several open-source tools aid in identifying and reducing biases in AI systems. Microsoft's Fairlearn and Themis-ML facilitate fairness assessments and modifications throughout model construction, whereas IBM's AI Fairness 360 (AIF360) offers a set of bias detection and mitigation methods [9, 29, 30]. While Microsoft's InterpretML incorporates glass-box models for transparency, tools such as LIME and SHAP provide interpretability by evaluating model predictions; thus, ensuring explainability in AI models is crucial [10, 14, 31].

Additionally, privacy-preserving AI methods are becoming increasingly popular. These include homomorphic encryption, which allows calculations on encrypted data without decryption; federated learning, which decentralizes model training; and differential privacy, which shields individual data from exposure. [14–16]. Robust methods such as IBM's Adversarial Robustness Toolbox (ART), SecML for security evaluation in machine learning models, and CleverHans for adversarial defensive benchmarking are necessary to address AI security risks [32–34].

Frameworks for ethical AI governance are essential for bringing AI systems into compliance with moral and legal requirements. Algorithmic Impact Assessments offer systematic assessments of AI's societal impact, IBM Watson OpenScale guarantees real-time model monitoring, and Google's PAIR program improves equity and human-AI interaction [19, 35, 36]. Together, these resources support the safe and responsible application of AI and strike a balance between creativity and morality.

Table 2 provides a comparative analysis of the major ethical AI tools based on key criteria, such as fairness, interpretability, privacy, and security.

Ethical AI in Practice and Case Studies

1. *Ethical AI in real-world applications:* Responsible AI adoption across industries depends on the use of ethical AI concepts. Although AI has greatly increased productivity and decision-making, practical implementations of technology raise moral questions about prejudice, privacy, and responsibility. This section examines the important industries in which the use of AI has produced both achievements and setbacks.
2. *Ethical AI in healthcare:* From illness diagnosis to treatment suggestions, AI-driven healthcare solutions sparked ethical questions about bias, data protection, and transparency. Despite its potential, IBM Watson for Oncology showed biased treatment recommendations because of training data restrictions, highlighting the importance of different datasets [37]. Similarly, Google DeepMind's partnership with the NHS came under fire for gaining unapproved access to patient data, underscoring the need for informed consent in medical AI [38].
3. *Ethical AI in finance:* Fairness is still a major issue in the finance industry despite the widespread use of AI for risk assessment, fraud detection, and credit scoring. Concerns regarding algorithmic discrimination were raised by the Apple Card scandal, which revealed gender prejudice in the AI-driven credit distribution [39]. However, ZestFinance showed how fairness-aware AI models can support financial inclusion and reduce prejudices [40].
4. *Ethical AI in hiring and human resources:* The goal of AI-powered employment tools is to expedite the hiring process, yet they frequently perpetuate past prejudice. The dangers of skewed training data were illustrated by Amazon's AI hiring system, which discriminated against female applicants after being educated on historical hiring trends [41]. To guarantee bias-free hiring procedures, Pymetrics uses fairness-aware machine learning algorithms [42].
5. *Ethical AI in law enforcement and surveillance:* Serious questions concerning responsibility, privacy, and prejudice are raised by the use of AI in surveillance and law enforcement. Stricter rules are required because studies have revealed that facial recognition systems from large tech companies have greater error rates for women and people of color [7]. Owing to biased crime data, predictive police models implemented in a number of US cities were discovered to

disproportionately target minority populations, underscoring the dangers of AI-driven law enforcement [43].

6. *Ethical AI in social media and content moderation:* Social media content is moderated by AI; however, there are still issues in striking a balance between the regulation of false information and the right to free speech. YouTube's AI-powered content filtering system has come under fire for unfairly harming some communities through biased enforcement and erroneous demonetization [44]. There are concerns about censorship and the moral ramifications of automated content regulation because Facebook's AI for fake news identification finds it difficult to distinguish between satire and false information [45].

FUTURE PROSPECTS OF ETHICAL AI

The development of ethical AI is influenced by improvements in public engagement, transparency, justice, and regulatory frameworks. Maintaining accountability and ethical alignment is essential as AI systems grow increasingly ingrained in society.

These real-world examples show how difficult it is to use ethical AI, and emphasize the necessity of transparent regulations, fairness-driven AI research, and openness**.**

Tables 3 and 4 provide a comparative analysis of key ethical AI case studies across different sectors.

Table 1. Comparative analysis of ethical AI frameworks.

Principle	Author's name	Description	Techniques/tools	Challenges
Accountability	Cannarsa [1]	Assigning responsibility for AI decisions	Human-in-the-loop, AI audits	Lack of liability frameworks
Inclusivity	OECD [2]	Ensuring AI benefits all demographics	Diverse datasets, stakeholder engagement	Data scarcity, regional biases
Fairness	Buolamwini and Gebru [7]	Ensuring AI does not discriminate	Pre-processing, In-processing, Post-processing, AI Fairness 360	Conflicting fairness metrics
Transparency	Ribeiro et al. [10]	Making AI decision-making understandable	SHAP, LIME, Counterfactual Explanations	Trade-off between explainability and accuracy
Privacy	Dwork [14]	Protecting user data from misuse	Differential Privacy, Federated Learning	Performance overhead, compliance with regulations
Security	McMahan et al. [15]	Preventing adversarial attacks and misuse	Adversarial training, ethical hacking	AI vulnerabilities, evolving threats

Table 2. Comparative analysis of ethical AI tools.

Tool	Author's name	Functionality	Fairness	Explainability	Privacy	Security
AIF360	IBM Research [9]	Bias detection and mitigation	✅	❌	❌	❌
LIME	Ribeiro et al. [10]	Model-agnostic interpretability	❌	✅	❌	❌
SHAP	Dwork [14]	Feature attribution analysis	❌	✅	❌	❌
Differential Privacy	Dwork [14]	Data privacy enhancement	❌	❌	✅	❌
Federated Learning	McMahan et al. [15]	Decentralized AI training	❌	❌	✅	❌
Fairlearn	Microsoft [29]	Fairness evaluation and improvement	✅	❌	❌	❌
ART	Nicolae [32]	Adversarial robustness	❌	❌	❌	✅
CleverHans	Papernot et al. [20]	Security benchmarking	❌	❌	❌	✅

Note: Now, the references are sorted in increasing numerical order. Let me know if you need further adjustment.
Legend: ✅; = Supported; ❌ = Not Supported

Table 3. Comparative analysis of ethical AI case studies.

Sector	Author's name	Case study	Ethical challenge	Key lessons learned
Law Enforcement	Buolamwini and Gebru [7]	Facial Recognition Bias	Racial discrimination in AI	Need for fairness testing and regulatory oversight
Healthcare	IBM [9]	IBM Watson for Oncology	Bias in medical recommendations	Need for diverse datasets and rigorous validation
Healthcare	Powles and Hodson [37]	Google DeepMind NHS	Privacy violations	Importance of data consent and transparency
Finance	Weinberg [39]	Apple Card Gender Bias	Gender discrimination	Need for fairness-aware AI models
Finance	ZestFinance [40]	ZestFinance Fair Lending	Fairness in credit scoring	Interpretable AI for ethical finance
HR	Dastin [41]	Amazon AI Hiring Tool	Gender bias in recruitment	Avoiding biased training datasets
HR	People [42]	Pymetrics Bias-Free Hiring	Fair hiring practices	Algorithmic fairness auditing
Law Enforcement	Angwin [43]	Predictive Policing	Bias in crime predictions	Ensuring transparency in AI-based policing
Social Media	YouTube [44]	YouTube AI Moderation	Content censorship bias	Improving AI Fairness and transparency
Social Media	Facebook [45]	Facebook AI for Fake News Detection	Misinformation control vs free speech	Ethical balancing in AI governance

Table 4. Comparative study of literature.

S.N.	Reference (author name)	Year	Key contribution	Merits	Demerits
1	Russell and Norvig [3]	2021	Comprehensive analysis of AI principles and ethics	Covers various AI applications and ethical concerns	Lacks focus on specific AI governance frameworks
2	Buolamwini and Gebru [7]	2018	Identified bias in commercial facial recognition systems	Raised awareness on fairness issues in AI	Limited to specific AI applications in facial recognition
3	Ribeiro et al. [10]	2016	Developed LIME for AI interpretability	Improves AI transparency and decision-making	Not always effective for complex deep learning models
4	Dwork [14]	2008	Introduced Differential Privacy	Enhances privacy while preserving data utility	Computationally expensive for large datasets
5	McMahan et al. [15]	2017	Developed Federated Learning	Improves data privacy and security	Requires high computational resources
6	Powles and Hodson [37]	2017	Criticized DeepMind's NHS patient data access	Raised privacy concerns in healthcare AI	Lack of informed consent before data collection
7	ZestFinance [40]	2021	AI-driven fair lending model	Promotes financial inclusion	Potential biases in algorithmic decision-making
8	Dastin [41]	2022	AI hiring tool exhibited gender bias	Highlighted risks of biased AI in recruitment	System was scrapped due to biased concerns
9	Angwin J [43]	2022	Identified racial bias in predictive policing AI	Showed limitations of biased crime data	Reinforced systemic biases in law enforcement
10	Setiawan [45]	2022	AI for misinformation detection on social media	Helps reduce misinformation spread	Struggles with differentiating satire from fake news

Table 5. Comparative analysis of future ethical AI trends.

Trend	Author's name	Key focus	Expected impact
Stronger AI Regulations	Cannarsa [1]	EU AI Act, US AI Bill of Rights	Legal enforcement of AI ethics
Explainable AI (XAI)	Borrego-Díaz et al. [46]	Self-interpretable models	Improved AI transparency and trust
AI Ethics Education	Dieterle [49]	AI ethics in academia and industry	Ethical AI adoption at scale
AI for Social Good	Rolnick et al. [55]	Climate change, humanitarian aid	AI-driven sustainability efforts
Autonomous AI Ethics	Goodall, Cath et al., Scharre [58–60]	Moral decision-making in AI systems	Human-centered AI governance
Public Participation in AI	Unver MB [61]	Citizen review panels, participatory design	Democratic AI decision-making

1. *Advancements in explainable AI (XAI):* Enhancing transparency and confidence in AI systems will be greatly aided by Explainable AI (XAI). To improve explainability in fields such as healthcare and finance, future advancements will concentrate on self-interpretable models, hybrid AI techniques that combine deep learning and symbolic reasoning, and industry-specific frameworks. [46–48].
2. *Integration of AI ethics in education and training:* Ethical AI education is included in corporate training programs, professional certification initiatives, and university curricula to encourage responsible AI development. The goal of these initiatives is to give organizations and developers the skills they need to create just responsible AI systems [47–51].
3. *Regulatory and policy developments in ai ethics:* To uphold moral principles, governments and international organizations are tightening AI legislation. The US AI Bill of Rights, the EU Artificial Intelligence Act, and UNESCO's AI ethics guidelines are important endeavors. It is anticipated that future rules will require stringent data protection controls, explainability standards for high-risk applications, and frequent AI audits [52–54].
4. *AI for social good and ethical ai research:* With current research concentrating on climate change mitigation, ethical AI in humanitarian aid, and fairness-aware algorithms to reduce bias, artificial intelligence is being increasingly used for social effects. These initiatives address ethical concerns while showcasing the AI's ability to promote constructive social change [55–57].
5. *Autonomous AI and ethical considerations:* There are new ethical issues with the development of autonomous AI systems, such as self-driving cars and AI-powered legal and military decision-making. Future AI governance will depend on ensuring human oversight in military AI applications, fairness in judicial AI, and moral decision-making in autonomous cars [56–60].
6. *The role of public participation in ethical AI:* With projects such as citizen AI review panels, participatory AI design, and decentralized AI governance models gaining pace, it is anticipated that public involvement in AI governance will increase. Through greater transparency and community involvement, these strategies seek to guarantee that AI is consistent with social values [61–63].

To develop AI systems that are just, open, and consistent with human values, the public, business, academia, and legislators will need to work together, as shown in Table 5.

CONCLUSION

To ensure justice, accountability, openness, and privacy, AI development must be ethical. Fairness-aware algorithms and audits are necessary to address the persistent issue of bias in AI models, which can be observed in Amazon's recruitment process and Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment. Transparency is improved by Explainable AI (XAI) methods, such as SHAP and LIME, although complete interpretability is still a problem. Differential privacy and federated learning are privacy-preserving techniques that safeguard user data without sacrificing functionality. The necessity for a unifying framework is highlighted by disparities in global AI governance, which include China's state-controlled approach, stringent EU laws, and US standards. Even if AI innovations, such as DeepMind's diagnostics, have promise, unethical behavior highlights

the need for cooperation between researchers, policymakers, and business executives to guarantee responsible AI. Sustainable AI development requires ethical AI that is not merely a goal.

REFERENCES

1. Cannarsa M. Ethics guidelines for trustworthy AI. In: DiMatteo LA, Janssen A, Ortolani P, de Elizalde F, Cannarsa M, Durovic M, editors. The Cambridge Handbook of Lawyering in the Digital Age. Cambridge: Cambridge University Press; 2021. p. 283–297. doi:10.1017/9781108936040.022.
2. Canton H. Organisation for Economic Co-operation and Development—OECD. In: Europa Publications, editor. The Europa Directory of International Organizations 2021. 23rd ed. London: Routledge; 2021. p. 677–687. doi:10.4324/9781003179900-102.
3. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Harlow: Pearson; 2021.
4. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. Minds Mach. 2020;30(1):99–120. doi:10.1007/s11023-020-09517-8.
5. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389–399. doi:10.1038/s42256-019-0088-2.
6. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. SSRN Electron J. 2020. doi:10.2139/ssrn.3518482.
7. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, editors. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research. Vol. 81. PMLR; 2018. p. 77–91. Available from: https://proceedings.mlr.press/v81/buolamwini18a.html
8. Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst. 2012;33(1):1–33. doi:10.1007/s10115-011-0463-8.
9. IBM. (2026). Artificial Intelligence (AI) Solutions. [online] IBM. Available from: https://www.ibm.com/solutions/artificial-intelligence
10. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. New York (NY): Association for Computing Machinery; 2016. p. 1135–1144. doi:10.1145/2939672.2939778.
11. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. SSRN Electron J. 2017;31:841. doi:10.2139/ssrn.3063289.
12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–215. doi:10.1038/s42256-019-0048-x.
13. European Parliament; Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Off J Eur Union. 2016 May 4;59(L 119):1-88.
14. Dwork C. Differential privacy: a survey of results. In: Agrawal M, Du D, Duan Z, Li A, editors. Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science. Vol. 4978. Berlin, Heidelberg: Springer; 2008. p. 1–19. doi:10.1007/978-3-540-79228-4_1.
15. McMahan B, Moore E, Ramage D, Hampson S, Aguera y Arcas B. Communication-efficient learning of deep networks from decentralized data. In: Singh A, Zhu J, editors. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research. 2017;54:1273–1282. Available from: https://proceedings.mlr.press/v54/mcmahan17a.html
16. Gentry C. Fully homomorphic encryption using ideal lattices. In: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing. ACM; 2009. p. 169–178. doi:10.1145/1536414.1536440.

17. Shahriari K, Shahriari M. IEEE standard review—ethically aligned design: a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, ON, Canada, 2017, Jul 21–23; Toronto, ON, Canada. 2017. p. 197–201. doi:10.1109/IHTC.2017.8058187.
18. Reisman D, Schultz J, Crawford K, Whittaker M. Algorithmic Impact Assessments: a Practical Framework for Public Agency Accountability. New York (NY): AI Now Institute; 2018. Available from: https://ainowinstitute.org/reports/ai-now-report-2018.pdf
19. Raji ID, Buolamwini J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society; 2019 Jan 27–28; Honolulu, HI, USA. New York (NY): Association for Computing Machinery; 2019. p. 429–435. doi:10.1145/3306618.3314244.
20. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA. 2016. p. 582–597. doi:10.1109/SP.2016.41.
21. Carlini N, Wagner D. Adversarial examples are not easily detected: bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security; 2017 Nov 3; Dallas, TX, USA. New York (NY): Association for Computing Machinery; 2017. p. 3–14. doi:10.1145/3128572.3140444.
22. OpenAI. (2022). Security and privacy at OpenAI. [online] OpenAI. Available from: https://openai.com/security-and-privacy/
23. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS 2016); 2016 Dec 5–10; Barcelona, Spain. Red Hook (NY): Curran Associates, Inc.; 2016. p. 3315–3323.
24. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: Dasgupta S, McAllester D, editors. Proceedings of the 30th International Conference on Machine Learning; 2013 Jun 17–19; Atlanta, GA, USA. Proceedings of Machine Learning Research. 2013;28(3):325–333. Available from: https://proceedings.mlr.press/v28/zemel13.html
25. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019. Available from: https://christophm.github.io/interpretable-ml-book/
26. Michael K. Editorial IEEE Transactions on Technology and Society editorial board profiles. IEEE Trans Technol Soc. 2024;5(2):119–148. doi:10.1109/TTS.2024.3423208.
27. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19); 2019; Atlanta, GA, USA. New York: Association for Computing Machinery; 2019. p. 59–68. doi:10.1145/3287560.3287598.
28. Barocas S, Hardt M, Narayanan A. Fairness and machine learning: Limitations and opportunities. Cambridge, MA: MIT Press; 2023.
29. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, et al. Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft; 2020. Available from: https://www.microsoft.com/en-us/research/wp-content/uploads/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf
30. Bantilan N. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. J Technol Hum Serv. 2018;36(1):15–30. doi:10.1080/15228835.2017.1416512.
31. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: A unified framework for machine learning interpretability. [Preprint]. 2019. arXiv:1909.09223. doi:10.48550/arXiv.1909.09223.
32. Nicolae MI, Sinn M, Tran MN, Buesser B, Rawat A, Wistuba M, et al. Adversarial robustness toolbox v1.0.0. [Preprint]. 2018. arXiv:1807.01069. doi:10.48550/arXiv.1807.01069
33. Lacour C, Massart P, Rivoirard V. Estimator selection: A new method with applications to kernel density estimation. Sankhya A. 2017;79(2):298–335. doi:10.1007/s13171-017-0107-5.
34. Melis M, Demontis A, Pintor M, Sotgiu A, Biggio B. SecML: A Python library for secure and explainable machine learning. [Preprint]. 2019. arXiv:1912.10013. doi:10.48550/arXiv.1912.10013.

35. Zhang D, Maslej N, Brynjolfsson E, Etchemendy J, Lyons T, Manyika J, et al. The AI Index 2022 annual report. [Preprint]. 2022. arXiv:2205.03468. doi:10.48550/arXiv.2205.03468.
36. Norouzi K, Ghodsi A, Argani P, Andi PA, Hassani H. Innovative artificial intelligence tools: exploring the future of healthcare through IBM Watson's potential applications. In: Nguyen TA, editor. Sensor Networks for Smart Hospitals. New York: Elsevier; 2025. p. 573–588. doi:10.1016/B978-0-443-36370-2.00028-1.
37. Powles J, Hodson H. Google DeepMind and healthcare in an age of algorithms. Health Technol. 2017;7(4):351–367. doi:10.1007/s12553-017-0179-1.
38. Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: Building foundations for knowledge and wisdom. BMC Med. 2019;17(1):143. doi:10.1186/s12916-019-1382-x.
39. Weinberg L. Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. J Artif Intell Res. 2022;74:75–109. doi:10.1613/jair.1.13196.
40. ZestFinance (2021). Zest AI Honored in Fast Company's 2021 Next Big Things in Tech Awards. [Online] PR Newswire. Available from: https://www.prnewswire.com/news-releases/zest-ai-honored-in-fast-companys-2021-next-big-things-in-tech-awards-301428185.html
41. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. In: Martin K, editor. Ethics of Data and Analytics: Concepts and Cases. New York: Auerbach Publications; 2022. p. 296–299. doi:10.1201/9781003278290-44.
42. Modgil S. (2018). How AI startup Pymetrics wants to make hiring bias free [Online]. People Matters Global. Available from: https://sea.peoplemattersglobal.com/article/hr-technology/how-ai-startup-pymetrics-wants-to-make-hiring-bias-free-20022
43. Angwin J, Larson J. Bias in criminal risk scores is mathematically inevitable, researchers say. In: Martin K, editor. Ethics of Data and Analytics: Concepts and Cases. New York: Auerbach Publications; 2022. p. 265–267. doi:10.1201/9781003278290-38.
44. Rock A, Jebaseeli TJ. A content moderation system for YouTube using hybrid deep neural networks. AIP Conf Proc. 2025;3297(1):090035. doi:10.1063/5.0286780.
45. Setiawan R, Ponnam VS, Sengan S, Anam M, Subbiah C, Phasinam K, et al. Certain investigation of fake news detection from Facebook and Twitter using artificial intelligence approach. Wirel Pers Commun. 2022;127(2):1737–1762. doi:10.1007/s11277-021-08720-9.
46. Borrego-Díaz J, Galán-Páez J. Explainable artificial intelligence in data science: From foundational issues towards socio-technical considerations. Minds Mach. 2022;32(3):485–531. doi:10.1007/s11023-022-09603-z.
47. Mehra A. Hybrid AI models: Integrating symbolic reasoning with deep learning for complex decision-making. J Emerg Technol Innov Res. 2024;11:f693–f695.
48. Chinnaraju A. Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. World J Adv Eng Technol Sci. 2025;14(3):170–207. doi:10.30574/wjaets.2025.14.3.0106.
49. Dieterle E, Dede C, Walker M. The cyclical ethical effects of using artificial intelligence in education. AI Soc. 2024;39:633–643. doi:10.1007/s00146-022-01497-w.
50. Martin K. Google research: who is responsible for ethics of AI? In: Martin K, editor. Ethics of Data and Analytics: Concepts and Cases. New York: Auerbach Publications; 2022. p. 434–446. doi:10.1201/9781003278290.
51. Shahriari K, Shahriari M. IEEE standard review – Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, ON, Canada. 2017. p. 197–201. doi:10.1109/IHTC.2017.8058187.
52. Butt J. Analytical study of the world's first EU artificial intelligence (AI) act. Int J Res Publ Rev. 2024;5(3):7343–7364.
53. House W. Blueprint for an AI Bill of Rights: Making automated systems work for the American people. Nimble Books; 2022.

54. Van Norren DE. The ethics of artificial intelligence, UNESCO and the African Ubuntu perspective. J Inf Commun Ethics Soc. 2023;21:112–128. doi:10.1108/JICES-04-2022-0037.
55. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling climate change with machine learning. ACM Comput Surv. 2022;55:1–96. doi:10.1145/3485128.
56. Efe A. A review on risk reduction potentials of artificial intelligence in humanitarian aid sector. İnsan ve Sosyal Bilimler Dergisi. 2022;5(2):184–205. doi:10.53048/johass.1189814.
57. Kidwai-Khan F, Wang R, Skanderson M, Brandt CA, Fodeh S, Womack JA. A roadmap to artificial intelligence (AI): Methods for designing and building AI-ready data to promote fairness. J Biomed Inform. 2024;154:104654. doi:10.1016/j.jbi.2024.104654.
58. Goodall NJ. Machine ethics and automated vehicles. In: Meyer G, Beiker S, editors. Road Vehicle Automation. Cham: Springer; 2014. p. 93–102. doi:10.1007/978-3-319-05990-7_9.
59. Borgesano F, De Maio A, Laghi P, Musmanno R. Artificial intelligence and justice: A systematic literature review and future research perspectives on Justice 5.0. Eur J Innov Manag. 2025;28(11):349–385. doi:10.1108/EJIM-01-2025-0117.
60. Scharre P, Lamberth M. Artificial intelligence and arms control. [Preprint]. 2022. arXiv:2211.00065. doi:10.48550/arXiv.2211.00065.
61. Unver MB. AI governance: Compromising democracy or democratising AI? SSRN Electron J. 2024. doi:10.2139/ssrn.4913658.
62. Gerdes A. A participatory data-centric approach to AI ethics by design. Appl Artif Intell. 2022;36(1):2009222. doi:10.1080/08839514.2021.2009222.
63. Hu B, Rong H, Tay J. Is decentralized artificial intelligence governable? Towards machine sovereignty and human symbiosis. 2025 Jan 9.

