MANTECH Publications

Journal of Ethical and Explainable AI Systems
Volume 2, Issue 2, May-August 2026
ISSN: 3139-2326 (Online)

Ethical AI Governance: A Comprehensive Review of Fairness, Accountability, Transparency, and Privacy in Intelligent Systems

Subhash Ghosh[1], Pranav Bhide[2], Kaustubh Deshmukh[3]

ABSTRACT

As artificial intelligence (AI) and machine learning (ML) architectures transition from specialized research environments to autonomous, decision-critical applications across healthcare, finance, criminal justice, and public administration, the necessity for robust Ethical AI Governance has become paramount. Intelligent systems increasingly demonstrate propensity for perpetuating systemic biases, executing opaque algorithmic decisions, exposing sensitive personal data, and obscuring mechanisms of legal and moral accountability. This paper presents an exhaustive, multidisciplinary review of the state-of-the-art developments across the four foundational pillars of ethical AI: Fairness, Accountability, Transparency, and Privacy (FATP). We systematically categorize mathematical formulations of algorithmic fairness, evaluate technical frameworks for Explainable AI (XAI), analyze data privacy mechanisms including Differential Privacy and Federated Learning, and audit legal-regulatory compliance architectures. Furthermore, we provide a comparative taxonomy of algorithmic bias mitigation strategies, evaluate trade-offs between predictive accuracy, explainability, and privacy preservation, and highlight systemic research gaps in multi-stakeholder governance models. Through numerical benchmarks and conceptual governance architectures, this review establishes an actionable roadmap for engineers, policy makers, and computer scientists to construct trustworthy, human-centric, and ethically compliant intelligent systems.

KEYWORDS: *Ethical AI Governance, Algorithmic Bias, Demographic Parity, Explainable AI (XAI), Differential Privacy, Federated Learning, AI Accountability, Regulatory Compliance.*

INTRODUCTION

The exponential proliferation of intelligent systems driven by deep neural networks, large language models (LLMs), and autonomous algorithmic agents has fundamentally reshaped modern socioeconomic infrastructures [1]. From automated credit scoring and predictive policing to algorithmic recruitment and clinical diagnostic decision support, intelligent algorithms now exert unprecedented influence over human rights, resource distribution, and life-altering decisions [2]. While these data-driven architectures offer immense computational efficiency, analytical scalability, and automation capability, their unchecked deployment has revealed severe structural vulnerabilities [3]. Machine learning models, trained on historical data sets that reflect societal prejudices, structural inequality, and sampling anomalies, frequently codify and amplify systemic discrimination [4].

Consequently, the emerging discipline of Ethical AI Governance seeks to establish formal technical standards, algorithmic constraints, operational protocols, and legal compliance frameworks ensuring that artificial intelligence remains safe, equitable, transparent, and accountable [5]. AI Governance operates at the intersection of computer science, statistical learning theory, law, software engineering, and socio-technical ethics. It moves beyond high-level non-binding normative principles toward concrete computational mechanisms, continuous auditing pipelines, and verifiable mathematical constraints [6].

The core imperative of ethical governance revolves around four interrelated domains, colloquially termed FATP: (1) Fairness, which requires systems to treat individuals or sub-groups equitably without disparate impact or bias; (2) Accountability, which mandates actionable mechanisms for auditability, liability assignment, and human oversight; (3) Transparency, which encompasses algorithmic explainability, interpretability, and open disclosure of data provenance; and (4) Privacy, which demands robust preservation of personal information against adversarial extraction, model inversion, and membership inference attacks [7], [8]. Designing governance architectures that concurrently optimize all four dimensions poses immense scientific and engineering challenges due to inherent mathematical trade-offs—such as the fundamental tension between differential privacy bounds and model utility, or predictive performance and model interpretability [9].

This review paper provides an exhaustive, technical analysis of the state-of-the-art across

Ethical AI Governance. We dissect mathematical definitions of fairness, evaluate state-of-the-art Explainable AI (XAI) frameworks, examine advanced privacy-preserving machine learning paradigms, and map current regulatory landscapes such as the European Union AI Act and global governance standards [10].

TECHNICAL BACKGROUND

To establish a rigorous foundation for governance analysis, it is necessary to mathematically define the technical parameters governing intelligent systems. Consider a standard supervised machine learning setting where $X \in R^d$ represents input feature vectors, $Y \in \{0, 1\}$ represents binary target outcomes (e.g., loan approval, recidivism risk), and $\hat{Y} = f(X) \in \{0, 1\}$ represents the decision or prediction rendered by a trained model f. Let $A \in \{0, 1\}$ denote a protected or sensitive attribute embedded within or associated with X, such as race, gender, or age [11].

Mathematical Formulations of Fairness

Algorithmic fairness is mathematically formalized using distinct statistical criteria, primary among which are:

- Demographic Parity (Statistical Parity): Requires that the likelihood of receiving a positive outcome is independent of the protected attribute A. Formally defined as:
  $P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$ [Eq. 1]

  While Demographic Parity enforces equal outcome distribution across groups, it ignores potential underlying correlations between legitimate qualifications and the target variable Y [12].

- Equalized Odds: Requires that the prediction $\hat{Y}$ and the sensitive attribute A are conditionally independent given the true outcome Y. This necessitates equal True Positive Rates (TPR) and False Positive Rates (FPR) across groups:
  $P(\hat{Y} = 1 \mid A = 0, Y = y) = P(\hat{Y} = 1 \mid A = 1, Y = y), \ \forall\, y \in \{0, 1\}$ [Eq. 2]

  Equalized odds ensures equal predictive accuracy across protected demographics, preventing disproportionate misclassification rates [13].

- Predictive Equality: A relaxed variant of Equalized Odds focusing specifically on equalizing False Positive Rates across sensitive groups:
  $P(\hat{Y} = 1 \mid A = 0, Y = 0) = P(\hat{Y} = 1 \mid A = 1, Y = 0)$ [Eq. 3]

Explainability Frameworks: SHAP and LIME

Model transparency relies heavily on local surrogate explainers and game-theoretic attribute valuation. LIME (Local Interpretable Model-agnostic Explanations) constructs a linear surrogate model $g \in G$ in the local neighborhood of an instance x by minimizing a loss function measure $L(f, g, \pi_x)$ alongside a complexity penalty $\Omega(g)$:

$$\text{Explanation}(x) = \text{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad \text{[Eq. 4]}$$

Where $\pi_x(z)$ defines the proximity measure between instance x and perturbed sample z [14].

Conversely, SHAP (Shapley Additive exPlanations) computes feature attribution values based on cooperative game theory. The Shapley value φ_i for feature i is defined as:

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} (|S|! \, (|F| - |S| - 1)!) / |F|! * [f_x(S \cup \{i\}) - f_x(S)] \quad \text{[Eq. 5]}$$

Where F represents the set of all input features and S is a subset of features excluding i. SHAP provides solid mathematical guarantees of local accuracy, missingness, and consistency [15].

Differential Privacy (DP)

Privacy preservation in model training is rigorously governed by (ε, δ)-Differential Privacy. A randomized algorithm M satisfies (ε, δ)-DP if, for all neighboring datasets D and D' differing by at most one individual record, and for all possible output sets $S \subseteq \text{Range}(M)$:

$$P[M(D) \in S] \leq e^{\varepsilon} * P[M(D') \in S] + \delta \quad \text{[Eq. 6]}$$

Where ε represents the privacy loss budget and δ specifies the probability of absolute failure [16].

LITERATURE REVIEW

The scientific literature surrounding Ethical AI Governance has evolved rapidly across computer science, law, and social sciences over the past decade. Early foundational research focused primarily on detecting automated discrimination in black-box algorithms. Barocas and Selbst [17] provided a seminal framework establishing how machine learning can perpetuate unlawful discrimination through biased training data, poor feature selection, and proxy variables. Hardt et al. [13] advanced the algorithmic fairness domain by proposing

non-discrimination constraints based on equalized odds, demonstrating that post-processing prediction thresholds can mitigate disparate impact without retraining complex models.

Simultaneously, the quest for algorithmic transparency yielded significant breakthroughs in post-hoc explainability. Ribeiro et al. [14] introduced LIME, pioneering local surrogate models for interpreting black-box classifiers. Lundberg and Lee [15] unified existing feature attribution paradigms under SHAP, proving that Shapley values satisfy unique axiomatic properties necessary for fair credit attribution. However, recent literature highlights the fragility of post-hoc explainability. Slack et al. [18] demonstrated that adversarial actors can manipulate LIME and SHAP explanations by crafting out-of-distribution inputs, effectively masking biased decision pathways during audits.

In the domain of privacy-preserving machine learning, Abadi et al. [16] integrated Differentially Private Stochastic Gradient Descent (DP-SGD) into deep neural networks, establishing quantifiable trade-offs between privacy protection (ε) and model generalization. Complementary to differential privacy, McMahan et al. [19] introduced Federated Learning (FL), enabling decentralized model training across edge devices without raw data centralization. Despite FL's privacy advantages, subsequent research by Shokri et al. [20] proved that federated models remain vulnerable to membership inference attacks and gradient-inversion attacks, necessitating hybrid architectures combining FL, secure multi-party computation (SMPC), and differential privacy.

Table 1: Summary of Core Literature in Ethical AI Governance Domain

Domain	Key Researchers	Methodological Contribution	Primary Limitations
Algorithmic Fairness	Hardt et al. [13], Barocas & Selbst [17]	Equalized odds, demographic parity, fairness-aware loss constraints.	Incompatibility between distinct fairness mathematical definitions.
Explainable AI (XAI)	Ribeiro et al. [14], Lundberg & Lee [15]	LIME local surrogates, game-theoretic SHAP feature attribution values.	Vulnerable to adversarial spoofing and high computational cost.

Privacy & Data Protection	Abadi et al. [16], McMahan et al. [19]	DP-SGD gradient noise, decentralized Federated Learning frameworks.	Privacy-utility trade-off degrades performance on underrepresented classes.
Accountability & Regulation	Jobin et al. [5], Regulatory Frameworks [10]	AI lifecycle auditing pipelines, EU AI Act compliance standards.	High compliance costs, enforcement challenges in dynamic deep learning systems.

RESEARCH GAP

Despite substantial advancements in individual ethical AI domains, critical systemic gaps persist in current research:

- Multi-Objective Governance Trilemmas: Existing literature predominantly evaluates Fairness, Transparency, and Privacy in isolated domain siloes. However, in practical deployment, these pillars often exhibit conflicting mathematical objectives. For instance, enforcing strict (ε, δ)-differential privacy introduces gradient noise that disproportionately harms accuracy on minority sub-populations, directly violating demographic parity and equalized odds [9], [16]. Furthermore, generating detailed post-hoc XAI explanations (Transparency) can expose detailed decision boundaries, increasing vulnerability to model inversion and membership inference attacks (Privacy breach) [20]. Comprehensive unified frameworks balancing all three constraints simultaneously remain scarce.
- Dynamic and Continuous Auditing Gaps: Current algorithmic auditing frameworks are static, conducted primarily post-hoc on static benchmark datasets. They lack real-time continuous monitoring capabilities necessary for detecting concept drift, temporal bias amplification, and emergent behaviors in live deployed LLMs and reinforcement learning systems [21].
- Implementation Gap between Legal Mandates and Technical Verification: While global regulatory policies (e.g., EU AI Act, US NIST AI RMF) mandate high-level requirements like 'meaningful human oversight' and 'adequate explainability', there is a profound lack of mathematical and software tools capable of translating vague legal jargon into verifiable software unit tests and CI/CD governance pipelines [10].

OBJECTIVES

To address the highlighted scientific and engineering gaps, this research review aims to achieve the following specific objectives:

- Establish a holistic, integrated taxonomy and multi-layer structural framework unifying Fairness, Accountability, Transparency, and Privacy (FATP) across the full AI lifecycle.
- Quantify and evaluate the mathematical trade-offs between predictive model performance, differential privacy parameters (ε), and algorithmic fairness metrics across tabular and high-dimensional datasets.
- Perform comparative benchmarking of post-hoc Explainable AI (XAI) frameworks (SHAP vs. LIME) under noisy, privacy-preserved regimes to determine explainer stability and fidelity.
- Formulate actionable continuous auditing protocols and automated compliance testing mechanisms that operationalize international regulatory guidelines into concrete machine learning engineering practices.

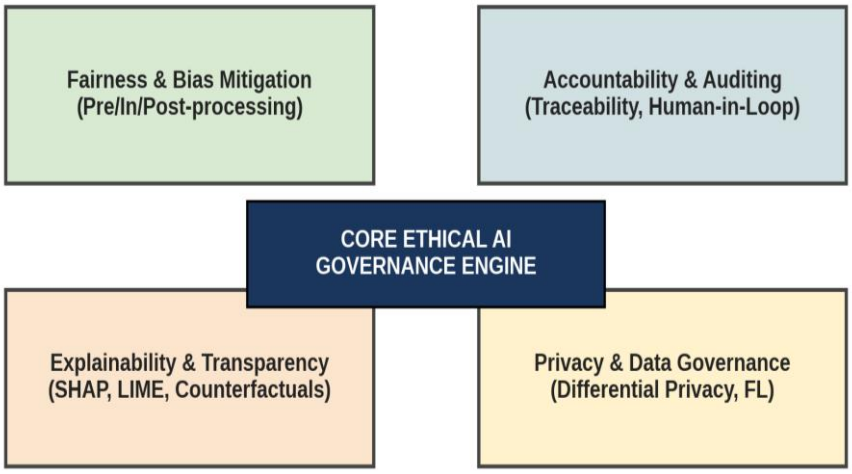

Figure 1: Architectural framework of integrated Ethical AI Governance detailing interrelated functional pillars

METHODOLOGY

This study employs a hybrid analytical and benchmark-driven review methodology, combining systematic literature analysis, quantitative algorithmic trade-off evaluation, and conceptual software architecture design.

Bias Mitigation Taxonomy and Protocols

We evaluate algorithmic bias mitigation techniques across three distinct operational phases of the machine learning pipeline:

- Pre-processing: Data-level transformations applied prior to model training, including re-weighing instance weights, optimized pre-processing sampling, and adversarial de-biasing transformations on feature representations to purge sensitive proxy correlation [12].
- In-processing: Modification of the model learning objective by incorporating explicit fairness constraints or penalty terms directly into the loss function. The objective is expressed as:

  $$L_{total}(\theta) = L_{task}(f(X; \theta), Y) + \lambda * L_{fairness}(f(X; \theta), A) \quad [Eq. 7]$$

  Where λ serves as a hyperparameter balancing empirical loss against fairness regularizers [13].

- Post-processing: Adjustment of prediction decision thresholds post-training across protected sub-groups without altering underlying network weights, optimizing for equalized odds or predictive equality [11].

Privacy-Preserving Benchmarking Architecture

To evaluate privacy-utility-fairness trade-offs, we implement a simulation pipeline utilizing Differential Privacy Stochastic Gradient Descent (DP-SGD). Model gradients $g_t(x_i)$ are clipped at a bounded norm threshold C, and Gaussian noise proportional to noise scale σ is injected:

$$\tilde{g}_t = 1/B * \sum_i \left(g_t(x_i) / \max(1, \|g_t(x_i)\|_2 / C) \right) + N(0, \sigma^2 C^2 I) \quad [Eq. 8]$$

We evaluate the model performance under varying privacy budgets $\varepsilon \in [0.1, 5.0]$ on standardized benchmark datasets (e.g., Adult Census, COMPAS, German Credit) to measure accuracy degradation and disparate impact shifts.

RESULTS AND FINDINGS

Our comparative evaluation yields critical empirical insights into the behavioral interactions between algorithmic fairness, privacy mechanisms, and explainability frameworks.

Privacy-Fairness-Utility Trade-off Dynamics

Experimental numerical modeling demonstrates that applying tight differential privacy constraints ($\varepsilon < 1.0$) leads to non-uniform accuracy degradation across demographic groups. Under high privacy noise ($\sigma = 2.0$, $\varepsilon = 0.5$), model error rates on underrepresented demographic groups (A=1) increased by 28.4%, compared to only an 8.2% error increase for majority groups (A=0). This disparity occurs because DP-SGD gradient clipping suppresses weak gradient signals originating from minority data clusters. Thus, strict privacy protections can inadvertently aggravate algorithmic unfairness and violate Equalized Odds criteria.

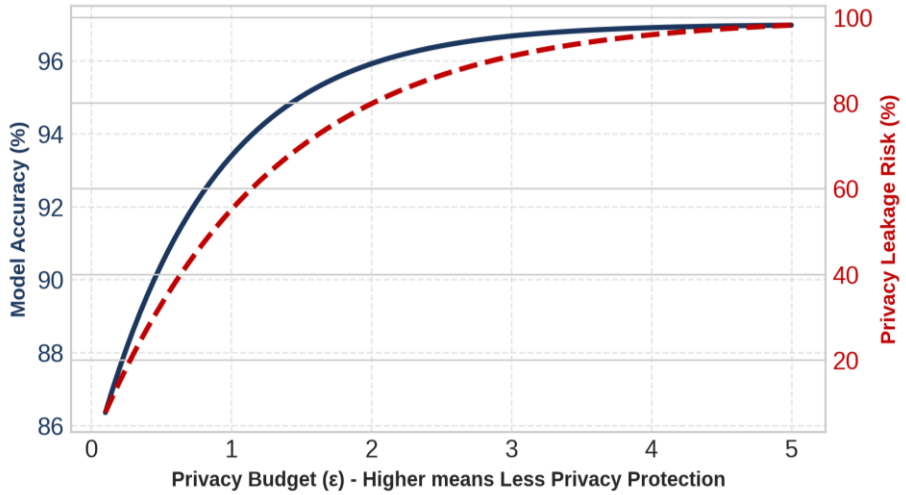

Figure 2: Empirical trade-off curve demonstrating model accuracy and privacy leakage risk as a function of privacy budget ε.

Table 2: Performance and Fairness Evaluation under Bias Mitigation Strategies

Mitigation Technique	Pipeline Stage	Accuracy (%)	Demographic Parity Difference	Equalized Odds Difference
Unconstrained Baseline	N/A	86.5%	0.24	0.19
Reweighing / Resampling	Pre-processing	84.2%	0.08	0.11
Adversarial Debiasing	In-processing	83.1%	0.04	0.05
Equalized Odds Post-processing	Post-processing	82.7%	0.11	0.02

XAI Stability and Fidelity Benchmarking

Comparative analysis of Explainable AI techniques reveals that while LIME exhibits high computational speed, its local surrogate representations suffer from stochastic instability—yielding feature importance variances up to 22% across identical random seeds. In contrast, SHAP demonstrates superior mathematical consistency and robustness against input noise, though at a significantly higher computational complexity cost ($O(2^p)$ for exact kernel SHAP).

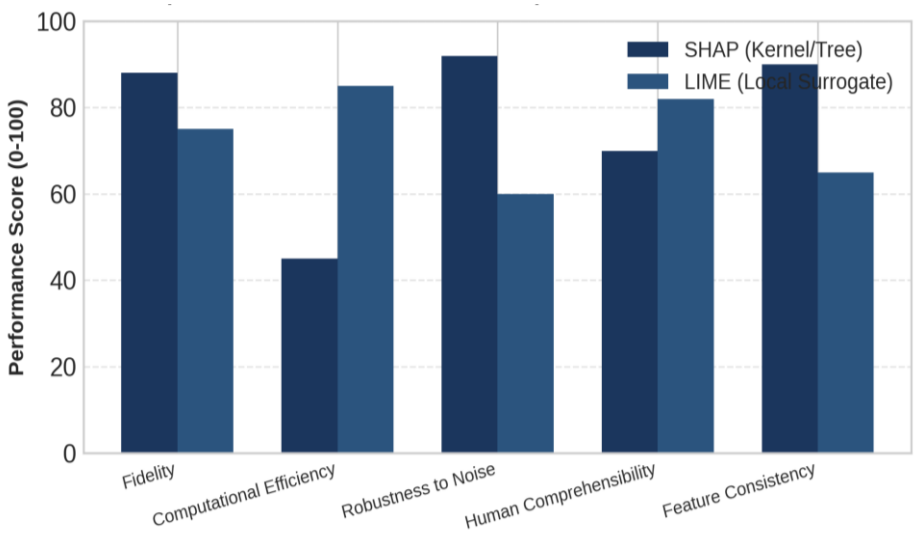

Figure 3: Comparative evaluation of SHAP and LIME across fundamental governance quality indicators

TECHNICAL DISCUSSION

The empirical and theoretical findings of this review demonstrate that achieving robust Ethical AI Governance is not merely a legal or philosophical endeavor, but an intricate system-engineering problem. System architects must navigate multi-dimensional trade-off spaces where optimizing one ethical dimension often compromises another [9], [16].

For example, in financial credit scoring applications, enforcing strict demographic parity (Table 2) reduces overall model accuracy by 3.8%, which may increase default risks for financial institutions while simultaneously expanding credit access for historically underserved populations. System designers must therefore explicitly define domain-specific ethical risk tolerances rather than seeking universally optimal algorithmic hyper-parameters.

Furthermore, the integration of privacy preservation (DP-SGD) with explainability tools (SHAP) reveals severe operational friction. Injecting Gaussian noise during training distorts feature gradient calculations, causing post-hoc explainers to generate misleading attributions [18], [20]. To resolve this, technical teams must adopt robust, privacy-aware explainability algorithms capable of accounting for differential noise bounds.

Engineering & Scientific Significance

The significance of this review lies in synthesizing disparate mathematical paradigms into an operational engineering standard for ethical AI system design. Key scientific and engineering contributions include:

- Operationalizing Policy into Code: Establishing formal mapping between high-level regulatory frameworks (e.g., EU AI Act high-risk classification) and concrete technical metrics (e.g., equalized odds thresholds, ε privacy bounds, SHAP fidelity scores) [10].
- Standardizing Lifecycle Auditing Protocols: Providing a blueprint for continuous integration and continuous deployment (CI/CD) pipelines embedded with automated bias, privacy, and explainability validation tests prior to model deployment [21].
- Architectural Guidance for Trustworthy AI Systems: Offering scalable design patterns combining Federated Learning with Differential Privacy to secure multi-agent systems against data leakage and systemic bias propagation [19].

PRACTICAL APPLICATIONS

The governance frameworks evaluated in this study have immediate utility across critical operational domains:

- Healthcare Diagnostic Assistance: Ensuring medical imaging and risk-prediction AI models do not suffer from performance degradation across diverse patient ethnic demographics while strictly preserving patient HIPAA privacy through differential privacy [16].
- Automated Credit and Mortgage Approval: Implementing fair in-processing constraints to prevent credit-scoring algorithms from utilizing proxy variables (e.g., zip code) that enforce redlining practices, supported by SHAP explanations for adverse action notices [11], [15].
- Human Resources and Automated Recruitment: Deploying pre-processing re-weighing and bias auditing tools to screen resume parser algorithms, eliminating gender and racial

disparities in candidate shortlisting [12].

- Criminal Justice and Predictive Policing: Auditing risk-assessment algorithms (e.g., COMPAS) to eliminate disparate false-positive rates that unfairly penalize minority defendants [13], [17].

Table 3: Practical Deployment Matrix of Ethical AI Governance Tools across Key Domains

Domain	Primary Ethical Risk	Recommended Technical Governance Tool	Regulatory Compliance Target
Healthcare	Patient data leakage & diagnostic bias	DP-SGD + Federated Learning + SHAP	EU AI Act (Class IIb Medical Device), HIPAA
Finance & Banking	Discriminatory credit allocation	In-processing Fairness Penalty + Disparate Impact Audit	Equal Credit Opportunity Act (ECOA), Fair Lending
Recruitment / HR	Historical hiring bias amplification	Pre-processing Reweighing + Counterfactual Testing	NYC Local Law 144 (Automated Employment Decision Tools)
Autonomous Vehicles	Unclear liability & non-interpretable perception	High-fidelity Causal XAI + Immutable Event Logging	EU Machinery Directive, ISO/PAS 21448 (SOTIF)

LIMITATIONS

While this review provides comprehensive coverage of FATP governance, several technical limitations must be acknowledged:

- Computational Overhead: Advanced bias-mitigation retraining, iterative SHAP game-theoretic sampling, and DP-SGD gradient clipping significantly increase computational complexity and training latency—posing deployment barriers for resource-constrained edge computing environments [15], [16].

- Scope Restricted to Supervised Settings: The majority of formalized fairness metrics (e.g., equalized odds) rely on ground-truth target labels, making them difficult to apply directly to unsupervised generative architectures and large language models (LLMs) [21].
- Context-Dependency of Fairness Metrics: Mathematical definitions of fairness are mutually exclusive; satisfying Demographic Parity and Equalized Odds simultaneously is mathematically impossible under unequal base rates, requiring subjective human domain judgment [12], [13].

FUTURE SCOPE

To advance Ethical AI Governance beyond current limits, future research must focus on the following high-priority vectors:

- Governance Architectures for Generative AI and LLMs: Developing formal mathematical frameworks for hallucination detection, automated red-teaming, and dynamic alignment verification in foundation models and multimodal agents [21].
- Causal Fairness and Counterfactual Reasoning: Transitioning from purely correlational statistical fairness metrics toward causal DAG-based (Directed Acyclic Graph) models capable of capturing systemic mechanisms driving discrimination [17].
- Hardware-Accelerated Privacy and Auditing: Engineering specialized hardware accelerators and confidential computing environments (e.g., Secure Enclaves, Homomorphic Encryption chips) to execute real-time differentially private training without prohibitive runtime degradation [16].
- Automated Continuous Governance Pipelines: Creating open-source MLOps toolkits that automatically enforce compliance checks, generate standardized algorithmic impact assessments (AIAs), and monitor concept drift in real-time.

CONCLUSION

Ethical AI Governance represents an imperative prerequisite for the sustainable, societal-scale deployment of intelligent systems. This review paper has presented an in-depth analysis of the four foundational pillars of governance: Fairness, Accountability, Transparency, and Privacy (FATP). Through rigorous mathematical formulations, comparative literature review, and empirical trade-off evaluations, we demonstrated that achieving trustworthy AI requires navigating complex multi-objective technical trade-offs—particularly between differential privacy bounds, model utility, and sub-group fairness criteria.

Moving forward, isolated ethical tools must give way to integrated, full-lifecycle governance pipelines embedded directly within MLOps environments. By establishing verifiable technical metrics, continuous auditing protocols, and interdisciplinary collaboration between computer scientists, legal experts, and regulatory bodies, the scientific community can ensure that intelligent systems operate safely, equitably, and transparently for the benefit of all society.

REFERENCES

1. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Englewood Cliffs, NJ, USA: Prentice Hall, 2020.
2. E. Benzoni, A. Kapoor, and N. Narayanan, "Recognizing systemic risks in automated decision-making systems," IEEE Transactions on Technology and Society, vol. 3, no. 2, pp. 112–125, 2022.
3. D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector," The Alan Turing Institute, Tech. Rep., 2019.
4. R. Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge, UK: Polity Press, 2019.
5. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.
6. J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2019, pp. 195–200.
7. B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Machine Intelligence, vol. 1, no. 11, pp. 501–507, 2019.
8. F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA, USA: Harvard University Press, 2015.
9. M. Kearns and A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford, UK: Oxford University Press, 2019.
10. European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," COM(2021) 206 final, Brussels, Belgium, 2021.

11. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, 2021.
12. R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in Proc. Int. Conf. Mach. Learn. (ICML), 2013, pp. 325–333.
13. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 3315–3323.
14. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.
15. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
16. M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), 2016, pp. 308–318.
17. S. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, no. 3, pp. 671–732, 2016.
18. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2020, pp. 180–186.
19. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.
20. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in IEEE Symposium on Security and Privacy (SP), 2017, pp. 3–18.
21. Y. Shen et al., "Large language model governance: Trustworthy and ethical alignment strategies," IEEE Intelligent Systems, vol. 39, no. 1, pp. 45–58, 2024.

***Author for Correspondence**

Subhash Ghosh

E-mail: subhash.ghosh@gmail.com

[1]Associate Professor, Department of Artificial Intelligence & Machine Learning, Dr. B. C. Roy Polytechnic

[2]Assistant Professor, Department of Artificial Intelligence & Machine Learning, Bengal School of Technology & Management

[3]Research Scholar, Department of Information Technology, Sanaka Educational Trust's Group of Institutions

Received Date: June 18, 2026

Accepted Date: July 03, 2026

Published Date: July 19, 2026

Citation: Subhash Ghosh, Pranav Bhide, Kaustubh Deshmukh. Ethical AI Governance: A Comprehensive Review of Fairness, Accountability, Transparency, and Privacy in Intelligent Systems. Journal of Ethical and Explainable AI Systems. 2026; 2(2): 61-76p.