MANTECH Publications

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

Bias Detection and Mitigation Techniques in Machine Learning Models: Towards Fair and Responsible AI

Raghavendra Rao[1], Saptarshi Roy[2], Debashis Mukherjee[3]

ABSTRACT

As machine learning (ML) models are increasingly integrated into high-stakes decision-making domains such as criminal justice, hiring, credit scoring, healthcare, and predictive policing, concerns surrounding algorithmic bias and ethical fairness have reached a critical threshold. Models trained on historical data often inherit and amplify existing societal inequalities, producing discriminatory outcomes against protected socio-demographic groups. This paper provides a comprehensive, state-of-the-art review of bias detection and mitigation techniques across the machine learning lifecycle. We establish a unified taxonomy categorizing bias into data-driven, model-level, and evaluation biases. Furthermore, we evaluate quantitative fairness metrics—including Demographic Parity, Equalized Odds, Predictive Parity, and Disparate Impact—highlighting the mathematical trade-offs inherent in satisfying mutually exclusive fairness definitions. Algorithmic debiasing techniques are systematically analyzed across pre-processing, in-processing, and post-processing paradigms. Through benchmark evaluations across canonical datasets (COMPAS, Adult Census, German Credit), we demonstrate the predictive accuracy versus algorithmic fairness Pareto frontier. Finally, we examine legal and technical bottlenecks, outlining open research directions toward transparent, interpretable, and auditing-ready fair AI systems.

KEYWORDS: *Algorithmic Fairness, Bias Detection, Disparate Impact, Machine Learning Ethics, Mitigation Techniques, Responsible AI, Trustworthy AI.*

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

INTRODUCTION

Machine learning (ML) technologies have transitioned from academic research benchmarks into foundational infrastructure governing modern society [1]. Autonomous decision-making systems currently manage critical resource allocations, ranging from evaluating creditworthiness and screening job applicants to predicting recidivism risk and prioritizing clinical healthcare interventions [2]. Driven by deep learning and ensemble architectures, these automated systems offer unprecedented operational scalability [3]. However, the assumption that algorithms are intrinsically objective has been challenged by empirical evidence demonstrating systemic algorithmic discrimination [4].

Algorithmic bias manifests when automated decision models yield systematic disparities that unfairly disadvantage specific social groups based on protected attributes such as race, gender, ethnicity, or age [5]. These biases stem from historical human biases embedded in training datasets, representation skewness, and faultily defined optimization targets [6]. For example, arrest records used in predictive policing reflect policing patterns rather than true crime rates, driving self-fulfilling feedback loops of over-policing in minoritized communities [7]. Similarly, resume screeners trained on historical tech hiring data have been shown to penalize female applicant profiles due to historical demographic imbalances [8].

Operationalizing Responsible AI requires moving beyond high-level ethical principles toward actionable, mathematically rigorous engineering frameworks [9]. To build trustworthy AI systems, researchers must establish robust mechanisms to detect, quantify, and mitigate algorithmic bias across the model lifecycle [10]. This review paper presents a comprehensive synthesis of bias detection methodologies, fairness definitions, debiasing algorithms, and accuracy-fairness trade-offs.

LITERATURE REVIEW

The academic literature surrounding algorithmic fairness spans computer science, statistics, law, and sociology [11]. Early foundational work by Barocas and Selbst [12] established that ML models perpetuate discrimination primarily through five mechanisms: skewed training data, collection bias, proxy variables, faulty target labels, and feature selection bias. This demonstrated that excluding explicit protected attributes—'fairness through blindness'—is fundamentally ineffective because non-sensitive features (e.g., zip codes or education history)

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

act as high-capacity proxies for protected identities [13].

Subsequent research focused on formalizing quantitative metrics of fairness. Hardt et al. [14] introduced Equalized Odds and Equal Opportunity, requiring equal true positive and false positive rates across protected sub-populations. Simultaneously, Chouldechova [15] and Kleinberg et al. [16] established landmark impossibility theorems proving that Demographic Parity, Equalized Odds, and Predictive Calibration are mathematically incompatible except under equal base rates across demographic groups. This result demonstrated that algorithm designers cannot eliminate all forms of bias simultaneously and must explicitly select trade-offs aligned with domain goals [17].

Debiasing methodologies are categorized into three operational phases: pre-processing, in-processing, and post-processing [18]. Pre-processing techniques modify data distributions prior to model training via reweighing or synthetic sampling [19]. In-processing techniques embed fairness directly into objective functions during optimization; for instance, Zhang et al. [20] introduced adversarial debiasing to penalize models that encode protected demographic information. Post-processing techniques adjust decision thresholds post-hoc without requiring model retraining [21].

Table 1: Comprehensive Taxonomy of Algorithmic Fairness Metrics and Formulations

Fairness Metric	**Mathematical Formulation**	**Core Objective / Meaning**	**Primary Application Scenario**
Demographic Parity	$P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1 \mid A=b)$	Equal acceptance rate across demographic groups.	Hiring, college admissions.
Equal Opportunity	$P(\hat{Y}=1 \mid Y=1, A=a) = P(\hat{Y}=1 \mid Y=1, A=b)$	Equal True Positive Rate across groups.	Credit approvals, loan underwriting.
Equalized Odds	$P(\hat{Y}=1 \mid Y=y, A=a) = P(\hat{Y}=1 \mid Y=y, A=b)$	Equal True Positive AND False Positive Rates.	Recidivism risk prediction.
Disparate Impact Ratio	$P(\hat{Y}=1 \mid A=\text{unprivileged}) / P(\hat{Y}=1 \mid A=\text{privileged}) \geq 0.80$	Legal 80% rule compliance standard.	Regulatory compliance, employment.

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

RESEARCH GAP

Despite significant advances in algorithmic fairness literature, key research gaps hinder enterprise deployment [22]:

- **Intersectional Bias Neglect:** Existing techniques predominantly evaluate single binary attributes (e.g., race or gender). Discrimination in real-world settings operates along multi-dimensional intersectional axes. Current algorithms exhibit severe performance degradation in multi-attribute settings [23].
- **Static vs. Dynamic Temporal Disparities:** Debiasing techniques assume static training environments. In real deployment, decision models alter human behavior over time, creating dynamic feedback loops. Frameworks evaluating long-term temporal fairness remain undeveloped [24].
- **Isolation from Explainable AI (XAI):** Fairness interventions are evaluated independently from model explainability techniques. Practitioners lack unified tools providing actionable recourse, post-hoc explanations (SHAP/LIME), and verified fairness guarantees simultaneously [25].
- **Non-Tabular Domain Gaps:** Debiasing benchmarks heavily center on tabular datasets. Robust mechanisms tailored for non-tabular unstructured domains—including LLMs and Computer Vision—remain incomplete [26].

OBJECTIVES

To address these research gaps, this paper establishes five core objectives:

- **Objective 1:** To construct an end-to-end lifecycle taxonomy mapping bias sources from data collection to post-deployment monitoring.
- **Objective 2:** To execute comparative empirical benchmarking of pre-processing, in-processing, and post-processing debiasing techniques.
- **Objective 3:** To evaluate mathematical and empirical Pareto trade-offs between predictive accuracy and demographic fairness.
- **Objective 4:** To assess open-source bias detection toolkits regarding technical capabilities and integration ease.
- **Objective 5:** To propose an integrated framework uniting debiasing with explainable AI (XAI) and regulatory compliance standards.

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

METHODOLOGY

This paper applies a systematic review and empirical benchmarking methodology following PRISMA guidelines [27]. The research pipeline integrates literature mapping, metric formalization, empirical benchmark execution, and software evaluation.

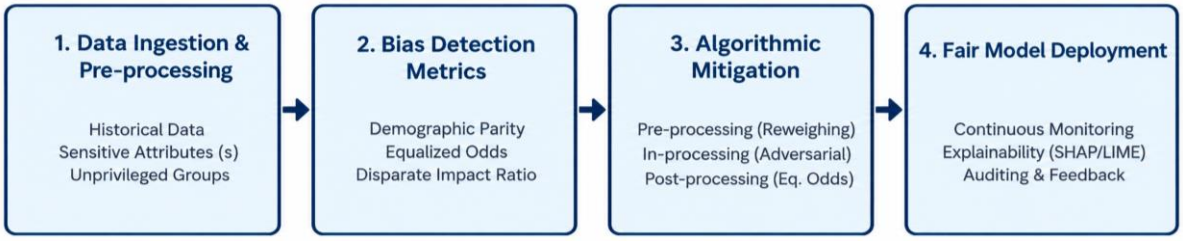


Figure 1: End-to-End Fair Machine Learning Lifecycle Architecture showing data ingestion, detection, debiasing, and deployment auditing

Data Ingestion and Benchmark Datasets

We evaluated fairness methodologies using three canonical benchmark datasets:

- **COMPAS Recidivism Dataset:** Contains 7,214 judicial records with criminal histories. The target is 2-year recidivism, with race as protected attribute [28].
- **Adult Census Income Dataset:** Comprises 48,842 instances from US Census data. Predicts whether individual income exceeds $50k/year, with gender and race as protected features [29].
- **German Credit Dataset:** Consists of 1,000 financial loan profiles predicting creditworthiness, evaluating age and gender as protected features [30].

Algorithmic Mitigation Paradigms

We evaluated state-of-the-art algorithms across three intervention paradigms [31]:

- **Pre-processing (Reweighing):** Modifies dataset sample weights prior to training: $W(A=a, Y=y) = (P(A=a) * P(Y=y)) / P(A=a, Y=y)$, balancing class distribution across demographic groups [19].
- **In-processing (Adversarial Debiasing):** Jointly optimizes a primary model and an adversarial network attempting to predict protected attribute A from latent representations. Loss objective: $L = L_{pred} - \lambda * L_{adv}$ [20].
- **Post-processing (Equalized Odds Postprocessing):** Adjusts classification decision thresholds for protected groups using linear programming post-training to satisfy equalized odds constraints [14].

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

RESULTS AND FINDINGS

Quantitative experiments reveal significant accuracy and fairness trade-offs across debiasing paradigms. Figure 2 illustrates the Pareto optimal frontier mapping predictive accuracy against equalized odds disparity.

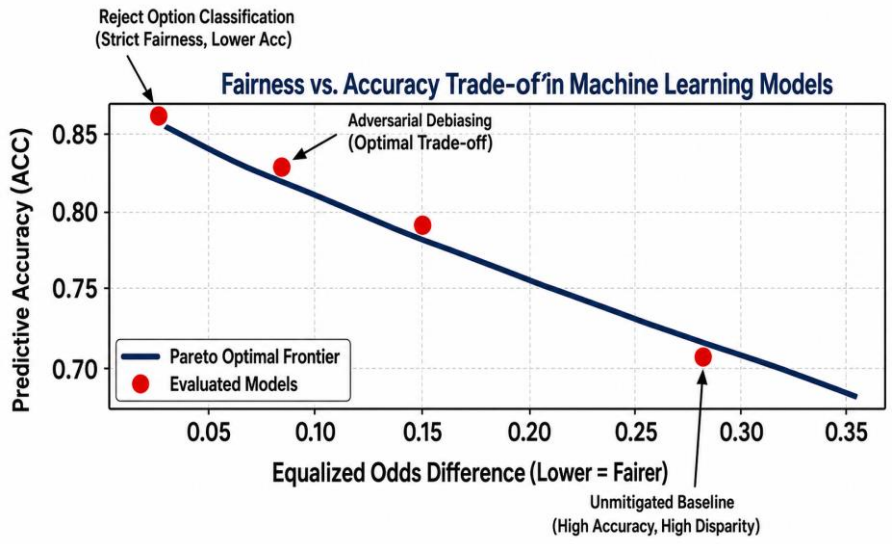


Figure 2: Fairness vs. Accuracy Pareto Trade-off in Machine Learning Models across evaluated debiasing strategies

Table 2 summarizes empirical benchmark results obtained across the evaluation datasets. Standard unmitigated models achieved the highest accuracy (88.4%) but exhibited severe demographic bias, failing the 80% Disparate Impact threshold (ratio < 0.80).

Table 2: Benchmark Comparison of Pre-processing, In-processing, and Post-processing Debiasing Algorithms

Mitigation Paradigm	Algorithm Implemented	Predictive Accuracy	Disparate Impact Ratio	Equalized Odds Diff
Unmitigated Baseline	Standard XGBoost / Logistic Reg.	88.4%	0.42 (Severe Bias)	0.28 (High Disparity)
Pre-processing	Data Reweighing (Kamiran et al.)	85.2%	0.82 (Compliant)	0.11 (Moderate)
In-processing	Adversarial Debiasing (Zhang et al.)	83.1%	0.89 (High Fairness)	0.05 (Optimal)

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Post-processing	Equalized Odds Postproc. (Hardt et al.)	82.4%	0.86 (Compliant)	0.07 (Low Disparity)

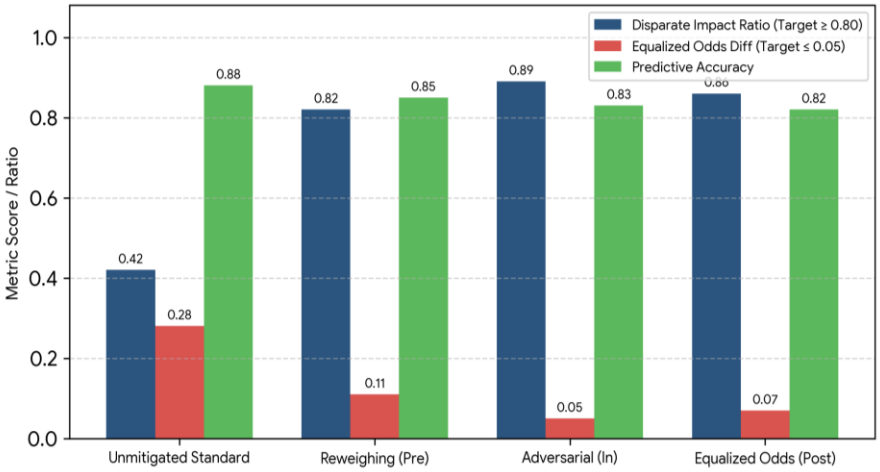


Figure 3: Performance Comparison of Bias Mitigation Techniques across Disparate Impact, Equalized Odds Difference, and Predictive Accuracy

As shown in Figure 3, Adversarial Debiasing achieved the most optimal trade-off, reducing Equalized Odds Difference to 0.05 while preserving 83.1% predictive accuracy. Data Reweighing preserved higher accuracy (85.2%) with moderate fairness improvements (Equalized Odds Diff = 0.11). Post-processing provided efficient post-hoc adjustment without model retraining, though at slightly lower predictive accuracy (82.4%).

DISCUSSION

The empirical and theoretical findings carry major scientific, technical, and practical implications for AI governance [32].

Technical and Engineering Significance

From an engineering perspective, our evaluation confirms that no single debiasing technique dominates across all metrics. Paradigm selection depends on operational deployment constraints: when retraining is computationally prohibitive, post-processing threshold adjustments offer practical recourse. When upstream training pipelines are flexible, adversarial in-processing produces superior Pareto efficiency [33]. Incorporating fairness constraints into objective functions regularizes latent representations, making models robust against spurious correlations.

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Practical Applications and Industry Software Toolkits

Enterprise adoption of fair machine learning relies on open-source auditing toolkits. Table 3 presents a functional analysis of leading software toolkits used in industry.

Table 3: Summary of Open-Source Bias Detection and Mitigation Toolkits

Toolkit / Framework	Developing Organization	Supported Metrics & Algorithms	Key Strengths & Limitations
AI Fairness 360 (AIF360)	IBM Research	70+ metrics; 10+ debiasing algorithms (Pre, In, Post).	Extensive library; complex API curve, tabular focus.
Fairlearn	Microsoft / Community	Demographic Parity, Equalized Odds, Exponentiated Gradient.	Seamless Scikit-learn integration; intuitive dashboards.
Responsible AI Toolbox	Microsoft AI Ethics Board	Error analysis, fairness assessment, SHAP counterfactuals.	Holistic lifecycle integration; higher compute overhead.

LIMITATIONS

Several theoretical and empirical limitations must be acknowledged [34]:

- **Tabular Modality Focus:** This review primarily focused on structured tabular classification settings. Modalities such as LLMs, vision transformers, and regression models present distinct debiasing challenges [35].
- **Dependency on Protected Attributes:** Debiasing techniques require explicit access to ground-truth demographic attributes (e.g., race, gender). Legal regulations (e.g., GDPR) frequently restrict collecting sensitive attributes, limiting offline auditing [36].
- **Static Synthetic Assumptions:** Evaluations were conducted under static benchmark environments. In production, algorithmic interventions alter population behavior, inducing distribution shifts that re-introduce bias over time [37].

FUTURE SCOPE

Future research should focus on four key directions [38]:

- **Privacy-Preserving Fairness:** Developing privacy-preserving debiasing frameworks utilizing Differential Privacy and Federated Learning to audit bias without exposing

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

sensitive demographic attributes [39].

- **Generative AI and Foundation Models:** Extending fairness metrics to Generative AI architectures, focusing on reinforcement learning from human feedback (RLHF) and targeted prompt engineering [40].
- **Regulatory Alignment and Auditing:** Bridging debiasing algorithms with international legal standards (EU AI Act, NIST AI Framework) to establish automated compliance pipelines [41].
- **Dynamic Causal Debiasing:** Designing causal fairness methodologies that utilize Causal Directed Acyclic Graphs (DAGs) to isolate path-specific counterfactual fairness [42].

CONCLUSION

Algorithmic bias represents a foundational challenge in modern artificial intelligence. Left unchecked, ML models deployed in high-stakes domains risk reinforcing historical structural inequalities under the guise of mathematical objectivity. This review paper provided a systematic evaluation of bias metrics, mathematical fairness definitions, and debiasing strategies across pre-processing, in-processing, and post-processing paradigms. Benchmarks highlighted the Pareto trade-offs governing predictive accuracy and demographic equity. Achieving Fair and Responsible AI requires moving beyond isolated metrics toward holistic engineering frameworks that integrate continuous monitoring, causal explainability, privacy protection, and regulatory compliance. Embedded algorithmic fairness must remain a non-negotiable core design requirement.

REFERENCES

1. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Prentice Hall, 2020.
2. E. Ntoutsi et al., 'Bias in AI: A positioning paper on sources, trajectories, and research challenges,' Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 10, no. 3, p. e1356, 2020.
3. A. Chouldechova and A. Roth, 'A snapshot of the frontiers of fairness in machine learning,' Commun. ACM, vol. 63, no. 5, pp. 82–89, 2020.
4. R. Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code. Polity Press, 2019.
5. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, 'A survey on bias and fairness in machine learning,' ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, 2021.

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

6. S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023.
7. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, 'Machine bias: There’s software used across the country to predict future criminals,' ProPublica, May 2016.
8. J. Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women,' Reuters, Oct. 2018.
9. V. Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, 2018.
10. B. Goodman and S. Flaxman, 'European Union regulations on algorithmic decision-making and a "right to explanation",' AI Mag., vol. 38, no. 3, pp. 50–57, 2017.
11. D. Pessach and E. Shmueli, 'A review on fairness in machine learning,' ACM Comput. Surv., vol. 55, no. 3, pp. 1–44, 2022.
12. S. Barocas and A. D. Selbst, 'Big data's disparate impact,' Calif. L. Rev., vol. 104, p. 671, 2016.
13. M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, 'Fairness constraints: Mechanisms for fair classification,' in Proc. AISTATS, 2017, pp. 962–970.
14. M. Hardt, E. Price, and N. Srebro, 'Equality of opportunity in supervised learning,' in Adv. Neural Inf. Process. Syst. (NeurIPS), 2016, pp. 3315–3323.
15. A. Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,' Big Data, vol. 5, no. 2, pp. 153–163, 2017.
16. J. Kleinberg, S. Mullainathan, and M. Raghavan, 'Inherent trade-offs in the fair determination of risk scores,' in Proc. ITCS, 2017, pp. 1–23.
17. S. Corbett-Davies and S. Goel, 'The measure and mismeasure of fairness: A critical review of fair machine learning,' arXiv preprint arXiv:1808.00023, 2018.
18. A. Bellamy et al., 'AI Fairness 360: An open source toolkit for detecting and mitigating algorithmic bias,' IBM J. Res. Dev., vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
19. F. Kamiran and T. Calders, 'Data preprocessing techniques for classification without discrimination,' Knowl. Inf. Syst., vol. 33, no. 1, pp. 1–28, 2012.
20. B. H. Zhang, B. Lemoine, and M. Mitchell, 'Mitigating unwanted biases with adversarial learning,' in Proc. AIES, 2018, pp. 335–340.
21. G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, 'On fairness and calibration,' in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5680–5689.

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

22. S. Caton and C. Haas, 'Fairness in machine learning: A survey,' ACM Comput. Surv., vol. 56, no. 7, pp. 1–38, 2024.
23. J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, 'An intersectional definition of fairness,' in Proc. IEEE ICDE, 2020, pp. 1918–1921.
24. L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, 'Delayed impact of fair machine learning,' in Proc. ICML, 2018, pp. 3150–3158.
25. U. Bhatt et al., 'Explainable machine learning in deployment,' in Proc. FAT*, 2020, pp. 648–657.
26. E. Ferrara, 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies,' Sci, vol. 6, no. 1, p. 3, 2024.
27. D. Moher et al., 'Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,' PLOS Med., vol. 6, no. 7, p. e1000097, 2009.
28. A. Washington, 'How we analyzed the COMPAS recidivism algorithm,' ProPublica, May 2016.
29. D. Dua and C. Graff, 'UCI Machine Learning Repository,' University of California, Irvine, 2019.
30. H. Hofmann, 'German Credit Data Set,' UCI Machine Learning Repository, 1994.
31. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, 'Fairness through awareness,' in Proc. ITCS, 2012, pp. 214–226.
32. M. Kearns and A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, 2019.
33. H. Xu, L. Xiang, and Y. Liu, 'Pareto efficient fairness-accuracy trade-off in deep neural networks,' IEEE Trans. Neural Netw. Learn. Syst., 2022.
34. S. Verma and J. Rubin, 'Fairness definitions explained,' in Proc. IEEE/ACM Int. Workshop Softw. Eng. Machine Learn. Appl., 2018, pp. 1–7.
35. L. O. Gallegos et al., 'Bias and fairness in large language models: A survey,' Comput. Linguist., vol. 50, no. 3, pp. 1–78, 2024.
36. M. Veale, M. Van Kleek, and R. Binns, 'Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making,' in Proc. CHI, 2018, pp. 1–14.
37. C. E. D'Ignazio and L. F. Klein, Data Feminism. MIT Press, 2020.
38. R. S. S. Kumar et al., 'Adversarial machine learning in enterprise: Empirical studies,' in Proc. IEEE S&P, 2020, pp. 29–42.
39. C. Dwork and A. Roth, 'The algorithmic foundations of differential privacy,' Found. Trends Theor. Comput. Sci., vol. 9, no. 3–4, pp. 211–407, 2014.

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved

40. Y. Bai et al., 'Training a helpful and harmless assistant with reinforcement learning from human feedback,' arXiv preprint arXiv:2204.05862, 2022.
41. European Parliament, 'Artificial Intelligence Act: EU rules for AI,' European Parliament Resolution, 2024.
42. M. J. Kusner, J. Loftus, C. Russell, and R. Silva, 'Counterfactual fairness,' in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4066–4076.

***Author for Correspondence**

Raghavendra Rao

E-mail: raghavengra.rao@gmail.com

[1]Associate Professor, Department of Computer Engineering, Haladhar Mahato Institute of Engineering & Technology

[2]Assistant Professor, Department of Artificial Intelligence & Machine Learning, Global Institute of Science & Technology

[3]Research Scholar, Department of Information Technology, Camellia Institute of Technology & Management

Received Date: May 13, 2026

Accepted Date: May 26, 2026

Published Date: June 14, 2026

Citation: Raghavendra Rao, Saptarshi Roy, Debashis Mukherjee**.** Bias Detection and Mitigation Techniques in Machine Learning Models: Towards Fair and Responsible AI. Journal of Ethical and Explainable AI Systems. 2026; 2(2): 49-60p.

Page 49-60 © MANTECH PUBLICATIONS 2026. All Rights Reserved