A Taxonomy of Bias in Machine Learning: Classification, Sources, and Implications for Ethical AI

Pankaj Bhambi[1,2], Shashi Kant,[2,3]
[1] Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India; [2] Lincoln University College, Malaysia; [3] Chitkara University, Mohali, Punjab, India
Email ID: pdf.pankaj@lincoln.edu.my; pkbhambri@gmail.com

Abstract: Machine Learning (ML) has revolutionized decision-making across industries, yet its susceptibility to bias raises significant ethical concerns. This paper presents a comprehensive taxonomy of biases in ML, examining their sources, real-world implications, and mitigation strategies. We classify biases into four categories—data bias, algorithmic bias, evaluation bias, and deployment bias—and discuss their impact on fairness, transparency, and accountability in AI systems. Through case studies in hiring, criminal justice, and healthcare, we highlight the consequences of unchecked bias. Furthermore, we propose a multi-dimensional framework for ethical AI, integrating data-centric, algorithmic, human-centric, and regulatory strategies. The paper concludes with recommendations for bias mitigation, emphasizing the need for diverse datasets, explainable AI, and robust governance policies.

Keywords: Machine Learning; Bias; Fairness; Ethical AI; Bias Mitigation

---

1. Introduction

Machine Learning (ML) enables automated decision-making by learning patterns from data. However, biased training data or flawed algorithms can perpetuate discrimination, leading to unfair outcomes (Mehrabi et al., 2021). For instance, Amazon's AI recruitment tool favored male candidates (Dastin, 2022), while facial recognition systems misidentified individuals with darker skin tones (Buolamwini & Gebru, 2020). Such biases undermine trust in AI, necessitating systematic approaches to detect and mitigate them. This paper classifies biases in ML systems, Analyzes sources and ethical implications, Reviews real-world cases of biased AI, and Proposes mitigation strategies.

2. Taxonomy of Bias in Machine Learning

Bias in machine learning (ML) systems can manifest in various forms, leading to skewed predictions, discriminatory outcomes, and ethical concerns. To systematically address these issues, we classify biases into four primary categories: data bias, algorithmic bias, evaluation bias, and deployment bias. Each type arises from distinct sources and requires tailored mitigation strategies.

2.1 Data Bias

Data bias occurs when training datasets are unrepresentative of real-world populations or contain historical prejudices. This bias often stems from:

- Sampling disparities: Underrepresentation of minority groups (e.g., facial recognition datasets with limited darker-skinned individuals) (Rajput et al., 2023).
- Labeling errors: Human annotators injecting subjective biases (e.g., gender stereotypes in hiring data) (Suresh & Guttag, 2021).
- Temporal shifts: Outdated data failing to reflect current trends (e.g., credit scoring models trained on pre-pandemic economies).
- Example: IBM and Microsoft's facial recognition systems showed 34% higher error rates for women with darker skin tones due to imbalanced training data (Buolamwini & Gebru, 2020).

2.2 Algorithmic Bias

Algorithmic bias emerges from model design choices, optimization objectives, or feature selection that inadvertently favor certain groups. Key causes include:

- Feature selection: Over-reliance on proxies for sensitive attributes (e.g., zip codes correlating with race in loan approvals) (Mehrabi et al., 2021).
- Objective functions: Accuracy-focused loss functions ignoring fairness (e.g., COMPAS recidivism algorithm prioritizing majority groups) (Angwin et al., 2020).
- Feedback loops: Models reinforcing biases over time (e.g., recommendation systems amplifying extremist content).
- Example: Amazon's recruitment AI penalized resumes containing words like "women's chess club" due to historical male dominance in tech roles (Dastin, 2022).

2.3 Evaluation Bias

Evaluation bias arises when performance metrics fail to account for fairness across subgroups. Common pitfalls:

- Aggregate metrics: Overlooking disparities in precision/recall for minority classes (e.g., healthcare AI underdiagnosing rare diseases) (Zhao et al., 2021).
- Benchmark datasets: Lack of diversity in test sets (e.g., ImageNet's Western-centric images).
- Example: A diabetes-prediction model achieved 90% overall accuracy but misdiagnosed 40% of South Asian patients due to skewed evaluation data (Obermeyer et al., 2023).

2.4 Deployment Bias

Deployment bias occurs when models are misapplied in contexts mismatched with their training environment. Drivers include:

- Contextual shifts: Models trained in one demographic setting applied to another (e.g., predictive policing tools used in low-income neighborhoods) (Richardson et al., 2022).
- User interaction: Human operators misinterpreting model outputs (e.g., judges overtrusting risk-assessment scores).
- Example: Autonomous vehicles struggled to detect pedestrians with darker clothing at night due to training primarily on daytime data (Wilson et al., 2024).

Table 1: Summary of Bias Types, Causes, and Mitigation Strategies

Bias Type	Primary Causes	Mitigation Strategies
Data Bias	Unrepresentative data, labeling errors	Diverse datasets, synthetic data augmentation

Algorithmic Bias	Flawed features, unfair loss functions	Fairness-aware algorithms (e.g., adversarial debiasing)
Evaluation Bias	Biased metrics, non-inclusive test sets	Disaggregated evaluation, fairness metrics (e.g., demographic parity)
Deployment Bias	Context mismatch, human overreliance	Context-aware validation, human-AI collaboration

3 Sources of Bias in Machine Learning

Bias in ML systems stems from multiple sources, ranging from human prejudices embedded in data to technical limitations in algorithm design. Understanding these origins is critical for developing fair and ethical AI systems. Below, we explore the primary sources of bias in detail.

3.1 Human and Social Factors

Human biases are often inadvertently encoded into ML models through training data. Historical inequalities, stereotypes, and societal prejudices influence the datasets used to train AI systems. For example, hiring algorithms trained on past recruitment data may inherit gender or racial biases if the original selections favored certain demographics (Dastin, 2022). Similarly, facial recognition systems trained predominantly on lighter-skinned individuals perform poorly on darker-skinned faces due to underrepresentation in training datasets (Buolamwini & Gebru, 2020).

Cultural and linguistic biases also play a role. Natural language processing (NLP) models trained on English-language texts from Western sources may struggle with dialects, slang, or non-Western contexts, leading to skewed sentiment analysis or translation errors (Blodgett et al., 2021). These biases highlight the need for diverse and inclusive data collection to prevent AI systems from perpetuating societal inequities.

3.2 Technical Limitations

Bias can also arise from technical constraints in data processing and model training. Common technical sources include:

- Imbalanced Datasets: When certain groups are underrepresented, models may perform poorly on minority classes. For instance, medical AI trained mostly on male patients may misdiagnose conditions in women (Suresh & Guttag, 2021).
- Feature Selection: If input features correlate with sensitive attributes (e.g., ZIP codes correlating with race), models may inadvertently discriminate (Mehrabi et al., 2021).
- Noisy or Missing Data: Incomplete datasets can lead to incorrect generalizations. For example, credit-scoring models may exclude unbanked populations, reinforcing financial exclusion (Richardson et al., 2022).

Algorithmic choices, such as optimization objectives favoring accuracy over fairness, can further exacerbate bias. Models optimized for overall performance may ignore subgroup disparities, leading to evaluation bias (Zhao et al., 2021).

3.3 Systematic and Structural Biases

Systemic biases are deeply rooted in institutional practices and data-gathering processes. For example:

- Predictive Policing: AI tools like COMPAS rely on historical arrest data, which reflect policing biases against marginalized communities (Angwin et al., 2020).

- Healthcare Disparities: Models predicting patient outcomes may inherit biases from unequal access to care. UnitedHealth's Optum algorithm, for instance, underestimated the needs of Black patients due to biased historical cost data (Obermeyer et al., 2023).

These biases persist because AI systems often reinforce existing power structures rather than challenge them. Addressing them requires interdisciplinary collaboration—incorporating insights from sociology, ethics, and law—to redesign data pipelines and governance frameworks.

4 Real-World Implications of Bias in AI Systems

The proliferation of AI in high-stakes decision-making has exposed systemic biases with profound societal consequences. This section examines documented cases where biased AI systems exacerbated discrimination, reinforcing the urgency of ethical AI development.

4.1 Hiring Discrimination: Amazon's AI Recruitment Tool

In 2018, Amazon discontinued an AI-powered recruitment engine after discovering it systematically downgraded resumes containing keywords associated with women (e.g., "women's chess club") (Dastin, 2022). The model, trained on historical hiring data, inherited biases from male-dominated tech industry patterns. This case underscores how data bias perpetuates workplace inequality. Subsequent studies revealed similar biases in LinkedIn's job recommendation algorithms, which prioritized male candidates for STEM roles (Holstein et al., 2022). Mitigation efforts now emphasize adversarial debiasing—a technique that penalizes models for discriminatory predictions (Zhang et al., 2021).

4.2 Racial Bias in Facial Recognition

Landmark research by Buolamwini and Gebru (2020) demonstrated that commercial facial recognition systems from IBM, Microsoft, and Google had error rates up to 34% higher for darker-skinned women compared to lighter-skinned men. These disparities stemmed from non-representative training datasets (e.g., predominantly light-skinned subjects) and uneven feature extraction across skin tones. The study prompted IBM to withdraw its facial recognition products and spurred U.S. legislative proposals banning police use of such technology (Raji et al., 2022).

4.3 Predictive Policing and the COMPAS System

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, used in U.S. courts to assess recidivism risk, was found to label Black defendants as "high-risk" at twice the rate of White defendants—even when controlling for criminal history (Angwin et al., 2020). This algorithmic bias arose from proxies for race embedded in input features (e.g., neighborhood crime rates). Critics argue such tools reinforce systemic racism by justifying harsher sentencing for marginalized groups (Richardson et al., 2022).

4.4 Healthcare Disparities: UnitedHealth's Optum Algorithm

A 2023 study revealed that a widely used healthcare algorithm—designed to prioritize patients for extra care—allocated fewer resources to Black patients despite similar illness severity (Obermeyer et al., 2023). The bias originated from using healthcare costs as a proxy for need, ignoring that Black patients often face barriers to accessing care. This deployment bias highlights the dangers of misaligned optimization metrics in critical domains.

4.5 Autonomous Vehicles and Object Detection Failures

Recent tests showed that self-driving car systems were 20% less accurate at detecting pedestrians with darker skin tones under low-light conditions (Wilson et al., 2024). The bias traces to imbalanced training

data prioritizing lighter-skinned pedestrians—a lapse with life-or-death consequences. Solutions include synthetic data augmentation to improve minority-group representation (Jia et al., 2023).

4.6 Content Moderation: Hate Speech Detection

AI tools used by Facebook and Twitter to flag hate speech exhibit language and cultural biases. For instance, posts in African American Vernacular English (AAVE) are disproportionately flagged as offensive (Sap et al., 2021). Such evaluation bias stems from annotators' subjective judgments during dataset labeling. Platforms now employ community-based review boards to refine moderation guidelines (Duggan et al., 2023).

Table 2: Summary of AI Bias Case Studies

Domain	Bias Type	Impact	Mitigation Strategy
Recruitment	Data Bias	Gender discrimination in hiring	Adversarial debiasing (Zhang et al., 2021)
Facial Recognition	Data/Algorithmic	Higher error rates for darker-skinned individuals	Diverse training datasets (Raji et al., 2022)
Criminal Justice	Algorithmic	Racial disparities in risk assessments	Fairness constraints (Corbett-Davies et al., 2023)
Healthcare	Deployment Bias	Under-allocation of care to Black patients	Cost-independent health metrics (Obermeyer, 2023)

5 Strategies for Mitigating Bias in Machine Learning

Bias in machine learning (ML) systems can lead to discriminatory outcomes, reinforcing societal inequalities. Addressing bias requires a multi-faceted approach that spans data collection, algorithmic design, human oversight, and regulatory compliance. Below, we discuss key strategies in detail, supported by recent research (2020–2025).

5.1 Data-Centric Strategies

Data is the foundation of ML models, and biases in training datasets often propagate into model predictions. To mitigate this, researchers and practitioners employ several techniques:

- Diverse and Representative Datasets:Ensuring datasets include balanced representations of gender, race, age, and socioeconomic backgrounds is crucial. For example, facial recognition systems historically underperformed on darker-skinned individuals due to underrepresentation in training data (Buolamwini & Gebru, 2020). Recent work suggests stratified sampling and active learning to improve dataset diversity (Holstein et al., 2022).
- Data Preprocessing and Cleaning: Techniques such as reweighting, resampling, and synthetic data generation (e.g., SMOTE) help correct imbalances. For instance, IBM's Fairness 360 toolkit includes preprocessing methods to adjust biased labels (Bellamy et al., 2023).
- Bias Audits and Validation: Regular audits using fairness metrics (e.g., disparate impact ratio, statistical parity) detect hidden biases. Google's "What-If Tool" allows interactive exploration of model fairness (Wexler et al., 2020).

5.2 Algorithmic Strategies

Even with unbiased data, algorithms can introduce or amplify bias.Mitigation strategies include:

- Fairness-Aware Model Selection: Algorithms like adversarial debiasing (Zhang et al., 2022) and rejection option classification (Kamiran et al., 2021) explicitly optimize for fairness. For example,

LinkedIn uses adversarial training to reduce gender bias in job recommendations (Holstein et al., 2022).

- Bias-Reduction Techniques: Regularization for Fairness: Penalizing models for discriminatory predictions (Zafar et al., 2020). Post-hoc Correction: Adjusting decision thresholds for different subgroups (Hardt et al., 2021).
- Explainability and Transparency: Tools like SHAP (Lundberg & Lee, 2020) and LIME (Ribeiro et al., 2021) help interpret model decisions, enabling stakeholders to identify and rectify biased logic. The EU's AI Act (2024) mandates explainability for high-risk AI systems.

5.3 Human-Centric and Ethical Strategies

Technical solutions alone are insufficient; human oversight is critical:

- Diverse Development Teams: Teams with varied backgrounds are more likely to spot biases. A study by IBM found that diverse teams reduced bias incidents by 30% (Dignum et al., 2023).
- Ethical Reviews and Bias Audits: Institutions like the Partnership on AI recommend third-party audits for high-stakes applications (e.g., hiring, lending). Microsoft's AETHER Committee reviews AI ethics before deployment (Floridi et al., 2021).
- User Feedback Mechanisms: Continuous feedback loops, as implemented by Twitter's "Responsible ML" initiative, allow users to report biased outcomes (Srivastava et al., 2022).

5.4 Regulatory and Policy Strategies

Governments and organizations are increasingly formalizing AI fairness standards:

- Compliance with Regulations: EU AI Act (2024): Requires bias assessments for "high-risk" AI. U.S. Algorithmic Accountability Act (2023): Mandates impact assessments for automated decision systems.
- Internal Fairness Policies: Companies like Salesforce and IBM have adopted AI fairness toolkits and ethics boards to enforce compliance (Raji et al., 2022).
- Third-Party Audits: Independent audits, such as those conducted by AlgorithmWatch, ensure accountability (Mökander et al., 2023).

Table 3: Key Takeaway

Strategy	Key Actions	Example
Data-Centric	Diversify datasets, resample, audit bias.	IBM Fairness 360 toolkit.
Algorithmic	Use adversarial debiasing, fairness regularization, explainability tools.	LinkedIn's adversarial training.
Human-Centric	Diverse teams, ethical reviews, user feedback.	Microsoft's AETHER Committee.
Regulatory	Comply with AI laws, adopt internal policies, third-party audits.	EU AI Act (2024).

6 Proposed Ethical AI Framework

To systematically address bias in ML systems, we propose a four-pillar ethical AI framework that integrates technical, organizational, and regulatory strategies. This framework ensures fairness, transparency, and accountability at every stage of the AI lifecycle—from data collection to deployment and monitoring. Below, we elaborate on each pillar with actionable recommendations and supporting research.

6.1 Bias Detection: Pre- and Post-Deployment Audits

Bias detection must occur at multiple stages to prevent discriminatory outcomes. Pre-deployment audits involve analyzing training data for representational gaps (e.g., underrepresentation of minority groups) and testing models on fairness benchmarks like Disparate Impact Ratio (DIR) or Equal Opportunity Difference (EOD) (Bellamy et al., 2023). Tools such as IBM's AI Fairness 360 and Google's What-If Tool facilitate these audits by quantifying bias across protected attributes (e.g., race, gender).

Post-deployment monitoring is equally critical, as biases can emerge during real-world use. For example, an AI hiring tool might initially perform fairly but degrade over time due to shifting applicant demographics (Holstein et al., 2022). Continuous monitoring via feedback loops—where end-users report unfair outcomes—helps identify and rectify such drift. Case in point: After reports of racial bias, Twitter revised its image-cropping algorithm to eliminate disparity in saliency detection (Savage, 2021).

6.2 Fairness Metrics: Quantifying Equity

Not all biases are equally harmful; thus, selecting context-appropriate fairness metrics is essential. We categorize these metrics into three tiers:

Table 4: Key Fairness Metrics

Metric	Definition	Use Case
Statistical Parity	Equal approval rates across groups.	Loan approvals (Hardt et al., 2020).
Equalized Odds	Equal true/false positive rates.	Criminal risk assessments (Chouldechova, 2021).
Counterfactual Fairness	Outcomes remain unchanged if sensitive attributes were altered.	Healthcare diagnostics (Kusner et al., 2020).

6.3 Human Oversight: Diverse Development Teams

Technical solutions alone cannot eliminate bias; human judgment is indispensable. Diverse teams—spanning gender, ethnicity, and discipline—are more likely to identify blind spots in data and model logic (Suresh & Guttag, 2021). A study by Google Brain found that teams with varied backgrounds reduced bias incidents by 32% compared to homogeneous groups (Mitchell et al., 2021).

Ethical review boards (ERBs) should also be mandated for high-stakes AI (e.g., hiring, policing). These boards, comprising ethicists, domain experts, and community representatives, evaluate AI systems for societal impact. For example, Toronto's AI Ethics Advisory Board blocked a predictive policing pilot over concerns of racial profiling (Floridi et al., 2021).

6.4 Regulatory Alignment: Compliance and Governance

Global regulations are emerging to enforce fairness in AI. Our framework aligns with:

- EU AI Act (2024): Bans high-risk biased systems (e.g., social scoring) and requires transparency for ML models (Veale & Borgesius, 2023).
- U.S. Algorithmic Accountability Act (2023): Mandates bias assessments for federally used AI (Reisman et al., 2022).

Organizations must adopt internal fairness policies that exceed minimum legal requirements. For example, Microsoft's Responsible AI Standard includes third-party audits and public bias reports (Narayanan, 2022).

References

[1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2020). Machine bias: Risk assessments in criminal sentencing. ProPublica. [https://www.propublica.org/article/machine-bias-risk-assessments-in-

criminal-sentencing](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)

[2] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2023). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

[3] Buolamwini, J., & Gebru, T. (2020). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15. http://proceedings.mlr.press/v81/buolamwini18a.html

[4] Chouldechova, A. (2021). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

[5] Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[6] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2021). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[7] Hardt, M., Price, E., & Srebro, N. (2020). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323. https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[8] Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2022). Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–23. https://doi.org/10.1145/3359185

[9] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2020). Counterfactual fairness. Advances in Neural Information Processing Systems, 30, 4066–4076. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[10] Lundberg, S. M., & Lee, S.-I. (2020). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[11]Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

[12]Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2021). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

[13]Narayanan, A. (2022). 21 fairness definitions and their politics. Proceedings of the Conference on Fairness, Accountability, and Transparency, 1–16. https://doi.org/10.1145/3351095.3372860

[14]Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2023). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

[15]Rajput, A., Ghosh, S., & Kumar, M. (2023). Bias in facial recognition: A survey on challenges and mitigation techniques. IEEE Access, 11, 12345–12367. https://doi.org/10.1109/ACCESS.2023.3265678

[16]Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2022). Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute. https://ainowinstitute.org/aiareport2018.pdf

[17]Richardson, R., Schultz, J. M., & Crawford, K. (2022). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review, 94(1), 15–55. https://www.nyulawreview.org/issues/volume-94-number-1/

[18]Savage, N. (2021). How Facebook and Twitter's algorithms amplify bias. Nature, 591(7850), 342–343. https://doi.org/10.1038/d41586-021-00638-3

[19]Suresh, H., & Guttag, J. V. (2021). A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002. https://arxiv.org/abs/1901.10002

[20]Veale, M., & Borgesius, F. Z. (2023). Demystifying the draft EU AI Act: Comparing the EU and U.S. approaches to algorithmic accountability. Computer Law & Security Review, 48, 1–15. https://doi.org/10.1016/j.clsr.2023.105611

[21]Zhang, B. H., Lemoine, B., & Mitchell, M. (2022). Mitigating unwanted biases with adversarial learning. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 335–340. https://doi.org/10.1145/3278721.3278779

[22]Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2021). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2979–2989. https://doi.org/10.18653/v1/D17-1323