



Reducing Bias with Artificial Intelligence: Algorithmic Fairness, Bias Mitigation, Data Diversity, and Human Oversight

Yash, David, Dr. Manju Rani^{a,*}

^aAssistant Professor, School of Behavioural Sciences, Apeejay Stya University, Gurugram, Haryana, India

ABSTRACT:

Artificial intelligence (AI) systems increasingly inform decisions in employment, education, healthcare, finance, criminal justice, and other high-impact domains. Although these systems can improve efficiency and consistency, they may reproduce or amplify existing inequalities when training data are incomplete, unrepresentative, historically biased, or measured inconsistently. This conceptual review examines how bias can enter the AI lifecycle and outlines approaches for identifying, evaluating, and mitigating unfair outcomes. It discusses historical, representation, measurement, aggregation, and evaluation bias; the risks associated with proxy variables; and commonly used fairness criteria, including demographic parity, equal opportunity, predictive equality, and individual fairness. It then considers mitigation strategies at the pre-processing, in-processing, and post-processing stages, alongside validation methods such as subgroup analysis, counterfactual testing, adversarial probing, and robustness assessment. The paper argues that no single metric or technical intervention can establish fairness across all contexts. Responsible AI therefore requires context-sensitive fairness objectives, representative and high-quality data, transparent evaluation, meaningful human oversight, and continuous monitoring after deployment. The review concludes that fairness should be treated as a lifecycle responsibility shared by developers, deploying organisations, domain experts, and affected communities.

Keywords: artificial intelligence; algorithmic fairness; machine learning; data diversity; bias mitigation; human oversight; responsible AI

Introduction

Artificial intelligence (AI) is increasingly used to support or automate decisions that affect people's opportunities, resources, and access to services. Applications include recruitment, credit assessment, educational support, healthcare prioritisation, and criminal justice. Although automated systems may appear objective because they apply mathematical procedures consistently, their outputs can still reflect unequal social conditions, biased labels, incomplete data, or unsuitable modelling choices (Mehrabi et al., 2021; Tabassi, 2023).

Machine-learning systems identify statistical patterns in the examples on which they are trained. If those examples encode historical discrimination or fail to represent the people affected by a system, a model may reproduce these patterns at scale. Bias can also arise when a variable that appears neutral functions as a proxy for a sensitive or socioeconomic characteristic. These risks cannot be addressed solely by changing the algorithm; they require attention to data collection, problem formulation, measurement, evaluation, deployment, and monitoring.

This paper provides a conceptual review of the main sources of AI bias and approaches used to evaluate and mitigate it. It focuses on five connected areas: the origins and forms of bias; fairness assessment; mitigation across the model-development pipeline; data diversity and validation; and human oversight and continuous monitoring. Its central argument is that responsible AI requires a context-sensitive combination of technical evaluation and accountable human judgement.

Understanding bias in artificial intelligence

In AI, bias refers to systematic patterns in data, model behaviour, or decision processes that can produce unfair or unjustified disadvantages for particular people or groups. Relevant groups may be distinguished by characteristics such as gender, race, age, disability, geography, or socioeconomic position. Bias does not require deliberate discriminatory intent: a model can learn and reproduce unequal patterns from historical records even when developers have not explicitly instructed it to treat groups differently.

For example, a recruitment model trained on past hiring decisions may learn to favour characteristics associated with previously selected applicants. If earlier hiring practices excluded qualified people from particular groups, the model may reproduce that pattern. Bias should therefore be assessed throughout the AI lifecycle, including data collection and labelling, feature selection, model training, evaluation, deployment, and subsequent monitoring (Mehrabi et al., 2021; Tabassi, 2023).

Data, learning, and proxy variables

Machine-learning models estimate patterns from examples and use those patterns to generate predictions for new cases. Model performance consequently depends not only on the algorithm but also on the quality, coverage, relevance, and provenance of the training data. Missing groups, inconsistent measurements, and labels shaped by historical decisions can distort the relationships learned by a model.

Variables that appear neutral may also act as proxies for characteristics that are socially sensitive or relevant to disadvantage. For example, a widely used healthcare algorithm predicted healthcare costs rather than illness; because spending differed across racial groups at comparable levels of need, the proxy contributed to under-identification of Black patients for additional care. Reformulating the target to better reflect health needs substantially changed the distribution of patients identified for additional help (Obermeyer et al., 2019). Removing an explicitly sensitive variable does not necessarily eliminate this risk because other variables may carry similar information.

Major forms of AI bias

Historical bias arises when data or outcomes reflect existing social inequalities. A model trained on such data may treat historically unequal outcomes as if they were neutral indicators of merit or future success.

Representation bias occurs when the data inadequately covers the people or situations in which the system will be used. A model may perform well for groups that dominate the training set but less accurately for underrepresented groups.

Measurement bias occurs when a variable or label does not validly or consistently measure the concept it is intended to represent. Measurement procedures may systematically disadvantage particular groups, leading the model to learn distorted relationships.

Aggregation bias occurs when a single model or set of assumptions is applied across groups or contexts where the relevant relationships differ. Overall performance may obscure these group-specific differences.

Evaluation bias arises when test datasets, benchmarks, or performance measures fail to represent the populations and conditions in which a system will operate. Consequently, a strong overall score may conceal weak performance for a particular subgroup. These forms of bias can overlap and should not be treated as mutually exclusive categories.

Why AI fairness matters

AI bias matters because automated predictions can influence access to employment, credit, education, healthcare, public services, and liberty. In high-impact settings, even a small disparity may have serious consequences when decisions are repeated at scale or when affected individuals have limited opportunities to challenge them.

Biased systems may also contribute to feedback loops. For example, a system that directs resources away from a group may generate fewer positive outcomes for that group; those outcomes may later be used as training data, reinforcing the original pattern. Fairness is therefore not merely a question of model accuracy. It also concerns the legitimacy of the decision, the consequences for affected people, the availability of explanation and appeal, and the accountability of the organisation using the system.

Detecting potential bias

A systematic bias assessment begins by identifying the purpose of the system, the people affected by its decisions, the outcomes it predicts, and the consequences of errors. Reviewers should examine who is represented in the training and evaluation data, how labels were produced, which variables are used, and what historical or institutional processes may have shaped the data.

Model outcomes and error rates can then be compared across relevant groups and, where data permit, across intersections of characteristics. Such comparisons can reveal disparities that are hidden by aggregate performance measures. However, a statistical difference does not by itself establish that a system is unfair, nor does the absence of a measured difference prove that it is fair. Findings require interpretation in relation to the task, the quality of the data, the applicable legal and ethical requirements, and the likely consequences of the decision.

Fairness metrics and assessment tools

Fairness metrics provide structured ways to examine how model outcomes differ across groups. Demographic parity compares the rates at which groups receive a positive outcome. Equal opportunity focuses on whether qualified individuals have comparable true-positive rates. Predictive equality compares false-positive rates across groups. Individual fairness concerns whether individuals who are similar in task-relevant ways receive similar treatment. These metrics operationalise different notions of fairness and should not be treated as interchangeable (Mehrabi et al., 2021).

Tools such as Aequitas, Fairlearn, AI Fairness 360, and the What-If Tool can support the analysis of group outcomes, error rates, and mitigation options. These tools assist evaluation; they do not determine which fairness objective is appropriate or whether a disparity is justified in a particular setting. Different fairness criteria may conflict, especially when groups have different base rates. The selected metrics should therefore be justified by the application's purpose, risks, stakeholder needs, and relevant governance requirements rather than chosen solely because they are easy to calculate.

Bias mitigation across the AI lifecycle

Bias mitigation is commonly organised into three stages. Pre-processing approaches modify or improve the data before model training. Examples include reviewing labels, correcting data-quality problems, reweighting observations, and improving coverage of underrepresented populations. Data augmentation may be useful in some settings, but synthetic or rebalanced data should be validated carefully because it can preserve or introduce unwanted patterns.

In-processing approaches incorporate fairness-related objectives or constraints into model training. These methods may reduce selected disparities, but their effectiveness depends on the chosen objective, the data, and the model. Post-processing approaches adjust predictions or decision thresholds after training. They may be useful when a model cannot be retrained, but they can create additional operational, legal, or interpretive considerations.

No mitigation method should be assumed to work automatically. Changes may affect predictive performance, calibration, other fairness measures, or the experience of affected groups. Mitigation should be evaluated empirically, documented, and reviewed against the system's intended use and risk level.

Human oversight and accountability

Fairness is partly a normative question: it depends on the purpose of the system, the context in which it is used, and the interests and rights of affected people. Technical metrics can identify differences, but they cannot independently determine which differences are acceptable or what remedy is appropriate. Human oversight is therefore necessary to define fairness objectives, investigate root causes, interpret evidence, and decide whether a system should be modified, restricted, or withdrawn.

Effective governance benefits from the involvement of technical specialists, social scientists, domain experts, ethicists, legal and compliance personnel, and representatives of affected communities. Human review should be meaningful rather than symbolic: reviewers need appropriate expertise, access to relevant evidence, authority to challenge model outputs, and clear procedures for escalation and appeal. Human decision-makers can also introduce their own biases, so oversight itself should be documented, evaluated, and subject to accountability.

Validation and stress testing

AI systems should be assessed for fairness and stability in addition to aggregate predictive performance. Subgroup analysis can reveal performance differences across relevant populations, while intersectional analysis may identify patterns obscured by broad categories. Counterfactual testing can examine whether a prediction changes when a protected or otherwise irrelevant characteristic is altered while task-relevant information is held constant. Such tests require careful assumptions because changing one characteristic may affect other variables in realistic settings.

Adversarial probing tests model behaviour on difficult, unusual, or deliberately constructed inputs. Robustness testing examines whether small, irrelevant changes in input lead to disproportionate changes in output. These methods can reveal weaknesses that standard test sets fail to capture. Validation should be conducted before deployment and repeated after material changes to the model, data, operating context, or affected population.

Data diversity and quality

Representative, relevant, and high-quality data are important foundations for fair AI. In a well-known audit of three commercial gender-classification systems, Buolamwini and Gebru (2018) reported maximum error rates of up to 34.7% for darker-skinned females and 0.8% for lighter-skinned males, illustrating how aggregate performance can conceal intersectional disparities. Data collection should reflect the population and situations in which the system will be used, while accounting for meaningful variation within groups.

Improving data quality may require targeted data collection, collaboration with affected communities, review of labelling practices, and careful use of synthetic data. These activities must be balanced with privacy, consent, data minimisation, and security requirements. The aim is not simply to collect more records but to ensure that the data provide an adequate and ethically defensible basis for the intended task.

Continuous monitoring after deployment

Fairness cannot be established once and assumed to remain stable. Data distributions, populations, institutional practices, and user behaviour may change after deployment. Such changes can affect both model performance and the distribution of errors across groups.

A monitoring plan should specify which outcomes and fairness measures will be tracked, how often they will be reviewed, who is responsible for the review, and what thresholds trigger investigation or intervention. The NIST AI Risk Management Framework recommends a structured approach to governing, mapping, measuring, and managing AI risks across the system lifecycle (Tabassi, 2023). Periodic independent audits, documentation of model changes, accessible complaint mechanisms, and the ability to pause or roll back a system can strengthen accountability.

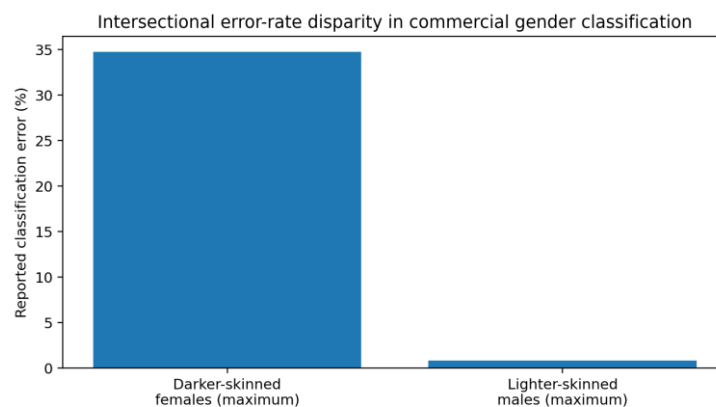
Empirical evidence from the literature (2016–2026)

Evidence published over the past decade demonstrates that AI bias is not merely theoretical. The examples below are selected studies and reviews rather than a complete systematic review. Because the sources use different domains, methods, and inclusion criteria, their sample sizes should be interpreted within each study and not pooled into a single time trend.

Table 1 - Selected evidence on AI bias and fairness, 2016–2026.

Year	Source / scope	Reported evidence	Implication for this paper
2016	Barocas & Selbst; legal analysis of big-data discrimination	Explains how data-driven decision systems can produce disparate impact even without explicit use of protected characteristics.	Bias assessment must consider indirect pathways and institutional context.
2018	Buolamwini & Gebru; commercial gender-classification systems	Maximum reported error: up to 34.7% for darker-skinned females versus 0.8% for lighter-skinned males.	Subgroup and intersectional evaluation can reveal disparities hidden by aggregate accuracy.
2019	Obermeyer et al.; healthcare risk algorithm	Changing the target from healthcare cost to health need increased Black patients identified for additional help from 17.7% to 46.5%.	Proxy labels can encode unequal access and distort model decisions.
2021	Mehrabi et al.; survey of bias and fairness in machine learning	Synthesizes sources of bias, fairness definitions, and mitigation approaches across machine learning.	Fairness is multidimensional; metric choice should be justified.
2022	Starke et al.; systematic review of empirical fairness perceptions	Synthesized 58 empirical studies and more than 33,000 observations of people's fairness perceptions.	Public perceptions and human-centred evaluation complement technical metrics.
2023	Tabassi / NIST AI RMF 1.0	Provides a voluntary framework organised around Govern, Map, Measure, and Manage AI risks.	Fairness should be governed across development and deployment.
2025	Liu et al.; scoping review of clinical AI fairness	467 papers included; 267 (57.2%) reported bias identification linked to mitigation; 200 (42.8%) focused on identification without mitigation.	Detection does not guarantee that a mitigation is implemented or evaluated.
2026	Zipperling et al.; systematic review of data-level interventions	Reviewed 79 papers; identified limited attention to data-extension strategies and inconsistent terminology.	Data-level interventions need stronger empirical evaluation and clearer reporting.

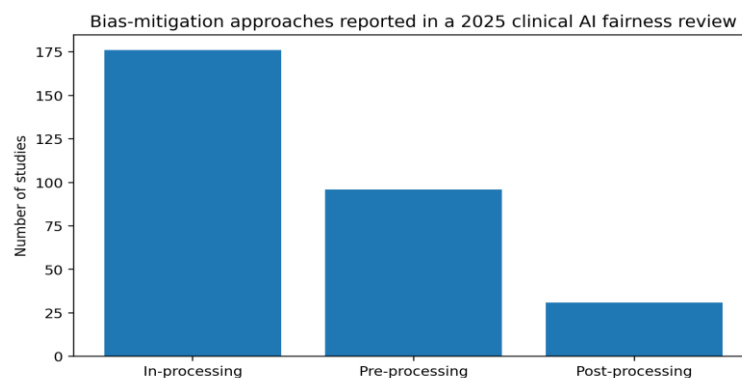
Note. The table is a purposive evidence summary, not a systematic review or an annual publication-count dataset. Reported values are reproduced from the cited sources.



Source: Buolamwini & Gebru (2018); reported maximum subgroup error rates.

Fig. 1 - Reported maximum subgroup error rates in commercial gender classification.

Note. Values are from Buolamwini and Gebru (2018); the comparison illustrates a specific audit and should not be generalized to all facial-analysis systems.



Source: Liu et al. (2025), 267 studies reporting mitigation; categories overlap.

Fig. 2 - Bias-mitigation approaches reported in a 2025 clinical AI fairness review.

Note. Counts are from Liu et al. (2025). Categories overlap because a paper could use more than one mitigation stage; the bars are not mutually exclusive.

Discussion

The review highlights that AI bias is a lifecycle concern rather than a defect confined to a particular algorithm. It can arise through historical inequalities, inadequate representation, flawed measurements, proxy variables, modelling assumptions, evaluation practices, and deployment conditions. These sources may interact, meaning that a single intervention is unlikely to address every pathway to unfairness.

Fairness metrics and software tools can provide important evidence, but their interpretation depends on the application and the consequences of decisions. Similarly, mitigation strategies can reduce a specified disparity without resolving the institutional conditions that produced it. A defensible approach combines data-quality review, context-appropriate fairness measures, empirically tested mitigation, subgroup and robustness testing, human oversight, documentation, and post-deployment monitoring. The responsibility for fairness is shared among developers, organisations that deploy AI, domain experts, regulators where applicable, and communities affected by automated decisions.

Limitations

This paper is an evidence-informed narrative review and does not report an original empirical dataset or a reproducible systematic search protocol. It therefore does not estimate the prevalence of particular forms of bias across all AI applications or establish that one intervention is universally more effective than another. The cited reviews cover different domains and inclusion criteria, so their counts should not be combined as if they represented one common dataset. Fairness definitions and applicable requirements vary across contexts, and a method suitable in one setting may be inappropriate in another. Future work should use transparent search and screening procedures and compare mitigation approaches across tasks, populations, and model types.

Future directions

Future research should improve methods for detecting proxy bias and hidden disparities, particularly where sensitive attributes are unavailable or data are sparse. Greater attention is also needed to intersectional evaluation, valid measurement, representative data collection, and the long-term effects of mitigation after deployment. Research on human–AI collaboration should examine how explanations, review interfaces, and appeal mechanisms can help people identify and challenge potentially unfair outcomes. Finally, organisations need stronger documentation, independent auditing, stakeholder participation, and accountability mechanisms so that fairness remains embedded in AI governance throughout the system lifecycle.

Conclusion

AI-generated decisions are not automatically fair because they are produced by mathematical systems. Models learn from data and design choices that may reflect existing social inequalities or omit important experiences. This paper has discussed historical, representation, measurement, aggregation, and evaluation bias, along with proxy variables, fairness metrics, mitigation strategies, validation, data quality, human oversight, and continuous monitoring.

The central conclusion is that reducing AI bias requires an ongoing, context-sensitive process. Representative and well-documented data, appropriate fairness measures, tested mitigation methods, robust validation, meaningful human review, and post-deployment monitoring each play a role. Fairness should therefore be treated as a shared lifecycle responsibility, with clear accountability and opportunities for affected people to question consequential decisions.

REFERENCES

- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. <https://doi.org/10.15779/Z38BG31>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Liu, M., Ning, Y., Teixayavong, S., Liu, X., Mertens, M., Shang, Y., Li, X., Miao, D., Liao, J., Xu, J., Ting, D. S. W., Cheng, L. T.-E., Ong, M. E. H., & colleagues. (2025). A scoping review and evidence gap analysis of clinical AI fairness. *npj Digital Medicine*, 8, Article 360. <https://doi.org/10.1038/s41746-025-01667-2>
- 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), Article 115, 1–35. <https://doi.org/10.1145/3457607>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). 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>
- Starke, C., Baleis, J., Keller, B., Marcinkowski, F., & others. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2). <https://doi.org/10.1177/20539517221115189>
- Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>

-
- Zipperling, D., Deck, L., Jessat, L., & Kühl, N. (2026). Data-centric algorithmic fairness: A systematic review of data-level interventions for algorithmic fairness. *Proceedings of the Fifth European Conference on Algorithmic Fairness, Proceedings of Machine Learning Research*, 350, 113–135. <https://proceedings.mlr.press/v350/zipperling26a.html>