WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Reducing Bias with Artificial Intelligence Algorithmic Fairness, Human Behaviour, and Responsible AI

***Akansha Bhardwaj[1], Prashant Singh[2], Dr. Manju Rani[3]***

**Department:** School of Behavioural Sciences
**Institution:** Apeejay Stya University, Gurugram, Haryana, India

**ABSTRACT :**

Artificial intelligence (AI) increasingly supports decisions in recruitment, education, healthcare, finance, and public services. Although these systems can improve speed and consistency, they may also reproduce or amplify existing inequalities when training data, labels, measurements, or evaluation procedures are biased. This paper examines AI bias as both a technical problem and a behavioural and social concern. It reviews common forms of bias, including historical, representation, measurement, aggregation, and evaluation bias, and discusses how biased automated decisions may affect opportunity, trust, perceived fairness, and the experience of people subject to such decisions. Drawing on established work on algorithmic fairness, machine-learning bias, and responsible AI, the paper outlines approaches to detecting and mitigating bias before model training, during model development, and after predictions are generated. It also considers the importance of representative data, transparent communication, meaningful human oversight, and ongoing monitoring. The paper uses a conceptual narrative-review approach and does not report an original experiment or new dataset. It argues that no single fairness metric or technical intervention can guarantee fairness in every setting; decisions must be evaluated in context and with attention to the people affected. Responsible AI therefore requires technical testing alongside behavioural insight, stakeholder participation, and accountability.

**Keywords: artificial intelligence; algorithmic bias; algorithmic fairness; human behaviour; trust; social inequality; bias mitigation; responsible AI**

## 1. Introduction

Artificial intelligence systems increasingly influence decisions that affect people's everyday lives. Automated tools may rank job applicants, recommend educational support, help prioritize healthcare, or assess applications for financial services. These tools are often presented as consistent and objective because they use computational rules. However, a model can produce unequal outcomes when it learns from incomplete, unrepresentative, or historically unequal data. As Barocas, Hardt, and Narayanan (2019) explain, fairness problems are connected to the data, the prediction task, and the social context in which a system is used.

AI bias matters beyond model accuracy. A decision can affect a person's access to opportunity, their confidence in an institution, and their belief that they have been treated fairly. If a person cannot understand or challenge an automated decision, the experience may also reduce trust and create a sense of powerlessness. These are behavioural and social consequences, not only technical defects.

This paper considers how AI bias develops, how it can be detected and reduced, and why human judgement and stakeholder perspectives are needed alongside technical tools. Its central argument is that responsible AI requires a lifecycle approach: suitable data, explicit fairness goals, testing across relevant groups, appropriate mitigation, human review, and continued monitoring.

## 2. Research Objectives

- To explain the major forms and sources of bias in AI systems.
- To examine why algorithmic bias matters for human behaviour, trust, perceived fairness, and access to opportunities.
- To review common approaches for detecting and mitigating AI bias.
- To identify the roles of data diversity, human oversight, transparency, and continuous monitoring in responsible AI.

## 3. Review Questions

- How can bias enter an AI system during data collection, model development, evaluation, and deployment?
- What kinds of harms can biased automated decisions create for people and communities?
- Which technical and organizational approaches can help detect and reduce these harms?
- Why are human oversight and context-sensitive fairness decisions necessary?

## 4. Methodology

This paper uses a conceptual narrative-review approach. It synthesizes established publications on algorithmic fairness, machine-learning bias, and responsible AI, including scholarly work by Barocas, Hardt, and Narayanan; Mehrabi and colleagues; Selbst and colleagues; and guidance from the National Institute of Standards and Technology. The sources are used to organize key concepts, identify recurring risks, and compare broad categories of mitigation. This is not a systematic review: the paper does not claim an exhaustive search, formal screening process, or quantitative meta-analysis. It also does not present original participant data, experiments, or statistical results. The discussion should therefore be read as a conceptual synthesis rather than as evidence from a new empirical study.

## 5. Understanding Bias in Artificial Intelligence

In this paper, AI bias refers to systematic patterns in a model or decision process that can disadvantage particular people or groups in a given context. Not every difference in model outcomes automatically proves unfair discrimination. Researchers must examine the task, the quality of the data, the consequences of errors, and the fairness goals relevant to the setting. Mehrabi et al. (2021) describe bias in machine learning as a broad problem that can arise from multiple sources and at different stages of system development.

### *5.1 Major forms of bias*

Historical bias: Past decisions or social inequalities are reflected in the data. A hiring model trained on records from a workforce that excluded certain groups may learn patterns that reproduce that exclusion.
Representation bias: Some populations or circumstances are missing or underrepresented. A system may work less reliably for people whose experiences were not adequately represented during training.
Measurement bias: The variables used to represent a concept do not measure it equally or accurately across groups. For example, a proxy for need or performance may be misleading in different social contexts.
Aggregation bias: A single model is applied to groups or contexts that have meaningfully different patterns, potentially concealing important differences in performance.
Evaluation bias: A test dataset or benchmark does not adequately reflect the people and conditions in which the system will be used. High overall accuracy can conceal poor performance for a subgroup.
Proxy bias: A variable that appears neutral, such as location or purchase history, may indirectly reflect social or economic characteristics and contribute to unequal outcomes.

## 6. Why AI Bias Matters for Behaviour and Society

Automated decisions can influence more than immediate outcomes. They may shape how people understand institutions, whether they feel respected, and whether they believe an appeal process is meaningful. A person who receives an unexplained rejection from an automated recruitment system may have difficulty knowing whether the decision reflected their qualifications, a data error, or an unfair pattern. When decisions are difficult to challenge, people may perceive the process as less transparent or less legitimate.
Biased systems can also reproduce existing inequalities. For example, if a model is trained on past decisions that favoured one group, its recommendations may continue to favour that group. Those recommendations may then influence future decisions and generate new records that appear to confirm the original pattern. This feedback loop can make an initial disparity harder to identify and correct.
The behavioural perspective also highlights that humans are not automatically free from bias when they use AI. Decision-makers may accept a model's recommendation because it appears technical or authoritative, a problem commonly discussed as automation bias. Conversely, people may reject useful recommendations without examining the evidence. Human oversight must therefore involve training, clear responsibilities, and the ability to question model outputs rather than simply approving them.

## 7. Detecting and Evaluating Bias

A fairness assessment should begin by identifying who may be affected, what decision the model supports, and what kinds of errors may cause harm. Developers and reviewers should examine data coverage, label quality, missing information, and the possibility that apparently neutral variables act as proxies for sensitive characteristics. They should also compare performance and outcomes across relevant groups and, where appropriate, intersections of characteristics.

### *7.1 Common fairness measures*

Demographic parity: compares the rate at which different groups receive a positive outcome.
Equal opportunity: examines whether qualified individuals across groups have comparable chances of receiving a positive outcome.
Predictive equality: compares a particular error rate, such as false-positive rates, across groups.
Individual fairness: asks whether people who are similar in task-relevant ways receive similar decisions.
These measures answer different questions and may conflict in some settings. No single metric can establish fairness in every context. Statistical results

need interpretation alongside the purpose of the system, the costs of different errors, legal and ethical requirements, and the views of affected stakeholders. Tools such as AI Fairness 360 and Fairlearn can support analysis, but a tool's output does not replace contextual judgement.

## 8. Approaches to Reducing AI Bias

### *8.1 Pre-processing*

Pre-processing methods change the data before model training. Examples include correcting data-quality problems, improving representation where appropriate, reweighting examples, and reviewing labels. These methods can help address some data-related problems, but additional data should not be collected without appropriate privacy safeguards, and synthetic data should not be assumed to represent real experiences accurately.

### *8.2 In-processing*

In-processing methods change how a model is trained. A developer may incorporate a fairness constraint or use an optimization approach intended to reduce a selected disparity. Such methods require careful testing because improving one measure may affect another measure or overall predictive performance.

### *8.3 Post-processing*

Post-processing methods adjust predictions or decision thresholds after a model has been trained. They may be useful in some applications, especially where the underlying model cannot easily be changed. Their use should be justified, documented, and evaluated against the selected fairness goal and applicable requirements.

### *8.4 Human review and appeal*

Human review is most useful when reviewers have relevant expertise, access to meaningful information, sufficient time, and authority to question or override an output. A review process should not merely rubber-stamp the model's recommendation. People affected by high-impact decisions should, where feasible, receive understandable explanations and a route to correct inaccurate information or request reconsideration.

## 9. Human Oversight, Transparency, and Accountability

Fairness involves social values as well as mathematical properties. A model may meet a chosen statistical target and still cause harm if the target does not reflect the real problem. Selbst et al. (2019) warn that fairness interventions can fail when technical systems are separated from the social settings in which they operate. Responsible AI therefore requires participation from technical specialists, domain experts, behavioural and social scientists, organizational decision-makers, and affected communities.

Transparency should make relevant aspects of a system understandable to the people responsible for its use and, where appropriate, to those affected by it. Accountability requires documentation of the system's purpose, known limitations, evaluation results, responsible personnel, and procedures for responding to complaints or newly identified disparities. The NIST AI Risk Management Framework similarly emphasizes governance and ongoing processes for mapping, measuring, and managing AI risks (NIST, 2023).

## 10. Practical Framework for Responsible AI

1. Define the context: Identify the decision, affected people, likely harms, and appropriate fairness goals.
2. Review the data: Check representation, quality, labels, missingness, and possible proxy variables.
3. Measure performance: Evaluate accuracy and relevant fairness measures across groups and conditions.
4. Mitigate identified problems: Choose a data, training, or post-processing intervention appropriate to the cause.
5. Conduct human review: Include qualified reviewers and provide routes for correction and appeal.
6. Monitor continuously: Reassess the system after deployment, document changes, and intervene when risks increase.

## 11. Discussion

The reviewed concepts indicate that AI bias is a lifecycle problem rather than a defect located only in an algorithm. Bias can enter through historical data, the choice of target variable, measurement practices, model assumptions, evaluation design, and the way a system is used. Technical tools can help reveal disparities, but the interpretation of those disparities requires an understanding of the application and the people affected.

A behavioural perspective strengthens the analysis by considering trust, perceived fairness, automation bias, and the ability of people to challenge decisions. Human involvement alone does not guarantee fairness, because human reviewers may also rely on stereotypes or defer too readily to automated outputs. Oversight should therefore be supported by clear procedures, documentation, reviewer training, and independent checks.

The practical implication is that organizations should not treat fairness as a one-time certification. Fairness goals and system performance should be reviewed as the population, data, and operating conditions change. The appropriate balance between accuracy, fairness, privacy, transparency, and other

values will depend on the use case and should be made explicit.

## 12. Limitations

This paper is a conceptual narrative review and does not report original experiments, interviews, surveys, or a new dataset. It does not systematically compare the performance of specific bias-mitigation algorithms. The discussion is limited by the selected conceptual and guidance sources, and the conclusions should not be interpreted as proof that any one intervention will work in every domain. Future empirical studies could compare methods using carefully documented datasets and examine how affected people experience AI-assisted decisions.

## 13. Future Research

- Compare bias-mitigation techniques across different domains and datasets, reporting both predictive performance and relevant fairness measures.
- Study how explanations, appeal mechanisms, and human review affect perceived fairness and trust in AI-supported decisions.
- Investigate how automation bias influences professionals who use AI recommendations in recruitment, education, healthcare, and public services.
- Develop practical monitoring approaches that detect changing disparities while protecting privacy.
- Include affected communities in defining fairness goals and evaluating whether interventions address real-world harms.

## 14. Conclusion

AI can support useful and efficient decisions, but it is not automatically objective or fair. Biased data, flawed measurements, inadequate evaluation, and inappropriate deployment can create systematic disadvantages for individuals and groups. These outcomes may also affect trust, perceived fairness, and access to important opportunities. Reducing bias requires a combination of representative and high-quality data, context-appropriate fairness evaluation, mitigation techniques, meaningful human oversight, transparency, and continuous monitoring. A behavioural and social perspective is essential because fairness is experienced by people within institutions, not only measured in model outputs. Responsible AI should therefore be judged not only by accuracy, but also by how carefully it identifies, prevents, explains, and responds to unfair outcomes.

**REFERENCES :**


- Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. Online book draft. https://fairmlbook.org/
- 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. https://doi.org/10.1145/3457607
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68). https://doi.org/10.1145/3287560.3287598
- Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. https://doi.org/10.1145/3465416.3483305