Open Access
Full Length Article

ORIGINAL ARTICLE

Mitigating Bias in AI Systems: A Comprehensive Review of Sources and Strategies

Ahmed Farhan Abbasi[1] Shah Murad Chandio[2]

[1] PhD Scholar Institute of Mathematics and Computer Science, University of Sindh, Jamshoro.
[2] Associate Professor, Institute of Mathematics and Computer Science, University of Sindh, Jamshoro.

ABSTRACT

Artificial intelligence (AI) is increasingly deployed in high-stakes environments such as healthcare, recruitment, finance and public safety. Although AI promises efficiency and accuracy, substantial evidence shows that it can reproduce or amplify societal inequalities through biased datasets, algorithmic design choices, human cognitive bias and automation bias. This study conducts a systematic literature review of 128 peer-reviewed publications from 2015 to 2025 to examine the sources of bias in AI systems and evaluate the effectiveness of mitigation strategies. Bias is shown to originate at multiple stages of the AI lifecycle, data, model development and real-world deployment, and tends to re-emerge when mitigation methods are applied in isolation. In response, this study introduces the Integrated Bias-Mitigation Framework (IBMF), a lifecycle-oriented model that aligns technical interventions with governance and continuous monitoring. The findings highlight the need for coordinated mitigation strategies rather than one-off technical solutions. The study concludes with policy, organizational and research recommendations to support sustainable fairness and accountability in AI.

Keywords: *Artificial Intelligence; Algorithmic Bias; Fairness; Bias Mitigation; Responsible AI; Automation Bias; Cognitive Bias*

Author`s Contribution
[1]Data analysis, interpretation, and manuscript writing, Active participation in data collection
[1,2]Conception, synthesis, planning of research. Interpretation and discussion

Address of Correspondence
Ahmed Farhan Abbasi
Email: sm.chandio@usindh.edu.pk

Article info.
Received: June,15, 6, 2024
Accepted: November 13, 2024
Published: December 30, 2024

Cite this article: *Jogi A S, Wadho A. Mitigating Bias in AI Systems: A Comprehensive Review of Sources and Strategies J. inf. commun. technol. robot. appl.2024; 14(1):45-55*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

AI systems have become a part of decisions that impact healthcare access, employment opportunities, the granting of financial credit, and the safety of the population. As AI is often regarded as objective and data-driven, there is increasing evidence that automated systems reproduce or exacerbate existing societal disparities [1], [2]. The discriminatory recruitment tools, racial disparities in clinical risk prediction, and biased predictive policing algorithms are high-profile examples that prove that biased AI may lead to actual harm, and vulnerable populations are disproportionately affected [3], [4], [5].

Despite the existence of numerous fairness-enhancing techniques, practical experience suggests that fairness tends to deteriorate once deployed, which raises the question of why the positive effects of fairness-enhancing

techniques implemented during the development process can be maintained. Nevertheless, more recent research has demonstrated that bias often recurs following deployment despite earlier mitigation [6], [7]This raises a fundamental question: how do we maintain AI fairness in real-world environments, rather than just in laboratory settings?

This study is a response to this challenge based on a systematic review of the academic literature and real-life cases. It considers the way prejudice arises throughout the AI lifecycle and assesses the efficiency of mitigation methods, such as the human-in-the-loop deployment and post-deployment monitoring.

This study will identify, analyze, and synthesize sources of bias in AI systems, and critically assess measures to reduce bias, particularly in high-stakes applications.

LITERATURE REVIEW

1.1| Research Objectives:

- To review academic literature and case evidence to identify sources of bias across AI lifecycle stages.
- To synthesize and classify mitigation strategies and evaluate their strengths and limitations.
- To compare mitigation strategies across sectors including healthcare, recruitment, finance and policing.
- To develop an integrated framework that connects sources of bias with mitigation strategies and governance implications.

Based on the limitation identified in the various phases of the AI lifecycle, this study proposes a combined bias-reduction system to address the following research questions.

1.2 | Research Questions

- What are the predominant sources of bias in AI systems, and how do they manifest across the AI lifecycle?
- What mitigation strategies have been proposed to counter bias in AI, and how effective are they?
- How do mitigation strategies differ across application domains?
- What ethical and governance implications arise from the use of bias mitigation techniques, and how can an integrated framework support sustainable fairness?

2.| Literature Review:

The study of AI fairness has grown rapidly due to concerns that machine learning applications replicate and exacerbate existing social disparities. Initial research associated differences in automated decision-making with the greater socio-technical infrastructures in which AI systems are executed [1], [8]. The following empirical reports revealed that the differences in performance statistics across demographic groups are not isolated anomalies, but a systematic phenomenon grounded in the data and institutional practice [3], [5]. Together, this body of research shows that algorithmic bias cannot be understood as a purely technical failure; instead, it reflects historical, social, and political conditions embedded in data and system design.

2.1|Sources of Bias in AI Systems:

Across the reviewed literature, four primary sources of bias appear consistently: data bias, algorithmic bias, cognitive bias and automation bias.

Data bias arises when the datasets used to train AI models are not representative of the population to which the model will be applied or when training data encodes historical inequalities. Under-representation of demographic groups and the use of problematic proxy variables (e.g., health cost as a measure of health need) produce systematic disparities in outcomes ([5], [9]. Such data imbalances are obvious in healthcare, hiring and lending domains, where ground-truth labels often reflect past human judgment rather than objective outcomes.

Algorithmic bias emerges even when datasets are improved. Feature selection, model optimization goals and fairness-accuracy trade-offs can embed discriminatory effects in the model's internal logic [10], [11], [12]. For example, fairness-constrained optimization improves equity metrics. However, it can reduce computational scalability, while adversarial debiasing provides substantial performance gains yet remains resource-intensive and difficult to deploy outside research environments.

Human cognitive bias persists when developers, annotators, or end users unintentionally transfer subjective judgment into system development. Annotation decisions and feedback-based labelling pipelines can reproduce individual and organizational prejudices, even when formal fairness goals are in place [4], [9], [13]. Bias therefore enters long before a model is trained, beginning with human interpretation of data.

Automation bias occurs after deployment when humans over trust system recommendations. Institutional pressures to increase efficiency, combined with inadequate oversight of models, create conditions in which AI outputs are accepted without critical scrutiny [2], [6], [7].). The resulting feedback loop reinforces biased decisions and often reintroduces disparities that earlier fairness interventions had reduced during development.

2.2|Synthesis of Prior Work:

Across the literature, a clear pattern emerges bias is not a static defect located at a single stage of the AI lifecycle, but an evolving property of socio-technical systems. Data corrections reduce disparities only temporarily, as bias reappears during model optimization or deployment. Likewise, post-deployment post-processing techniques are only effective for a brief period if organizational oversight mechanisms are weak. Recent scholarship thus increasingly advocates for lifecycle-wide approaches that integrate technical and governance-level mitigation [12], [13], [14] .This need for holistic strategies provides the conceptual foundation for the Integrated Bias-Mitigation Framework (IBMF) developed in later sections of this study.

RESEARCH METHODOLOGY

This study adopted a Systematic Literature Review (SLR) to synthesize current evidence on the sources of bias in AI systems and the effectiveness of mitigation strategies. The SLR approach was chosen because existing research on AI fairness is fragmented across disciplines and application domains, and a structured review offers the most rigorous way to consolidate knowledge at scale [15], [16]). Academic databases, including Scopus, IEEE Xplore, ACM Digital Library, Web of Science, and Google Scholar, were searched using combinations of keywords such as AI bias, algorithmic fairness, bias mitigation, ethical AI, and responsible AI. The search timeframe (20152025) was adequate to capture the most recent findings in fairness studies and avoid outdated or irrelevant research that does not relate to fairness algorithms.

The included studies had to address AI bias or fairness, make an empirical or conceptual contribution, and be written in English. Exclusion criteria included articles that focused on improving accuracy without considering fairness, lacked methodological transparency, or were editorial or opinion pieces. The PRISMA 2020 guidelines were used for screening, and the final sample comprised 128 publications. In all databases, 1,742 records were identified (Scopus = 620, Web of Science = 410, IEEE Xplore = 260, ACM Digital Library = 210, Google Scholar = 200, and 42 other records through cross-referencing). There were 1,430 records to undergo title and abstract screening after eliminating duplicate records (312). Of this number, 1,178 were omitted as irrelevant. The remaining 252 full-text articles were evaluated for eligibility, with 124 articles being excluded (48 were not empirical studies, 39 lacked direct relevance to fairness/AI bias, 22 lacked an adequate methodology description, and 15 were not written in English). In the end, 128 articles were included in the final synthesis, meeting all the requirements as per PRISMA 2020. Extracted data captured: type of bias examined, domain of application, mitigation strategy used, evaluation metrics and reported outcomes. Thematic synthesis (Braun & Clarke, 2006) was employed to reveal cross-study patterns on how bias emerges, how mitigation techniques perform and what challenges persist across deployment contexts.

4.|Findings:

The synthesis identified three dominant patterns, the first showing that fairness degrades when models encounter population shifts highlighting the limits of static dataset correction Studies in healthcare, recruitment and criminal justice repeatedly demonstrate that underrepresented groups experience higher error rates even when the same model performs well on majority groups [5], [9], [13].. Despite the increase in fairness in controlled settings where dataset balancing and synthetic

augmentation were applied, the effects were minimal when applied to other populations, suggesting that dataset correction is insufficient to maintain long-term fairness.

One of the best-documented instances in the financial industry illustrates how the erosion of fairness occurs when credit-scoring models are exposed to changing population dynamics [18], [19]. Research on a credit model that had been trained on past repayment statistics initially passed fairness tests, but when released, it became biased against younger and lower-income borrowers. The training data for this model would become increasingly skewed as new lending decisions were influenced by the distribution of approved and rejected borrowers. This feedback mechanism resulted in the decreasing fairness measures with each repeat, showing that the methods of fairness correction that remain constant fail to maintain fairness in financial transactions in the real world [10].

Second, the difference may increase when algorithmic design decisions are made with improved datasets. Sensitive factors (gender, race, or socioeconomic background) can be indirectly encoded using feature selection, optimization goals, and proxy variables [12], [20]Fairness-constrained and adversarial debiasing in-processing methods have been reported to be successful in research studies. Nevertheless, their practical implementation remains negligible due to computational costs, domain-specific optimization, and limited access to sensitive demographic information.

Third, bias often recurs post-deployment, driven by feedback loops, automation bias, and institutional incentives. End users can also adopt over-trust model suggestions once AI systems are installed in the decision-making environment and align with organizational objectives such as efficiency or throughput [2], [6], [21]Research has shown that systems designed with fairness protections can still be discriminatory when used without continuous monitoring, calibration, or human supervision.

The examples of predictive policing systems in the public safety field demonstrate that bias can grow as these systems are used [22], [23]. A generally applied crime-forecasting instrument generated more risk estimates in the neighborhoods with historical high levels of policing, even though the real rates of crimes were also stable or decreasing. The more law enforcement acted in accordance with these predictions, the more arrests occurred, which supports the model's assumptions. This reinforcing loop led to a statistically significant increase in the over-classification of minority groups as risky, highlighting how feedback loops can contribute to the degradation of fairness despite the initial calibration of the model [24].

Combined, these tendencies indicate that prejudice is not a fixed attribute of models but a dynamic quality of socio-technical systems. Most mitigation methods are only effective temporarily, as they focus on individual pipeline stages and fail to account for lifecycle interactions among datasets, model behavior, and deployment environments.

Table 1. Table 1. Summary of Key Bias Patterns Across the AI Lifecycle

Lifecycle Stage	Dominant Source of Bias	Reason Fairness Fails Over Time	Representative Studies
Data	Sampling imbalance ; historical discrimination	Fairness gains weaken under population shift	Obermeyer et al. (2019) ; Chen et al. (2023)
Algorithms	Proxy features; optimization goals; metric conflict	Difficult to scale; limited access to sensitive data	Kleinberg et al. (2017); Ferrara (2024)
Deployment	Automation bias; institutional incentives ; weak auditing	Fairness degradation over time	Raji et al. (2020); Gray (2024); Jarrahi (2025)

This table summarizes the dominant patterns of fairness degradation identified in the literature and specifies the lifecycle stages in which these issues occur.

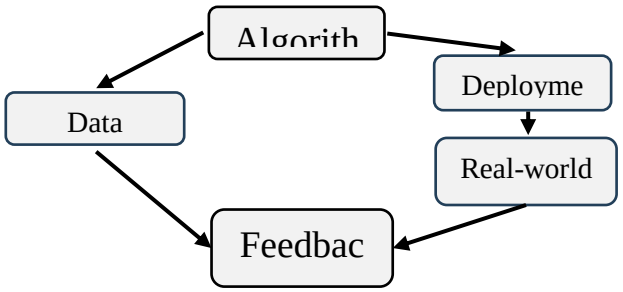


Figure 1: Bias propagation across the AI lifecycle:

In this figure, the sources of bias are illustrated and addressed at every phase of the AI lifecycle, including data collection, preprocessing, model development, deployment, and monitoring. The loop of feedback on the results of deployment affects further data collection and re-training of models, leading to the fairness loss even with mitigation. The significant reinforcement directions of bias are underlined to demonstrate the ineffectiveness of single-point mitigation.

4.1|Integrated Bias-Mitigation Framework (IBMF):

The results of this review have shown that creating AI systems free of bias is unlikely to be achieved when interventions are implemented at a single point in the development pipeline. Pre-processing identifies discrepancies in the datasets, though it may be influenced by population drift. In-processing improves equity in model formation, but it is computationally expensive and difficult to operate. Post-processing minimizes disparities in a short period of time, but not the causes of inequality. Likewise, governance-based strategies encourage accountability but cannot compensate for technical flaws in design. All these limitations highlight the importance of a lifecycle-based, holistic approach. This study, in turn, suggests the Integrated Bias-Mitigation Framework (IBMF) as a multi-tiered framework that unites technical and organizational procedures to ensure sustainable fairness throughout the lifecycle of AI.

The IBMF comprises five overlapping layers that are not operational in a sequence, although they work continuously. It begins with data governance, which reduces bias during data collection and curation, and analyses representativeness, the validity of the guidelines applied to annotate data, proxy variable audit, and provenance documentation [9], [13]. On this foundation, the model development stage ensures impartiality in the training procedure through the adoption of fairness-constrained optimization, adversarial debiasing, and counterfactual reasoning, and by treating fairness as a model objective rather than a defect correction [11], [12].

Fairness performance should then be justified through intensive evaluation and stress testing, based on subgroup performance measures, intersectional analysis, and scenario-based robustness tests. This step elucidates the fairness/accuracy/computational cost trade-offs and enables informed decision-making rather than the unintended trade-off of fairness [20], [25]. However, in practice, development does not necessarily translate into equitable outcomes. This is why IBMF focuses on human-in-the-loop governance during deployment to ensure that automated predictions do not substitute for important judgment and organizational responsibility. Mechanisms of governance include algorithmic audit processes, escalatory channels for model-based harms, multidimensional scrutiny councils, and transparent ownership of equity results [4], [26].

Post-deployment monitoring and system adaptation are the final elements of IBMF, as fairness is a dynamic attribute that must be sustained over the long term. As soon as they are released, AI systems receive feedback on changing user behavior, social organization and the environment. Ongoing auditing, fairness-tracking dashboards, subgroup error rates, user feedback mechanisms, and model recalibration can be used to avoid fairness degradation and to strengthen discrimination [6], [14]. This perpetual monitoring is feedback to data governance; thus, the framework is an iterative protocol as opposed to a linear one.

Combined, the IBMF's bias mitigation idea is a long-term commitment that requires organization on both technical and organizational levels. Fairness is thus not to be understood as an absolute goal but as a dynamic commitment that extends beyond model design to actual decision-making scenarios. The framework also serves AI

fairness scholarship by integrating disparate mitigation strategies into a unified lifecycle model that offers practical advice to data scientists, developers, organizations, and policymakers in high-stakes sectors.

Table 2. Conceptual Summary of the IBMF

IBMF Layer	Role in Fairness	Typical Mechanisms
Data governance	Preventing bias at source	Dataset representativeness checks; annotation protocols; proxy audits
Fairness-aware model development	Embed fairness into training	Fairness-constrained optimization; adversarial debiasing; explainability
Evaluation & stress-testing	Identify biases prior to deployment	Subgroup metrics; intersectional testing; scenario analysis
Governance & deployment oversight	Ensure responsible model use	Algorithmic audits; escalation pathways; human-in-the-loop review
Post-deployment monitoring	Sustain fairness over time	Continuous auditing; model recalibration; end-user feedback

The proposed lifecycle approach is summarized visually in Figure 2, which integrates data governance, fairness-aware model development, evaluation, deployment oversight and post-deployment monitoring into a closed iterative system.

Figure 2. Distribution of fairness degradation across AI lifecycle stages

This number represents an overview of the percentage of studies that were reviewed and found to exhibit such degradation of fairness in data-related processes, algorithm design, or post-deployment feedback. It provides a comparative perspective on which areas of fairness the issues are most often found and the phases where bias recurrence is most likely.

4.2|Comparative Evaluation of Mitigation Strategies:

Although a wide range of techniques has been proposed to mitigate bias in AI, their performance varies widely across data environments, domain characteristics, and deployment contexts. The literature shows that bias mitigation cannot be divorced from practical constraints, fairness objectives or the socio-technical conditions in which AI systems operate. When viewed collectively, mitigation strategies fall into four overarching categories: pre-processing, in-processing, post-processing and governance mechanisms. Each category has advantages, limitations and typical conditions under which it is most effective.

Pre-processing methods, such as dataset reweighting, resampling and synthetic augmentation, have demonstrated strong improvements when dataset imbalance is the primary source of discrimination [13], [25].). These techniques are attractive because they do not require altering a model's internal architecture and can be applied to vendor-supplied systems. However, their impact is likely to decrease when models are applied to populations different from the training sample. Healthcare and recruitment: Case studies demonstrate that even acquired benefits of fairness, obtained through rebalancing, often fail when demographic distributions of a population change or when other groups are brought into the data pipeline [7], [9] Pre-processing is therefore a robust base, yet not a solution on its own.

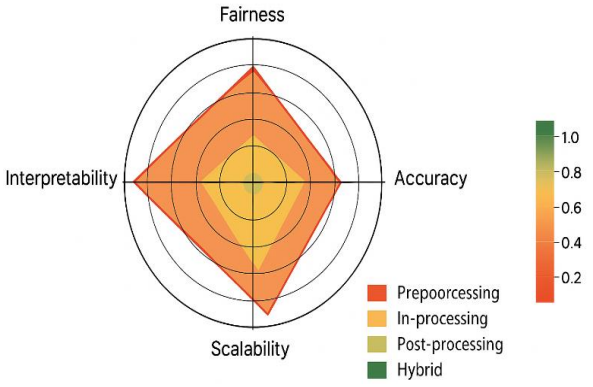
The theoretically most promising approaches to removing algorithmic bias in training are in-processing methods, such as fairness-constrained optimization, representation learning and adversarial debiasing [11],

[12]. These methods reduce structural-level disparities and do not use post-hoc adjustments. However, the actual usage is not widespread. Numerous real-world organizations lack access to protected attributes required to compute fairness loss functions due to legal or ethical constraints. Moreover, adversarial training can be

computationally expensive and, in practice, may not be feasible for real-time decision-making (e.g., insurance pricing or financial scoring systems). Consequently, in-processing is still primarily focused on academic research rather than on standard industrial practice.

Post-processing techniques, including equal opportunity correction and threshold correction, are standard in the industry because they do not require retraining the model and can be quickly compared with operational systems [10]. These methods have shown quantifiable short-term improvements in fairness outcomes in lending, hiring, and risk-scoring applications [4], [9]. They, however, do not address the underlying data or structural sources of bias, since these only occur at the model output stage. The fairness degradation across several deployments revealed that, after a few months of operation, post-processing alone cannot guard against feedback loops or the reintroduction of disparities by institutional incentives[2], [7].

Mechanisms driven by governance hold another position, as they are aimed at human control and the institute's responsibility, rather than altering the model's behavior. These processes involve algorithm auditing, assigning accountability, maintaining fairness risk registries, and establishing model-related harm escalation routes [6], [26].). Government policies have been proven effective (high stakes, such as criminal justice or healthcare) when there are a dedicated monitoring units and a mechanism to step in and enforce procedures. However, there is no way that governance can make up for erroneous modelling choices or biased datasets. In



cases of implementation where there is no robust technical infrastructure, oversight is reactive rather than preventive.

A comparative synthesis of these methods is a testament to the fact that no category of mitigation would ensure fairness in the long term. All the techniques address a distinct level of injustice and fail under various circumstances. The most significant overall evidence across the research is that mitigation effectiveness hinges on the compatibility of a variety of plans, and not on the discovery of an optimal method. This insight directly supports the IBMF approach, where data governance, fairness-aware development, evaluation, oversight and post-deployment monitoring reinforce one another rather than act independently.

Table 3. Comparative Evaluation of Bias-Mitigation Strategies Across Lifecycle Stages

Strategy	Strengths	Limitations	Best-fit Conditions
Pre-processing	Corrects dataset imbalance early	Weak under domain/population shift	Minority under-representation in structured datasets
In-processing	Strongest fairness gains	High computational and data-access demands	Controlled research environments; sensitive attribute access
Post-processing	Fast deployment; no retraining	Treats symptoms, not causes	Deployed vendor-supplied systems needing rapid fairness improvement
Governance	Human oversight and accountability	Cannot fix technical weaknesses	High-stakes public-impact settings requiring risk control

The table compares the advantages and disadvantages of the key mitigation strategies and identifies which stages of the lifecycle each strategy is most effective for.

Figure 3. Sources and drivers of fairness degradation in AI systems

This diagram categorizes the primary sources of fairness degradation observed during data review, including data drift, demographic imbalance, lack of recalibration, feedback loops, and lack of human control. The figure illustrates the interaction between these factors to determine the long-term outcome of fairness and model performance.

4.3|Case Studies:

To evaluate how sources and mitigations of AI bias manifest across real-world environments, evidence from the reviewed literature was synthesized across four high-stakes domains: healthcare, recruitment and hiring, finance, and public safety. These domains offer representative contexts in which AI decision-making directly influences human welfare, resource allocation and life opportunities. Across all four cases, the success of mitigation depended not only on technical interventions but also on institutional decision structures, deployment practices and post-deployment monitoring.

In healthcare, diagnostic and risk-prediction models repeatedly demonstrated disparities in treatment recommendations and triage outcomes due to the under-representation of specific demographic groups and the use of historical cost data as a proxy for health needs [5], [9]. Although dataset balancing and synthetic augmentation improved equity in experimental evaluations, fairness often deteriorated during deployment due to shifting population distributions and inadequate model recalibration [7]. These findings reflect the need for continuous monitoring and lifecycle fairness rather than one-off dataset correction.

Recruitment and hiring applications illustrate how apparently neutral features can encode harmful social patterns. Automated résumé-screening systems trained on historical hiring decisions replicated gender and racial disparities because feature embeddings, such as job tenure, language use or educational background, served as proxies for protected attributes [4], [9]. In-processing fairness limits minimized group differences in training, but when labor-market conditions varied or post-deployment supervision was lax, fairness degradation. These findings indicate that the development governance of fairness-aware models should be accompanied by periodic auditing and deployment.

The same scenario could not be said of the finance industry. Structural inequalities are often incorporated into credit risk models and insurance algorithms through behavioral and socioeconomic characteristics [1], [2]. Post-processing methods, including threshold manipulation, were widely used because they quickly balanced acceptance rates and did not require retraining. Their temporary success, however, was indicative of larger problems, and the fairness degradation would often reappear as economic circumstances changed. Real-world fairness needs more than score adjustments and requires governance that will grow along with the environment.

The most worrying evidence of reinforcement cycles comes from the context of public safety and policing. Predictive policing, based on crime and arrest history, favored directing law-enforcement resources to already over-policed neighborhoods, with self-reinforcing feedback loops[1], [6]Although algorithmic fairness corrections were made in advance of deployment, biases re-emerged because human decision-makers were inclined to accept algorithmic risk scores, and organizational incentives favored efficiency over fairness. The cases demonstrate that governance and post-deployment auditing are effective in avoiding bias and reducing fairness concerns over time.

In these areas, there is a regularity of failures of fairness that are regular and seldom due to a technical malfunction of any kind, but instead to the conjunction of data, algorithms and deployment circumstances. Effective mitigation in any field would entail technical changes as well as organizational accountability for consistent monitoring and supervision. This strengthens the IBMF model, which views fairness as a lifecycle responsibility that must be incorporated into the operational environment.

RESULTS AND DISCUSSION

As this review shows, bias in AI systems is not a consequence of technical malfunctions, but an emergent feature of socio-technical settings in which data, algorithms, and institutional practices interact. The asymmetry of information and historical inequity had

always led to failures of fairness, which were later compounded by the model development process that optimized decisions and proxy factors. Such failures were evident at the time of deployment due to the feedback loop, automation bias, and the absence of governance. This evolving tendency points to one of the most limiting aspects of modern mitigation research: solutions to restore fairness are likely to target individual lifecycle phases, yet bias persists throughout the lifecycle. Consequently, real-life outcomes are not always linked to fair gains made during development.

The existing literature informs the results and warns against conceptualizing fairness as a technical optimization problem (Noble, 2018; Crawford, 2021). Most thoughtfully designed fairness-promoting optimization approaches, most notably fairness-constrained optimization and adversarial debiasing, can substantially improve fairness in an experimental setting but lack scalability, especially when sensitive attributes are unknown or real-time predictions are required (Zhang, Lemoine and Mitchell, 2018; Ferrara, 2024). Meanwhile, operationally attractive strategies such as post-processing threshold choices appear to be effective in small-scale applications but degrade when data distributions change or when organizational motivation is based on efficiency rather than equity (Crawford, 2021; Grey, 2024). Governance-based tools, such as algorithms, audits, and human-in-the-loop oversight, cannot work without technical sound models and high-quality datasets to ensure they lead to public accountability. All these trends, in turn, confirm the hypothesis that fairness is not a quality of a model that is determined, but a dynamically changing performance requirement that depends on technical and organizational factors.

IBMF is a combination of these lessons and conceptualizes fairness as a lifecycle requirement and not an engineering process. Unlike the disjointed mitigation strategies, the IBMF integrates data governance, fair and rigorous model development, deployment control, and post-deployment monitoring into a cohesive system. This holistic framework makes two important contributions to the fairness literature. Theoretically, it resolves the tension between technical and governance-oriented perspectives by treating them as complementary rather than competing domains. In practice, it provides implementable guidance for organizations that deploy AI in high-stakes settings, clarifying who is responsible for fairness at each lifecycle stage and when different types of mitigation should be activated.

Although the review advances understanding of AI fairness, several limitations point toward future research opportunities. First, the review focused on peer-reviewed English-language studies, which may exclude non-English surveillance, public-policy and industry work. Second, the heterogeneity of fairness metrics and deployment conditions prevented quantitative meta-analysis; future work may examine domain-specific aggregations. Third, empirical evaluation of governance and post-deployment monitoring remains limited, partly because organizations rarely allow external auditing of proprietary models. Longitudinal field research is crucial to evaluate whether lifecycle-based frameworks, such as the IBMF, prevent fairness degradation over time. Lastly, even though the IBMF is domain-independent, future research can investigate domain-specific adaptations in which fairness can rely on domain-specific priorities, such as patient safety in the medical field or due-process protections in law enforcement.

This study suggests that sustainable fairness in AI should have a change of focus to model-centric interventions and structural and iterative mitigation approaches that are operational during the entire lifecycle of AI systems. The IBMF helps drive this change in attitude by showing that the best bias-reduction strategies are not individualized but integrated systems of governance and technical practice. The solution to AI bias is, thus, a long-term commitment to operations and not a short-term adaptability of models that strengthens the aspect of fairness as an institutional duty

CONCLUSION

The work is a synthesis of evidence from 128 peer-reviewed articles that explored sources of bias in AI systems and the effectiveness of mitigation strategies in high-stakes domains. The review has shown that bias is introduced at several stages of the AI lifecycle, including data collection, model construction, and real-world

application, and persists when interventions are applied only to individual parts of the socio-technical system rather than the system. Even such sophisticated mitigation measures cannot be effective when implemented without organizational controls and post-implementation follow-up. The results prompt a reevaluation of the belief that fairness can be realized when technical optimization is applied separately and introduce bias alleviation as a continuous operational task.

The Integrated Bias-Mitigation Framework (IBMF) presented in this study is expected to overcome this issue by organizing data governance, model development that considers fairness, stress testing and assessment, deployment monitoring, and post-deployment monitoring. It contributes by incorporating technical and organizational mechanisms into a lifecycle framework that adapts to changes in user behavior, institutional incentives, and environmental conditions. Rather than viewing fairness as a unitary process to be addressed at the model training stage, according to IBMF, a balanced approach to equity in AI can be ensured through technical, procedural, and ethical safeguards that mutually reinforce one another in the long term.

Several implications can be drawn from this synthesis. For AI practitioners and data scientists, fairness and interpretability requirements should be established in advance, not applied in the latter stages of implementation. For organizations that apply AI, accountability, audit procedures, and allocate resources to address AI, a need to prevent fairness degradation and automation bias. To policymakers and regulators, compliance frameworks should not just remain transparent but also continuously include checks for fairness, external auditing, and effective redress mechanisms. These proposals point to the fact that fairness is not a technical but a socio-technical process that should be sustained on the institutional level.

There are limitations to the research. The research is based on peer-reviewed English articles, which may ignore key non-academic and non-English sources. In addition, heterogeneity in the fairness measure prevented quantitative pooling of study results. These limitations suggest possible future directions for research, in particular, longitudinal studies of fairness in real deployments and the creation of sector-specific adaptations of the lifecycle fairness models. The relevance of such work is to determine whether the fairness degradation can be avoided using the holistic methods, such as the IBMF, in operational settings.

Overall, the evidence analyzed in this study suggests that long-term fairness in AI does not require algorithmic solutions. Unbiased mitigation is an iterative, structural and governance-conditioned effort. The IBMF offers conceptual and practical advice to AI practitioners, organizations and regulators seeking to deploy AI responsibly, equitably and in line with social values by melding the technical and organizational aspects of fairness into a single lifecycle model.

REFERENCE

[1] Afreen, J., Mohaghegh, M. and Doborjeh, M. (2025), 'Systematic literature review on bias mitigation in generative AI', AI and Ethics, vol. 5, no. 5, pp. 4789–4841. https://doi.org/10.1007/s43681-025-00721-9.

[2] Barocas, S. and Selbst, A.D. (2016), 'Big data's disparate impact', California Law Review, vol. 104, no. 3, pp. 671–732.https://doi.org/10.2139/ssrn.2477899

[3] Braun, V. and Clarke, V. (2006), 'Using thematic analysis in psychology', Qualitative Research in Psychology, vol. 3, no.2,pp.77101.https://doi.org/10.1191/1478088706qp063oa.

[4] Buolamwini, J. and Gebru, T. (2018), 'Gender shades: Intersectional accuracy disparities in commercial gender classification', in Friedler, S.A. and Wilson, C. (eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT), Proceedings of Machine Learning Research, vol. 81, pp. 77–91. Available at: https://proceedings.mlr.press/v81/buolamwini18a.html

[5] Calmon, F.P., Wei, D., Natesan Ramamurthy, K. and Varshney, K.R. (2017), 'Optimized data pre-processing for discrimination prevention', in Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). https://doi.org/10.48550/arXiv.1704.03354

[6] Chen, Z. (2023), 'Ethics and discrimination in artificial intelligence-enabled recruitment practices', Humanities and Social Sciences Communications, vol. 10, article 567. https://doi.org/10.1057/s41599-023-02079-x

[7] Crawford, K. (2021), Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, Yale University Press, New Haven, CT. ISBN 9780300209570.

[8] Ferrara, E. (2024), 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and

mitigation strategies', Sci, vol. 6, no. 1, article 3. https://doi.org/10.3390/sci6010003

[9] Gray, M., Samala, R., Liu, Q., Skiles, D., Xu, J., Tong, J. and Wu, L. (2024), 'Measurement and mitigation of bias in artificial intelligence: A narrative literature review for regulatory science', Clinical Pharmacology & Therapeutics, vol. 115, no. 4, pp. 687–697. https://doi.org/10.1002/cpt.3117

[10] Hardt, M., Price, E. and Srebro, N. (2016), 'Equality of opportunity in supervised learning', in Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS), pp. 3315–3323. (Preprint) https://doi.org/10.48550/arXiv.1610.02413

[11] Jarrahi, M.H., Karami, A., Conway, P., Memariani, A. and Lutz, C. (2025), 'Navigating the muddy waters of bias in artificial intelligence research: Understanding divergent meanings and conceptions', Technology in Society, vol. 84, article 102437 (online first).

[12] Kitchenham, B. and Charters, S. (2007), Guidelines for Performing Systematic Literature Reviews in Software Engineering, EBSE Technical Report, Keele University and Durham University Joint Report.

[13] Noble, S.U. (2018), Algorithms of Oppression: How Search Engines Reinforce Racism, NYU Press, New York, NY. https://doi.org/10.2307/j.ctt1pwt9w5

[14] Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019), 'Dissecting racial bias in an algorithm used to manage the health of populations', Science, vol. 366, no. 6464, pp. 447–453. https://doi.org/10.1126/science.aax2342

[15] O'Neil, C. (2016), Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, New York, NY.

[16] Page, M.J. et al. (2021), 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', Journal of Clinical Epidemiology, vol. 134, pp. 178–189. https://doi.org/10.1016/j.jclinepi.2021.03.001

[17] Raghavan, M., Barocas, S., Kleinberg, J. and Levy, K. (2020), 'Mitigating bias in algorithmic hiring: Evaluating claims and practices', in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20), ACM, Barcelona, pp. 469–481. https://doi.org/10.1145/3351095.3372828

[18] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020), 'Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing', in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20), ACM, Barcelona, pp. 33–44. https://doi.org/10.1145/3351095.3372873

[19] Suresh, H. and Guttag, J.V. (2021), 'A framework for understanding sources of harm throughout the machine learning life cycle', in Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), ACM. https://doi.org/10.1145/3465416.3483305

[20] Zhang, B.H., Lemoine, B. and Mitchell, M. (2018), 'Mitigating unwanted biases with adversarial learning', in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), ACM, pp. 335–340. https://doi.org/10.1145/3278721.3278779