The American Journal of Engineering and Technology

ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 192-201
DOI 10.37547/tajet/Volume07Issue05-19

Mitigating Algorithmic Bias in Predictive Models

Tamanno Maripova
Data Analyst New York, USA

OPEN ACCESS

SUBMITED 21 March 2025
ACCEPTED 24 April 2025
PUBLISHED 29 May 2025
VOLUME Vol.07 Issue 05 2025

CITATION
Tamanno Maripova. (2025). Mitigating Algorithmic Bias in Predictive Models. The American Journal of Engineering and Technology, 7(05), 192–201. https://doi.org/10.37547/tajet/Volume07Issue05-19.

COPYRIGHT
© 2025 Original content from this work may be used under the terms of the creative commons attributes 4.0 License.

Abstract: This article considers the issue of systematic errors in predictive machine-learning models generating disparate outcomes for different social groups and proposes a holistic approach to its mitigation. The risks and increasing legal requirements, along with corporate commitments to ethical AIs, drive the relevance of this study. The work herewith attempts to develop a bias-source taxonomy at data collection and annotation, proxy-feature selection, model training, and deployment stages; also, it tries to compare pre-, in-, and post-processing methods' effectiveness on representative datasets measured by demographic parity, equalized error rates, and disparate impact. This article is unprecedented in undertaking a two-level approach: first, a systematic review of regulatory definitions (NIST, IBM) and case studies (COMPAS, healthcare-service prediction, face recognition) that identified key bias factors from sample imbalance to feedback loops; second, an empirical comparison of Reweighing, adversarial debiasing, threshold post-processing techniques alongside flexible multi-objective strategies—YODO (via AI Fairness 360 and Fairlearn libraries)—considering acceptable accuracy losses. The root source of unfairness remains data bias; hence, pre-processing must be undertaken (rebalancing, synthetic oversampling), while in- and post-processing can essentially harmonize group metrics at some cost in accuracy reduction Furthermore, without continuous online monitoring and documentation (datasheets, model cards), the balanced model risks losing fairness due to dynamic feedback effects. Bringing together technical fixes with rules and making the audit process official ensures the ability to copy and openness, which is key for long-term faith in AI systems. This article will help machine-learning builders, AI-responsibility experts, and checkers find ways to find, gauge, and lessen algorithmic bias in live models.

Keywords: algorithmic bias, fairness, pre-processing, in-processing, post-processing, demographic parity, equalized odds, disparate impact, AI Act, NIST AI RMF, model cards.

Introduction: Algorithmic bias refers to the systematic error of running a machine-learning model in such a way that causes members of different social groups to receive drastically different predictions or decisions. It develops when a model takes on historic inequities in the data, amplifies them through its training process, or applies them to new situations where implicit correlations stand in for causal connections. Consequently, some groups are perceived as "risk-neutral" by default and others as "high risk" a priori, though in reality, event probabilities are equal. This is precisely how the National Institute of Standards and Technology (NIST) describes the problem—as a consequence of quantitative methods "flattening" rich social context into numerical categories, creating an illusion of objectivity—while IBM defines it as "systematic errors that produce unfair outcomes" [1, 2].

Concurrent legal, reputational, and social effects confirm the significance of this issue for business and society. The legal risk is evident: under the EU AI Act, violations of the prohibition against discriminatory practices may incur fines of up to €35 million or 7% of a company's global annual turnover [2]. Reputational damage is measured in lost trust: 86% of surveyed organizations believe customers prefer brands that transparently apply ethical principles to their AI systems [3]. The social cost manifests in concrete human lives: analysis of the COMPAS tool showed that non-recidivist Black defendants were almost twice as likely as white defendants (45% vs. 23%) to be incorrectly classified as "high risk" of reoffending [4]. Together, these facts demonstrate that ignoring bias not only exacerbates existing inequalities but also creates direct financial losses and legitimacy threats for companies in the eyes of society.

MATERIALS AND METHODOLOGY

The materials and methodology of this study are based on a critical review of 29 publications from academic journals, industry reports, and regulatory documents. The theoretical foundation employs definitions of algorithmic bias from NIST and IBM, emphasizing systematic errors that lead to unfair outcomes [1, 2] and an empirical analysis of the COMPAS tool demonstrating real cases of discrimination in judicial predictions [4]. To detect data biases, we analyzed model performance across groups. Specifically, we compared top-5 classification accuracy on ImageNet for images from regions with different income levels [5] and examined gender-recognition errors in commercial systems across "race–gender" combinations [6].

To mitigate bias, three classes of technical strategies were considered. The first line of defense comprises pre-processing methods, such as Reweighing, that adjust the weights of training-set instances without altering the algorithm, achieving a disparate impact of 1.0 on the Adult dataset [18, 19]. The second class includes in-processing techniques that embed fairness constraints directly into the loss function: adversarial debiasing achieved equalized odds parity with no more than a 2% reduction in overall accuracy [8]. The third line entails post-processing algorithms that adjust model output probabilities via threshold optimization to balance group error rates [20].

The legal and regulatory justification of the approach is ensured by mapping these technical practices to the requirements of the EU AI Act (mandatory dataset audit and discrimination checks, Art. 10) [10], NIST AI RMF recommendations (category "harmful bias") [11, 27], Canada's Algorithmic Impact Assessment [12], the ICO's GDPR and AI guidance [13], Singapore's Model AI Governance Framework [14], the UK Financial Conduct Authority's directives for the financial sector [15], and the ISO/IEC 42001:2023 standard on continuous fairness monitoring [16]. These documents draw upon the OECD principles for eliminating unfair bias in AI systems [17].

RESULTS AND DISCUSSION

Algorithmic bias almost always begins with data bias: if individual countries, income brackets, or social groups are underrepresented in the training set, the model inevitably absorbs the statistical skew. Analysis [5] showed that for six popular ImageNet classifiers, top-5 accuracy on objects from households with monthly incomes below USD 50 is on average 10% lower than on images from the wealthiest categories, and the gap widens for scenes from non-Western regions, as shown in Fig. 1 [5].

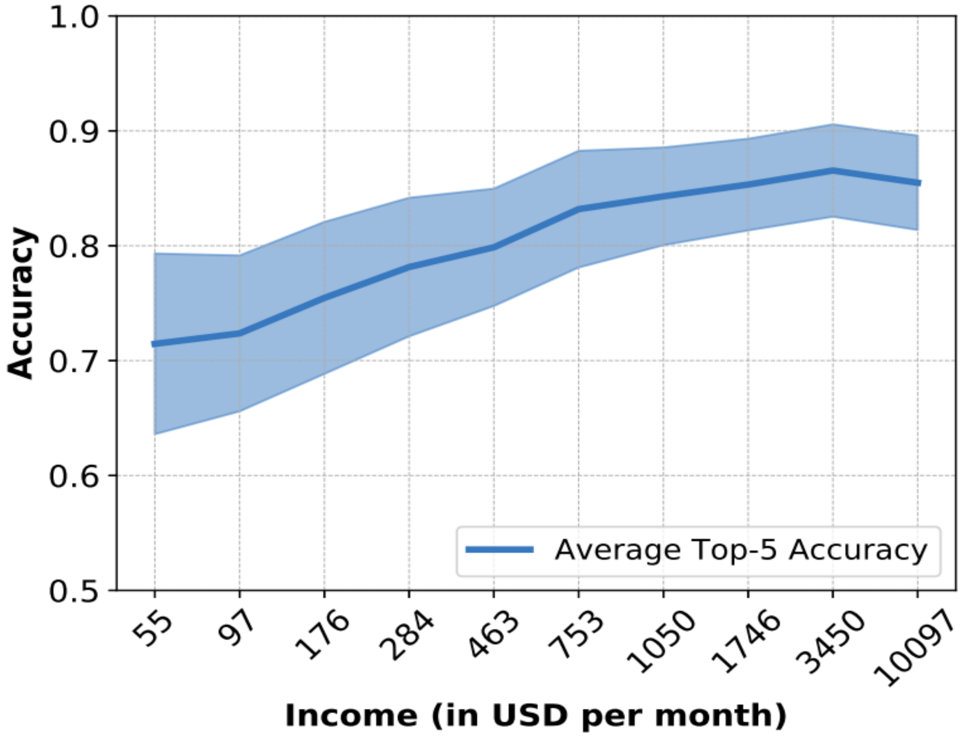

Fig. 1. Top-5 Accuracy by Income [5]

Such “blind spots” are not accidental: they reflect a historical research focus on English-language Internet content and commercially attractive markets. If these imbalances are not counteracted by rebalancing, synthetic oversampling, or causal justification of features, subsequent development stages can only mitigate rather than eliminate the root cause.

A model can err even with a formally balanced sample due to measurement distortions. A classic example is sensor inaccuracies or manual annotation errors that correlate with appearance. In study [6], commercial gender-recognition systems misclassified dark-skinned women in 34.7% of cases, whereas for light-skinned men the error was only 0.8%. Because the algorithm “sees” incorrect or noisy labels as truth, subsequent training merely entrenches these differential errors, transforming them into systematic discrimination.

Another important source of bias is the choice of objective function and evaluation metrics. In the widely used patient stratification algorithm studied in [7], healthcare cost was used as a proxy for health status. The metric that optimally reflected costs proved poorly correlated with actual care needs, and even a perfect model under this formulation inevitably produces a biased outcome.

Even with a correctly specified task, the model architecture and hyperparameter settings influence the error distribution. Overly aggressive regularization or skewed class-weight coefficients can shift the decision boundary so that gains in overall accuracy come at the expense of a higher false-negative rate for the vulnerable group. Study [8] showed that post-processing a single decision by selecting differentiated thresholds can equalize false-positive and true-positive rates between groups at the cost of a moderate loss in overall accuracy of a couple of percentage points. This underscores that fairness concerns must be addressed in the data and the very “wiring” of the algorithm.

Finally, feedback loops can quickly bias even a perfectly calibrated model after deployment. In [9], the PredPol predictive-policing system, after only a few iterations, began directing police almost exclusively to neighborhoods where arrests had already been recorded, amplifying the divergence between observed and actual crime activity. Since the model’s actions generate the subsequent training set, even a slight initial bias accumulates exponentially. Such dynamic effects require online monitoring and active “continuous” debiasing methods; otherwise, any static fairness assessment rapidly becomes outdated.

Regulatory efforts to reduce algorithmic bias form a multilayered system in which supranational norms set minimal requirements, and sectoral and national documents refine them for specific risks. Today, this system's center is Regulation (EU) 2024/1689, the EU AI Act. For “high-risk” systems, it introduces a mandatory dataset audit, discrimination checks before deployment, and a requirement to maintain detailed documentation on data collection and annotation processes, enshrined in Art. 10 “Data Governance” [10].

Suppose the European approach relies on strict enforcement in the United States. In that case, the voluntary but widely adopted NIST AI Risk Management Framework serves as a “de facto standard.” Since its publication on January 26, 2023, the document has defined a risk matrix in which “harmful bias” is highlighted as one of five key categories; by summer 2024, NIST had added a separate profile for generative models, identifying even more new risks, including erroneous content personalization [11].

Several governments and sectoral regulators are building their complementary mechanisms. In Canada, all federal algorithms are subject to a mandatory Algorithmic Impact Assessment: 51 risk-related questions and 34 mitigation measures allow systems to be classified into four impact levels, with proportional bias-handling requirements [12]. In March 2023, the UK Information Commissioner’s Office updated its “AI and Data Protection” guidance, detailing how to assess and mitigate bias at every stage of the model lifecycle and permitting the processing of sensitive data for discrimination testing [13]. In May 2024, Singapore released the “Model AI Governance Framework for Generative AI,” dedicating chapters to data provenance and independent testing, and recognizing bias mitigation as one of nine pillars of “trust” [14]. In 2024, the UK Financial Conduct Authority integrated the risk of unfair outcomes into its overall oversight of credit-scoring models [15], and the ISO/IEC 42001:2023 international standard proposed a managerial “overlay” for all AI processes, including mandatory fairness-metric monitoring [16].

For international alignment, these frameworks draw on the OECD principles, which since 2019 have emphasized the need to eliminate “unfair bias” and by May 2023 had inspired over 1,000 policies across 70 jurisdictions [17]. The common thread is a risk-based approach: the higher the potential social harm, the more detailed the data checks, model transparency, and legal safeguards must be. As a result, companies operating globally effectively climb a unified “compliance ladder”: from NIST’s voluntary metrics and industry guides through ISO 42001 certification to the legally binding requirements of the EU AI Act. This evolutionary logic reduces regulatory fragmentation and shifts the fight against algorithmic bias from ethical declarations to measurable, verifiable obligations.

The regulatory requirement to measure and mitigate bias moves the issue from abstract ethics into practical engineering solutions, so developers rely on three classes of technical strategies. The first line of defense is pre-processing methods that correct the data before training the model. In practice, this may be simple weight rebalancing: for the Adult Income dataset, the initial disparate-impact ratio between men and women was 0.36, and after applying Reweighing, it became 1.0—that is, statistically “discrimination-free” [18]. In clinical prediction of postpartum depression, the same technique raised disparate impact from 0.31 to 0.79 and almost eliminated the difference in true-positive rates between racial groups while preserving model accuracy [19]. A comparison of bias metrics on the test dataset—using a baseline model, a race-blind model, a model debiased via Reweighing, and a model debiased via Prejudice Remover (logistic regression)—is shown in Fig. 2. Such methods require no algorithmic changes. Still, their efficacy is limited to cases where bias resides entirely in the data.

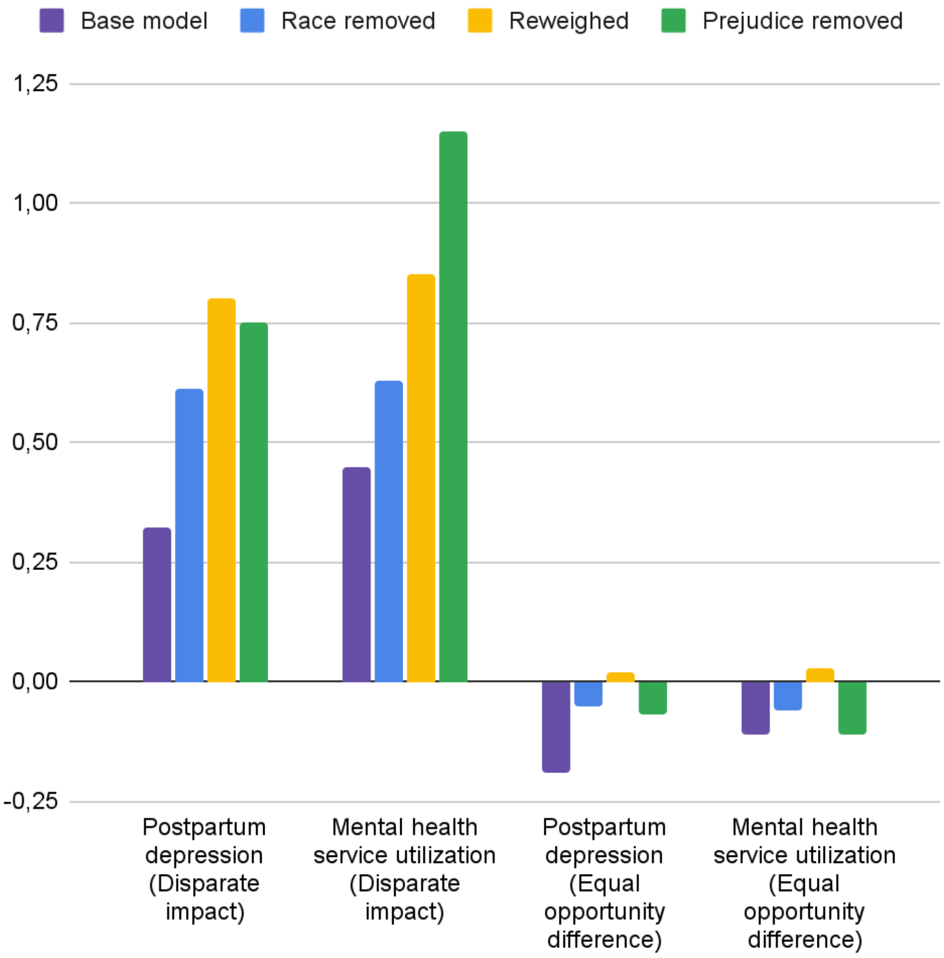

Fig. 2. Comparison of bias metrics [19]

If reweighing proves insufficient, one moves to in-processing techniques that embed fairness directly into the loss function. The most popular approach is adversarial debiasing: alongside the primary predictor, a discriminator is trained to infer the protected attribute, and the predictor's objective is to make accurate forecasts while obfuscating valuable information to the discriminator. On the Adult Income dataset, this scheme improved disparate impact and reduced average-odds difference to nearly zero with only a 2% drop in overall accuracy [18]. Adversarial methods provide the most incredible group parity but require gradient access to the model and can be unstable without careful tuning.

The third line comprises post-processing algorithms that modify the obtained predictions without retraining the model. A classic example is a linear program that adjusts predicted probabilities to equalize false-positive and false-negative rates between privileged and vulnerable groups while leaving test power almost unchanged [20]. This "black-box" approach is especially valuable when the original model is proprietary or frozen, but it is limited to binary classification tasks and sensitive to threshold choices.

Specialized libraries exist to facilitate the rapid integration of all three tactics. IBM AI Fairness 360 implements ten mitigation algorithms covering the full pre-, in-, and post-processing spectrum. It provides 70 metrics for evaluating group and individual fairness, making it the most comprehensive open platform [21]. A lighter but actively developed alternative is Microsoft's Fairlearn. Thus, a developer can execute Reweighing or Adversarial Debiasing in AIF360 with a few lines of code, compare results with Fairlearn metrics, and document the trade-off between accuracy and fairness, thereby ensuring compliance with both regulatory minima and internal corporate-responsibility standards.

Classic fairness metrics—demographic parity, equalized odds, and predictive calibration—measure statistical dependencies but ignore causal links between features and protected attributes. Consequently, satisfying them all simultaneously is mathematically impossible outside trivial cases; the "fairness impossibility theorem" proves that the three most popular criteria cannot be achieved simultaneously, necessitating a trade-off in real-world data [22]. In healthcare, the choice of a "convenient" proxy label illustrated how external measures can mislead: the algorithm optimized for treatment costs enrolled only 17.7% of Black patients into additional support instead of the clinically justified 46.5% (Fig. 3), because historically less was spent on Black patients [7].

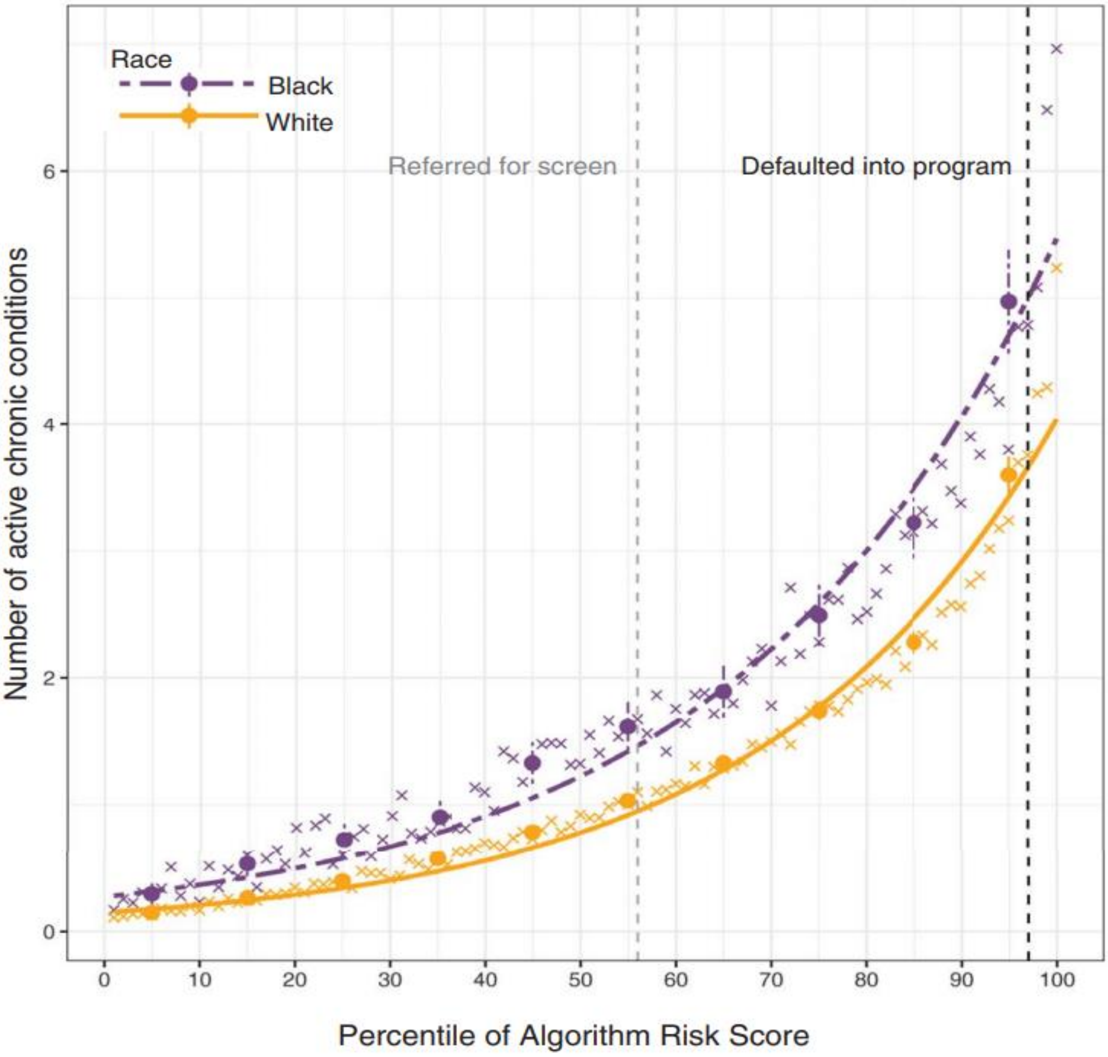

Fig. 3. Number of chronic illnesses versus algorithm-predicted risk, by race [7]

To move beyond purely correlational criteria, counterfactual fairness is employed: a decision is deemed fair if it would remain unchanged in a hypothetical world where the individual belongs to a different group under the same risk factors. Formalized via structural causal models, this approach "removes" group-only associations. In a classical experiment predicting law-student performance, the counterfactually fair "Fair Add" model reduced root-mean-square error from 0.873 to 0.918 (i.e., lost about 5%)—but eliminated prediction dependence on race: when the protected attribute was swapped, grade distributions coincided completely, whereas the baseline model exhibited a systematic shift [23]. This example demonstrates that a small accuracy cost can radically reduce hidden discrimination.

Organizations aim to navigate the accuracy–fairness trade-off rather than fix a single configuration. Modern methods combine both criteria into a unified loss function or treat them as a multi-objective problem. The You Only Debias Once (YODO) approach trains the model simultaneously on two extremes—accuracy-optimum and fairness-optimum—and finds in weight space a “line” of solutions along which the balance can be adjusted at inference time. For the ACS-E dataset, generating one hundred Pareto points took 3.53 s instead of 425 s for training one hundred separate models, with each solution remaining on the same “error ↔ demographic parity” front [24]. Combined or multi-objective optimizations do not override theoretical limits but provide managers with a transparent navigation tool, allowing the selection of a point acceptable to business, legal, and social-responsibility requirements simultaneously.

Once a company has identified and mapped bias sources to regulatory requirements, the next task is to formalize a reproducible bias-management process. In practice, this begins with a full-scale audit and data profiling: technical specialists verify sample representativeness, assess annotation quality, and identify signs of historical bias, while internal auditors record checkpoints. Regulators and professional communities already consider such an audit standard: ISACA defines algorithmic audit as a key method for detecting bias at “all points of the model lifecycle” [25]. However, real-world adoption remains limited: only 47.2% of organizations working with generative AI conduct regular checks [26]. These figures indicate that a missing audit stage remains the most significant “gap” in discrimination protection.

The next step is to define which model use cases are critical and which metrics fairness will be measured. NIST AI RMF recommendations propose starting with a harm map: first, describe which groups may be harmed, and only then select a statistical criterion—demographic parity, equalized odds, or individual fairness—that best reflects that risk [27]. This sequence helps avoid optimizing a “convenient” metric unrelated to social harm. The data team then conducts a series of controlled experiments: applying pre-, in-, and post-processing methods to the baseline model, with results displayed on the accuracy–fairness trade-off surface. Integration via AIF360 and Fairlearn reduces the “hypothesis → evaluation” cycle to minutes, enabling product managers to make decisions based on a complete picture of trade-offs.

When an acceptable configuration is found, the results are documented. For datasets, datasheets are created describing provenance and limitations; model cards present group-specific metrics and safe-use recommendations for models. Such documents are already hailed as a “selection tool” for AI transparency, and their use is piloted by large corporations and industry consortia [28]. A standardized card greatly simplifies internal reviews and regulator interactions: all key assumptions and tests are collected in one place.

The final stage is deployment and continuous online monitoring. Uber’s practical experience showed that without automated tracking of data shifts and spikes in group-specific error rates, incorrect decisions accumulate unnoticed until reaching a crisis threshold [29]. Thus, the fight against bias transitions from one-off initiatives to ongoing operations: the model, documentation, and monitoring form a unified control chain in which a failure at any link is quickly detected and remedied.

Thus, algorithmic bias permeates all stages of model development—from unevenly represented data and distorted metrics to architectural decisions and feedback loops in production—and requires a comprehensive approach. On one hand, at the level of regulatory governance (from the EU AI Act to national frameworks and ISO standards), mandatory audits, transparency requirements, and accountability measures have already been established; on the other, technical methods (pre-, in-, and post-processing, causal and multi-objective optimization) enable minimization of imbalances both during development and after deployment. Finally, introducing systematic checks, “harm maps,” datasheets, and model cards transforms the struggle against bias from a mere declaration into an ingrained process that ensures reproducibility and accountability. In the conclusion, we will articulate key recommendations for creating truly fair and reliable predictive models.

CONCLUSION

In conclusion, it has been demonstrated that algorithmic bias is a complex issue permeating every stage of a model’s lifecycle: from data collection and annotation through the selection of target metrics, architectural configuration, and deployment in a production environment. Sources of bias may include historical imbalances in the data and annotation noise that become entrenched during training, as well as improperly chosen proxy variables and metrics that fail to account for the actual needs of protected groups. Moreover, even a correctly trained model is susceptible to feedback-loop effects in production, which amplify the initial bias in the absence of continuous monitoring.

Achieving fairness requires diverse technical techniques: pre-, in-, and post-processing methods, each addressing a specific subtask. Pre-processing reduces data skew; embedding fairness constraints into the loss function enables explicit consideration of equity requirements during training; and post-processing provides a "black-box" mechanism for balancing errors when access to the model's internal parameters is limited. However, none of these approaches offers a universal solution: a trade-off between accuracy and fairness is inevitable, and the specific business and social context must determine the optimal balance.

Equally important is the incorporation of auditing, documentation, and continuous control processes: from preliminary dataset profiling and metric selection to the publication of datasheets and model cards that record assumptions and test outcomes for different groups. Only a formalized, reproducible process will allow regulators and internal auditors to verify compliance with bias-mitigation obligations and enable organizations to respond promptly to emerging deviations in fairness metrics.

Finally, international, regional, and sectoral regulatory frameworks establish minimal requirements and create a "compliance ladder" ranging from NIST's voluntary recommendations to the mandatory audits under the EU AI Act. This evolutionary structure reduces fragmentation and facilitates the shift from ethical declarations to measurable, verifiable commitments.

Thus, an effective strategy for combating algorithmic bias must integrate technical mitigation methods, auditing and documentation processes, continuous monitoring, and regulatory compliance mechanisms. These elements will ensure the reliability and fairness of predictive models over the long term.

REFERENCES

R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, vol. 1270, no. 1270, Mar. 2022, doi: https://doi.org/10.6028/nist.sp.1270.

A. Jonker and J. Rogers, "What is algorithmic bias?" *IBM*, Sep. 20, 2024. https://www.ibm.com/think/topics/algorithmic-bias (accessed Apr. 18, 2025).

A. Davison, "AI ethics tools," *IBM*, Sep. 03, 2024. https://www.ibm.com/think/insights/ai-ethics-tools (accessed Apr. 19, 2025).

J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed Apr. 20, 2025).

T. Devries, I. Misra, and C. Wang, "Does Object Recognition Work for Everyone?," The CVPR. Accessed: Apr. 21, 2025. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/papers/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.pdf

J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research*, vol. 81, no. 1, pp. 1–15, 2018, Accessed: Apr. 22, 2025. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, Accessed: Apr. 03, 2025. [Online]. Available: https://www.ftc.gov/system/files/documents/public_events/1548288/privacycon-2020-ziad_obermeyer.pdf

M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *Arxiv*, Oct. 07, 2016. https://arxiv.org/abs/1610.02413v1 (accessed Apr. 23, 2025).

D. Ensign, S. Friedler, S. Neville, C. Scheidegger, S. Venkatasubramanian, and C. Wilson, "Runaway Feedback Loops in Predictive Policing," *Proceedings of Machine Learning Research*, vol. 81, 2018, Accessed: Apr. 23, 2025. [Online]. Available: https://proceedings.mlr.press/v81/ensign18a/ensign18a.pdf

European Parliament, *P9_TA(2024)0138 Artificial Intelligence Act*. 2024. Accessed: Apr. 23, 2025. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

["Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," *NIST*, 2024, doi: https://doi.org/10.6028/nist.ai.600-1.

"Algorithmic Impact Assessment Tool," *The Government of Canada*, May 30, 2024. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html (accessed Apr. 24, 2025).

"Guidance on AI and data protection," *ICO*, Jun. 13, 2023. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ (accessed Apr. 24, 2025).

"Model AI Governance Framework for Generative AI," *AI Verify Foundation*, May 30, 2024. https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf (accessed Apr. 25, 2025).

J. Rusu, "AI Update," FCA. Accessed: Apr. 24, 2025. [Online]. Available: https://www.fca.org.uk/publication/corporate/ai-update.pdf

R. P. Grubenmann, "ISO/IEC 42001: The latest AI management system standard," *KPMG*, 2024. https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html (accessed Apr. 24, 2025).

OECD, "AI Principles," *OECD*, 2024. https://www.oecd.org/en/topics/ai-principles.html (accessed Apr. 25, 2025).

H. Mahmoudian, "Reweighing the Adult Dataset to Make it 'Discrimination-Free,'" *Medium*, Apr. 14, 2020. https://medium.com/data-science/reweighing-the-adult-dataset-to-make-it-discrimination-free-44668c9379e8 (accessed Apr. 26, 2025).

Y. Park *et al.*, "Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression," *JAMA Network Open*, vol. 4, no. 4, p. e213909, Apr. 2021, doi: https://doi.org/10.1001/jamanetworkopen.2021.3909.

P. Awasthi, M. Kleindessner, and J. Morgenstern, "Equalized odds postprocessing under imperfect group information," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, PMLR, 2020. Accessed: Apr. 28, 2025. [Online]. Available: https://proceedings.mlr.press/v108/awasthi20a/awasthi20a.pdf

"Understand and mitigate bias in ML models," *AI Fairness 360*. https://ai-fairness-360.org/ (accessed Apr. 29, 2025).

B. Hsu, R. Mazumder, P. Nandy, and K. Basu, "Pushing the limits of fairness impossibility: Who's the fairest of them all?" *36th Conference on Neural Information Processing Systems*, 2022, Accessed: May 18, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/d3222559698f41247261b7a6c2bbaedc-Paper-Conference.pdf

M. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017, Accessed: May 04, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

X. Han, T. Chen, K. Zhou, Z. Jiang, Z. Wang, and X. Hu, “You Only Debias Once: Towards Flexible Accuracy-Fairness Trade-offs at Inference Time,” *Arxive*, Mar. 10, 2025. https://arxiv.org/pdf/2503.07066 (accessed May 06, 2025).

[25] V. Prasad, “AI Algorithm Audits: Key Control Considerations,” *ISACA*, Aug. 02, 2024. https://www.isaca.org/resources/news-and-trends/industry-news/2024/ai-algorithm-audits-key-control-considerations (accessed May 08, 2025).

H. Dhaduk, “State of Generative AI in 2024,” *Simform*, Apr. 02, 2024. https://www.simform.com/blog/the-state-of-generative-ai/ (accessed May 09, 2025).

“AI Risk Management Framework,” *NIST*, Jan. 2023, doi: https://doi.org/10.6028/nist.ai.100-1.

“Datasheets for Datasets: Impact and Adoption Across Academic and Industry Sectors,” *Hackernoon*, Jun. 11, 2024. https://hackernoon.com/datasheets-for-datasets-impact-and-adoption-across-academic-and-industry-sectors (accessed May 12, 2025).

J. Le, “Datacast Episode 67: Model Observability, Ai Ethics, And Ml Infrastructure Ecosystem With Aparna Dhinakaran,” *James Le*, Jun. 28, 2021. https://jameskle.com/writes/aparna-dhinakaran (accessed May 18, 2025).