Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Algorithmic Bias in AI Systems: Examining Its Impact on Fairness and Decision-Making

Dharma Patel

dharmapatel016@gmail.com

Marwadi University https://orcid.org/0009-0003-6629-1170

Systematic Review

Keywords: algorithmic bias, fairness in AI, machine learning, automated decision-making, AI regulation, discriminatory algorithms, AI ethics

Posted Date: June 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10083731/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Algorithmic decision-making is no longer a theoretical concern: it operates across healthcare, criminal justice, employment, and financial services, and the disparities it produces are well documented. This paper examines how bias enters machine learning pipelines, tracing it to skewed training data, poorly chosen proxy variables, and the design of optimization objectives. Four case studies provide the empirical basis: Obermeyer et al.'s analysis of a commercial healthcare risk-scoring tool [3], Buolamwini and Gebru's audit of commercial facial analysis systems [4], ProPublica's investigation of the COMPAS recidivism tool [5], and the abandonment of Amazon's automated hiring system [6]. Each case shows that bias in AI is neither accidental nor self-correcting; it follows from design choices that encode existing social inequalities into automated outputs. The paper also examines competing formal definitions of fairness, showing that these definitions are mathematically incompatible and that no single algorithm can satisfy all of them simultaneously [8]. Technical mitigation strategies [12] are reviewed alongside the regulatory frameworks established by the EU AI Act [15], New York City's Local Law 144 [17], and related governance developments. The paper concludes that technical fixes alone are not sufficient and that durable fairness requires institutional accountability, representative data governance, and enforceable legal standards.

IMPACT STATEMENT

Artificial intelligence systems are increasingly used to support decisions in areas such as medical care, criminal justice, employment, and finance. As these systems influence opportunities, resources, and results for individuals, concerns about algorithmic bias have become an important and significant difficulty for responsible AI deployment. This paper contributes to the ongoing discussion by examining how bias emerges throughout the machine learning lifecycle and how it influences fairness in automated decision-making.

The significance of this work lies in bringing together evidence from broadly cited real-world instances that demonstrate the pragmatic results of biased AI systems. By analyzing failures in medical care risk assessment, facial recognition, criminal justice prediction, and automated hiring, the paper shows that algorithmic bias is frequently rooted in data collection practices, model design choices, and organizational decision-making procedures. The study furthermore highlights the limitations of existing fairness metrics, highlighting that distinct definitions of fairness can conflict with one another and may need tough and arduous trade-offs.

Beyond identifying issues, the paper reviews present and existing technical reduction techniques and emerging regulatory responses. The results suggest that enhancing fairness cannot depend entirely on algorithmic modifications. Effective management necessitates clear evaluation, representative data practices, organizational responsibility, and enforceable legal safeguards. By integrating technical, moral, and regulatory viewpoints, this paper offers a complete and thorough overview of algorithmic bias and provides insights that can support researchers, practitioners, policymakers, and organizations seeking to develop more equitable AI systems.

I. INTRODUCTION

A recidivism prediction tool that assigns a Black defendant a higher risk score than a white defendant who later reoffends is not producing random noise [5]. A healthcare algorithm that routes fewer resources to Black patients despite equivalent or greater medical need is not malfunctioning [3]. A facial recognition system that misidentifies a dark-skinned woman while correctly classifying a light-skinned man is not an outlier [4]. Each of these outcomes has been documented in peer-reviewed research and investigative journalism. Together they describe a consistent pattern: AI systems trained on historical data reproduce historical inequalities, and in some cases widen them.

Research on algorithmic bias has grown substantially since the mid-2010s. Computer scientists, lawyers, sociologists, and philosophers have reached a shared conclusion: automated systems described as objective frequently produce discriminatory outcomes, even without any discriminatory intent [1], [2]. That convergence has produced an extensive technical literature on fairness-aware machine learning and a growing legal literature on whether existing anti-discrimination frameworks apply to algorithmic outputs.

The shift from isolated complaints to a coherent research agenda reflects three developments occurring together. First, machine learning moved from academic benchmarks into consequential infrastructure: lending, hiring, policing, and clinical triage now rely routinely on statistical predictions rather than solely on human judgment. Second, journalists and academic auditors gained the technical means to interrogate proprietary systems from the outside, producing the disaggregated performance evidence that motivates much of this review. Third, legislators in multiple jurisdictions began treating biased automated decisions as a regulatory problem rather than a reputational one, creating obligations that did not exist a decade earlier. Each of these developments is traceable in the evidence base on which this review draws: a literature that spans peer-reviewed computer science, investigative journalism, and primary legal text in roughly the proportions shown in Fig. 1.

AI & Society has long situated these concerns within their broader cultural and political context, attending to how computational tools interact with social structures, institutional power, and the conditions under which knowledge is produced. This paper contributes to that tradition by examining the mechanisms of algorithmic bias with enough technical detail to be actionable, while maintaining the social-scientific framing that separates questions of fairness from questions of accuracy.

The review is organized to serve three audiences without privileging any one of them: computer scientists seeking a consolidated account of where formal fairness work has and has not resolved practical disputes; legal and policy scholars seeking a technically grounded summary of what current regulation does and does not require; and practitioners responsible for building or procuring AI systems who need to know, concretely, where in a development pipeline bias is most likely to enter and what evidence would reveal it. Each subsequent section returns to this three-way audience explicitly where the technical, legal, and practical threads diverge.

Section II defines algorithmic bias. Section III traces the pathways through which bias enters AI systems, including the data-representation imbalances illustrated in Fig. 2. Section IV presents four case studies

substantiated with quantitative evidence (Figs. 3–5) and situates them within a broader timeline of documented incidents and regulatory responses (Fig. 6). Section V examines competing formal definitions of fairness and the mathematical reasons they cannot all be satisfied simultaneously. Section VI reviews technical mitigation strategies. Section VII surveys the regulatory landscape, including the phased schedule by which the European Union's AI Act enters into force (Fig. 7). Section VIII identifies unresolved challenges, including the bias risks specific to large generative models. Section IX concludes.

II. DEFINING ALGORITHMIC BIAS

[1] define bias in machine learning as a systematic error in decision-making that produces unfair outcomes. The definition requires two conditions: the error must be systematic rather than occasional, and its consequences must be measurable against some standard of fairness. The second condition is immediately complicated by the fact that fairness is not a single property but a family of related and sometimes conflicting ones.

For this paper, algorithmic bias refers to machine learning systems whose outputs systematically disadvantage individuals or groups on the basis of protected characteristics: race, gender, age, disability, or socioeconomic status. The disadvantage may be direct, where the protected characteristic is used as an input or closely proxied, or indirect, where facially neutral inputs correlate strongly enough with a protected characteristic to produce disparate impact without disparate treatment [1].

Several distinctions matter here. Algorithmic bias differs from random error. Random errors distribute unpredictably across a population, while systematic errors consistently harm identifiable groups. It differs from model inaccuracy: a model can be inaccurate without being socially biased, and a highly accurate model can still discriminate if its errors concentrate among particular subgroups. It also differs from intentional discrimination, though it can produce comparable or worse outcomes. Most documented cases involve systems that were not designed to discriminate but whose training data, feature selection, or objective functions produced discrimination regardless [2].

[2] distinguishes three categories: data bias, arising from unrepresentative or historically contaminated training sets; algorithm bias, arising from design choices including the optimization objective, feature selection, and class-imbalance handling; and human decision bias, arising when downstream users over-rely on system outputs, a pattern sometimes called automation bias.

Each category has a distinct intervention point. Data bias is addressed by changing what a system learns from: collecting more representative samples, auditing historical records for the social conditions that produced them, or reweighting underrepresented groups before training. Algorithm bias is addressed by changing how a system learns: the loss function it minimizes, the constraints placed on that minimization, and the way subgroup performance is monitored during development. Human decision bias is addressed by changing how a system's output is used: limiting the weight given to an automated score, training decision-makers to recognize automation bias, and preserving a genuine option to override the system. The case studies in Section IV illustrate failures at each of these points: training data in the healthcare and facial-

recognition cases, objective design in the criminal-justice case, and organizational reliance on a flawed score in the employment case.

It is useful to state the problem more formally. Let X denote the observable features available to a model, A a protected attribute such as race or sex, Y the true outcome of interest, and $\hat{y} = f(X)$ the model's prediction. A model is unbiased with respect to A, in the weakest sense relevant to this review, if no function of $\hat{y}$ depends on A once X is held fixed in a way that violates an agreed fairness criterion. The qualification matters: A is rarely available to the model directly, but X frequently contains variables that are statistically dependent on A, so that $\hat{y}$ can depend on A indirectly even when A is excluded as an input. Section III traces how this indirect dependence arises in practice, and Section V returns to the question of which fairness criterion the dependence should be judged against.

The remainder of this review follows the structure these distinctions imply. Section III examines the data-bias and algorithm-bias categories in mechanical detail; Section IV grounds both categories, together with human decision bias, in four documented cases; Sections V and VI return to the normative and technical questions those cases raise; and Sections VII and VIII examine, respectively, how legal systems and open research questions have responded to the same underlying problem.

III. HOW BIAS ENTERS AI SYSTEMS

A. Historical Bias in Training Data

Machine learning systems learn from data. If that data encodes historical discrimination, the system will reproduce it [1]. This is not a failure of learning. The system has accurately captured the statistical regularities in its training set; the problem is that those regularities reflect unjust social arrangements rather than facts about the world.

[1] document this across several domains. Credit scoring datasets carry the imprint of lending practices that excluded minority applicants. Criminal justice datasets carry the imprint of policing practices that disproportionately surveilled and arrested minority communities. Medical datasets carry the imprint of disparities in healthcare access and utilization. A model trained to predict outcomes from historical data will reproduce the inequities that shaped that data.

Representation bias is a related problem. When training data comes predominantly from one demographic group, the model performs well on that group and poorly on others. [4] found that the benchmark datasets used to train commercial facial analysis systems contained between 79.6% and 86.2% lighter-skinned subjects, which accounts directly for the accuracy gaps their audit found. Figure 2 shows this imbalance for the two benchmarks Buolamwini and Gebru audited: IJB-A, compiled largely from official photographs of senior government figures, and Audience, compiled from social-media images; both are dominated by lighter-skinned subjects by a margin of roughly four to one or greater.

The skew documented in Fig. 2 is not incidental to how these datasets were assembled. IJB-A was built substantially from photographs of public officials, a population that, at the time of collection, was disproportionately male and light-skinned in the jurisdictions sampled; Audience drew on publicly shared

social-media images, which are not demographically representative of the global population of faces a deployed system will encounter. Neither dataset was built with skin-type balance as a design criterion, and neither vendor evaluated by Buolamwini and Gebru reported subgroup accuracy prior to the audit, so the imbalance went undetected until an external party measured it directly [4].

This pattern is not specific to face data. Hospital records used to train clinical prediction models are denser for patients who interact more often with the health system, which itself correlates with insurance status and access; criminal-history databases are denser for neighborhoods subject to more intensive policing, independent of the underlying rate of offending; resume corpora used to train hiring tools are denser for the demographic groups historically over-represented in a given industry. In every case, the density of data is itself a record of unequal access to institutions, not a neutral sampling artifact, so correcting for representation requires more than collecting additional examples; it requires understanding why the original collection was uneven.

B. Proxy Variables and Label Choice Bias

Even without protected characteristics as explicit inputs, models can produce discriminatory outcomes through proxy variables: features that correlate strongly with protected characteristics and effectively substitute for them. In the United States, zip code correlates with race due to residential segregation; prior arrest record correlates with race due to differential policing; healthcare cost correlates with race due to structural disparities in insurance and care access [3].

The proxy-variable mechanism has a long history in anti-discrimination law that predates machine learning. [25] trace how facially neutral criteria long used in lending, housing, and employment, such as zip code or credit history, can produce disparate impact on protected groups even when no discriminatory intent is present, and how United States disparate-impact doctrine developed specifically to reach this kind of harm. Machine learning does not introduce a new legal category of problem; it multiplies the number and subtlety of proxies a system can exploit, often without any human deliberately selecting them. A linear regression with a handful of hand-chosen features makes its proxies relatively easy to inspect. A gradient-boosted model or neural network trained on hundreds of correlated features can encode a proxy relationship implicitly, through interactions among variables that no individual feature reveals on its own.

[3] make this mechanism visible in their analysis of a commercial healthcare algorithm. The tool was designed to identify patients who would benefit from care management programs. It predicted future healthcare costs rather than health status directly, on the assumption that cost and health correlate. They do, but imperfectly: because Black patients in the United States have historically used less healthcare than white patients with equivalent needs, due to financial barriers, lack of insurance, and distrust of medical institutions, the algorithm systematically underestimated Black patients' health burden. Choosing healthcare cost as a proxy, rather than a direct measure of illness, introduced racial bias without any reference to race.

[3] estimated that correcting this disparity would raise the share of Black patients receiving additional help from 17.7% to 46.5%, a change driven entirely by adjusting how health needs were measured.

C. Feature Selection and Model Architecture

Bias also enters through model design decisions: which features to include, how to handle missing data, which loss function to minimize, how to weight performance across subgroups. These choices appear technical but carry normative content. They encode assumptions about what matters, what can be traded off, and whose errors are acceptable [1], [2].

A model that optimizes for aggregate accuracy will perform best on the majority group, because minimizing overall error means attending most closely to the largest population segments. Without explicit attention to subgroup performance, high aggregate accuracy can coexist with substantially unequal treatment across demographic groups.

Consider a binary classifier trained on a population in which the minority subgroup makes up fifteen percent of the training examples. A loss function that simply minimizes the overall error rate will, all else equal, fit the majority subgroup's decision boundary more closely, because eighty-five percent of the gradient signal during training originates there. The model can reach high aggregate accuracy while the minority subgroup's error rate remains substantially higher, and because aggregate accuracy is the metric most commonly reported, the disparity can persist through several rounds of model development without being noticed. The same dynamic affects threshold selection: a single decision threshold applied uniformly across groups optimizes overall accuracy but not necessarily equalized error rates, a tension formalized in Section V.

D. Feedback Loops

Algorithmic systems do not only predict conditions; they also shape them. A predictive policing tool that recommends increased surveillance in a neighborhood generates more arrests there, which then feeds back into training data confirming the original prediction. A credit scoring system that denies loans in lower-income areas reduces those residents' access to capital, which affects their financial histories and future scores. [7] described this in the context of predictive policing as a feedback loop capable of producing runaway over-policing of minority communities regardless of the underlying crime distribution.

Feedback loops are particularly difficult to detect with conventional validation, because each iteration of the loop is internally consistent: the model's predictions and the data used to retrain it agree with one another by construction, even as both drift further from the underlying reality the system was meant to measure. [7] modeled this dynamic for predictive policing and found that even a small initial disparity in recorded crime rates between neighborhoods can compound, over repeated deployment cycles, into an enforcement gap far larger than any difference in actual offending. The same structure recurs wherever a system's outputs influence the conditions it later measures: a hiring tool that down-ranks candidates from a particular university reduces that university's representation in the company's future training data; a credit model that denies loans in a neighborhood reduces recorded repayment histories there, which the next model version then treats as evidence of risk.

IV. CASE STUDIES

A. Healthcare: The Obermeyer et al. Algorithm Study

In 2019, Obermeyer, Powers, Vogeli, and Mullainathan published a study in *Science* documenting racial bias in a commercial healthcare algorithm used to manage the care of roughly 200 million patients annually in the United States [3]. The algorithm assigned risk scores that determined referrals to care management programs. The researchers compared the health status of Black and white patients assigned the same risk score at a large academic medical center and found that Black patients at any given score were substantially sicker than white patients at the same score, as measured by the number of active chronic conditions.

The mechanism, described in Section III-B, was the use of healthcare cost as a proxy for health status. Because structural inequalities meant less was spent on Black patients with equivalent health needs, the algorithm translated that spending gap into a risk-score gap. Retraining the algorithm to use direct health measures rather than cost would have nearly tripled the proportion of Black patients receiving additional care [3].

Figure 3 summarizes the headline result of the correction Obermeyer and colleagues proposed: replacing the cost-based label with a label constructed from a combination of health prediction targets nearly tripled the share of Black patients flagged for additional care management, from 17.7 percent to 46.5 percent, without changing the algorithm's architecture or the data available to it. The entire disparity, in other words, traced to a single labeling decision made early in the system's design, long before any model was trained.

The case is instructive because the algorithm performed well by standard accuracy metrics: it predicted healthcare costs accurately for both groups. The bias was undetectable until researchers asked whether equivalent scores corresponded to equivalent levels of need across racial groups. They did not.

The study is also notable for what it did not find. The algorithm was not trained on race, did not have access to race as a feature, and showed no evidence of having learned a racial classification internally. Auditing it for the presence of a protected attribute, the conventional first step in a fairness review, would have found nothing. The bias was visible only once researchers compared outcomes across groups conditional on the score, which requires the kind of subgroup-disaggregated evaluation that, as Section VI discusses, is not yet standard practice for proprietary clinical software. The vendor's own validation reported the cost-prediction accuracy that motivated the original design choice and did not report this comparison, which is precisely the comparison that revealed the problem.

B. Criminal Justice: The COMPAS Recidivism Tool

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a risk assessment tool used by courts and probation officers across the United States to estimate the probability of reoffending. In 2016, ProPublica analyzed the tool using data on more than 7,000 individuals arrested in Broward County, Florida [5]. The analysis found that Black defendants who did not go on to reoffend were nearly twice as likely as white defendants to be falsely flagged as high risk, while white defendants who did reoffend were more likely than Black defendants to be incorrectly classified as low risk.

Figure 4 reports the two error rates at the center of the dispute: a false positive rate of 44.9 percent for Black defendants against 23.5 percent for white defendants, and a false negative rate of 28.0 percent for

Black defendants against 47.7 percent for white defendants. The asymmetry runs in opposite directions for the two error types, which is exactly the pattern Chouldechova's proof predicts whenever the underlying base rate of reoffending differs between the two groups being compared, as it did in the Broward County sample, where the two-year reoffending rate was higher among Black defendants than white defendants in this dataset. Neither ProPublica's nor Northpointe's numbers were in error; they were measuring different, individually coherent, jointly incompatible properties of the same score.

Northpointe, the tool's developer, responded that COMPAS was calibrated: the probability of actually reoffending, given a high-risk score, was approximately 60% for both Black and white defendants. Both claims are correct. [8] proved this formally, showing that when the base rate of the predicted outcome differs between groups, as it does in virtually all real applications, no algorithm can simultaneously satisfy calibration, equal false positive rates, and equal false negative rates. The ProPublica-Northpointe dispute was therefore not a factual disagreement; it was a disagreement about which definition of fairness should govern.

The use of COMPAS in sentencing has drawn legal attention. In *State v. Loomis* [23], the Wisconsin Supreme Court held that using an algorithmic risk score in sentencing did not violate due process, provided the score was not the sole determinative factor. Critics have argued that this holding does not address the opacity of proprietary algorithms or the structural inequalities encoded in their inputs.

The Wisconsin court's reasoning in *Loomis* rested on the fact that the trial judge had access to other information besides the COMPAS score and was not required to follow it, which the court treated as sufficient procedural protection. That reasoning does not resolve the deeper problem raised by the case: the defendant was not permitted to examine the algorithm's internal weighting because Northpointe asserted trade-secret protection over it, so neither the defendant nor, on appeal, the reviewing court could verify how the score had been produced or whether the specific risk factors used were themselves correlated with race. Subsequent state and federal cases have cited *Loomis* for the narrow proposition it decided without resolving the broader question of what disclosure a defendant is owed when a proprietary score contributes to a liberty-affecting decision.

C. Facial Recognition: The Gender Shades Study

[4] audited three commercial facial analysis systems, measuring performance across gender and skin tone using the Fitzpatrick Skin Type classification. The error rate for lighter-skinned male subjects was as low as 0.8%. For darker-skinned female subjects, the error rate reached 34.7%, a gap of more than 43 percentage points. The worst-performing systems on dark-skinned women made errors over 34 times more often than on light-skinned men.

Figure 5 breaks the result down by vendor and by the four intersectional subgroups Buolamwini and Gebru defined. IBM's system showed the widest spread, misclassifying 34.7 percent of darker-skinned female faces against 0.3 percent of lighter-skinned male faces, a 34.4 percentage-point gap; Microsoft showed a comparable 20.8-point gap between the same two subgroups. Darker-skinned female faces were the worst-served subgroup for every vendor, but the ranking of the remaining three subgroups was not consistent

across systems: Face++, unlike the other two vendors, performed marginally better on darker-skinned male faces than on lighter-skinned female faces, a reminder that gender and skin-type effects do not combine in a uniform or strictly additive way, and that auditing only the marginal distributions, gender alone or skin tone alone, would have missed exactly the interaction the intersectional breakdown was designed to reveal.

The explanation was the training data: between 79.6% and 86.2% of subjects in the benchmark datasets were lighter-skinned. Systems trained predominantly on lighter faces perform worse on darker faces. The problem lay not in the algorithms but in the data used to train them and in the absence of any requirement to test performance across demographic subgroups before deployment [4].

A larger-scale follow-up using government-operated face recognition algorithms reached a more qualified conclusion. The U.S. National Institute of Standards and Technology evaluated upwards of one hundred algorithms submitted by vendors worldwide for one-to-one verification and one-to-many identification tasks and found demographic effects on false match and false non-match rates across most algorithms tested, but with a wide spread: the most accurate algorithms in the evaluation showed demographic differences close to the limits of measurement, while many lower-performing algorithms showed differences of an order of magnitude or more [26]. The finding complicates a simple narrative in which facial recognition is uniformly biased: it indicates instead that demographic disparity and overall accuracy are correlated, so that the regulatory and procurement question is not only whether a system has been tested for subgroup disparities but which algorithm, among many available, a deploying agency selects.

The study had practical effects. IBM updated its facial recognition service and later discontinued the program in 2020, citing concerns about racial bias and mass surveillance. The study also showed what independent auditing can find: the disparities it documented had not been disclosed by the companies and could not be inferred from aggregate performance figures [13].

D. Employment: Amazon's Automated Hiring Tool

In 2018, Reuters reported that Amazon had abandoned an automated hiring tool it had been developing since 2014 [6]. The system was trained to rank job applicants using ten years of previously submitted resumes, a dataset that was predominantly male, reflecting the gender imbalance in the technology industry.

By 2015, Amazon's engineers had found that the system penalized resumes containing the word "women's" (as in "women's chess club") and gave lower scores to graduates of two all-women's colleges [6]. The company adjusted the system to neutralize those specific terms but could not confirm that it was gender-neutral in other respects and eventually scrapped it. As [18] noted, tools of this kind do not remove human bias from hiring; they encode it into the selection process.

The Amazon case is frequently cited because the company itself, rather than an outside auditor, identified the problem and chose not to deploy the system; it is one of the few documented instances in which an organization's internal testing caught a bias problem before public deployment rather than after. That outcome should not be read as evidence that internal testing is generally sufficient. Amazon engineers found the gendered penalty only because they happened to test the system's behavior on specific, manually

chosen terms; a less conspicuous proxy, one not tied to an obviously gendered word, could plausibly have passed the same review undetected. The episode illustrates a more general point about hiring algorithms: a resume corpus reflects not only who applied but who was hired, promoted, and retained under decades of prior practice, so a model trained to predict "successful" past hires risks learning to reproduce whichever pattern of advantage already prevailed, across gender, race, age, or institutional pedigree, even when no single feature is intended to capture any of them.

Figure 6 situates these four cases on a common timeline alongside the regulatory and policy milestones discussed in Section VII. Read together, the two curves show documented bias incidents accumulating from 2016 through 2019, the period in which most of the foundational audits in this review were published, followed by a regulatory response that did not begin in earnest until 2021 and accelerated sharply only from 2023 onward. The roughly five-year lag between documentation and regulation recurs across all four cases: each was identified by researchers or journalists working independently of any legal requirement to test for bias, and in each case the organization responsible changed its practice, if at all, only after the finding became public.

V. THE PROBLEM OF DEFINING FAIRNESS

A core finding of the fairness literature is that many intuitive fairness criteria are mathematically incompatible. [8] proved that when the prevalence of a predicted outcome differs between groups, no algorithm can simultaneously satisfy calibration, equal false positive rates, and equal false negative rates. Because prevalence differs between groups in most real applications, algorithm designers must choose which fairness criterion to satisfy. That choice is not technical; it is normative.

A. Group Fairness Criteria

Group fairness criteria evaluate a model's outputs in aggregate across two or more groups defined by a protected attribute, without examining any individual prediction in isolation. Statistical parity, also called demographic parity, requires that the proportion of individuals receiving a positive outcome, a loan, an interview, a low-risk classification, be equal across groups: $P(\hat{y}=1 \mid A = a) = P(\hat{y}=1 \mid A = b)$ for any two groups a and b. The criterion is easy to audit and easy to explain, but it makes no reference to the true outcome Y, so a model can satisfy statistical parity by being equally inaccurate, or equally generous, toward both groups regardless of their actual qualification.

Equalized odds [9] corrects for this by conditioning on the true outcome: it requires equal true positive rates and equal false positive rates across groups, $P(\hat{y}=1 \mid Y = 1, A = a) = P(\hat{y}=1 \mid Y = 1, A = b)$, and the corresponding equality for $Y = 0$. This is the criterion ProPublica's COMPAS analysis implicitly applied when it compared false positive and false negative rates across racial groups (Section IV-B).

Calibration takes a third benchmark: the predicted score itself. A score is calibrated across groups if, for any score value s, the proportion of individuals who actually experience the outcome is the same regardless of group membership, $P(Y = 1 \mid \hat{y}=s, A = a) = P(Y = 1 \mid \hat{y}=s, A = b)$. Calibration is the criterion Northpointe invoked

in defending COMPAS, and it is a natural requirement wherever a score is meant to be interpreted as a probability rather than merely a ranking.

B. Individual and Counterfactual Fairness

Group criteria can be satisfied while still treating two similar individuals very differently, provided the discrepancies wash out in aggregate. [27] proposed individual fairness as a response: similar individuals should receive similar predictions, formalized as a Lipschitz condition on the model with respect to a task-specific similarity metric, so that the distance between two individuals' predictions is bounded by their distance under that metric. The approach shifts the central difficulty from choosing among aggregate statistics to specifying the similarity metric itself, a metric that must encode, for a given decision, which differences between people are relevant and which are not, a judgment that [27] note is itself a normative rather than purely statistical one.

Counterfactual fairness asks a related but distinct question: would this individual have received the same prediction had their protected attribute been different, holding fixed everything causally independent of that attribute? The criterion requires a causal model of how features relate to the protected attribute and to one another, not merely a statistical association, which makes it more demanding to implement than the group criteria above but arguably closer to the ordinary moral intuition that a decision should not depend on a person's race or sex. In the COMPAS context, a counterfactual fairness analysis would ask whether a defendant's risk score would change if their race were different but their actual history of offending, employment, and residence remained the same, a question the available data cannot answer directly because race correlates with many of the other features being held fixed.

C. The Impossibility of Simultaneous Satisfaction

These criteria are not merely different; under realistic conditions they are mutually exclusive. [10] and, independently, [8] proved that calibration and equal false positive and false negative rates cannot all hold across two groups unless either the model is a perfect predictor or the base rate of the outcome, $P(Y = 1 \mid A = a)$, is identical across the groups being compared. The proof is a direct consequence of how the relevant error rates are related through Bayes' theorem once the base rate is fixed; it does not depend on any property of a specific algorithm or dataset, which is why it applies equally to COMPAS, to the healthcare algorithm in Section IV-A, and to any other binary classifier deployed across groups with different outcome prevalence.

The practical consequence is unavoidable: whenever the base rates genuinely differ, which they do in recidivism, in healthcare cost, and in most other applications discussed in this review, a designer must choose which property to sacrifice. Optimizing for calibration accepts unequal error rates across groups; optimizing for equal error rates accepts that the same score will mean different things, in terms of actual risk, for different groups. [11] frame this choice as resting on a prior judgment about whether the observed difference in base rates reflects a genuine underlying difference, an interpretation that treats unequal calibration as the lesser harm, or whether it reflects the downstream effect of historical injustice operating through the very features used to measure the outcome, an interpretation that treats unequal error rates as

the lesser harm. The Obermeyer healthcare case (Section IV-A) is again instructive: the “base rate” difference in healthcare cost between Black and white patients was not a fact about underlying health needs but an artifact of unequal past spending, which is precisely why the appropriate response was judged to be changing the prediction target rather than accepting the trade-off as fixed.

D. Choosing Among Definitions in Practice

[1] catalogue more than twenty distinct fairness criteria proposed in the machine learning literature, considerably more than the five discussed above, reflecting both the genuine conceptual difficulty of the concept and the diversity of contexts in which AI systems operate. Fairness in medical triage, where a missed high-risk patient and an unnecessarily flagged low-risk patient carry very different costs, is not the same concept as fairness in advertising delivery, where the harms of unequal exposure are diffuse and the relevant comparison group is often unclear. Each context shapes which error matters more and to whom, which is why no single criterion has displaced the others despite a decade of technical work on the question.

In practice, the choice of criterion is frequently made implicitly, by whichever metric a development team happens to report, rather than explicitly, through a deliberate judgment about which kind of error the deployment context makes more costly. Section VI discusses tools that allow a chosen criterion to be enforced once selected; this section’s conclusion is that no tool can substitute for making the selection in the first place, and that the selection should be documented and justified rather than left as an implicit consequence of which open-source library a team adopts.

VI. TECHNICAL MITIGATION STRATEGIES

A. Pre-Processing Approaches

Pre-processing techniques address bias in the training data before a model is built. Methods include reweighting training examples, resampling to improve demographic balance, transforming features to remove correlation with protected attributes, and constructing more representative benchmark datasets. The AI Fairness 360 toolkit from IBM Research [12] implements a range of these methods and provides fairness metrics for measuring disparate impact before deployment.

Reweighting works by assigning each training example a weight inversely proportional to its group’s representation relative to the outcome it experienced, so that a gradient-based learner receives an equally strong corrective signal from a rare combination of group and outcome as from a common one. Resampling achieves a similar effect by duplicating or removing examples rather than reweighting them, which is computationally simpler but can discard information when applied aggressively to a severely imbalanced dataset. Feature transformation methods go further, altering the feature space itself, for instance by learning a representation that retains predictive information about the outcome while removing information that would allow the protected attribute to be reconstructed from the remaining features, a goal closely related to the adversarial techniques described in Section VI-B.

The Gender Shades audit directly prompted pre-processing investment. Several companies expanded their facial analysis training sets after its publication, increasing the proportion of darker-skinned subjects. Follow-up audits found meaningful reductions in accuracy gaps, confirming that the intervention worked, even if disparities were not fully eliminated [13].

B. In-Processing Approaches

In-processing techniques build fairness constraints into the training procedure itself. A classifier can be trained to minimize loss subject to a constraint, for example that the difference in false positive rates between demographic groups stays below a specified threshold. Adversarial debiasing adds a second network trained to predict the protected attribute from the classifier's outputs; the classifier is then trained to perform well on the primary task while making it difficult for the adversary to recover the protected attribute.

Constrained optimization approaches add the fairness criterion directly to the training objective, typically as a penalty term whose weight controls the trade-off between predictive accuracy and the chosen fairness measure. This allows a practitioner to trace out the achievable trade-off curve rather than accepting a single fixed point on it. Adversarial debiasing pursues the same goal through a different mechanism: a second network is trained concurrently to predict the protected attribute from the primary model's internal representation or output, and the primary model is penalized whenever the adversary succeeds, which pushes the primary model toward representations from which the protected attribute cannot easily be recovered. Both approaches require access to the protected attribute during training, even when it is excluded from the model's inputs at deployment, which raises the data-governance tension discussed in Section VII-A.

[9] proposed setting different decision thresholds for different demographic groups after training, such that equalized odds is achieved. The approach shows that accuracy and fairness are not always in conflict: accepting a modest reduction in aggregate accuracy can produce substantial improvements in fairness across groups.

C. Post-Processing Approaches

Post-processing techniques adjust model outputs after training to reduce disparate impact. They are useful when the model cannot be modified, as when a proprietary third-party system is in use. Calibration can be performed separately for each demographic group, so that predicted probabilities reflect actual outcomes equally across groups. Decision thresholds can be adjusted to equalize error rates.

Group-specific calibration recalibrates the model's score separately within each group so that a given score corresponds to the same empirical outcome rate across groups, directly targeting the calibration criterion defined in Section V-A; group-specific thresholds instead set a different cutoff for each group so that the resulting positive rate, true positive rate, or false positive rate is equalized, directly targeting whichever of the group criteria in Section V-A the deploying organization has selected. Because [10] and [8] establish that these two goals are jointly unsatisfiable when base rates differ, post-processing forces an explicit choice of one over the other, where pre- and in-processing approaches can sometimes obscure that the choice is

being made at all. Using a different threshold for different groups also raises a separate legal question, addressed further in Section VII-C: in some jurisdictions, deliberately adjusting a decision rule by protected group, even to improve fairness, can itself be challenged as a form of disparate treatment, leaving organizations to navigate between two different legal theories of discrimination that point in opposite directions.

Post-processing has real limits. Adjusting thresholds after training does not change the model or the data that produced it; it manages the effects of bias rather than the causes. When group membership is unavailable or legally prohibited as a variable, post-processing options narrow further [8].

D. Algorithmic Auditing

Independent third-party auditing is increasingly recognized as a necessary complement to internal testing. The Gender Shades study showed that companies had not identified, let alone disclosed, the accuracy gaps in their own systems [4]. [14] developed an audit framework modeled on established practices in accounting and safety engineering: defining objectives, gathering evidence, applying consistent evaluation criteria, and publishing results.

[14] structure an internal audit around five stages: scoping the system's intended use and affected populations, mapping the specific harms that could arise from that use, collecting evidence of the system's actual behavior across subgroups, evaluating that evidence against a predetermined threshold of acceptability agreed before the evidence was collected, and reporting the result in a form that a non-specialist reviewer, such as a company's legal or executive leadership, can act on. The ordering matters: evaluating against a threshold fixed in advance, rather than one chosen after seeing the results, is what distinguishes an audit from a post-hoc justification of whatever the system already does.

External audits face the additional difficulty of accessing a system at all. The Gender Shades and Actionable Auditing studies were possible because the systems audited were public application programming interfaces that could be queried directly with a constructed test set [4], [13]; a system that does not expose this kind of access, or that monitors and restricts automated querying, is far harder for an independent researcher to evaluate without the cooperation, or legal compulsion, of the organization that controls it.

Standardization efforts have begun to address the methodological gap. The International Organization for Standardization's ISO/IEC 42001 management-system standard and a growing set of sector-specific audit checklists give organizations a common vocabulary for describing what an audit covered, though none yet specifies the statistical thresholds an audit should apply, leaving that normative judgment, the same one Section V-D identifies as unavoidable, to each auditor or regulator individually. The result is that two audits described as compliant with the same standard can apply substantively different fairness criteria, which limits how directly their conclusions can be compared across organizations or industries.

Auditing faces practical barriers. Proprietary systems may be closed to outside researchers. Audit methodologies are not standardized. Companies may have legal or competitive reasons to resist disclosing findings. Closing these gaps has become a central focus of AI governance research and, increasingly, of regulation [14].

VII. THE REGULATORY LANDSCAPE

A. The European Union AI Act

The EU Artificial Intelligence Act [15] (Regulation (EU) 2024/1689), which entered into force in August 2024, is the most comprehensive AI-specific regulatory framework enacted to date. It uses a risk-based classification with four tiers: unacceptable risk (prohibited), high risk (subject to mandatory requirements), limited risk (transparency obligations), and minimal risk. AI systems used in employment, education, law enforcement, credit scoring, and access to essential services are classified as high risk.

High-risk systems must be trained on sufficiently representative datasets, produce technical documentation that allows post-hoc investigation, be monitored for bias after deployment, and include human oversight mechanisms. Article 10 specifically requires that training, validation, and testing datasets be examined for biases capable of producing discriminatory outputs. High-risk systems must pass a conformity assessment before deployment and be registered in an EU-wide database [15].

Figure 7 traces the cumulative entry into force of these obligations. Prohibited-practice rules and a baseline AI-literacy requirement applied first, from February 2025; obligations for general-purpose AI models and the governance bodies responsible for enforcement followed in August 2025; the bulk of the high-risk system requirements described above, together with the Article 50 transparency rules for AI-generated content, are scheduled to apply from August 2026; and a final tranche covering AI embedded in products already regulated under other EU product-safety law is scheduled for August 2027. As of this writing, the European Commission has proposed, through a Digital Omnibus package tabled in November 2025, to defer the August 2026 high-risk obligations for stand-alone Annex III systems to December 2027 and the obligations for product-embedded systems to August 2028, citing the unavailability of the harmonized technical standards that high-risk providers need in order to demonstrate conformity; that proposal had not been formally adopted at the time this review was prepared, so Fig. 7 shows the schedule as originally legislated.

Non-compliance with the high-risk obligations carries penalties of up to fifteen million euros or three percent of an undertaking's global annual turnover, whichever is higher, with steeper penalties of up to thirty-five million euros or seven percent reserved for violations of the prohibited-practices tier; the scale of these fines, well above the equivalent thresholds under the General Data Protection Regulation in percentage terms, signals that the Act's drafters intended bias-relevant obligations such as Article 10's data-governance requirement to be treated as a compliance priority rather than a best-practice recommendation.

The Act also addresses a practical tension between bias correction and data protection. Correcting for demographic bias requires demographic data, but collecting and using sensitive personal data is heavily restricted under the General Data Protection Regulation. Article 10(5) allows the processing of special-category data for bias monitoring and correction, subject to specified safeguards. As [16] has noted, the interaction between these provisions and the GDPR remains legally unsettled. [24] provides a detailed analysis of this relationship.

B. United States Regulatory Developments

The United States has no comprehensive federal AI regulation, but several targeted measures address algorithmic bias. New York City's Local Law 144 [17], which took effect in 2023, requires employers using automated employment decision tools to commission annual independent bias audits and to publish results. It defines a bias audit as an impartial evaluation for evidence of adverse impact based on sex/gender and race/ethnicity. Narrow in scope, it is nonetheless the first US law to mandate independent auditing of algorithmic hiring tools and has prompted similar proposals elsewhere.

Early implementation experience with Local Law 144 illustrates the gap that can open between a disclosure requirement and its enforcement. An academic study that sent 155 trained investigators to check the websites of 391 covered employers found that only eighteen had posted an audit report and only thirteen had posted the separate transparency notice the law also requires, a result the authors term "null compliance" because the law gives employers broad discretion to determine for themselves whether a given tool falls within its scope, so that the absence of a posted audit cannot be distinguished from a good-faith determination that no audit was required [29]. A subsequent audit by the New York State Comptroller's office reached a similar conclusion from the regulator's side: reviewing the same employers the city's own enforcement agency had checked, the Comptroller identified at least seventeen instances of potential non-compliance where the agency's review had identified only one, and found that the agency's enforcement relied almost entirely on complaints, of which it had received just two during the two-year audit period [30]. Taken together, the two findings suggest that a disclosure mandate without a well-resourced, proactive enforcement mechanism may establish a legal obligation without materially changing what most covered employers actually do.

The Biden administration's Executive Order 14110 on Safe, Secure, and Trustworthy AI (October 2023) directed more than fifty federal agencies to develop policies on algorithmic bias. The Trump administration revoked it in January 2025, and the federal direction on AI regulation is now less clear.

The Federal Trade Commission has indicated that discriminatory algorithmic outputs may be actionable under existing authority over unfair or deceptive practices. The Consumer Financial Protection Bureau has taken the position that lenders using AI in credit decisions remain subject to the Equal Credit Opportunity Act and must be able to explain credit denials, which creates compliance difficulties for opaque machine learning models.

C. The Limits of Legal Frameworks

Most anti-discrimination law was written with human decision-makers in mind and applies awkwardly to algorithmic systems. Disparate impact doctrine requires plaintiffs to identify a specific practice and demonstrate its adverse impact. When a machine learning model is a system of many interacting components, isolating the specific practice responsible for a disparate outcome is both technically and legally difficult. Proprietary systems may be shielded from discovery, preventing the level of examination a claim would require.

The discovery problem identified in Section IV-B recurs throughout this area of law. Anti-discrimination statutes generally presume that a plaintiff, once a claim is filed, can obtain the information needed to

evaluate the practice being challenged; trade-secret and intellectual-property protections asserted over proprietary models complicate that presumption in ways courts have not yet resolved consistently. Some jurisdictions have begun experimenting with structural responses, such as requiring a regulator rather than an individual plaintiff to hold audit access, precisely to avoid relitigating the discovery question in every individual case.

Coverage is also uneven. US regulation is largely sector-specific, leaving many applications outside its scope. The EU AI Act [15] is broader but applies only within the EU and to systems deployed to EU residents. Systems that would be prohibited or strictly regulated in one jurisdiction can be deployed in another with fewer constraints.

The jurisdictional patchwork extends well beyond the United States and the European Union. The United Kingdom has so far relied on its existing Equality Act and on sector regulators rather than AI-specific legislation; Canada's proposed Artificial Intelligence and Data Act, introduced in 2022 as part of a broader bill, did not survive the prorogation of Parliament in January 2025 and would need to be reintroduced before taking effect, leaving Canada without dedicated federal AI legislation as of this writing; and several jurisdictions in Asia and the Gulf have to date prioritized AI investment and infrastructure policy over bias-specific regulation. The practical consequence, returned to in Section VIII-D, is that a system found to produce discriminatory outputs in one jurisdiction can often continue to operate, with little modification, in another that has not yet adopted a comparable requirement.

One frequently proposed remedy borrows directly from environmental law: an algorithmic impact assessment, conducted and published before a high-stakes system is deployed, modeled on the environmental impact statements long required before major infrastructure projects. Canada has required a version of this for federally procured automated decision systems since 2019, and a small number of US states have adopted comparable requirements for specific sectors, though typically without the citizen-suit enforcement mechanism that gives environmental impact statements much of their practical force. Whether an algorithmic equivalent would fare better than Local Law 144's disclosure-without-robust-enforcement model, discussed in Section VII-B, is an open empirical question rather than one this review can resolve.

VIII. UNRESOLVED CHALLENGES

A. Explainability and the Black Box Problem

Many high-performing models, particularly deep neural networks and gradient-boosted ensembles, do not produce explanations that affected individuals or oversight bodies can use. This poses practical problems for fairness. A person who receives an adverse decision cannot understand or contest it. A regulator cannot verify that the decision process complies with applicable law. A developer cannot easily identify which features or interactions are producing discriminatory outputs.

The field of explainable AI (XAI) has produced tools including LIME (locally interpretable model-agnostic explanations) and SHAP (SHapley Additive explanations) that generate post-hoc explanations for individual

predictions. These tools are useful, but they produce approximations rather than true explanations. A SHAP value gives the marginal contribution of a feature to a particular prediction; it does not describe the underlying computation. Whether explanations of this kind satisfy legal requirements for meaningful justification and due process is still unsettled [15].

The limitation is not merely technical. Even a complete and faithful explanation of how a model reached a particular score does not by itself answer the question a denied applicant or court actually wants answered, which is normative rather than mechanical: was this decision fair, and what would have needed to be different for the outcome to change? Counterfactual explanations, which describe the smallest change to an applicant's record that would have flipped the decision, come closer to answering that question, but they inherit the same proxy problem described in Section III-B: if income, zip code, or credit history are correlated with a protected attribute, a counterfactual phrased in terms of those features may implicitly tell an applicant to change something tightly bound to their race or sex, which is not a remedy a regulator is likely to accept as meaningful.

A complementary line of work focuses on documentation rather than post-hoc explanation: model cards and datasheets that accompany a system's release with a structured description of its intended use, training data provenance, and measured subgroup performance. These documents do not explain any individual prediction, but they make the kind of subgroup-disaggregated evidence that revealed the bias in each of Section IV's case studies available before deployment rather than only after an external audit forces its disclosure. Their main limitation is adoption rather than design: nothing in current US or EU law requires a vendor to publish one, so coverage remains voluntary and uneven across the industry.

B. Intersectionality

Most fairness evaluations examine one protected characteristic at a time. The Gender Shades study [4] was influential in part because it evaluated gender and skin tone together, finding that darker-skinned women were far worse served than either a gender-only or a skin-tone-only analysis would have suggested. This reflects the intersectionality framework developed by [19] and [20], which holds that individuals occupy multiple social positions simultaneously and that discrimination along different axes interacts rather than simply adding together.

The Gender Shades follow-up discussed in Section IV-C reinforces this point in an unusually direct way: the Face + + system's comparatively good performance on darker-skinned male faces, worse than the system's best subgroup but still better than its performance on lighter-skinned female faces, would have been invisible to an audit that tested skin tone and gender as separate, marginal characteristics rather than as a joint distribution. The same limitation applies beyond two dimensions. A health-risk model might perform acceptably when evaluated by race alone and by age alone while still performing poorly for a specific combination, such as older Black women, that neither marginal analysis would isolate, and the number of such combinations to check grows exponentially with the number of attributes considered, at the same time as the number of individuals available to estimate each combination's error rate shrinks.

Evaluating fairness across all intersections of protected characteristics becomes computationally intractable as the number of dimensions grows, and small sample sizes in intersectional subgroups make statistical estimates unreliable. Both problems are manageable but require deliberate effort that most deployed systems do not make. [1] identify intersectional bias as one of the most underaddressed areas in the machine learning literature.

C. Dynamic Environments and Distributional Shift

Models are trained on historical data and deployed in present conditions. As circumstances change, the relationship between features and outcomes may shift in ways that invalidate the model's predictions. This is a distributional shift, or concept drift. When that shift affects demographic groups differently, a model that was reasonably fair at deployment can become increasingly biased over time, without any automatic detection. Continuous monitoring of fairness metrics after deployment is technically feasible but is not yet standard practice in most sectors [1].

A concrete illustration arose during the COVID-19 pandemic, when healthcare utilization patterns shifted abruptly and unevenly across demographic groups as routine and elective care was deferred at different rates in different communities; a healthcare risk model trained on pre-pandemic utilization data, of the kind discussed in Section IV-A, would have faced a version of the same proxy problem that motivated the original correction, but driven by a sudden external shock rather than a slow accumulation of historical practice. Few of the fairness-monitoring pipelines described in Section VI-D were designed with this kind of abrupt, demographically uneven shock in mind, and post-deployment monitoring that checks aggregate accuracy on a rolling basis can fail to flag a shift that affects subgroups unevenly even while overall performance appears stable.

Distributional shift interacts with the feedback loops described in Section III-D in a way that can mask deterioration rather than reveal it. If a model's own deployment changes the population it is later evaluated against, for instance because a credit model's past decisions shaped who applied for credit in the first place, then a fairness metric computed on current data may show stability not because the model has remained fair but because the population available to measure it against has been reshaped by the model's own prior outputs. Detecting this requires comparing current outcomes against a counterfactual population that was never exposed to the system, a comparison that is rarely available outside a controlled experiment and almost never available for a system already in widespread use.

D. Global Disparities

Most algorithmic bias research has been conducted in North American and European contexts. The systems being deployed globally were predominantly trained on data from these contexts and may perform differently, or in qualitatively distinct ways, when applied to populations with different demographic compositions, languages, social structures, and historical conditions.

[21] observed that the integration of large language models into consequential decisions raises privacy, discrimination, and accountability concerns that are especially acute where regulatory infrastructure is underdeveloped. [22] documented how systems trained mainly on data from high-income countries can

reproduce and deepen inequalities in lower-income settings, partly because local populations are too underrepresented in training data to influence system behavior.

Research infrastructure itself exhibits a comparable imbalance. The conferences and journals that publish most fairness work, including the ACM FAccT conference from which several of this review's key citations are drawn, draw their program committees, reviewers, and authors overwhelmingly from North American and European institutions. This means that the problems that get formalized, the datasets that get assembled as benchmarks, and the metrics that become standard in evaluation all reflect the legal and social contexts of those institutions. A system that performs fairly by the measures a US or European research community developed may still perform poorly in contexts those communities did not anticipate, and the communities best positioned to detect and document such disparities are also those with the fewest resources to run the kind of large-scale audits that, in the cases reviewed in Section IV, were necessary to make the problem visible.

Underrepresentation compounds across the layers identified in Section III: a population underrepresented in a system's training data is also typically underrepresented among the auditors, regulators, and journalists positioned to detect the resulting disparities, and is least likely to have a comparable Local Law 144 or EU AI Act available to provide legal recourse even where a disparity is detected. The generative AI systems discussed further compound this concern further, because the languages and cultural contexts best represented in their training corpora are concentrated in exactly the high-income regions already over-represented in the bias literature surveyed throughout this review.

E. Generative AI and Foundation Models

Most of the bias literature surveyed above concerns discriminative models trained for a narrow, well-defined prediction task: a risk score, a classification, a ranking. Large generative models, including the large language models now embedded in search, customer service, and increasingly in consequential decisions such as resume screening and benefit-eligibility correspondence, introduce bias risks of a different character. [28] argue that the environmental and financial costs of training ever-larger language models fall disproportionately on communities least likely to benefit from the resulting systems, and that the scale of the training corpora involved makes it practically infeasible to audit that corpus for the same kind of representational imbalance documented for the much smaller benchmark datasets discussed in Section III-A; a language model's training data is typically scraped from a large fraction of the publicly available internet, which is not demographically or linguistically representative of the global population any more than the facial-recognition benchmarks in Fig. 2 were, but is several orders of magnitude larger and far less feasible to inspect example by example.

[21] extend this concern to the deployment context that motivates this review: the integration of generative models into decisions about people, rather than merely into open-ended text generation, raises the same privacy, discrimination, and accountability questions raised by the discriminative models in Section IV, compounded by the fact that a generative model's output is harder to characterize statistically than a single risk score, making the group-fairness criteria formalized in Section V difficult to apply directly. A model asked to draft a recommendation letter, summarize a candidate's work history, or explain a benefits denial

can encode the same historical and representational biases documented throughout this review, but because its output is free text rather than a number, the disparity is harder to measure, harder to audit using the methods described in Section VI-D, and, for the reasons given in Section VIII-A, harder to explain to the person it affects. [22] document a related effect in lower-income settings, where systems trained predominantly on data from high-income countries reproduce and at times deepen existing inequalities precisely because local populations are too underrepresented in the training data, in any language, to meaningfully shape system behavior. None of the mitigation strategies catalogued in Section VI was designed with a generative, open-ended output space in mind, which is among the reasons this review treats generative AI bias as unresolved rather than as a fifth case study alongside Section IV's four.

IX. CONCLUSION

The cases examined here follow the same structure. An AI system was deployed to make high-stakes decisions affecting access to healthcare [3], liberty [5], employment [6], or other essential resources [4]. The system produced outputs that systematically disadvantaged members of already-marginalized groups. In every case, the disparity traced back to design choices: the selection of training data, the use of a proxy variable, the choice of optimization objective, and decisions about what to measure and audit. None of these choices required discriminatory intent. All reproduced existing social inequalities.

Two conclusions follow. The first is that algorithmic bias cannot be resolved at the technical level alone [1], [2]. Pre-processing, in-processing, and post-processing methods [12] can reduce bias and should be used. But they cannot override the normative choices embedded in the selection of a fairness criterion [8], [11], and they cannot address the structural conditions, including residential segregation, differential healthcare access, and long-standing discrimination in lending and hiring, that produce the skewed historical data from which biased models learn.

The second is that regulatory frameworks need to match the scale of automated decision-making. The EU AI Act's requirements for representative training data, conformity assessment, and post-deployment monitoring are a substantive step [15]. New York City's mandatory audit requirement for algorithmic hiring tools shows that enforceable accountability is workable in principle [17], even where, as Section VII-B shows, the enforcement actually delivered so far has fallen short of that principle. Both are partial, and internationally coordinated governance remains absent from most of the world.

Several practical implications follow for the different audiences this review addresses. For practitioners building or procuring AI systems, the case studies in Section IV show that aggregate accuracy is not a sufficient validation criterion; subgroup-disaggregated evaluation, of the kind that revealed the bias in each of the four cases, should be a standard step rather than an optional one, and the choice of fairness criterion discussed in Section V should be made explicitly and documented, not left as an implicit consequence of whichever metric a development team happens to report. For auditors and researchers, Section VI-D's account of the access barriers facing external review suggests that mandated application programming interface access for qualified researchers, not merely a published summary report, would do more to close the gap Fig. 6 documents between when bias is typically discovered and when it is typically regulated. For policymakers, the enforcement experience reviewed in Section VII-B indicates that a disclosure requirement

without a correspondingly resourced enforcement mechanism may produce compliance on paper without changing what organizations actually do in practice.

This review has its own limitations, consistent with the evidence base summarized in Fig. 1. It draws predominantly on cases documented in the United States and the European Union, for the reasons discussed in Section VIII-D, and predominantly on discriminative rather than generative systems, for the reasons discussed in Section VIII-E. Both choices reflect where the published evidence is currently concentrated rather than a judgment that other jurisdictions or system types raise less serious concerns; if anything, Section VIII-D suggests the opposite may be true precisely where the evidence is thinnest.

What durable fairness in AI actually requires is institutional: representative governance of data collection and curation; independent, standardized, and enforceable auditing [14]; legal mechanisms giving affected individuals real recourse; and research that evaluates systems not only on aggregate accuracy but on how errors are distributed across social groups. The generative systems discussed in Section VIII-E suggest that this requirement will need to extend, and soon, to a class of systems whose outputs are far harder to characterize statistically than the risk scores and classifications examined in this review's four founding case studies. None of this is technically infeasible. It is what a society that takes equality seriously would require of the automated systems it uses to make consequential decisions about people's lives.

Declarations

Conflict of interest:

The authors have no conflicts of interest to declare.

Ethical approval:

Not applicable.

Funding:

The authors declare that this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability:

This is a review and analytical paper. No original datasets were collected.

References

1. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM-CSUR 54(6):1–35. https://doi.org/10.1145/3457607
2. Ferrara E (2023) Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci 6(1):3. https://doi.org/10.3390/sci6010003
3. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342
4. Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR, vol 81, pp 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html
5. Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. ProPublica, 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
6. Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 10 October 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
7. Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x
8. Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 5(2):153–163. https://doi.org/10.1089/big.2016.0047
9. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Adv Neural Inf Process Syst 29. https://arxiv.org/abs/1610.02413
10. Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). https://arxiv.org/abs/1609.05807
11. Friedler SA, Scheidegger C, Venkatasubramanian S (2021) The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. Commun ACM 64(4):136–143. https://doi.org/10.1145/3433949
12. Bellamy RKE, Dey K, Hind M et al (2018) AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. https://arxiv.org/abs/1810.01943
13. Raji ID, Buolamwini J (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 429–435. https://doi.org/10.1145/3306618.3314244
14. Raji ID, Smart A, White RN et al (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency, pp 33–44. https://doi.org/10.1145/3351095.3372873
15. European Commission (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

16. European Parliament Research Service (EPRS) (2025) Algorithmic discrimination under the AI Act and the GDPR. European Parliament, PE 769509. https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/769509/EPRS_ATA(2025)769509_EN.pdf
17. New York City (2021) Local Law 144 of 2021 (Int 1894-A): Automated employment decision tools. New York City Council. https://www.nyc.gov/site/dca/workers/workersrights/automated-employment-decision-tools.page
18. American Civil Liberties Union (ACLU) (2018) Why Amazon's automated hiring tool discriminated against women. https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against
19. Crenshaw K (1989) Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. Univ Chic Legal Forum 1989(1):139–167
20. Crenshaw K (1991) Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Rev 43(6):1241–1299. https://doi.org/10.2307/1229039
21. Shams S et al (2023) Bias and fairness in generative AI. (See also: Ferrara 2023 [2]; Mehrabi. 2021 [1].)
22. Frontiers in Research Metrics and Analytics (2024) Navigating algorithm bias in AI: Ensuring fairness and trust in Africa. https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2024.1486600/full
23. State v (2016) Loomis, 881 N.W.2d 749 (Wis. Wisconsin Supreme Court
24. Meding K (2025) It's complicated: The relationship of algorithmic fairness and non-discrimination regulations for high-risk systems in the EU AI Act. https://arxiv.org/abs/2501.12962
25. Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104(3):671–732
26. Grother P, Ngan M, Hanaoka K (2019) Face recognition vendor test (FRVT) part 3: Demographic effects. NIST Interag Rep NISTIR 8280. https://doi.org/10.6028/NIST.IR.8280
27. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12), pp 214–226. https://doi.org/10.1145/2090236.2090255
28. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), pp 610–623. https://doi.org/10.1145/3442188.3445922
29. Wright L, Muenster RM, Vecchione B, Qu T, Cai P et al (2024) Null compliance: NYC Local Law 144 and the challenges of algorithm accountability. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), pp 1701–1713. https://doi.org/10.1145/3630106.3658998
30. Office of the New York State Comptroller (2025) Enforcement of Local Law 144 – Automated employment decision tools. Audit Report 2025-N-2, December 2025. https://www.osc.ny.gov/state-agencies/audits/2025/12/02/enforcement-local-law-144-automated-employment-decision-tools

Figures

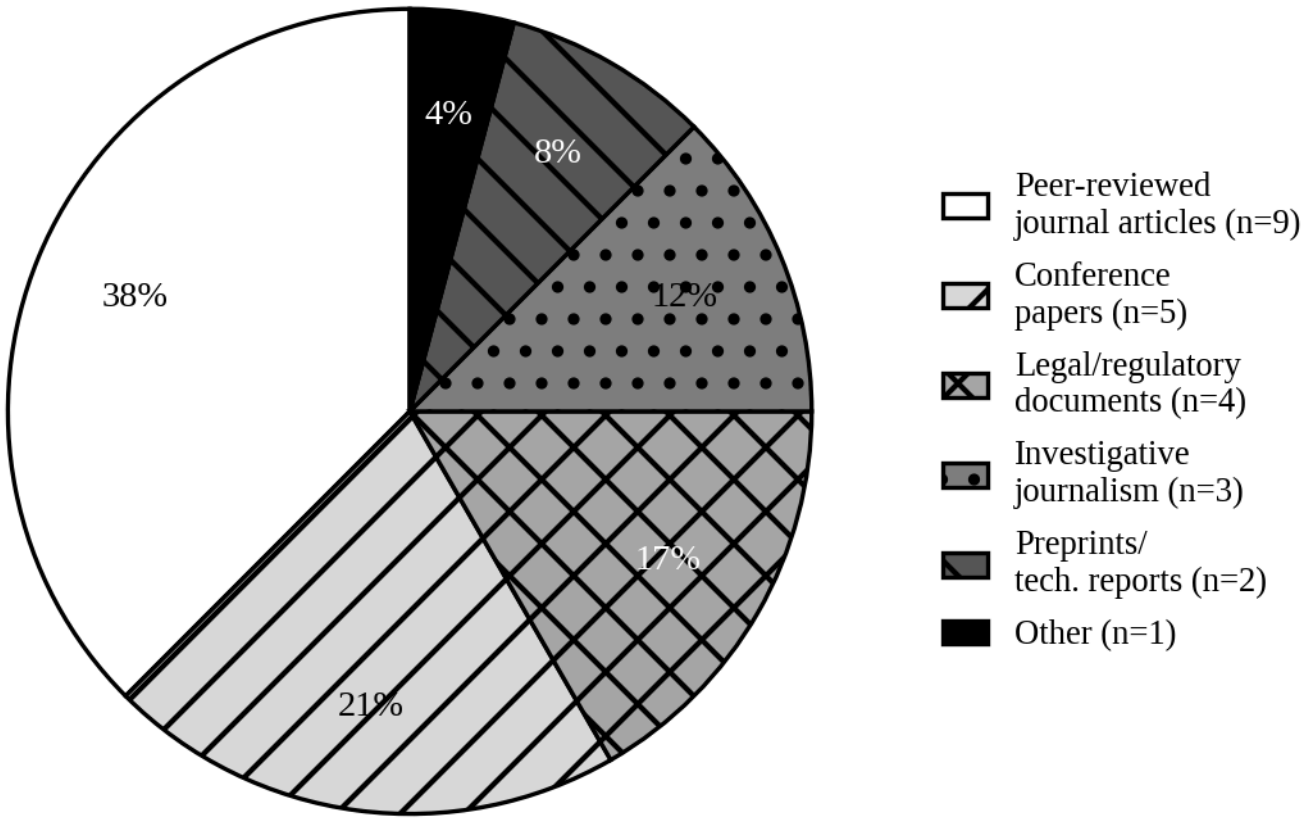


Figure 1

Composition of this review's cited evidence base by source type (n=24 references in the original literature base, by source type).

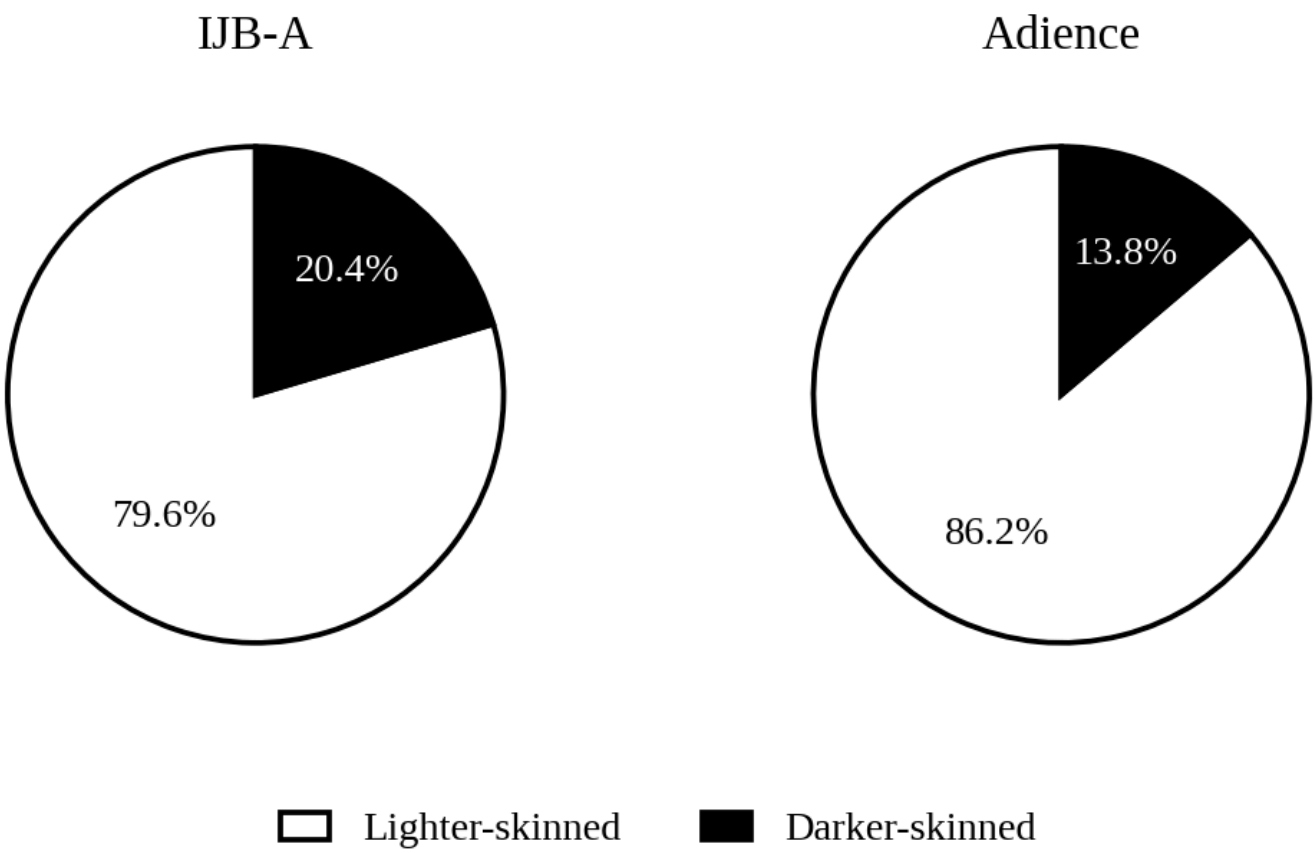


Figure 2

Skin-tone composition of two facial-recognition benchmark datasets audited in [4].

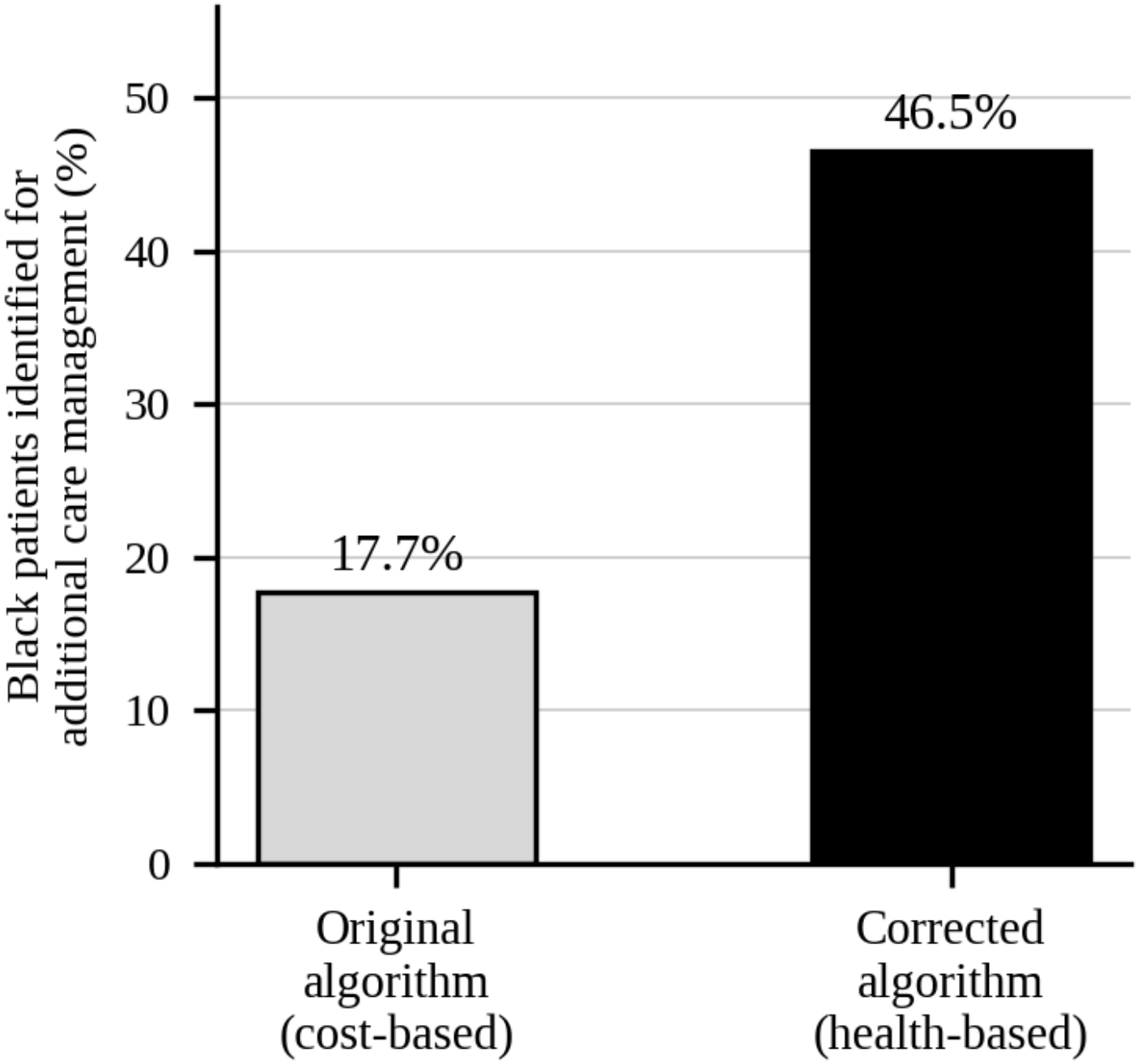


Figure 3

Share of Black patients identified for additional care management before and after correcting the algorithm's prediction target [3].

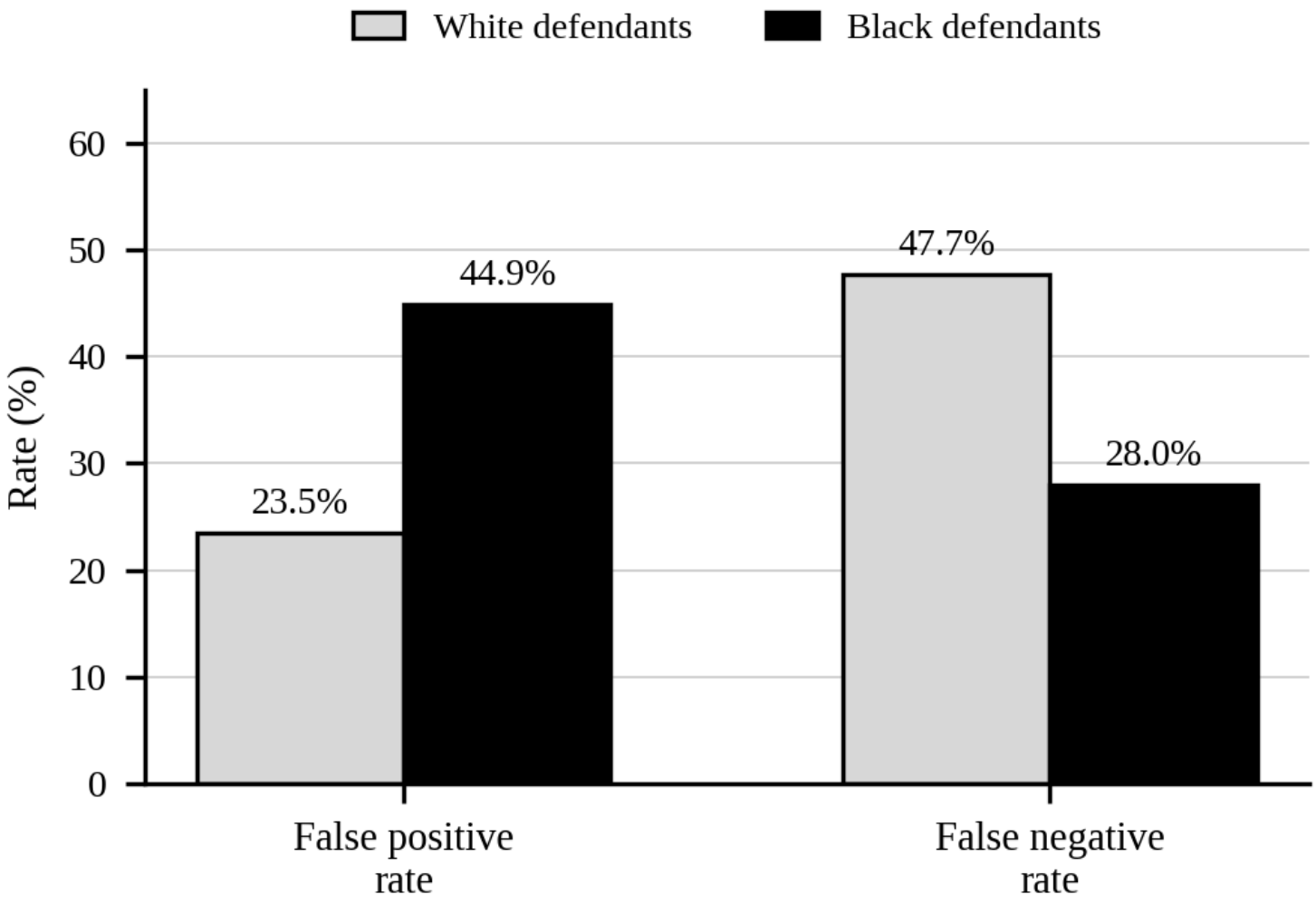


Figure 4

False positive and false negative rates of the COMPAS risk score, by defendant race, in ProPublica's Broward County analysis [5].

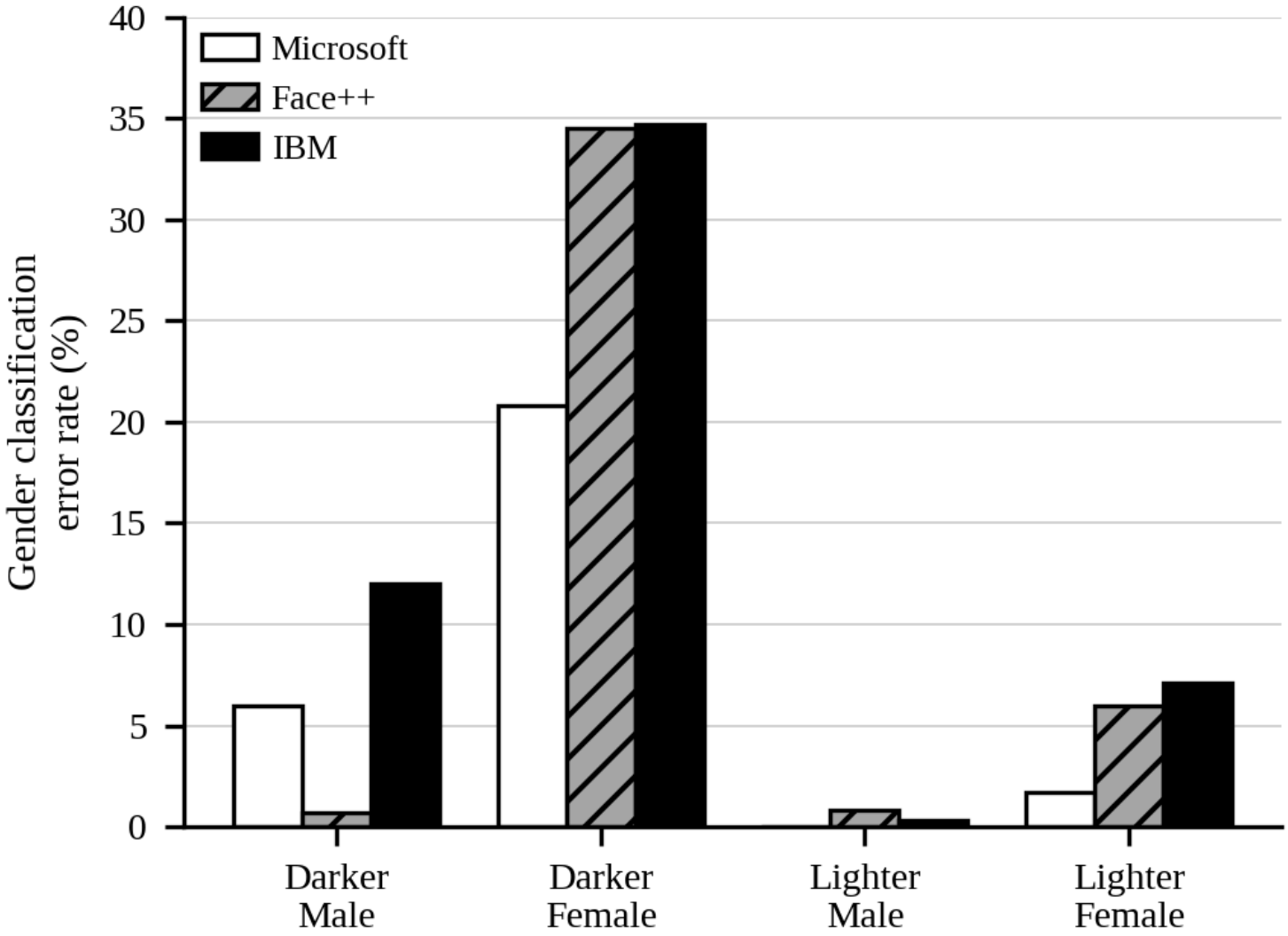


Figure 5

Gender classification error rate by intersectional subgroup for three commercial systems audited in [4].

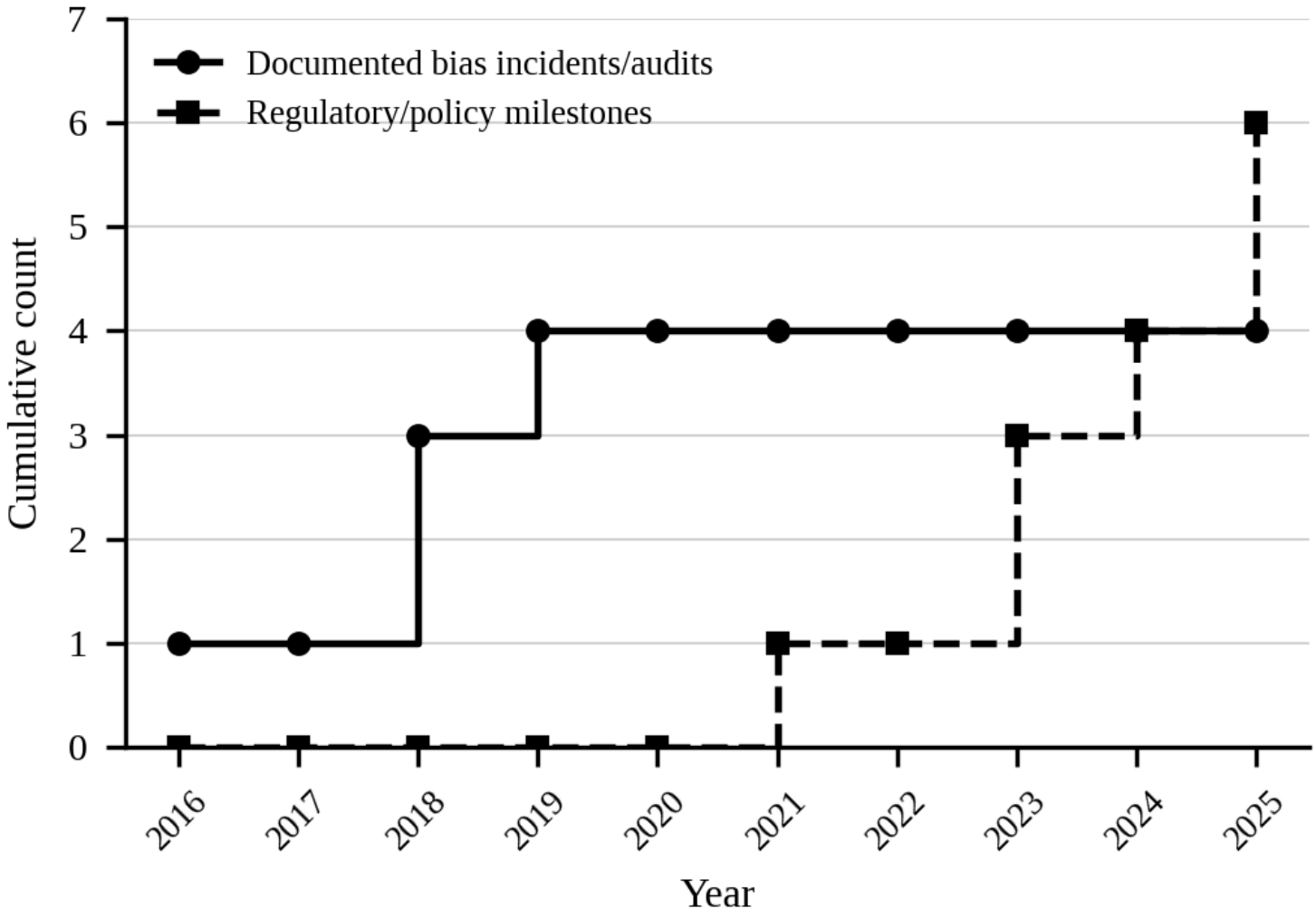


Figure 6

Cumulative documented bias incidents and audits versus cumulative regulatory and policy milestones, 2016–2025.

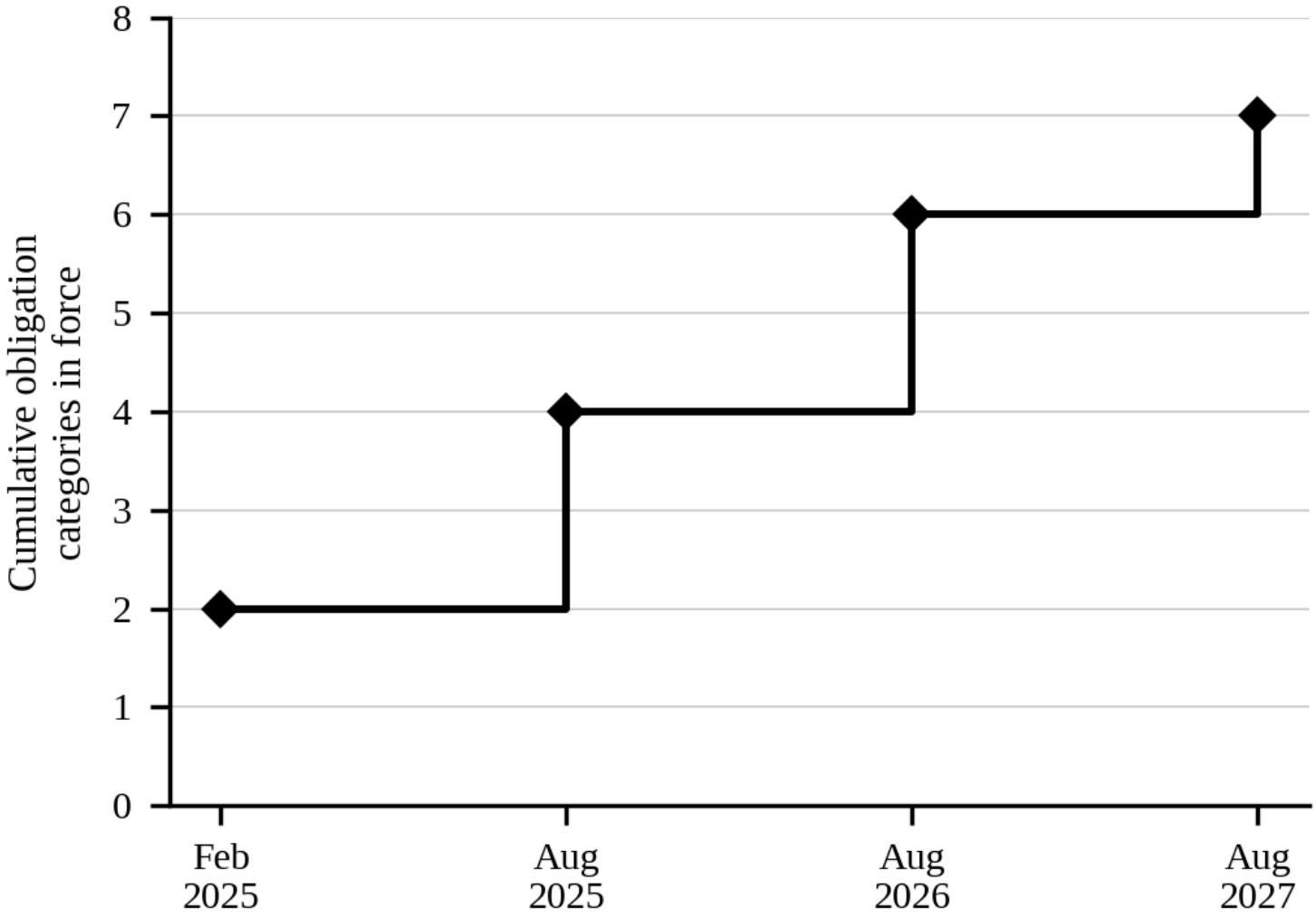


Figure 7

Cumulative entry into force of EU AI Act obligation categories, as originally legislated.