DOI: https://doi.org/10.5281/zenodo.20563720 | Vol 4. No. 1, 2026 | EISSN: 3121-9055

Research Paper EISSN: 3121-9055 Open Access

The Ethics of Algorithmic Bias in Generative AI: How Large Language Models and Image Generators Perpetuate Societal Bias

[1] **Prof S. Akin Ola**
solom202@yahoo.co.uk

[2] **Oladejo Samuel Adetunji**
oladejosa@live.com

[1] **Department of Computer, University of Ibadan**

[2] **Department of Computer and Information Science, Lead City University**

Corresponding Author:
Oladejo Samuel Adetunji
oladejosa@live.com

The rapid proliferation of generative artificial intelligence, encompassing large language models (LLMs) such as GPT-4 and Gemini as well as text-to-image generators including DALL·E and Stable Diffusion, has introduced unprecedented capabilities alongside profound ethical hazards. This paper investigates how societal biases encoded in training corpora are systematically perpetuated, amplified, and reified by these systems. Drawing on three domain case analyses, namely automated hiring screening, algorithmic credit scoring, and predictive law-enforcement tools, we synthesise documented evidence of disparate impact across race, gender, and socioeconomic strata, and foreground intersectional harm as a distinct and frequently under-measured dimension of that impact. We are explicit that these analyses are narrative syntheses of secondary sources rather than original longitudinal data collection. We then propose a tripartite governance framework comprising (1) pre-ingestion dataset auditing, (2) in-training fairness constraints via adversarial debiasing, and (3) post-deployment algorithmic auditing protocols informed by Explainable AI (XAI) techniques. We situate these technical proposals within an explicit ethical analysis, applying Rawlsian fairness principles and Sen's capability approach to specific cases, and situating these within the political economy of AI governance. Our central argument is that no single debiasing technique is sufficient, that several techniques are themselves in tension with privacy and with one another, and that meaningful fairness demands coordinated intervention across the model lifecycle and the wider sociotechnical system. We conclude with policy recommendations for regulators, model developers, and civil society stakeholders, and with reflection on how these dynamics manifest beyond the Global North.

Keywords: *Generative AI, Algorithmic Bias, Fairness in Machine Learning, Intersectionality, Ethical AI, Large Language Models, Dataset Bias, Explainable AI (XAI), Algorithmic Auditing, Political Economy of AI, Debiasing*

1. Introduction

Artificial intelligence systems have become deeply embedded in consequential social institutions, from the algorithms that rank job applicants to the risk-score tools that influence bail decisions. The emergence of generative AI over the past half-decade has accelerated this integration dramatically, introducing models capable of producing human-quality text, realistic imagery, and synthetic data at scale. Yet this capability is not neutral. Every generative model is, at its core, a statistical distillation of its training data; when that data reflects historical inequities, as human-generated data almost invariably does, the model learns, reproduces, and often amplifies those inequities.

Bias in machine learning is not a novel concern. Scholars such as Barocas, Hardt, and Narayanan (2023) have long documented how predictive models can encode discriminatory patterns, and regulatory bodies in the European Union and the United States have begun to acknowledge the problem in their legislative agendas. What is new with generative AI is the scale, opacity, and pervasiveness of the systems involved. A biased hiring algorithm affects candidates who apply to a particular company; a biased LLM embedded in a recruitment platform can affect millions of candidates across thousands of employers simultaneously.

Critically, algorithmic bias is not merely a technical artefact to be patched. Benjamin (2019) characterises the way ostensibly neutral technical systems can encode and obscure racial hierarchy as the "new Jim Code," while Noble (2018) demonstrates how the commercial logic of search and recommendation engines actively reproduces oppression rather than passively reflecting it. These accounts locate the problem not only in flawed data but in the institutions, incentives, and power asymmetries that produce, deploy, and profit from such systems. We adopt this sociotechnical and political-economic framing throughout.

This paper makes four primary contributions. **First**, it provides a structured taxonomy of bias sources in generative AI pipelines, distinguishing historical, representation, measurement, aggregation, feedback, automation, and intersectional biases. **Second**, it presents three domain case analyses, covering hiring, lending, and law enforcement, that synthesise

documented evidence of disparate impact, with explicit attention to intersectional harm. **Third**, it proposes a tripartite framework for bias mitigation and examines the practical tensions, notably between debiasing and privacy, that constrain it. **Fourth**, it situates these technical findings within a substantive ethical analysis, applying Rawlsian fairness principles and Sen's capability approach to specific cases, and within the political economy of algorithmic governance.

2. Background and Related Work

2.1 Defining Algorithmic Bias, and Why "Fairness" Is Contested

Algorithmic bias refers to systematic and repeatable errors in a computational system that create unfair outcomes for particular groups. In the context of machine learning, bias can manifest as differential accuracy (a model performing worse for one demographic group than another), differential exposure (a model amplifying certain viewpoints over others), or differential impact (a model's outputs producing real-world harm asymmetrically across groups).

Mehrabi et al. (2021) identify over twenty distinct types of algorithmic bias, ranging from historical bias embedded in training labels to evaluation bias arising from benchmark datasets that fail to represent all populations. For generative AI, the most salient categories are representation bias (certain groups being under- or mis-represented in training data) and stereotyping bias (associations between group membership and attributes being systematically skewed).

It is essential to establish at the outset that "fairness" in machine learning is not an intuitive or singular concept but a contested and plural one. Researchers have formalised fairness in mutually incompatible ways: Dwork et al. (2012) articulate individual fairness, that is, the principle that similar individuals should be treated similarly, a criterion that can diverge sharply from group fairness criteria such as demographic parity or error-rate balance. Chouldechova (2017) later proved that several standard group-fairness criteria cannot be satisfied simultaneously when base rates differ across groups, a result developed further in Section 6.1. We foreshadow this "impossibility" here so that readers approach the remainder of the paper understanding that to call a system "fair" is already to have made a value-laden choice among competing, formally irreconcilable definitions.

A further conceptual commitment frames our analysis: intersectionality. Crenshaw (1989) demonstrated that discrimination experienced at the intersection of multiple marginalised identities, as experienced for instance by Black women, cannot be understood as the simple sum of separately measured race and gender effects. As we show, the most severe documented harms of generative AI are frequently intersectional, yet standard auditing practice measures bias along one protected attribute at a time and therefore systematically under-detects them.

2.2 Generative AI and the Bias Amplification Problem

Generative models differ from traditional discriminative classifiers in that they learn the full joint distribution of their training data, making them susceptible to reinforcing patterns across an enormous output space. Bender et al. (2021), in their influential "Stochastic Parrots" paper, argue that LLMs trained on web-scale corpora do not merely reflect societal biases but actively amplify them, because dominant cultural narratives are statistically more frequent and thus receive disproportionate reinforcement during training. Weidinger et al. (2021) provide a systematic taxonomy of the ethical and social risks of LLMs, situating discrimination, exclusion, and toxicity as a primary risk area alongside information hazards and misinformation, and emphasising that mitigations are most effective when targeted at a harm's point of origin in the lifecycle. This principle directly motivates the staged framework presented in Section 5.

Image generation models face analogous challenges. Cho et al. (2023) demonstrated that DALL-E 2, when prompted with neutral occupational descriptors such as "a CEO" or "a nurse", produced images skewed dramatically toward White male representations for high-status roles and female representations for lower-status roles, mirroring and amplifying occupational segregation patterns found in stock-image training corpora.

2.3 Regulatory, Ethical, and Political-Economic Frameworks

The European Union's AI Act (European Parliament, 2024) classifies AI systems used in hiring, credit, and law enforcement as "high-risk" and mandates conformity assessments, human oversight, and transparency obligations. In the United States, the Equal Credit Opportunity Act and the Fair Housing Act impose anti-discrimination requirements that courts have increasingly held apply to algorithmic decision-makers.

Philosophically, several frameworks bear on AI fairness. Rawlsian justice demands that social and economic inequalities be arranged so as to benefit the least-advantaged members of society (the difference principle); utilitarian calculus requires that aggregate welfare gains not systematically exclude minority groups; and Sen's capability approach insists that institutions preserve individuals' real freedom to function as full social participants. We do not merely invoke these frameworks: Section 6.4 applies the Rawlsian difference principle and Sen's capability approach to specific cases analysed in this paper.

Finally, fairness cannot be analysed in abstraction from the political economy of AI. The systems examined here are commercial products whose deployment is driven by cost reduction and competitive advantage, whose harms fall on populations with little market or political power, and whose developers are frequently insulated from liability by trade-

secret protections and diffuse accountability. Benjamin (2019) and Noble (2018) show that these power asymmetries are not incidental to algorithmic bias but constitutive of it. We return to this in Section 6.3.

3. A Taxonomy of Bias Sources in Generative AI Pipelines

We organise bias sources across the stages of the generative AI lifecycle. To the conventional categories we add an explicitly cross-cutting row on intersectional bias, because, as Section 2.1 argued, harms compounded across multiple identities are routinely missed by single-attribute auditing.

Table 1. Taxonomy of Bias Sources Across the Generative AI Lifecycle

Stage	Bias Type	Description & Example
Data Collection	Historical Bias	Training data encodes past discrimination; e.g., resume datasets reflecting decades of gendered hiring decisions.
Data Collection	Representation Bias	Minority groups under-represented; e.g., medical-imaging models trained predominantly on data from lighter-skinned patients.
Pre-processing	Measurement Bias	Proxy variables correlate with protected attributes; e.g., ZIP code as a credit-risk proxy encodes residential racial segregation.
Training	Aggregation Bias	A single model trained across heterogeneous subgroups flattens within-group variation; e.g., sentiment models performing poorly on African American Vernacular English.
Fine-tuning / RLHF	Feedback Bias	Human raters carry their own biases; e.g., crowdworkers rating "helpful" responses may prefer outputs conforming to majority cultural norms.
Deployment	Automation Bias	Human reviewers over-rely on model outputs; e.g., hiring managers accepting AI rankings without scrutiny.
Cross-cutting	Intersectional Bias	Harms compounded at the intersection of multiple protected attributes that single-attribute audits miss; e.g., error rates highest for darker-skinned women (Buolamwini & Gebru, 2018; Wilson & Caliskan, 2024).

4. Domain Case Analyses: Bias in High-Stakes Applications

The three analyses that follow are narrative case reviews: structured syntheses of journalistic investigations, civil-society audits, regulatory findings, and peer-reviewed research. They are not original longitudinal data collection, and we do not present them as such. Their purpose is to establish, from the documented record, the patterns of bias that generative systems are now reproducing and amplifying.

4.1 Automated Hiring Screening

It is important to be precise about a foundational example. Amazon's experimental hiring tool, developed between 2014 and 2017 and reported by Dastin (2018), was a supervised machine-learning classifier, not a generative AI system. Trained on a decade of resumes reflecting the male-dominated composition of the technology industry, it penalised resumes containing the word "women's" and down-ranked graduates of all-women's colleges; Amazon discontinued it in 2017. We include the case not as direct evidence of generative AI bias but as an illustration of a failure mode, namely training on historically biased hiring records, that generative systems are now replicating and amplifying at far greater scale.

Recent peer-reviewed evidence demonstrates that this failure mode persists in LLM-based screening. Wilson and Caliskan (2024), in an audit using over 500 real resumes and systematically varied names, found that LLM-based retrieval models favoured White-associated names 85% of the time and female-associated names only 11% of the time. Strikingly, they never favoured Black male-associated names over White male-associated names. This is a paradigmatically intersectional result: the disadvantage to Black men is not predicted by, and is more severe than, the marginal race and gender effects considered separately. Haim, Salinas, and Nyarko (2024), auditing state-of-the-art models including GPT-4, similarly found that advice systematically disadvantaged names associated with racial minorities and women, with names associated with Black women receiving the least advantageous outcomes, a pattern that single-axis analysis would systematically obscure.

These findings echo the canonical intersectional result in computer vision: Buolamwini and Gebru (2018) found commercial gender-classification error rates up to 34.7 percentage points higher for darker-skinned women than for lighter-skinned men. Where such facial-analysis components are embedded in video-interview screening, that intersectional disparity is imported directly into hiring. The through-line across vision and language systems is consistent: harm concentrates at the intersection of marginalised identities, and masking a single attribute such as removing names is insufficient, because subtler proxies including word choice, institutions attended, and neighbourhood cues continue to leak protected-attribute information (Wilson & Caliskan, 2024).

4.2 Algorithmic Credit Scoring and Lending

The deployment of algorithmic tools in consumer lending has produced documented disparate-impact outcomes. In the most rigorous peer-reviewed study of the issue, Bartlett, Morse, Stanton, and Wallace (2022) used the pricing model of the U.S. government-sponsored enterprises as an identification strategy and found that risk-equivalent Latinx and African-American borrowers paid roughly 7.9 basis points more on purchase mortgages and 3.6 basis points more on refinancing. These disparities are estimated to cost minority borrowers on the order of US$450-765 million annually. Notably, they found that fintech algorithmic lenders did not eliminate this discrimination, consistent with models trained on historical lending data shaped by decades of redlining, which effectively treat geography as a racialised proxy and encode it as a negative signal.

More recent concern has centred on LLM-assisted loan-officer tools that generate recommendations from unstructured applicant narratives. Because information about an applicant's neighbourhood or community affiliations is unlikely to be fully excised from free text, such narratives can trigger associative reasoning that disadvantages applicants from historically redlined areas even in the absence of explicit race indicators, constituting a contemporary generative-era instance of the proxy discrimination that Bartlett et al. document in conventional algorithmic underwriting.

4.3 Predictive Law Enforcement and Recidivism Scoring

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism-prediction tool has been subject to extensive empirical scrutiny. ProPublica's analysis (Larson, Mattu, Kirchner, & Angwin, 2016) found that Black defendants were nearly twice as likely as White defendants to be falsely classified as future criminals (a false-positive rate of roughly 45% versus 23%), while White defendants were more often falsely classified as low-risk. Corbett-Davies, Pierson, Feller, Goel, and Huq (2017) showed that this asymmetry follows, in part, from the mathematical impossibility of simultaneously equalising calibration and error rates across groups with different base rates, a result developed further in Section 6.1, so that the dispute is less about a coding error than about which conception of fairness COMPAS should have satisfied. Dressel and Farid (2018) added a distinct and sobering finding: COMPAS's predictive accuracy was no better than that of untrained humans and could be matched by a simple two-feature linear model, undermining the premise that the tool offered any specialised predictive value to justify its risks.

The emergence of LLM-assisted sentencing and parole-recommendation tools represents a new frontier. Preliminary evidence indicates that models trained on legal corpora can reproduce documented racial sentencing disparities and may produce harsher recommendations for defendants with names statistically associated with Black identity. By extension, the most acute risks are likely to fall intersectionally on defendants who are marginalised along more than one axis. The opacity of such systems compounds these concerns: courts in several jurisdictions have ruled that COMPAS's proprietary algorithm need not be disclosed to defendants, raising acute due-process questions alongside the substantive fairness issues.

5. A Tripartite Framework for Bias Mitigation

5.1 Pre-Ingestion Dataset Auditing

The first pillar targets the data pipeline before model training begins. Dataset auditing involves systematic assessment of training corpora for representational imbalances, historically encoded stereotypes, and proxy variables correlated with protected attributes. We propose the following minimum standards for organisations training or fine-tuning generative models for high-risk applications. Demographic representation audits should quantify the proportion of training examples associated with each demographic group, compare these against population benchmarks, and document under-representation thresholds that trigger mandatory remediation. Crucially, these audits should be conducted at the level of intersectional subgroups (for example, darker-skinned women), not only along single attributes. As Section 4 demonstrated, single-axis audits systematically miss the most severe harms. Label-quality audits should scrutinise ground-truth labels for evidence of historical discrimination; hiring decisions used as training labels, for instance, should be tested for adverse impact before use. Counterfactual sensitivity analysis should test whether outputs change when protected attributes or their proxies are varied while other features are held constant.

A practical obstacle must be confronted directly: much of the above presupposes access to demographic labels, which organisations frequently do not hold. For legal, institutional, or commercial reasons, the very data needed to diagnose discrimination, including race, ethnicity, and disability status, may be absent. Veale and Binns (2017) set out approaches for mitigating discrimination without collecting sensitive data, including trusted-third-party arrangements and the use of proxies under privacy safeguards. Standardised documentation, specifically datasheets for datasets as proposed by Gebru et al. (2021), should accompany every high-risk corpus, recording provenance, composition, and known representational gaps so that downstream auditors can reason about bias even where individual-level demographic labels are unavailable.

5.2 In-Training Fairness Constraints and the Debiasing-Privacy Tension

The second pillar introduces fairness objectives during training. Adversarial debiasing, introduced by Zhang, Lemoine, and Mitchell (2018), trains a model jointly with an adversary whose goal is to predict protected attributes from the model's representations; the primary model is penalised for making the adversary's task easy, incentivising representations that are invariant to protected attributes. Fairness-aware reinforcement learning from human feedback (RLHF) extends this to fine-

tuning, providing raters with rubrics that explicitly incorporate fairness criteria and assigning negative reward to fairness-violating outputs. Reweighting techniques up-weight examples from under-represented groups to reduce the statistical dominance of majority patterns.

A significant practical tension constrains all of these methods, and Section 5.1's auditing standards share it: most debiasing and fairness-measurement techniques require demographic labels, yet such labels are themselves sensitive personal data. Under the EU General Data Protection Regulation, race, ethnicity, health, and sexual orientation are "special category" data (Article 9) whose collection and processing are tightly restricted; equivalent regimes apply elsewhere. Organisations therefore face a genuine dilemma: collecting the data needed to make a model fair may itself create privacy and re-identification risks and legal exposure. This is not a reason to abandon debiasing but a reason to treat the tension as a first-class design constraint. Privacy-preserving fairness methods, including the trusted-third-party and proxy-based approaches of Veale and Binns (2017) alongside techniques such as differential privacy and secure multi-party computation for fairness auditing, offer partial resolutions, allowing disparity to be measured and constrained without exposing individual sensitive attributes. Each technique also involves performance trade-offs (adversarial debiasing can reduce task accuracy; reweighting depends on reliable labels), making careful empirical evaluation essential.

5.3 Post-Deployment Algorithmic Auditing

The third pillar addresses bias that emerges or evolves after deployment. Disparate-impact monitoring should track production outputs across demographic groups, including intersectional subgroups, using statistical tests such as the four-fifths rule and standardised risk-difference metrics, with automated alerts when thresholds are exceeded. Explainability-assisted auditing leverages XAI tools including LIME (Ribeiro, Singh, & Guestrin, 2016) and SHAP to identify which input features drive disparate outputs, enabling targeted remediation (subject to the important limitations discussed in Section 6.2). Model cards (Mitchell et al., 2019) should accompany each deployed system, reporting performance disaggregated by demographic group and intended and out-of-scope uses. Finally, participatory auditing engages affected communities in structured evaluations, incorporating lived-experience knowledge that technical audits alone cannot capture; civil-society organisations including the Algorithmic Justice League and Data & Society have developed participatory-audit methodologies compatible with commercial-confidentiality constraints.

6. Discussion

6.1 The Limits of Technical Debiasing and the Structural Critique

A central finding of this analysis is that technical debiasing interventions, however sophisticated, are insufficient in isolation. Chouldechova (2017) demonstrated mathematically that three common fairness criteria, namely calibration, false-positive-rate parity, and false-negative-rate parity, are mutually incompatible when base rates differ between groups, which they typically do in domains shaped by historical discrimination. This "impossibility theorem" implies that developers and deployers must make explicit value judgements about which conception of fairness to prioritise, a decision that is inherently ethical and political in character rather than a purely technical one.

A deeper, structural critique follows. Selbst, Boyd, Friedler, Venkatasubramanian, and Vertesi (2019) warn that abstracting fairness into a technical optimisation problem can legitimise and entrench systems whose existence ought instead to be questioned. Selbst et al. describe this risk as the "formalism trap" and the "solutionism trap." Applied to Section 4.3, this critique is pointed: if a recidivism tool cannot satisfy competing fairness criteria simultaneously (Chouldechova, 2017; Corbett-Davies et al., 2017), and if its accuracy is no better than untrained humans (Dressel & Farid, 2018), then the appropriate response may not be to deploy a marginally fairer version but to ask whether algorithmic bail or recidivism prediction is an appropriate technology at all. A debiased tool that lends spurious objectivity to a fundamentally questionable practice may do more harm than an obviously crude one. Effective governance must therefore retain the option of non-deployment, not merely the obligation to optimise.

Furthermore, debiasing a model does not debias the broader sociotechnical system in which it is embedded. A debiased hiring model deployed by an organisation whose recruiters exhibit automation bias may produce less fair outcomes than a cruder model whose outputs are critically reviewed. Fairness requires governance of the entire system, including human-AI interaction design, organisational culture, and accountability structures.

6.2 The Role and Limits of Explainable AI

XAI techniques play a dual role in our framework: they serve as auditing instruments that expose discriminatory feature use, and as transparency mechanisms enabling affected individuals to contest decisions. The EU AI Act's requirement that high-risk systems offer "meaningful explanations" aligns with this dual function. However, the quality of explanations matters greatly. Post-hoc methods such as LIME and SHAP provide local approximations that may misrepresent global model behaviour, and their outputs can be manipulated by adversarially motivated actors, as Slack, Hilgard, Jia, Singh, and Lakkaraju (2020) demonstrated.

A more fundamental limitation must be acknowledged. Rudin (2019) argues that for high-stakes decisions, organisations should use inherently interpretable models rather than explaining black boxes after the fact, because post-hoc explanations

are approximations that can mislead. Most importantly for the present argument, XAI tools can describe how a model behaves but cannot determine whether that behaviour is just. An explanation that a loan was denied "because of postal code and income" is faithful and yet says nothing about whether relying on postal code is legitimate. This gap is of direct relevance to the EU AI Act's "meaningful explanation" requirement: a technically accurate explanation can satisfy a transparency mandate while leaving the substantive injustice entirely intact. Explainability is therefore necessary but not sufficient for fairness; it must be paired with normative evaluation of the features and objectives a model is permitted to use.

6.3 The Political Economy of Algorithmic Governance

Technical and ethical analysis is incomplete without attention to the incentives and power relations that determine whether biased systems are deployed at all. Three features of the political economy of AI are decisive. First, commercial incentives systematically favour deployment: automated screening and scoring reduce labour costs and increase throughput, and these gains accrue to deploying organisations while the costs of error fall on applicants, borrowers, and defendants who are typically dispersed and politically weak. Second, liability structures insulate developers and deployers from accountability: trade-secret protections shield proprietary models from scrutiny, as the COMPAS disclosure rulings illustrate, and responsibility is diffused across vendors, integrators, and end-users such that no single party bears clear legal exposure. Third, there is a profound power asymmetry between the organisations that build and deploy these systems and the communities subjected to them, a dynamic that Benjamin (2019) and Noble (2018) place at the centre of their respective analyses.

These conditions have a direct corollary for policy: algorithmic governance is politically contested terrain, not a neutral technical exercise. Major AI firms command substantial lobbying resources and routinely shape the scope, stringency, and enforcement of regulation, for instance by favouring self-administered bias audits with wide vendor discretion over independent third-party scrutiny. Any realistic reform agenda must reckon with this asymmetry rather than assume that well-designed standards will be adopted on their technical merits alone.

6.4 Applying Ethical Frameworks to the Cases

We now deliver the ethical analysis promised in Section 1 by applying two frameworks to the cases above. Rawls's difference principle holds that inequalities are justifiable only where they benefit the least-advantaged members of society. Applied to algorithmic lending (Section 4.2), the principle condemns the disparities documented by Bartlett et al. (2022): a system that extracts higher rates from already-disadvantaged Latinx and African-American borrowers worsens the position of the least well-off and cannot be justified by efficiency gains that accrue elsewhere. Applied to recidivism scoring (Section 4.3), it implies that the burden of predictive error should not fall most heavily on the group already most disadvantaged by the criminal-justice system, yet that is precisely what the false-positive asymmetry produces.

Sen's capability approach shifts the evaluative focus from outcomes to people's real freedom to achieve valued functionings. Applied to hiring (Section 4.1), the LLM screening biases documented by Wilson and Caliskan (2024) and Haim et al. (2024) constitute a capability deprivation: they foreclose the freedom to compete for employment on equal terms, and they do so most severely for those at intersectional margins, compounding existing constraints on their life chances. On this view the harm is not simply a statistical disparity in a hiring metric but a substantive diminution of freedom, which is why purely technical parity adjustments, absent attention to the lived capabilities of affected groups, constitute an inadequate response.

7. Policy Recommendations

On the basis of the foregoing analysis, we advance the following recommendations. We note at the outset that each will be contested: as Section 6.3 argued, algorithmic governance is shaped by powerful commercial interests, and recommendations that rely on industry self-assessment are particularly vulnerable to dilution. We therefore emphasise independent scrutiny and the retained option of non-deployment.

Table 2. Stakeholder-Specific Policy Recommendations

Stakeholder	Recommendation
Regulators	Mandate pre-deployment algorithmic impact assessments for high-risk AI, modelled on the EU AI Act but extended to generative systems used in hiring, lending, healthcare, and criminal justice. Require independent third-party audits at least annually, and require that audits report disparities at the intersectional-subgroup level, not only along single attributes. Preserve an explicit power to prohibit deployment where a practice cannot be made fair.
Model Developers	Publish datasheets for datasets and model cards for all publicly released models, documenting known limitations, demographic representativeness, and benchmark performance disaggregated by demographic group and key intersectional subgroups.
Deploying Organisations	Conduct ongoing disparate-impact monitoring in production, including intersectional monitoring, and establish genuine human-review procedures for high-stakes decisions.

	Address the debiasing-privacy tension by adopting privacy-preserving fairness methods where demographic labels cannot lawfully be collected. Adopt participatory auditing involving affected communities.
Civil Society	Advocate for algorithmic-transparency legislation enabling independent researchers and affected communities to audit deployed systems, including proprietary models, and counter the lobbying asymmetry that favours weak self-regulation.
Standards Bodies	Develop internationally harmonised standards for fairness metrics, audit methodologies (including intersectional and privacy-preserving methods), and disclosure requirements, building on IEEE P7003 and ISO/IEC TR 24368.

8. Conclusion

Generative AI systems are not mere mirrors of human knowledge; they are active producers of information, recommendations, and decisions that shape life chances at scale. The evidence reviewed in this paper indicates that these systems, as currently designed and deployed, perpetuate and amplify societal biases, frequently most severely at intersectional margins, with measurable and harmful consequences for marginalised populations in employment, credit, and the justice system.

The tripartite framework proposed here, comprising pre-ingestion dataset auditing, in-training fairness constraints, and post-deployment algorithmic auditing, provides a structured approach that acknowledges both technical and governance dimensions, while recognising the real tensions (notably between debiasing and privacy) that constrain it. No single intervention is sufficient. Fairness in generative AI demands coordinated action across the entire model lifecycle and across the sociotechnical and political-economic system in which models are embedded. In some cases, as the structural critique makes clear, the appropriate response is not to develop a fairer version of a given system but to forgo deployment altogether.

Finally, a limitation of scope deserves explicit acknowledgement. The cases analysed here are drawn overwhelmingly from the United States and Europe, where most documented audits have been conducted. This is a meaningful constraint. In many African and other Global South contexts, including the South African setting from which this work is written, the dynamics may be considerably more acute. Regulatory and audit capacity is often thinner, local-language and local-population data are scarce, and AI systems are frequently imported wholesale from the Global North, trained on populations they will not serve and deployed without local validation. A biased system exported into such a context may face neither the regulatory scrutiny nor the civil-society auditing infrastructure that constrained the cases examined above. Establishing the empirical extent of algorithmic bias in these settings, and building the institutional capacity to audit and contest it, is an urgent agenda that this paper can only flag and that we commend to future research.

We conclude with a normative claim: the deployment of biased generative AI in high-stakes domains is not merely a technical failure but an ethical and political one. Redressing it requires not only better algorithms but stronger institutions, more inclusive participatory processes, a willingness to refuse technologies that cannot be made just, and a fundamental commitment to the equal dignity of all persons affected by these systems.

References

Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30-56.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT '21, 610-623.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1-15.

Cho, J., Zala, A., & Bansal, M. (2023). DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. Proceedings of ICCV 2023.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. Proceedings of KDD 2017, 797-806.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 1989(1), 139-167.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of ITCS 2012, 214-226.

European Parliament. (2024). Artificial Intelligence Act. Official Journal of the European Union, L 2024/1689.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daum'e III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Haim, A., Salinas, A., & Nyarko, J. (2024). What's in a name? Auditing large language models for race and gender bias (arXiv:2402.14875). arXiv.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of FAccT 2019, 220-229.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

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.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of KDD 2016, 1135-1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of FAccT 2019, 59-68.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of AIES 2020, 180-186.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 1-17.

Weidinger, L., et al. (2021). Ethical and social risks of harm from language models (arXiv:2112.04359). arXiv.

Wilson, K., & Caliskan, A. (2024). Gender, race, and intersectional bias in resume screening via language model retrieval. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. Proceedings of AIES 2018, 335-340.