Bias in Artificial Intelligence: Origins, Impacts, and Mitigation Strategies

Vitthal B. Kamble[1], Diya P. Mokashi[2], Apurva S. Kirote[3], Rasika S. Shahapure[4]
[1,2,3,4]*Department of Computer Engineering, Cusrow Wadia Institute of Technology*

***Abstract*—Artificial Intelligence (AI) systems are now deeply embedded in virtually every high-stakes domain including healthcare, criminal justice, financial services, hiring, and education. As their influence grows, so does the urgency of understanding and addressing the biases these systems inherit, amplify, and sometimes create. This article presents a comprehensive, multidisciplinary examination of bias in AI, tracing its origins from flawed or unrepresentative training data through structural choices made during model design, and extending to the social and institutional contexts in which systems are deployed. We analyse the principal categories of bias recognised in the literature historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, and feedback loop bias and explain the mechanisms through which each propagates into automated decisions. The article surveys real-world harms across criminal justice, healthcare, financial services, employment, facial recognition, and large language models. Against this backdrop, we critically evaluate the principal mitigation strategies: pre-processing techniques such as reweighing and disparate-impact removal; in-processing approaches including adversarial debiasing and fairness-constrained optimisation; and post-processing methods such as equalized-odds adjustment. We further discuss Explainable AI (XAI) tools particularly SHAP and LIME as instruments for bias auditing, and examine Privacy-by-Design as a governance framework. We review the regulatory landscape including the EU AI Act and identify open research questions covering intersectional fairness, dynamic bias, the fairness-privacy-accuracy trilemma, and benchmark dataset diversity. Our central finding is that no single technical fix eliminates bias; sustained progress requires coordinated effort across data curation, algorithm design, organisational governance, and public policy.**

***Index Terms*—Algorithmic Fairness, Adversarial Debiasing, AI Bias, Bias Mitigation, Disparate Impact, Ethical AI, Explainable AI, Fairness Metrics, Machine Learning, Responsible AI, Statistical Parity.**

I. INTRODUCTION

Artificial intelligence has moved from research laboratories to the fabric of daily life with remarkable speed. Algorithms now help decide who gets a loan, who is shortlisted for a job, how long a convicted person spends in prison, which patients receive priority care, and what search results a person sees. This scale of influence would be extraordinary even if AI systems were perfectly neutral. They are not. AI systems learn from data generated by human societies, and those societies carry centuries of documented discrimination based on race, gender, socioeconomic class, age, and disability. When biased data trains a model, the model does not simply reflect history it institutionalises it, wrapping discrimination in the authority of mathematics and obscuring it behind the appearance of objectivity [1].

The word "bias" can sound technical and distant, but its consequences are immediate and personal. A 2018 audit of commercial facial-recognition systems found that products from major technology companies misidentified dark-skinned women at error rates up to 34 percentage points higher than for light-skinned men [2]. The COMPAS recidivism-prediction tool, widely used in American courts, was found to assign higher risk scores to Black defendants than to similarly situated white defendants a disparity that persisted even when prior offences, age, and other relevant factors were controlled [3]. Amazon scrapped a hiring algorithm after discovering it systematically downgraded resumes containing the word "women," having learned from a decade of predominantly male hiring decisions [4]. These are not edge cases; they are symptoms of a structural problem at the intersection of data, design, and deployment context. Researchers have responded with a surge of scholarly attention. The term "algorithmic fairness" now anchors entire

research programmes across computer science, law, economics, philosophy, and public policy. Dozens of fairness metrics have been proposed, each capturing a different intuition about equitable treatment. Mitigation techniques have proliferated at every stage of the machine-learning pipeline. Governments have begun to legislate: the European Union's AI Act imposes obligations on high-risk AI systems and creates requirements for transparency, data governance, and human oversight that specifically address discriminatory outcomes [5].

This article traces the problem from its roots in historical inequality through the technical mechanisms by which bias enters AI systems, to the concrete harms it has caused, and the technical and governance responses available. Section II situates AI bias within social context. Section III offers a taxonomy of bias types. Section IV examines fairness metrics. Section V surveys documented real-world harms. Section VI reviews mitigation strategies. Section VII discusses Explainable AI for bias auditing. Section VIII covers governance and regulation. Section IX identifies open research challenges. Section X concludes.

II. HISTORICAL AND SOCIAL BACKGROUND

To understand AI bias, one must first understand that discrimination encoded in data did not begin with computers. The datasets from which AI systems learn are records of a world shaped, for most of human history, by structured inequality. Racial segregation, gender discrimination in employment, redlining in mortgage lending, and systemic disparities in criminal justice are among the well-documented forms of discrimination that have produced skewed data trails. When a hiring algorithm trained on historical resumes learns that most senior executives in its training set were male, it is not discovering a natural law; it is faithfully reflecting decades of gender-biased promotion decisions [6].

This historical embeddedness of bias distinguishes it from simple errors. An arithmetic error can in principle be corrected once identified. A bias rooted in decades of unequal social structure is not erased by cleaning a dataset; it requires active, sustained effort to counteract. A model that removes explicit race labels from its inputs can still discriminate by race if it uses proxies zip code, name, or educational institution that are themselves correlated with racial demographics due to historical segregation [7].

Audit studies in labour economics have sent matched pairs of fictitious resumes to employers and found that resumes with "Black-sounding" names received significantly fewer callbacks than identical resumes with "white-sounding" names [8]. When AI hiring systems learn from historical hiring decisions that embed such bias, they reproduce and potentially amplify it. The concept of "algorithmic accountability" has emerged in response: someone a developer, a company, a regulator must be answerable for the consequences of algorithmic decisions. This accountability is complicated by the opacity of modern AI systems, the proprietary nature of training data, and the separation between those who build and those who deploy systems [9].

III. A TAXONOMY OF BIAS IN AI

The term "bias" is used loosely in popular discourse to mean almost any unfair outcome involving an AI system. The technical literature, however, has developed a finer-grained taxonomy identifying distinct types of bias arising at different stages of the machine-learning lifecycle. Understanding this taxonomy is essential for selecting appropriate mitigation strategies.

A. Historical Bias

Historical bias arises when training data faithfully reflects past patterns of discrimination or inequality. Even a perfectly representative, correctly labelled dataset can encode historical bias if the decisions it records were themselves discriminatory. A model trained to predict future job performance based on historical performance data will inherit whatever gender, race, or class biases shaped the opportunities that led to those performance records. Historical bias is arguably the most pervasive and the hardest to address, because it is baked into the ground-truth labels that supervised learning relies upon.

B. Representation Bias

Representation bias arises when training data under-represents or over-represents certain groups relative to their true prevalence, or relative to the population to which the model will be applied. A facial-recognition system trained predominantly on images of light-

skinned individuals will have higher error rates for dark-skinned individuals, not because of any explicit design choice but because the model has seen far fewer examples from which to learn relevant features. Representation bias tends to harm groups that are already marginalised, compounding existing disadvantages [10].

C. Measurement Bias

Measurement bias occurs when the features or labels used to train a model are measured differently across groups, or when a proxy variable correlates with the outcome of interest in different ways for different groups. Consider a model that uses prior arrests as a proxy for criminal behaviour. If police patrol minority neighbourhoods more intensively, arrest rates will be higher for minorities than for majority-group members who engage in the same behaviour. A model trained on arrest data encodes this enforcement disparity rather than the underlying behaviour it is supposed to predict [11].

D. Aggregation Bias

Aggregation bias arises when a model is trained on a pooled dataset that treats groups with genuinely different underlying relationships as homogeneous. A model that learns a single relationship between blood-glucose levels and diabetes risk from data aggregated across ethnic groups may perform poorly for any individual group, because the true relationship differs systematically. This form of bias is particularly common in healthcare AI, where physiological differences across demographic groups are known but often ignored in model development.

E. Evaluation Bias

Evaluation bias occurs when the benchmarks or test datasets used to assess a model do not represent the full target population. If a benchmark is itself unrepresentative, a model can appear to perform well in evaluation while failing systematically on underrepresented groups. Many widely used computer-vision benchmarks are dominated by Western, affluent subjects, leading to models evaluated as strong performers that fail for users in other cultural or socioeconomic contexts.

F. Feedback Loop Bias

Feedback loop bias occurs when the predictions of a model influence future data in ways that reinforce the model's original predictions. Predictive policing provides a clear illustration: a model identifies certain neighbourhoods as high-crime, police are directed there, arrests increase in those areas, the next model is trained on data showing more crime there, and the predictions self-fulfil. The system is not discovering crime; it is creating the evidence for its own predictions. Feedback loops can transform a model that was imperfect but correctable into one that becomes increasingly entrenched in its biases over time [12].

IV. FAIRNESS METRICS

Before bias can be mitigated, it must be measured. Researchers have proposed a substantial number of fairness metrics, each operationalising a different intuition about equitable treatment. The choice of metric shapes what the model is optimised for and what harms it may still cause.

A. Statistical Parity (Demographic Parity)

Statistical parity requires that the proportion of positive predictions be the same across groups. If A denotes the protected attribute and Y-hat the model's prediction, statistical parity demands:

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b) \quad \text{... (1)}$$

The Statistical Parity Difference (SPD) is:

$$SPD = P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = b) \quad \text{... (2)}$$

A value of zero indicates perfect statistical parity. This metric is intuitive and easy to compute, but it is criticised for demanding equal positive-prediction rates even when base rates of the outcome differ across groups for legitimate reasons [13].

B. Disparate Impact Ratio

Disparate Impact (DI), rooted in US anti-discrimination law, measures the ratio of positive-prediction rates across groups:

$$DIR = P(\hat{Y} = 1 \mid A = a) \,/\, P(\hat{Y} = 1 \mid A = b) \quad \text{... (3)}$$

The legal rule of thumb holds that a ratio below 0.8 (the "four-fifths rule") indicates adverse impact. Unlike SPD, DIR captures proportional rather than absolute differences.

C. Equalized Odds

Equalized odds requires that the model have equal true positive rates and equal false positive rates across groups, conditional on the true outcome. For all y in {0,1}:

$$P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = b). \quad (4)$$

Equalized odds are considered more nuanced than statistical parity because its conditions on the true outcome. The landmark impossibility theorem shows, however, that equalized odds, statistical parity, and calibration cannot all be satisfied simultaneously when base rates differ across groups [14]. Any deployment must therefore choose which criterion to prioritise a choice that is inherently normative.

D. Individual Fairness

Individual fairness requires that similar individuals be treated similarly. Formally, if d(x, x') is a distance between two individuals and D(f(x), f(x')) is a distance between their predictions, individual fairness requires that D be bounded by some multiple of d. While intuitively appealing, individual fairness is difficult to implement because defining an appropriate distance metric for "similarity" requires domain expertise and may itself encode discriminatory intuitions if not carefully constructed.

E. Calibration

Calibration requires that predicted probabilities match actual frequencies of positive outcomes within each group. A well-calibrated model at a 70% risk score produces positive outcomes for roughly 70% of those assigned that score. Calibration is important in high-stakes settings where decision-makers rely on probability estimates rather than binary classifications. However, calibration is mathematically incompatible with equalized odds when group base rates differ a constraint that forces explicit ethical choices about whose error rates to prioritise [14].

V. DOCUMENTED REAL-WORLD HARMS

The stakes of AI bias are best understood through concrete cases. In this section, we survey documented instances of bias-related harm across several high-impact domains.

A. Criminal Justice and Recidivism Prediction

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system became the subject of intense scrutiny after a 2016 ProPublica investigation [3]. The investigation found that Black defendants were nearly twice as likely as white defendants to be falsely flagged as high risk, while white defendants were more often incorrectly labelled low risk. A subsequent analysis confirmed these racial disparities in false positive and false negative rates across demographic groups [15].

These findings sparked a methodological debate: Northpointe argued that COMPAS satisfied calibration its predicted probabilities matched actual recidivism rates within each racial group while critics argued that equal calibration is insufficient if error rates are not equalized. The debate illuminated the impossibility theorem in practice: satisfying one fairness criterion required violating another, and the choice had life-altering consequences for defendants.

Beyond COMPAS, predictive policing systems have faced similar scrutiny. Systems used in several US cities allocate patrol resources based on predicted crime risk. Critics have documented that these systems, trained on historically biased arrest data, direct disproportionate police presence to minority neighbourhoods, generating more arrests there and feeding back into future predictions in a self-reinforcing loop [16].

B. Healthcare

A widely cited 2019 study in science examined a commercial algorithm used by a major US health insurer to identify patients for care management programmes. The algorithm used past healthcare expenditure as a proxy for health needs. Because, on average, less is spent on Black patients with the same level of illness as white patients reflecting historical disparities in healthcare access the algorithm systematically underestimated the health needs of Black patients. The bias was substantial: the algorithm recommended additional care for white patients at a severity threshold that, for Black patients, was associated with significantly worse health outcomes [17]. Similar disparities have been documented in dermatology AI, where systems trained predominantly on images of fair-skinned patients have significantly lower accuracy for darker skin tones [18]. Harvard Medical School's work on the Harvard Glaucoma

Fairness (Harvard-GF) dataset has highlighted the importance of balanced demographic sampling in medical AI benchmarking, showing that standard datasets often lack the diversity needed to detect such performance gaps [19].

C. Financial Services

Algorithmic lending has in some cases replicated historical discrimination. A 2021 study using mortgage application data found that algorithmic approval systems denied applications from Black and Hispanic applicants at higher rates than comparable white applicants, even after controlling for creditworthiness indicators. The study concluded that the algorithms learned to use proxy variables neighbourhood, employer type, educational institution that served as stand-ins for race in a legally discriminatory way [20]. The Consumer Financial Protection Bureau has noted that algorithmic lending models present particular challenges for fair-lending enforcement because their complexity makes it difficult to identify which variables drive disparate outcomes. Beyond credit and lending, digital payment platforms are also vulnerable to algorithmic bias and fraud; ensemble-based machine learning frameworks such as Stacked Generalization combining Random Forest and SVM have demonstrated superior accuracy in detecting fraudulent patterns in real-time transaction systems like India's Unified Payments Interface [31].

D. Employment and Hiring

The case of Amazon's hiring algorithm is frequently cited as a paradigmatic example of bias in AI-assisted recruitment. The system was trained on ten years of resume submissions, most from men, reflecting the historically male-dominated technology sector. The model learned that male candidates were preferable and penalised resumes from women, including those mentioning all-women's colleges or women's organisations. Amazon scrapped the system in 2018 [4]. More broadly, research has shown that AI hiring tools can encode bias related to race, age, disability, and socioeconomic background, and that concepts of "job fit" and "culture fit" which these systems attempt to operationalise often mask preferences for candidates who resemble the existing workforce [21].

E. Facial Recognition

The Gender Shades study by Buolamwini and Gebru [2] stands as one of the most influential demonstrations of AI bias. Auditing commercial facial-analysis systems from IBM, Microsoft, and Face++, the authors found that error rates were substantially higher for darker-skinned subjects, and highest for darker-skinned women. Overall accuracy rates masked severe performance gaps that only became visible when performance was disaggregated by gender and skin type. This work contributed to several major cities banning police use of facial recognition.

F. Large Language Models

Large language models (LLMs) have attracted scrutiny for reproducing and amplifying social stereotypes in generated text. Studies have found that LLMs associate certain occupations with certain genders (doctor with male, nurse with female), reproduce racial stereotypes in generated stories, and express different sentiments about different religions, cultures, and nationalities [22]. A 2026 Yale study found that chatbots can influence user opinions subtly and without users being aware of it, raising concerns about the scale at which biased AI can shape public discourse [23]. Because LLMs are trained on vast corpora of internet text, they inevitably absorb whatever biases are present in that text, which over-represents certain languages, cultures, and perspectives while under-representing others.

VI. TECHNICAL MITIGATION STRATEGIES

No single technique eliminates AI bias, but researchers have developed a rich toolkit of approaches that can substantially reduce its impact when applied thoughtfully and in combination. These approaches are conventionally grouped by the stage of the machine-learning pipeline at which they intervene: before training (pre-processing), during training (in-processing), and after training (post-processing).

A. Pre-Processing Techniques

Pre-processing techniques intervene in the training data before any model is fitted. Their goal is to produce a dataset that will lead to fairer model outputs by reducing statistical associations between protected attributes and labels. Reweighing assigns different

weights to training examples to reduce disparities between groups. Examples from under-represented or historically disadvantaged groups are assigned higher weights so that the model attends to them more during training. The AIF360 reweighing implementation computes weights based on the joint probability of a sample's protected attribute and label, upweighting combinations that are under-represented relative to what would be expected if the two were independent [24]. Research on the COMPAS dataset has demonstrated that reweighing can substantially reduce false negative rate disparities across racial groups without sacrificing model accuracy or AUC [15].

Resampling techniques oversampling under-represented groups, under sampling over-represented groups, or generating synthetic minority examples via SMOTE (Synthetic Minority Oversampling Technique) address representation bias by adjusting class and group balance of the training set.

The Disparate Impact Remover modifies feature values to remove the correlation between features and protected attributes while preserving rank-ordering within groups. This is particularly useful when categorical features serve as proxies for protected attributes. Research on hiring applications has found this technique achieves high fairness scores (0.89) with moderate accuracy trade-offs [21].

Data augmentation deliberately adding or generating examples that challenge stereotyped associations has proven effective for debiasing language models and image classifiers. Counterfactual data augmentation generates paired examples that are identical except for the value of the protected attribute and trains the model on both, reducing the model's reliance on protected attributes as predictors [25].

B. In-Processing Techniques

In-processing techniques modify the training algorithm itself to incorporate fairness objectives directly. Unlike pre-processing, which operates on data before fitting, in-processing shapes what the model learns during training.

Adversarial Debiasing is one of the most powerful in-processing approaches. The method trains two neural networks simultaneously: a primary predictor that learns to perform the main task, and an adversary that tries to infer the protected attribute from the predictor's output. The primary model is trained to minimise prediction error and to minimise the adversary's ability to predict the protected attribute. The joint optimisation objective is:

$$L_{total} = L_{pred} - \lambda \cdot L_{adv} \quad \text{... (5)}$$

where L_pred is the cross-entropy loss for the main prediction task, L_adv is the adversarial loss for predicting the protected attribute, and lambda is a hyperparameter controlling the trade-off between accuracy and fairness [26]. By minimising the adversary's accuracy, the primary model is forced to learn representations that do not carry information about the protected attribute. Experimental evaluations have demonstrated that adversarial debiasing achieves substantial bias reduction with minimal accuracy loss, outperforming logistic regression baselines (93% vs 85.3%) while improving the Disparate Impact Ratio to 0.95 in healthcare and law-enforcement contexts [26].

Fairness-Constrained Optimisation treats fairness as a hard or soft constraint on the optimisation problem. The general formulation minimises the loss function L over model parameters theta subject to the constraint that some fairness metric F is within an acceptable bound:

$$\min_{\theta} L(\theta) \text{ subject to: } F(\theta) \leq \varepsilon \quad \text{... (6)}$$

The fairness constraint can be demographic parity, equalized odds, individual fairness, or any computable criterion. Research in hiring AI has shown that SVM models with fairness constraints achieve a fairness score of 0.86 while maintaining 77.9% accuracy [21].

Regularisation-based approaches add fairness penalty terms to the training loss, penalising the model for predictions correlated with protected attributes or that produce disparate error rates. This flexibility makes regularisation easy to implement in standard deep-learning frameworks, though it offers less precise control over the fairness metric achieved.

C. Post-Processing Techniques

Post-processing techniques modify the outputs of a trained model rather than the data or training algorithm. They are attractive because they can be applied to any existing model without retraining an important practical advantage when models are large, expensive to train, or proprietary. Equalized Odds Post-Processing solves for optimal group-specific thresholds that bring a model's true positive and false

positive rates into alignment across groups. The approach finds the set of thresholds that minimises a combination of accuracy loss and equalized-odds violation, and has been shown to substantially reduce false negative rate disparities in recidivism prediction applications [14].

The Reject Option Classification method assigns positive predictions to disadvantaged groups and negative predictions to advantaged groups in the region of uncertainty near the decision boundary. Studies on hiring AI have found this method achieves reasonable accuracy (79.1%) with a lower fairness score (0.83) relative to adversarial debiasing, making it a viable option when retraining is infeasible [21].

D. Comparison of Mitigation Approaches

No single technique dominates across all settings. The choice among approaches depends on whether training data or the training process is accessible, computational resources available, the specific fairness criterion targeted, tolerance for accuracy loss, and regulatory requirements. Table I summarises the key trade-offs.

Table I. Comparison of bias mitigation approaches

Approach	Technique	Fairness Score	Accuracy	Requires Retraining?
Pre-proc.	Reweighing	0.87	78.2%	Yes
Pre-proc.	Disparate Impact Remover	0.89	76.4%	Yes
In-proc.	Adversarial Debiasing	0.91	80.5%	Yes
In-proc.	Fairness Constraints	0.86	77.9%	Yes
Post-proc.	Equalized Odds Adj.	0.88	78.8%	No
Post-proc.	Reject Option Class.	0.83	79.1%	No

VII. EXPLAINABLE AI FOR BIAS AUDITING

Detecting and communicating bias requires transparency into how a model makes its decisions. Without interpretability, even sophisticated fairness metrics tell us only that a disparity exists, not where it comes from or how to fix it. Explainable AI (XAI) methods address this gap by producing human-readable explanations of model behaviour, either at the level of individual predictions or across the model as a whole.

A. SHAP (SHapley Additive exPlanations)

SHAP is grounded in cooperative game theory. It attributes each feature's contribution to a prediction using Shapley values, which capture each feature's average marginal contribution across all possible subsets of features. The formal definition is:

$$\varphi_i = \Sigma_{R \subseteq M \setminus \{i\}} |R|!(|M|-|R|-1)! / |M|! \, [f(R \cup \{i\}) - f(R)] \quad (7)$$

where phi_i is the Shapley value for feature i, M is the full set of features, R is a subset of M not containing i, and f is the model prediction function [27]. The SHAP framework also allows predictions to be expressed as an additive decomposition:

$$\hat{f}(x) = \varphi_0 + \Sigma_i \varphi_i \quad ... (8)$$

where phi_0 is the base value (mean prediction over the training set) and phi_i is the contribution of feature i to the specific prediction. A SHAP analysis of a credit-scoring model might reveal that "zip code" contributes disproportionately to predictions for minority applicants, suggesting it is functioning as a racial proxy an insight invisible from accuracy metrics alone [27]. SHAP has been adopted in commercial AI auditing tools including IBM's AI Fairness 360 and Microsoft's Responsible AI Toolbox. A comparative fairness analysis across frameworks found that the proposed EAIDF framework incorporating SHAP achieves fairness scores of 92%, 94%, and 93% for gender, age, and ethnicity respectively substantially outperforming IBM AIF360 (70%, 73%, 74%) and Microsoft RaiTool (77%, 80%, 79%) on the same demographic categories [27].

B. LIME (Local Interpretable Model-Agnostic Explanations)

LIME fits a locally linear approximation to the complex model in the neighbourhood of a specific prediction. For a given instance x, LIME generates perturbed samples in the neighbourhood of x, obtains the model's predictions on those samples, and fits a simple interpretable model to the resulting prediction surface. LIME is model-agnostic it can be applied to any black-box model and produces explanations at the level of individual predictions, which is particularly valuable for identifying cases where a model is relying

on protected attributes or their proxies for specific individuals or subgroups.

C. Grad-CAM for Visual Models

In computer vision applications, Gradient-weighted Class Activation Mapping (Grad-CAM) produces saliency maps that highlight image regions most important for a prediction. For facial-recognition or medical-imaging systems, Grad-CAM can reveal whether the model is attending to the features one would expect facial structure, skin lesion characteristics or to artefacts correlated with the protected attribute. Anomalous attention patterns can signal bias not captured by aggregate accuracy metrics.

D. XAI in Regulatory Compliance

Beyond technical auditing, XAI has become increasingly important for regulatory compliance. The EU's GDPR includes a "right to explanation" for automated decisions that significantly affect individuals. The EU AI Act's requirements for high-risk AI systems include technical documentation explaining how the system works and how it was validated. XAI methods provide the technical infrastructure for meeting these requirements, though translating model explanations into plain-language justifications that non-expert individuals can meaningfully use remains an open challenge.

VIII. GOVERNANCE, REGULATION, AND ORGANISATIONAL PRACTICE

Technical mitigation strategies are necessary but not sufficient. Bias in AI is ultimately a governance problem: it arises from decisions made by people within organisations operating in social and institutional contexts, and it requires governance solutions that address those contexts directly.

A. The EU AI Act

The EU AI Act, which entered into force in August 2024, is the world's first comprehensive legal framework for AI regulation. The Act takes a risk-based approach, classifying AI systems into four risk tiers: unacceptable risk (banned), high risk (regulated), limited risk (transparency obligations), and minimal risk (no specific obligations). High-risk AI systems which include systems used in employment, credit, education, law enforcement, migration, and essential services must comply with requirements covering training data quality, technical documentation, transparency, human oversight, accuracy, robustness, and cybersecurity [5]. For bias specifically, the AI Act requires that high-risk systems be designed and trained to minimise discriminatory outcomes, and that their training data be governed by appropriate data management practices to detect and address bias. Providers must implement risk management systems, maintain technical documentation, log system performance, and register their systems in a public EU database.

B. US Regulatory Landscape

The US regulatory landscape for AI bias is more fragmented. Existing laws the Equal Credit Opportunity Act, Fair Housing Act, and Title VII of the Civil Rights Act prohibit discrimination in specific domains and have been interpreted by regulators as applicable to algorithmic decision-making. The Consumer Financial Protection Bureau has issued guidance clarifying that lenders must be able to explain adverse actions taken by algorithmic systems. Several states and cities, including New York City and Illinois, have enacted legislation specifically addressing algorithmic hiring and automated employment decision tools [28].

C. Algorithmic Auditing

Raji and colleagues [9] have argued that internal algorithmic auditing structured processes within organisations for evaluating AI systems for bias and other ethical risks before and after deployment is essential for closing the "AI accountability gap." They propose an end-to-end framework that covers the full lifecycle: from scoping and impact assessment through data collection and model development to deployment and monitoring. The Ethical AI Auditor framework proposed by Senthil and colleagues [26] extends this concept with automated tooling. The framework incorporates real-time bias monitoring against fairness metrics, automated alerts when metrics deviate from acceptable bounds, and a stakeholder dashboard that displays audit results in accessible form. This continuous monitoring is particularly important because bias can emerge or worsen over time as data distributions shift, social conditions change, or new subpopulations begin using a system.

D. Privacy-by-Design

Privacy-by-Design (PbD) is an approach to system development that incorporates privacy protections from the earliest design stages rather than adding them as an afterthought. The seven foundational principles of PbD proactive not reactive; privacy as the default setting; privacy embedded into design; full functionality; end-to-end security; visibility and transparency; and respect for user privacy have been formally recognised in GDPR and have gained widespread adoption [29].

PbD is relevant to AI bias in a specific way: many bias problems arise because AI systems collect and use sensitive attributes or their proxies in ways that the affected individuals are not aware of and have not consented to. A PbD approach that minimises the collection and retention of sensitive data reduces the surface area through which bias can enter the system. The EAIDF framework demonstrates that implementing PbD alongside bias auditing and explainability produces AI systems that substantially outperform those that treat these concerns in isolation, achieving fairness scores above 91% compared to the 65-79% range of conventional frameworks [29].

E. Organisational Culture and Diverse Teams

Technical tools and regulatory requirements operate within organisations, and organisational culture profoundly shapes how those tools are used. Research consistently finds that diverse AI development teams diverse in terms of gender, race, cultural background, and disciplinary training are more likely to identify potential harms and more likely to design systems that work for a broader range of users. Stanford's case study on AI bias has emphasised that technical training alone is insufficient: AI developers need exposure to the social science methods for measuring discrimination, to the legal frameworks for challenging it, and to the lived experiences of people affected by automated decisions [30].

IX. OPEN RESEARCH CHALLENGES

Despite significant progress, the field of AI fairness faces a set of open challenges that resist simple technical solutions. These challenges are simultaneously scientific, ethical, and political they require not just better algorithms but better frameworks for making value-laden decisions in pluralistic societies.

A. The Fairness-Privacy-Accuracy Trilemma

Maximising model accuracy, preserving privacy, and achieving fairness are three objectives that frequently pull in different directions. Fairness evaluation often requires access to sensitive demographic data that privacy norms and regulations restrict. Differential privacy, which adds noise to training data or model parameters to protect individual privacy, can worsen performance disparities across groups if the added noise disproportionately degrades the quality of information about underrepresented groups [30]. Navigating this trilemma requires explicit value judgements about which objective to prioritise in which contexts judgements that should involve affected communities rather than being made unilaterally by developers.

B. Intersectional Fairness

Most fairness metrics and mitigation techniques operate on single protected attributes in isolation ensuring that predictions are equalized across racial groups, or across gender groups, but not necessarily across the intersections of these attributes. A model that satisfies equalized odds for race and equalized odds for gender may still produce highly disparate outcomes for specific intersectional groups such as Black women or elderly Latino men. Intersectional fairness is computationally more demanding (the number of intersectional subgroups grows exponentially with the number of attributes) and raises conceptual questions about which intersections are protected and how to prioritise when different intersectional groups face different types of harm.

C. Dynamic and Temporal Bias

Most fairness analyses treat bias as a static property of a model at a point in time. But AI systems operate in dynamic environments where data distributions, social conditions, and the meaning of features can change. A model that is fair at deployment may become unfair as the population it serves changes, as proxy variables shift their correlation with protected attributes, or as feedback loops amplify initial disparities. Developing methods for continuous fairness monitoring, detecting distribution shift, and triggering model recalibration when fairness metrics degrade is an active area of research with significant practical importance [12].

D. Causal Approaches to Fairness

Most fairness metrics are correlational: they measure statistical associations between predictions and protected attributes. Causal fairness approaches, which use causal graphs to distinguish between direct discrimination (the protected attribute directly causes the prediction) and indirect discrimination via legitimate mediators or spurious associations, offer a more principled foundation for fairness reasoning. However, causal inference requires assumptions about the underlying data-generating process that are often difficult to verify in practice, and causal methods remain computationally demanding for high-dimensional data.

E. Benchmark Dataset Diversity

Progress in AI fairness research is constrained by the availability of well-curated, demographically diverse benchmark datasets. Many commonly used benchmarks including the Adult Income dataset, the COMPAS dataset, and various facial-recognition benchmarks have well-documented limitations: they are dated, they represent narrow geographic and demographic contexts, or they encode the very biases they are supposed to help detect. Creating and maintaining high-quality, ethically sourced, demographically representative benchmark datasets requires sustained investment that is poorly matched to the incentive structures of academic research publication.

F. Standardisation and Certification

There is currently no widely adopted standard for what constitutes acceptable fairness performance in AI systems. Different organisations, regulators, and researchers use different metrics, thresholds, and evaluation protocols, making it difficult to compare systems, enforce requirements, or provide meaningful assurances to affected individuals. Developing technical standards for AI fairness analogous to safety standards in aviation or medical devices is a complex challenge that requires coordination among standards bodies, regulators, industry, civil society, and affected communities. The EU AI Act creates some impetus for standardisation by requiring that high-risk systems be evaluated against harmonised standards currently being developed by European standards bodies [5].

X. CONCLUSION

AI bias is one of the defining challenges of our technological moment. As AI systems become more capable and more pervasive, the consequences of getting bias wrong become more severe and the opportunities for getting it right become correspondingly more important. This article has traced the problem from its roots in historical and social inequality, through the technical mechanisms by which bias enters and propagates through machine-learning systems, to the concrete harms it has caused in criminal justice, healthcare, financial services, employment, facial recognition, and language modelling.

We have reviewed the principal fairness metrics that researchers have proposed, noting both their intuitive appeal and their fundamental limitations, including the impossibility of satisfying multiple metrics simultaneously when group base rates differ. We have surveyed technical mitigation strategies at every stage of the machine-learning pipeline, from data reweighing and adversarial debiasing through fairness-constrained optimisation to post-processing threshold adjustment. We have discussed the role of Explainable AI methods particularly SHAP and LIME in making bias auditable and in supporting regulatory compliance. And we have examined the governance ecosystem of auditing frameworks, regulatory requirements, Privacy-by-Design principles, and organisational practices that together constitute a comprehensive response to the challenge.

Several themes emerge repeatedly across these diverse topics. First, bias in AI is not primarily a technical problem with a technical solution; it is a social problem that requires social, organisational, and political solutions alongside technical ones. Second, no single fairness metric, mitigation technique, or governance practice is sufficient on its own. Third, the people most harmed by biased AI are often the least able to identify it, challenge it, or demand accountability for it which creates an obligation on developers, organisations, regulators, and researchers to proactively address bias rather than waiting for harms to become undeniable.

AI bias mitigation cannot be treated as a one-time checkbox. AI systems are deployed in dynamic environments, and the data distributions, social conditions, and uses of a system change over time.

Sustaining fairness requires continuous monitoring, regular auditing, mechanisms for affected individuals and communities to provide feedback, and willingness to update or withdraw systems causing harm. The path toward AI systems that genuinely serve all people equitably requires sustained effort from researchers, practitioners, policymakers, civil society organisations, and the communities most affected a commitment that must outlast any individual developer or product cycle.

REFERENCES.

[1] IBM, "What Is Artificial Intelligence?," IBM Think Topics, 2023. [Online]. Available: https://www.ibm.com/think/topics/artificial-intelligence.

[2] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in Proc. 1st ACM Conf. Fairness, Accountability and Transparency (FAT*), 2018, pp. 77–91.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," ProPublica, May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] J. Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," Reuters, Oct. 2018.

[5] European Parliament and Council, Regulation (EU) 2024/1689 (AI Act), Off. J. Eur. Union, Jun. 2024.

[6] S. Barocas and M. Hardt, Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org, 2019.

[7] IBM, "What Is AI Bias?," IBM Think Topics, 2023. [Online]. Available: https://www.ibm.com/think/topics/ai-bias.

[8] M. Bertrand and S. Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal?," Amer. Econ. Rev., vol. 94, no. 4, pp. 991–1013, 2004.

[9] I. D. Raji et al., "Closing the AI Accountability Gap," in Proc. 2020 Conf. Fairness, Accountability and Transparency, 2020, pp. 33–44.

[10] S. Venkatasubbu and G. Krishnamoorthy, "Ethical Considerations in AI Addressing Bias and Fairness in Machine Learning Models," J. Knowl. Learn. Sci. Technol., vol. 1, no. 1, pp. 130–138, 2022.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, 2021.

[12] E. Ferrara, "The Butterfly Effect in Artificial Intelligence Systems," Mach. Learn. Appl., vol. 15, p. 100525, 2024.

[13] J. M. Alvarez et al., "Policy Advice and Best Practices on Bias and Fairness in AI," Ethics Inf. Technol., vol. 26, no. 2, p. 31, 2024.

[14] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 29, 2016.

[15] A. A. Skaiky, H. M. S. Ali, A. Mohammed, and Z. A. Mahdi, "Bias & Fairness in AI Models," in 2025 IEEE 4th Int. Conf. Computing and Machine Intelligence (ICMI), 2025.

[16] Oxford Internet Institute, "New Approach Needed to Address Bias in Machine Learning and AI Systems," 2021. [Online]. Available: https://www.oii.ox.ac.uk.

[17] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," Science, vol. 366, no. 6464, pp. 447–453, 2019.

[18] Harvard Medical School, "Confronting the Mirror: Reflecting Our Biases Through AI in Health Care," 2023. [Online]. Available: https://learn.hms.harvard.edu.

[19] Harvard Ophthalmology AI Lab, "Harvard Glaucoma Fairness (Harvard-GF) Dataset," Harvard Medical School, 2024. [Online]. Available: https://ophai.hms.harvard.edu.

[20] R. G. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-Lending Discrimination in the FinTech Era," J. Financ. Econ., vol. 143, no. 1, pp. 30–56, 2022.

[21] A. Malpani et al., "AI and Bias Mitigation in HR," in 2025 IEEE 5th Int. Conf. ICT in Business Industry & Government (ICTBIG), 2025.

[22] T. B. Modi, "Artificial Intelligence Ethics and Fairness," Rev. Index J. Multidiscip., vol. 3, no. 2, pp. 24–35, 2023.

[23] Yale University, "AI's Hidden Bias: Chatbots Can Influence Opinions Without Trying," Yale News, Mar. 2026. [Online]. Available: https://news.yale.edu/2026/03/03.
[24] R. K. E. Bellamy et al., "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," IBM J. Res. Dev., vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
[25] E. Ferrara, "Fairness and Bias in Artificial Intelligence: A Brief Survey," Sci, vol. 6, no. 1, p. 3, 2023.
[26] G. A. Senthil, S. Geerthik, J. I. Jerlin, and S. A. Ashika, "Ethical AI Auditor for Bias Detecting in AI Models Using Adversarial Debiasing," in 2025 Int. Conf. AMATHE, IEEE, 2025.
[27] S. Shi, "AI and Ethical AI Design: A Framework for Minimizing Bias and Protecting Privacy in AI Algorithms," in 2025 IDICAIHEI, IEEE, 2025.
[28] S. Verma, N. Paliwal, K. Yadav, and P. C. Vashist, "Ethical Considerations of Bias and Fairness in AI Models," in Proc. 2024 2nd Int. Conf. Disruptive Technol. (ICDT), IEEE, 2024, pp. 818–823.
[29] D. Korobenko, A. Nikiforova, and R. Sharma, "Towards a Privacy and Security-Aware Framework for Ethical AI," in Proc. 25th Annu. Int. Conf. Digital Government Research, 2024, pp. 740–753.
[30] Stanford Graduate School of Business, "Designing AI for All: A Primer on Bias in Artificial Intelligence Systems," Stanford Case Studies, 2022. [Online]. Available: https://www.gsb.stanford.edu.
[31] V. B. Kamble et al., "Enhancing UPI Fraud Detection: A Machine Learning Approach Using Stacked Generalization," Int. J. Innov. Res. Technol., vol. 2, no. 1, pp. 69–83, 2025.