Pages: 1-13
Volume-IV, Issue-III (2023)
https://globalknowledgeacademy.com/

Algorithmic Accountability in High-Stakes Decision Systems: A Review of Fairness Metrics, Bias Mitigation Techniques, and Transparency Requirements across Healthcare, Criminal Justice, and Financial Services

Jasmin Lumacad

Xavier University,USA

Corresponding E-mail: jasminmalmislumacad@gmail.com

Abstract:

Artificial intelligence systems deployed in high-stakes decision contexts, including healthcare diagnosis, criminal justice risk assessment, credit underwriting, and employment screening, exert profound influence over individual life outcomes. The progressive deployment of these systems since the mid-2010s has revealed systematic patterns of discriminatory behavior rooted in biased training data, biased algorithm design, and inadequate transparency mechanisms that prevent affected individuals from understanding, contesting, or seeking redress for algorithmically mediated decisions that affect them adversely. This review surveys the body of knowledge accumulated on algorithmic fairness, bias mitigation, and AI transparency from the foundational contributions of the 2016 to 2022 period through recent practical implementation work. We systematically review the major fairness metric families, including group fairness, error-rate parity, calibration, and individual fairness, characterize the impossibility results that constrain their simultaneous satisfaction, and survey preprocessing, in-processing, and post-processing bias mitigation techniques documented in the peer-reviewed literature. We analyze transparency and explainability mechanisms including model cards, datasheets for datasets, and post-hoc explanation methods, assessing their adequacy against the accountability requirements of GDPR Article 22, the EU AI Act high-risk system provisions, and the NIST AI Risk Management Framework. Sector-specific analyses of healthcare, criminal justice, and financial services reveal that domain characteristics, including the nature of the protected attributes at risk, the reversibility of adverse decisions, and the regulatory instruments applicable, require tailored fairness and accountability frameworks rather than generic deployments of domain-agnostic metrics. Recent applied implementation work demonstrates that these frameworks are operationally deployable rather than theoretically aspirational.

Keywords: algorithmic fairness; bias mitigation; AI transparency; explainable AI; model accountability; demographic parity; equalized odds; fairness impossibility; healthcare AI; criminal justice AI

I. Introduction

The deployment of machine learning systems as decision-support tools in domains that materially affect individual welfare has accelerated from experiment to production infrastructure with a speed that has substantially outpaced the development of the governance frameworks, technical methods, and legal instruments needed to ensure that these systems behave equitably toward all affected populations. By the

late 2010s, machine learning systems were routinely determining which job applications would be reviewed by human recruiters, which individuals would be flagged for enhanced scrutiny by financial crime prevention systems, which defendants would receive higher risk scores in pre-trial release and sentencing recommendations, and which patients would receive additional preventive care resources from healthcare risk stratification tools. Each of these applications carries the potential for systematic disparate impact against groups defined by race, gender, age, disability status, and other protected characteristics — an impact that is both ethically problematic and, in many jurisdictions, legally impermissible [1].

The research community's engagement with algorithmic fairness grew dramatically following a series of high-profile exposés of discriminatory AI behavior in the mid-2010s. The ProPublica investigation demonstrating racial disparity in the COMPAS recidivism prediction tool, the documentation of gender discrimination in Amazon's internal resume screening system, and a landmark study in Science revealing that a widely deployed healthcare utilization management algorithm systematically underserved Black patients collectively drove a decade of intensive work on the theoretical foundations of algorithmic fairness, the taxonomy of bias sources, and the methods available to detect and mitigate discriminatory behavior [2]. By 2022, the field had produced a rich body of peer-reviewed work spanning fairness metrics and their relationships, impossibility theorems constraining their simultaneous satisfaction, and a toolkit of preprocessing, in-processing, and post-processing mitigation techniques [3].

Alongside fairness research, a parallel literature developed on AI transparency and explainability, motivated by the recognition that a fair AI system that cannot explain its decisions provides no recourse to individuals who receive adverse outcomes. The introduction of standardized transparency artifacts including model cards for model reporting and datasheets for datasets established documentation practices that translate the concept of AI accountability from a policy aspiration into an engineering discipline [4]. Explainable AI methods, including LIME, SHAP, and counterfactual explanation generation, provided post-hoc interpretation tools that can surface the features driving individual predictions without requiring access to the model's internal architecture [5]. The implementation of these fairness and transparency techniques in practice requires the systematic engineering approach to fair and transparent AI design described in Section 5.

The purpose of this review is to consolidate this body of knowledge for practitioners and researchers who need to deploy accountable AI systems in high-stakes domains, and to identify the gaps between the theoretical toolkit and the practical requirements of domain-specific accountability. Section 2 surveys AI bias sources and their taxonomy. Section 3 reviews fairness metrics and their relationships. Section 4 presents the impossibility results that constrain fairness metric satisfaction. Section 5 surveys bias mitigation techniques. Section 6 reviews transparency and explainability mechanisms. Section 7 analyzes domain-specific accountability requirements. Section 8 discusses governance frameworks and regulatory alignment. Section 9 concludes with research priorities.

II. Taxonomy of Bias in AI Systems

2.1 Sources of algorithmic bias

Bias in AI systems originates from multiple sources across the machine learning pipeline, and understanding the source of bias is prerequisite to selecting appropriate mitigation techniques. The most comprehensive taxonomy of bias sources in the machine learning context identifies historical bias,

representation bias, measurement bias, aggregation bias, evaluation bias, and deployment bias as the primary categories, noting that multiple bias types often co-occur in production systems and that their interactions can produce discriminatory outcomes even when each individual bias source appears minor in isolation [6].

Historical bias arises when training data encodes the discriminatory patterns of historical human decision-making. A hiring model trained on past recruitment decisions inherits the gender and racial biases of the human recruiters whose decisions constitute its training labels. A credit scoring model trained on historical loan performance inherits the neighborhood-level discrimination of historical lending practices. Historical bias is particularly insidious because it is invisible in the training data statistics: the data accurately represents historical outcomes, but those outcomes were themselves biased [7]. The Amazon recruitment system failure exemplifies historical bias at scale: the system, trained on a decade of predominantly male hire decisions, learned to downweight features associated with female candidates as implicitly predictive of non-selection.

Representation bias arises when the training data underrepresents population subgroups, causing the model to learn poor-quality representations for underrepresented groups. Facial recognition systems trained predominantly on lighter-skinned faces exhibit substantially higher error rates for darker-skinned individuals, a finding with significant implications for criminal justice applications where facial recognition is used for suspect identification [8]. Medical diagnostic systems trained on patient populations from major academic medical centers may perform poorly on rural, low-income, and minority populations whose disease presentation patterns, comorbidity profiles, and healthcare utilization histories differ from the training population.

2.2 The measurement bias problem

Measurement bias arises when the proxy variables used to operationalize the target concept in training labels are systematically less accurate for some groups than others. The healthcare utilization study in Science provides the canonical example: a risk stratification algorithm used healthcare cost as a proxy for healthcare need, but because Black patients face greater barriers to accessing care, they systematically incur lower costs for the same level of medical need. The algorithm therefore systematically underestimated the care needs of Black patients — not because it treated race as a predictor, but because the proxy variable it was trained to optimize was a biased measure of the underlying concept it was intended to predict [9]. This form of bias is particularly difficult to detect because it does not manifest as differential treatment of a protected attribute; it manifests as systematic outcome disparity produced by an apparently neutral technical choice.

III. Fairness Metrics and Their Relationships

Figure 1 illustrates the four-stage accountability framework connecting bias sources to fairness metrics, mitigation pipeline, and accountability outputs.

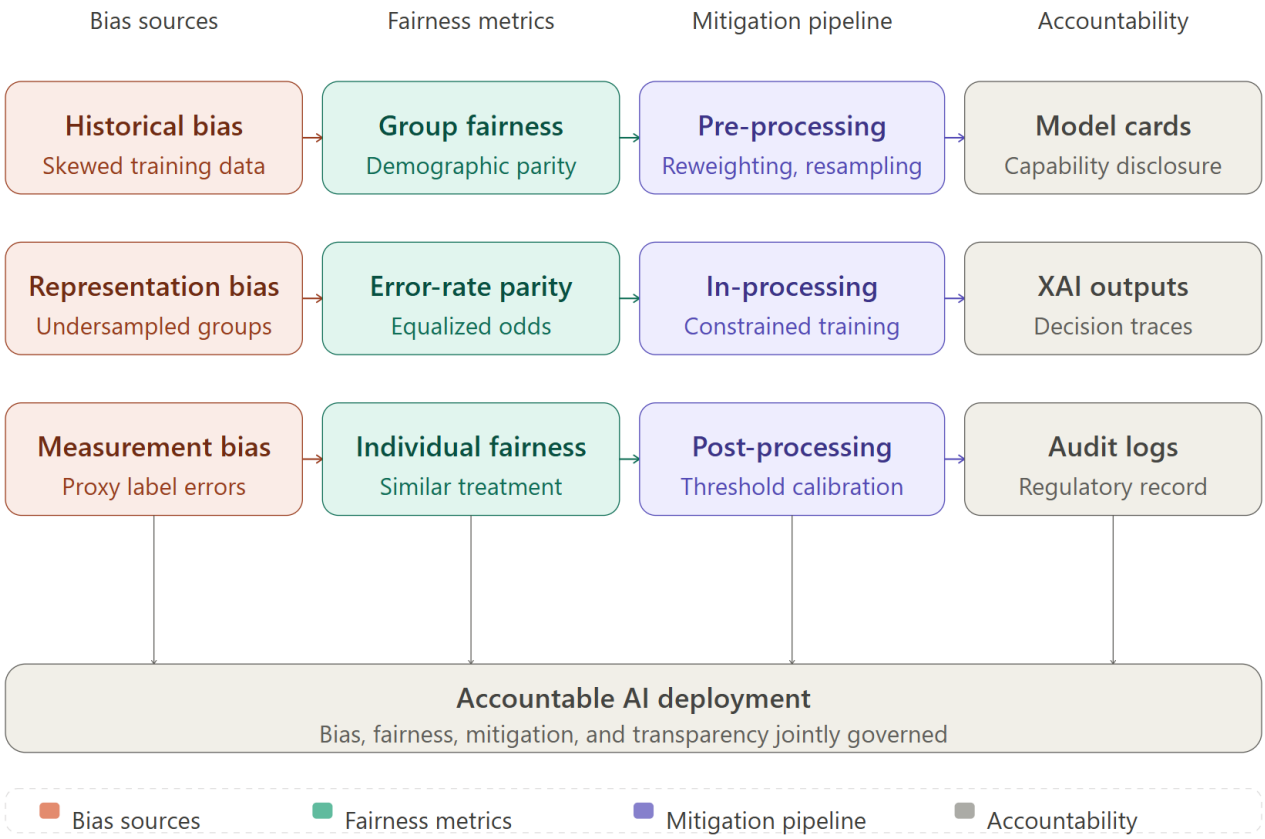

Figure 1. Algorithmic accountability framework showing the relationship between bias sources (coral), fairness metrics (teal), the three-stage mitigation pipeline (purple), and accountability outputs (gray). Bias source identification drives metric selection, which drives mitigation strategy, converging on accountable deployment.

Table 1 presents a systematic comparison of five major fairness metric families, specifying each metric's formal definition, the condition under which it is satisfied, the conditions under which it conflicts with other metrics, and its domain-specific relevance.

Table 1. Fairness Metric Comparison: Definitions, Compatibility, and Domain Relevance

Metric	Definition	Satisfied when	Incompatible with	High-stakes domain relevance
Demographic parity	Equal positive prediction rate across groups	P(Y_hat=1|A=0) = P(Y_hat=1|A=1)	Calibration when base rates differ	Hiring and credit: prevents systematic exclusion of groups

Equalized odds	Equal true positive and false positive rates across groups	TPR and FPR equal for each group A	Demographic parity when base rates differ	Criminal justice: equal recall and false alarm rate across racial groups
Equal opportunity	Equal true positive rate across groups	TPR equal for each group A	Equalized odds (relaxes FPR constraint)	Healthcare: equal detection rate for disease across demographic groups
Calibration	Predicted probability matches actual positive frequency	P(Y=1|Y_hat=p,A=a) = p for all a	Equalized odds when base rates differ	Finance: credit score equally predictive across groups
Counterfactual fairness	Decision unchanged when protected attribute counterfactually altered	Y_hat(A=a) = Y_hat(A=a') for same individual	Group fairness (protects individuals not groups)	Healthcare and legal: individual-level protection from attribute-based decisions

3.1 Group fairness metrics

Group fairness metrics, also called statistical fairness or demographic fairness metrics, require that some statistical measure of the model's predictions be equal across groups defined by protected attributes. Demographic parity requires that the positive prediction rate be equal across groups, without reference to the actual label. Equalized odds requires both the true positive rate and the false positive rate to be equal across groups. Equal opportunity is a relaxation of equalized odds that requires only the true positive rate to be equal, accepting that false positive rates may differ. These three metrics represent a progression from the weakest (demographic parity) to the strongest (equalized odds) group fairness requirement [10].

3.2 Individual fairness

Individual fairness, formalized in the foundational contribution of Dwork and colleagues, requires that similar individuals be treated similarly, where similarity is defined over the task-relevant feature space rather than the protected attribute space [11]. Individual fairness addresses the limitation of group fairness metrics that they can be satisfied by a model that treats individuals within each group arbitrarily as long as the group-level statistics are equal. A group-fair model might, for example, achieve demographic parity by randomly promoting low-scoring minority group members while demoting high-scoring majority group members, a form of treatment that is fair to neither individual. Individual fairness prevents this by requiring the model's output to be a Lipschitz function of a task-relevant similarity metric.

IV. Impossibility Results and Their Implications

One of the most significant contributions of the algorithmic fairness literature is the formal proof that certain fairness metrics cannot be simultaneously satisfied when the base rates of the positive outcome differ across groups — a condition that characterizes virtually all real-world deployment scenarios in which algorithmic fairness is relevant. The impossibility result demonstrates that demographic parity, calibration, and equalized odds cannot all be satisfied simultaneously except in the degenerate case where the classifier achieves perfect accuracy or the base rates are equal across groups [12].

The practical implications of these impossibility results are profound and have been the source of significant controversy in the deployment of algorithmic decision systems. The designers of the COMPAS recidivism prediction tool argued that their system satisfied calibration, meaning that a predicted risk score of 7 out of 10 corresponded to the same actual recidivism probability for Black and white defendants. The ProPublica analysis demonstrated that the same system violated equalized odds, producing substantially higher false positive rates for Black defendants than for white defendants. Both analyses were correct: the impossibility theorem guarantees that both cannot be achieved simultaneously when base rates differ, and the choice between them is a value judgment about which form of error is more ethically costly [13].

The impossibility results do not mean that fairness is unachievable, but they do mean that fairness cannot be achieved through technical optimization alone. The choice of fairness metric to optimize encodes a substantive value judgment about the relative weight of different types of error and the appropriate unit of analysis for fairness assessment. This value judgment should be made transparently by the organizations deploying these systems, in consultation with affected communities, and subject to democratic accountability rather than delegated to data scientists making quietly consequential technical choices in the course of model development [14].

V. Bias Mitigation Techniques

Table 2 surveys six primary bias mitigation techniques, organized by their position in the machine learning pipeline — preprocessing, in-processing, and post-processing — and characterizes each by its mechanism, the fairness metrics it addresses, and its primary limitations.

Table 2. Bias Mitigation Technique Survey: Mechanisms, Target Metrics, and Limitations

Technique	Stage	Mechanism	Fairness metric addressed	Limitation
Reweighting	Pre-processing	Assigns higher loss weight to underrepresented groups	Demographic parity, equal opportunity	May reduce accuracy for majority group; requires protected attribute access
Resampling	Pre-processing	Oversamples minority groups or undersamples majority	Demographic parity	Creates duplicate training instances; does not address

				structural bias in labels
Adversarial debiasing	In-processing	Adds adversarial component that penalizes discriminatory predictions during training	Demographic parity, equalized odds	Computationally expensive; adversarial objective may be unstable
Fairness constraints	In-processing	Adds fairness metric as constraint in optimization objective	Any differentiable metric	May be infeasible when fairness constraints conflict with each other
Calibrated equalized odds	Post-processing	Adjusts classification thresholds separately per group to equalize TPR/FPR	Equalized odds	Requires access to group membership at inference; may not satisfy other metrics
Reject option classification	Post-processing	Abstains from predicting near decision boundary for protected group members	Counterfactual fairness (approximately)	Reduces system coverage; requires fairness-accuracy trade-off calibration

5.1 Preprocessing approaches

Preprocessing bias mitigation techniques intervene in the training data before model fitting, addressing bias at its source rather than in the learned model. Reweighting approaches assign higher importance weights to underrepresented groups in the training objective, effectively instructing the model to prioritize accuracy for historically underserved populations. Resampling approaches achieve a similar effect by oversampling minority group instances or undersampling majority group instances to produce a more balanced training distribution. Both approaches have been validated in the IBM AI Fairness 360 toolkit, which provides standardized implementations for a range of preprocessing, in-processing, and post-processing methods [15]. The primary limitation of preprocessing approaches is that they address distributional imbalance without addressing the structural causes of bias in the labeling process itself, leaving measurement bias and historical bias in the labels unaddressed.

5.2 In-processing and post-processing approaches

In-processing techniques integrate fairness objectives directly into the model training process, either through adversarial components that penalize discriminatory predictions or through constrained

optimization formulations that add fairness metric satisfaction as a training constraint. Adversarial debiasing, in which a discriminator network attempts to predict the protected attribute from the classifier's internal representation, has demonstrated effectiveness across multiple fairness metrics and model architectures, though at the cost of increased training complexity and potential instability in the adversarial objective [16]. Post-processing techniques adjust model outputs after training, typically by setting separate decision thresholds for different groups to equalize a chosen metric across them. The calibrated equalized odds method provides a principled approach to threshold calibration that satisfies equalized odds without retraining the underlying model, making it applicable to black-box systems where training access is unavailable [17]. The engineering implementation of these methods in production-grade systems requires careful attention to the operational requirements of fairness by design, as documented in practical deployment frameworks [18].

VI. Transparency and Explainability Mechanisms

6.1 Documentation standards

The transparency of AI systems has been conceptualized along two complementary dimensions: process transparency, concerning how the system was built, and outcome transparency, concerning how specific decisions are generated. Model cards for model reporting provide a standardized framework for process transparency documentation, specifying a template that captures intended uses and out-of-scope uses, evaluation results broken down by demographic subgroups, and ethical considerations relevant to the model's deployment context [4]. Datasheets for datasets apply a parallel approach to training and evaluation data, requiring documentation of the data's provenance, composition, collection process, preprocessing steps, and known limitations [19]. Together, these documentation standards create the information infrastructure necessary for downstream auditors, regulators, and affected communities to evaluate whether an AI system's design and training are consistent with the accountability requirements of its deployment context.

6.2 Post-hoc explanation methods

Post-hoc explanation methods generate human-interpretable representations of the factors driving individual model predictions, addressing the outcome transparency dimension of AI accountability. LIME generates locally faithful linear approximations of complex model behavior in the neighborhood of individual predictions, producing feature importance scores that indicate which input features most strongly influenced the prediction for a specific instance [20]. SHAP provides a game-theoretic foundation for feature attribution that satisfies several desirable properties including consistency, local accuracy, and dummy player treatment, producing Shapley values that quantify each feature's marginal contribution to the prediction across all possible feature subsets [21]. Counterfactual explanation methods generate the minimal change to input features that would reverse the model's prediction, directly addressing the GDPR Article 7 right to explanation by showing the individual what they would need to change to receive a different outcome [22]. Explanations generated by these methods should be consistent with the model's actual decision-making process and not merely a post-hoc rationalization, a concern that has motivated research on explanation faithfulness and robustness [5].

VII. Domain-Specific Accountability Requirements

Table 3 analyzes domain-specific accountability requirements across four high-stakes deployment domains, identifying the primary fairness concern, the most relevant fairness metric, the applicable regulatory instrument, and the relevant implementation precedent in each domain. The table includes the contribution of Gupta's practical implementation framework for fair and transparent AI systems, which provides the applied engineering complement to the theoretical fairness literature reviewed in preceding sections [18].

Table 3. Domain-Specific Accountability Requirements Across High-Stakes Decision Domains

Domain	Primary fairness concern	Most relevant metric	Regulatory instrument	Implementation precedent
Healthcare	Differential diagnostic accuracy across demographic groups	Equal opportunity (equal TPR across groups)	FDA guidance on AI/ML-based software as medical device (2021)	Racial bias in healthcare utilization algorithm (Science, 2019)
Criminal justice	Differential recidivism prediction error rates across racial groups	Equalized odds (equal TPR and FPR)	No federal standard; state-level disparate impact law	COMPAS recidivism tool analysis (ProPublica, 2016)
Financial services	Differential credit approval and pricing across protected classes	Calibration with demographic parity	Equal Credit Opportunity Act; Fair Housing Act; CFPB guidance	Amazon recruitment tool bias against women (2018)
Employment	Differential candidate screening outcomes across gender and race	Equal opportunity	EEOC guidance on employer use of AI in hiring (2023)	Amazon recruitment tool; multiple hiring tool audits

7.1 Healthcare

Healthcare AI presents the highest stakes fairness challenges of any domain because errors in algorithmic medical decision-making can directly cause physical harm. The landmark study published in Science in 2019 demonstrated that a healthcare utilization management algorithm deployed by a major health insurer to prioritize high-risk patients for care management programs systematically underserved Black patients due to measurement bias in the proxy variable used to operationalize healthcare need [9]. At equal levels of measured risk, Black patients had substantially greater actual healthcare needs than white patients because systematic underaccess to care meant their conditions were less frequently treated, generating lower costs. The algorithm, trained to optimize future cost as a proxy for need,

therefore recommended additional resources for white patients with lower actual need while denying them to Black patients with greater actual need.

Correcting this bias required identifying and replacing the biased proxy variable — a structural intervention in data pipeline design rather than a parameter adjustment in the model. This case illustrates the insufficiency of post-processing fairness adjustment for addressing measurement bias, and the necessity of the kind of systematic fairness-by-design approach documented in applied implementation frameworks for AI transparency and accountability [18]. The FDA's guidance on AI and machine learning-based software as a medical device establishes the regulatory context for algorithmic accountability in healthcare, requiring transparency about intended use populations, known performance disparities across demographic subgroups, and mechanisms for monitoring real-world performance against pre-deployment validation results [23].

7.2 Criminal justice

The criminal justice domain presents the fairness challenge in its starkest form because the adverse outcomes of algorithmically mediated decisions include pretrial detention, longer sentences, and lifetime collateral consequences from criminal records. The deployment of risk assessment instruments in pretrial release, bail setting, sentencing, and parole decisions has accelerated substantially since 2010, driven by genuine interest in reducing both incarceration and reoffending through evidence-based decision support rather than purely punitive sentencing [13]. The fairness controversies that followed the ProPublica investigation of COMPAS established the impossibility theorem's implications in concrete policy terms: calibration and equalized odds cannot both be satisfied when base rates differ across racial groups, and the choice between them requires a substantive value judgment about the relative acceptability of different error types that the judicial system had not publicly articulated before deploying the tools embodying that choice.

The criminal justice context also raises the individual fairness challenge most acutely. A defendant whose risk score reflects statistical patterns associated with their demographic group rather than their individual circumstances is treated not as a person but as a statistical representative of a category, raising deep concerns about the compatibility of algorithmic risk assessment with due process and equal protection guarantees. Individual fairness methods that require the model's output to be a function of task-relevant features rather than protected attributes provide a technical framework for addressing this concern, but their implementation requires explicit specification of the task-relevant feature space that courts and legislatures have not yet provided.

7.3 Financial services

Financial services AI applications including credit scoring, underwriting, fraud detection, and investment advice are subject to the most developed regulatory framework of any algorithmic accountability domain, reflecting a century of fair lending law development. The Equal Credit Opportunity Act and the Fair Housing Act prohibit disparate treatment of protected classes in credit and housing transactions and extend their coverage to algorithmic systems through disparate impact doctrine, which holds that facially neutral practices that produce discriminatory outcomes are unlawful regardless of discriminatory intent [3]. The Consumer Financial Protection Bureau has issued guidance indicating that the ECOA's requirement for adverse action notices, which inform applicants of the specific reasons for a credit denial, applies to algorithmic credit decisions and cannot be satisfied by

generic reference to model score, requiring lenders to provide the individual feature-level explanation that SHAP and LIME methods can generate [22].

VIII. Governance Frameworks and Regulatory Alignment

The governance frameworks for AI accountability have developed in parallel with the technical fairness literature, producing regulatory instruments at the national, supranational, and international levels. The GDPR's Article 22 establishes the right not to be subject to solely automated decisions that significantly affect individuals, requiring that organizations deploying such systems provide meaningful information about the logic of the decision and implement suitable measures to safeguard individual rights, including the right to obtain human review of the decision and to contest it [24]. The GDPR's Article 25 data protection by design requirement establishes a positive obligation to implement privacy and fairness safeguards in AI system design rather than as compliance overlays applied after development [24].

The EU AI Act, which entered into force in August 2024, establishes a risk-stratified governance framework that subjects AI systems in high-risk application domains, including employment, credit, healthcare, criminal justice, and critical infrastructure, to mandatory requirements for technical documentation, human oversight, transparency, robustness, and accuracy [25]. The Act's high-risk requirements create a comprehensive accountability mandate that subsumes and extends the requirements of prior sector-specific regulation, requiring high-risk AI providers to implement risk management systems, conduct conformity assessments, and register their systems in a public EU-wide database. The NIST AI Risk Management Framework provides a voluntary but widely adopted governance structure in the US context, organizing accountability requirements around govern, map, measure, and manage functions that align closely with the fairness, bias mitigation, transparency, and monitoring requirements surveyed in this review [26].

Conclusions and Research Priorities

This review has surveyed the body of knowledge on algorithmic fairness, bias mitigation, and AI transparency developed between the foundational contributions of the 2016 to 2022 period and the practical implementation frameworks that have since translated theory into deployable engineering practice. The review has established four principal conclusions.

First, algorithmic bias is structurally embedded in the data, labeling processes, and optimization objectives of AI systems rather than an incidental artifact of poor technical practice. Addressing it requires systematic fairness-by-design approaches that intervene at bias sources rather than post-hoc adjustments to discriminatory systems, as the engineering implementation frameworks developed in recent applied work make clear [18]. Second, the impossibility results that constrain simultaneous satisfaction of competing fairness metrics are not technical failures to be overcome but structural features of the fairness problem that require explicit value judgments about the relative acceptability of different types of error — judgments that must be made transparently by accountable human institutions rather than quietly by technical teams. Third, transparency and explainability mechanisms, including model cards, datasheets for datasets, SHAP, LIME, and counterfactual explanation methods, provide the information infrastructure necessary for AI accountability but are not sufficient for it: accountability requires that the information they provide actually be used in oversight, audit, and redress mechanisms with genuine consequences for failures. Fourth, domain-specific fairness and accountability

requirements differ substantially across healthcare, criminal justice, and financial services in ways that mae domain-agnostic metric optimization an insufficient response to the governance challenge.

Three research priorities follow. First, the development of fairness-by-design methodologies that address measurement bias and historical bias at the labeling stage, where generic preprocessing and in-processing techniques are insufficient. Second, empirical evaluation of the behavioral impact of transparency mechanisms on affected individuals' ability to contest algorithmic decisions, which is the practical accountability test that the GDPR right to explanation was designed to satisfy but that has been rarely evaluated in real deployment contexts. Third, interdisciplinary frameworks for resolving the value conflicts embodied in fairness metric impossibility, drawing on legal, philosophical, and social science expertise alongside the technical methods reviewed here, to support the transparent, democratically accountable governance choices that algorithmic accountability ultimately requires.

References:

[1] B. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, no. 3, pp. 671-732, 2016.

[2] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human decisions and machine predictions," Quarterly Journal of Economics, vol. 133, no. 1, pp. 237-293, 2018.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1-35, 2021.

[4] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proc. ACM Conf. on Fairness, Accountability, and Transparency (FAccT), 2019, pp. 220-229.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[6] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023.

[7] V. Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, 2018.

[8] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Proc. ACM Conf. on Fairness, Accountability, and Transparency (FAT*), 2018, pp. 77-91.

[9] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447-453, Oct. 2019.

[10] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2016, vol. 29, pp. 3315-3323.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in Proc. ACM Innovations in Theoretical Computer Science (ITCS), 2012, pp. 214-226.

[12] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," Big Data, vol. 5, no. 2, pp. 153-163, Jun. 2017.

[13] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," Science Advances, vol. 4, no. 1, eaao5580, Jan. 2018.

[14] F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, 2015.

[15] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," arXiv preprint arXiv:1810.01943, 2018.

[16] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in Proc. AAAI/ACM Conf. on AI, Ethics, and Society (AIES), 2018, pp. 335-340.

[17] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, vol. 30, pp. 5680-5689.

[18] S. Gupta, "Designing Fair and Transparent AI Systems with Implementation of Algorithms," International Research Journal of Engineering and Technology (IRJET), vol. 10, p. 1867, 2023.

[19] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86-92, 2021.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135-1144.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, vol. 30, pp. 4765-4774.

[22] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard Journal of Law and Technology, vol. 31, no. 2, pp. 841-887, 2018.

[23] U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning-Based Software as a Medical Device Action Plan. FDA, 2021.

[24] European Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Official Journal of the European Union, L 119, 2016.

[25] European Parliament. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 2024.

[26] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. U.S. Department of Commerce, Jan. 2023.