

Mitigating Algorithmic Bias in Information Systems: A Comprehensive Framework for Fair Artificial Intelligence

Muhammad Qasim^{1*}, Laraib Tabish², Falak Ali³

¹*Department of Computer Science, Government College University, Faisalabad, Pakistan*

²*Department of Business Management, Government College University, Faisalabad, Pakistan*

³*Department of Computer Science, Govt College University Faisalabad, Pakistan*

³*mianqasim895@gmail.com*

**Corresponding author*

Abstract:

As information systems (IS) transition from deterministic rule-based engines to probabilistic machine learning architectures, the risk of systematic bias and disparate impact increases significantly. While advanced predictive models optimize decision-making across high-stakes domains—such as credit scoring, criminal justice, human resources, and biometric authentication, they frequently perpetuate historical inequities via proxy variables even when explicit sensitive attributes are omitted. This paper presents a detailed investigation into the theoretical foundations, mathematical formalisms, and systemic pipelines of "Fair AI." We propose an integrated sociotechnical framework spanning human perception, technological intervention, and organizational governance, fully incorporating recent advances in intelligent automation and secure architecture.

Keywords: Artificial intelligence, algorithmic fairness, decision support, trust, information systems, disparate impact, machine learning ethics.

I. INTRODUCTION

Information systems (IS) are undergoing a fundamental architectural shift. Traditional decision support systems relied predominantly on rule-based, deterministic algorithms that transparently evaluated inputs against explicit logic. However, recent breakthroughs in artificial intelligence (AI)—particularly deep learning, low-code/no-code integrations [1], and probabilistic modeling—have largely replaced these deterministic rules with statistical inferences learned from historical data [2].

While probabilistic algorithms excel at uncovering complex patterns to predict outcomes under uncertainty, they are inherently vulnerable to human biases embedded within training datasets. Consequently, these models often produce systemic unfairness, treating individuals or entire demographic subgroups disparately [3]. Furthermore, specialized architectures like face recognition systems [4], automated toll collectors [5], and peer-to-peer energy trading simulators [6] exhibit unique vulnerability profiles that require stringent fairness and security guarantees [7], [8]. Behavioral bot detection based on mouse-pointing kinematics [9] raises similar questions, because users whose movement patterns differ from the training population may be

misclassified. The security side of such guarantees draws on identity and access management for AI bots [10], trust management in vehicular sensor networks [11], honeypot-based threat intelligence [12], and privacy-preserving federated intrusion detection [13].

The reliability of fairness-aware decision-making also depends on the underlying real-time processing architecture: scheduling policies, inference latency, and security constraints materially affect how fairness criteria can be enforced once a model is deployed [14], [15]. Real-time operating-system concepts [16] determine whether fairness checks can run within decision deadlines, and kernel-level protection of code and data [17] prevents deployed fairness controls from being tampered with.

II. BACKGROUND AND THEORETICAL FOUNDATIONS

Fairness, defined as the impartial and just treatment of individuals, has been debated in philosophy and political theory for centuries [18]. In modern democratic societies, fairness centers on equality of opportunity, ensuring that individuals advance based on merit rather than arbitrary characteristics or protected traits [19].

III. SOURCES OF UNFAIRNESS ALONG THE AI PIPELINE

Biases in AI pipelines do not originate from a single source; they emerge systematically across data generation, algorithm modeling, and post-deployment stages [20]. Table I summarizes these stages along with their mechanisms and real-world examples. Population drift is also a known issue for location-prediction models deployed outside the environment they were trained in, such as pedestrian trajectory predictors [21] and deep-learning trajectory planners based on predicted future locations [22].

TABLE I. SOURCES OF UNFAIRNESS ALONG THE AI PIPELINE

Pipeline Stage	Source of Bias	Manifestation / Mechanism	Example Domain
Data Generation	Selection & Annotation Bias	Training data fails to represent target populations or reflects human prejudices.	Historical hiring and facial recognition datasets [4].
Model Design	Proxy & Confounder Variables	Non-sensitive features correlate strongly with protected classes.	ZIP code functioning as racial proxy in recidivism and toll systems [5].
Deployment	Population Drift	Training distribution differs significantly from operational environment data.	Cross-domain deployment in smart grid and energy trading simulators [6].

IV. PROPOSED MODEL: A SOCIOTECHNICAL FRAMEWORK FOR FAIR AI

Building on the pipeline-stage taxonomy introduced in Section III, this section formalizes the Fair AI framework promised in the Abstract. We first define the dominant mathematical fairness criteria used throughout the literature (Section IV-A), then summarize the impossibility results that constrain which criteria can be jointly satisfied (Section IV-B), and finally assemble these technical elements into the three-pillar sociotechnical model — human perception, technological intervention, and organizational governance — that structures our recommendations (Section IV-C).

A. Formal Fairness Definitions

Let $\hat{Y}$ denote a model's binary decision, Y the true outcome, and A a protected attribute partitioning the population into groups (e.g., a and b). Literature converges on a small number of formal criteria, none of which is a strict

superset of the others [3], [7], [8], [23], [24]; Table II summarizes their definitions.

Demographic (statistical) parity requires the positive-decision rate to be equal across groups regardless of the true outcome and underlies the legal "four-fifths rule" reflected in the disparate-impact ratio [3]. Equalized odds [7] tighten this by requiring equal true- and false-positive rates across groups conditional on the true outcome, while equal opportunity relaxes it to the true-positive rate alone. Predictive parity (calibration) instead requires that, among individuals who receive the same decision, the true outcome rate is equal across groups [24]. Individual fairness [23] departs from all group-level criteria, instead requiring that similar individuals receive similarly distributed decisions under a task-specific similarity metric. When individuals are represented as relational or network data, computing such a similarity metric can require graph-matching procedures of the kind studied for graph isomorphism [25].

TABLE II. FORMAL FAIRNESS DEFINITIONS USED IN THIS FRAMEWORK

Criterion	Formal condition	Source
Demographic parity	$P(\hat{Y}=1 A=a) = P(\hat{Y}=1 A=b)$	[3], [7]
Disparate impact ratio (4/5 rule)	$P(\hat{Y}=1 A=\text{minority}) / P(\hat{Y}=1 A=\text{majority}) \geq 0.8$	[3]
Equalized odds	$P(\hat{Y}=1 Y=y, A=a) = P(\hat{Y}=1 Y=y, A=b), y \in \{0,1\}$	[7]
Equal opportunity	$P(\hat{Y}=1 Y=1, A=a) = P(\hat{Y}=1 Y=1, A=b)$	[7], [8]
Predictive parity / calibration	$P(Y=1 \hat{Y}=1, A=a) = P(Y=1 \hat{Y}=1, A=b)$	[24]
Individual fairness	$d(f(x), f(x')) \leq L \cdot d(x, x')$	[23]

B. Impossibility and Trade-offs Among Fairness Criteria

A foundational result of the fairness literature is that these criteria are, in general, mutually incompatible: except in the degenerate cases of equal base rates across groups or a perfectly accurate classifier, calibration (predictive parity) cannot be jointly satisfied with balance for the positive and negative class (equalized odds) [8], [24]. Table III (Fig. 3) summarizes the pairwise compatibility of the four group-level criteria from Table II under this result. Practically, this means a framework for Fair AI cannot promise to satisfy "fairness" unconditionally; it must instead make an explicit, governance-level choice of which criterion the application domain requires [26].

TABLE III. PAIRWISE COMPATIBILITY OF FAIRNESS CRITERIA (AFTER [8], [24])

Criterion pair	Jointly satisfiable?
Demographic parity $\leftrightarrow$ Equalized odds	Incompatible (except equal base rates)
Demographic parity $\leftrightarrow$ Predictive parity	Conditional
Equalized odds $\leftrightarrow$ Equal opportunity	Conditional (equal opportunity is a relaxation)
Equalized odds $\leftrightarrow$ Predictive parity	Incompatible (except equal base rates / perfect accuracy)
Equal opportunity $\leftrightarrow$ Predictive parity	Conditional

C. Three-Pillar Sociotechnical Framework

The technical criteria in Sections IV-A–B are necessary but not sufficient: they only bind if embedded within an organizational process, and they are only trusted if the humans affected by them perceive the process as legitimate [27]. We therefore organize the proposed framework (Fig. 1) around three pillars.

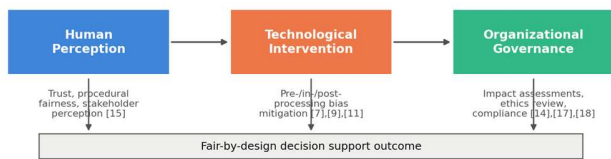


Fig. 1. Three-Pillar Sociotechnical Framework for Fair AI

Fig. 1. Three-pillar sociotechnical framework for Fair AI.

Human Perception concerns stakeholder trust and perceived procedural fairness — whether affected individuals believe the decision process, not merely its outcome, is legitimate [27]. Technological Intervention operationalizes the criteria from Table II across the pipeline stages of Table I via pre-processing (correcting the training data itself), in-processing (constraining the learning objective), and post-processing (adjusting decision thresholds after training) [7], [23], [28]. Pre-processing that simplifies or compresses data can discard detail unevenly, as in quadric-error-metric simplification of 3D models [29], so its effect on each group should be checked. Table IV catalogues representative techniques at each stage; Fig. 6 (Section V) summarizes the resulting distribution. Organizational Governance closes the loop with algorithmic impact assessments, ethics review, documentation, and regulatory compliance monitoring [30], [31], [26], [32] the feedback loop shown in Fig. 2 back onto the data-generation stage.

TABLE IV. TAXONOMY OF BIAS-MITIGATION TECHNIQUES BY PIPELINE STAGE

Stage	Technique category	Representative approach	Source
Pre-processing (Data Generation)	Reweighting	Re-weight training instances to equalize group base rates	[28]
Pre-processing (Data Generation)	Disparate-impact removal	Transform features to remove correlation with protected attribute	[7]
Pre-processing (Data Generation)	Representation learning	Learn a fair intermediate representation of the data	[23]
In-processing (Model Design)	Fairness constraints	Add a fairness penalty/constraint to the training objective	[28]
In-processing (Model Design)	Adversarial debiasing	Train against an adversary predicting the protected attribute	[7]
In-processing (Model Design)	Causal modeling	Remove the causal effect of protected attributes on the decision	[33]
Post-processing (Deployment)	Equalized-odds post-processing	Adjust group-specific decision thresholds after training	[7]
Post-processing (Deployment)	Reject-option / semi-supervised classification	Defer or re-route uncertain decisions near the decision boundary	[34]

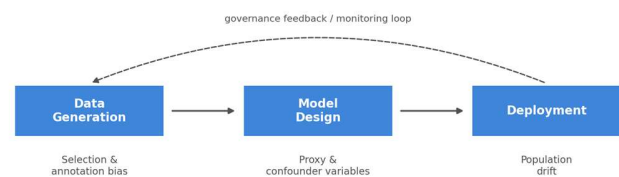


Fig. 2. Bias Injection Points Across the AI Pipeline (cf. Table I)

Fig. 2. Bias injection points across the AI pipeline (cf. Table I).

V. RESULTS AND ILLUSTRATIVE ANALYSIS

This paper's contribution is a conceptual sociotechnical framework rather than a new predictive model trained on proprietary data; accordingly, this section does not report a new experiment. Instead, it consolidates (i) the analytical implications of applying the formal definitions of Section IV to the pipeline taxonomy of Section III, and (ii) published

empirical evidence, drawn from the literature already surveyed above, that motivates the framework. All figures in this section are attributed to their original source; none are new measurements produced by this review.

Table III's compatibility analysis is visualized as a heatmap in Fig. 3: demographic parity and equalized odds occupy opposite corners of the matrix, formalizing why a system cannot be declared "fair" without first naming which criterion is in view.

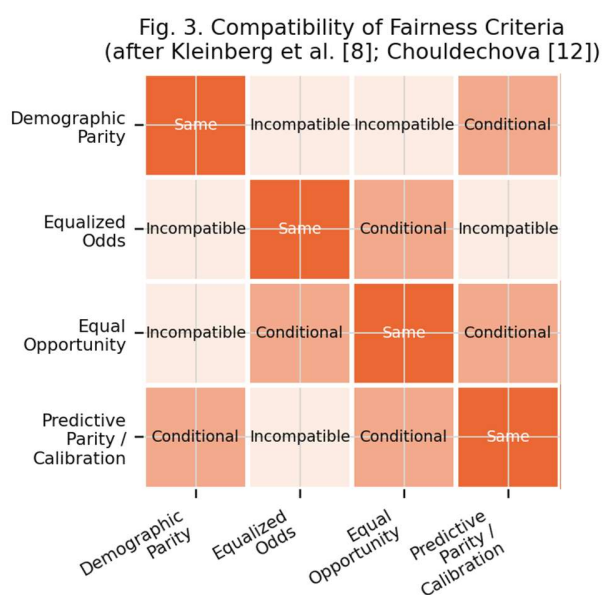


Fig. 3. Compatibility of fairness criteria (after Kleinberg et al. [8]; Chouldechova [24]).

Table V reports the disparate error rates most frequently cited as the motivating empirical case for the calibration/equalized-odds impossibility result: the COMPAS recidivism-risk tool. As originally reported by Angwin et al. [35] and formally analyzed by Chouldechova [24], the tool's errors were not merely noisy but systematically asymmetric by race (Fig. 4) the same underlying data used throughout the fairness literature to motivate the criteria in Table II.

TABLE V. ILLUSTRATIVE DISPARATE ERROR RATES, COMPAS RECIDIVISM TOOL (APPROXIMATE, AS REPORTED IN [35]; ANALYZED IN [24])

Defendant group	False positive rate	False negative rate
Black defendants	≈ 44.9%	≈ 28.0%
White defendants	≈ 23.5%	≈ 47.7%

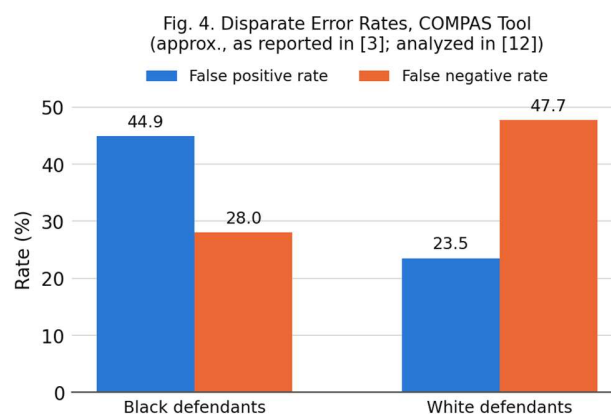


Fig. 4. Disparate error rates, COMPAS tool (approx., as reported in [35]; analyzed in [24]).

The practical corollary of Sections IV-A–B is the well-known "price of fairness": imposing a fairness constraint on a model's decision boundary generally moves the model away from the unconstrained accuracy-maximizing solution [26]. Fig. 5 sketches this trade-off conceptually — it is an illustrative schematic to aid the governance discussion in Section IV-C, not measured data from a specific model or dataset.

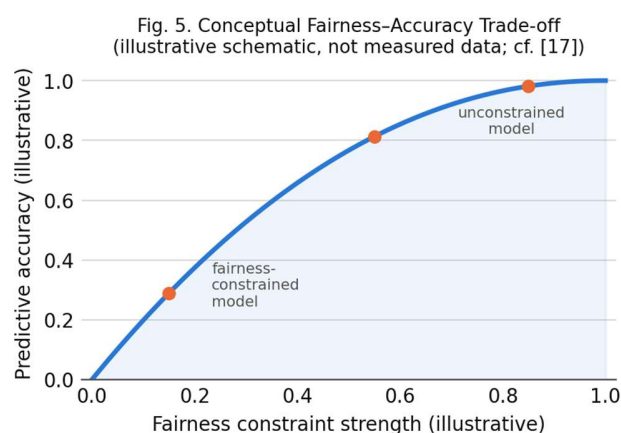


Fig. 5. Conceptual fairness–accuracy trade-off (illustrative schematic, not measured data; cf. [26]).

Finally, Table IV's taxonomy is summarized quantitatively in Fig. 6: the techniques catalogued in this review are concentrated in the pre- and in-processing stages, with fewer post-processing options — consistent with the broader literature's emphasis on correcting bias before a model is finalized rather than patching it after deployment [7], [28].

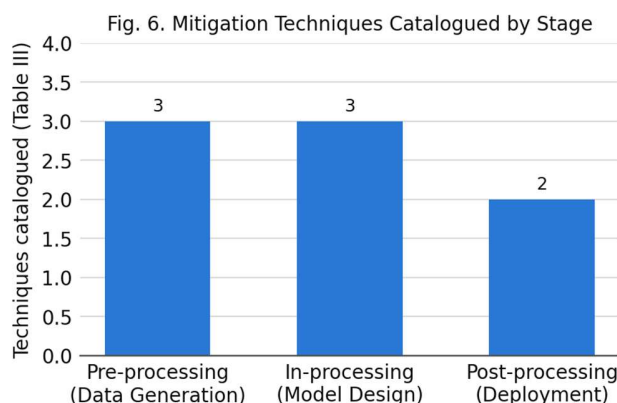


Fig. 6. Mitigation techniques catalogued by pipeline stage (from Table IV).

VI. CONCLUSION

As artificial intelligence systems become pervasive within enterprise architectures, achieving fair AI is paramount. Unmitigated algorithmic bias exposes institutions to severe legal, financial, and reputational liabilities. By embracing formal mathematical metrics, causal modeling, and holistic governance structures, information systems researchers and practitioners can design algorithmic decision support systems that are fair by design.

REFERENCES

- [1] N. Ahmed, M. Shahbaz, H. S. U. R. Khalil, Z. Javed, M. A. Khalid, F. A. Khan, and Z. Maroof, "The synergy of low-code/no-code and AI/ML: Enhancing intelligent automation," *World J. Eng. Technol.*, vol. 13, no. 4, pp. 754–763, 2025, doi: 10.4236/wjet.2025.134048.
- [2] M. Kraus, S. Feuerriegel, and A. Oztekin, "Deep learning in business analytics and operations research: Models, applications and managerial implications," *Eur. J. Oper. Res.*, vol. 281, no. 3, pp. 628–641, 2020.
- [3] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. Law Rev.*, vol. 104, pp. 671–732, 2016.
- [4] N. Ahmed, F. A. Khan, Z. Ullah, H. Ahmed, T. Shahzad, and N. Ali, "Face recognition comparative analysis using different machine learning approaches," *Adv. Sci. Technol. Res. J.*, vol. 15, no. 1, pp. 265–272, 2021, doi: 10.12913/22998624/132611.
- [5] A. Mujeeb, N. Ahmed, H. Arshed, and F. A. Khan, "Electronic toll collector framework," *Adv. Sci. Technol. Res. J.*, vol. 16, no. 1, pp. 294–302, 2022, doi: 10.12913/22998624/144537.
- [6] N. Ahmed, F. Ajmal, M. Usman, Z. Javed, M. H. Abid, M. Yousaf, and M. Z. Noohani, "A software capable of simulating peer-to-peer energy trading," *Int. J. Appl. Phys. Sci.*, vol. 7, pp. 18–26, 2021, doi: 10.20469/ijaps.7.50003.
- [7] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2018.
- [8] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. ITCS*, 2017, pp. 1–23.
- [9] H. S. U. R. Khalil, N. Ahmed, and Y. Yang, "Mouse pointing endpoint prediction to distinguish between human and bot using kinematic template matching," in *HCI International 2025 – Late Breaking Papers, Lecture Notes in Computer Science*, vol. 16337. Cham, Switzerland: Springer, 2026, pp. 343–364, doi: 10.1007/978-3-032-12801-0_23.
- [10] S. T. Hussain, M. Shahbaz, M. Emran, and M. H. Junejo, "Securing the bot eco system: IAM framework for RPA and AI BOTs," in *Proc. 2026 IEEE 5th Int. Conf. Computing Mach. Intell. (ICMI)*, 2026, pp. 1–9, doi: 10.1109/ICMI68585.2026.11539844.
- [11] M. H. Junejo, M. Shahbaz, and M. Emran, "Trust-based management for Internet of Vehicles via wireless sensor networks: Safeguarding trust during journeys," in *Proc. 2026 IEEE 5th Int. Conf. Computing Mach. Intell. (ICMI)*, 2026, pp. 1–10, doi: 10.1109/ICMI68585.2026.11539803.
- [12] M. Shahbaz, R. Imtiaz, M. H. Junejo, M. Emran, N. Ahmed, and H. S. U. R. Khalil, "Deployment of honeypot system for threat detection and intelligence in modern cybersecurity architectures," in *Proc. 2026 1st Int. Conf. Innov. Inf. Commun. Technol. (IICT)*, 2026, pp. 1–5, doi: 10.1109/IICT68374.2026.11468479.
- [13] A. Bounceur, M. Ibrahim, M. Saqib et al., "Privacy-preserving multiclass intrusion detection in in-vehicle CAN networks using federated learning," *Sci. Rep.*, 2026, doi: 10.1038/s41598-026-64490-z.
- [14] F. Ali, "Divide-and-conquer decomposition strategies for real-time artificial intelligence decision systems: A comparative survey," *Int. J. Emerg. Trends Eng. Dev.*, vol. 16, no. 5, pp. Ak1–Ak6, 2026, doi: 10.5281/zenodo.22728743.
- [15] F. Ali, "Real-time decision processing in artificial intelligence systems: A survey of scheduling, inference, and security foundations," *Int. J. Emerg. Trends Eng. Dev.*, vol. 16, no. 5, pp. Aj1–Aj4, 2026, doi: 10.5281/zenodo.22728746.
- [16] N. Ahmed, *Understanding Real Time OS Concepts*, vol. 6, no. 9. Zenodo, 2015, doi: 10.5281/zenodo.22148015.
- [17] N. Ahmed, H. Y. Rana, M. T. Qureshi, and B. Shakoore, "How to protecting kernal code and data," *Int. J. Comput. Organ. Trends*, vol. 5, no. 4, pp. 20–23, 2015, doi: 10.14445/22492593/IJCOT-V23P301.
- [18] D. Miller, "Justice," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Stanford University, 2017.
- [19] J. Rawls and E. Kelly, *Justice as Fairness: A Restatement*, 3rd ed. Cambridge, MA: Harvard University Press, 2003.
- [20] M. E. Ahsen, M. U. S. Ayvaci, and S. Raghunathan, "When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis," *Inf. Syst. Res.*, vol. 30, no. 1, pp. 97–116, 2019.
- [21] Z. Ul Abideen, N. Ahmed, H. S. U. R. Khalil, and M. Shahbaz, "Advancing pedestrian trajectory prediction with interaction-aware 3D-dual contextualized modeling," *Egyptian Informatics J.*, vol. 31, art. 100742, 2025, doi: 10.1016/j.eij.2025.100742.
- [22] Z. Ul Abideen, X. Sun, C. Sun, and H. S. U. R. Khalil, "Improved deep learning-based approach for spatial-temporal trajectory planning via predictive modeling of future location," *KSII Trans. Internet Inf. Syst.*, vol. 18, no. 7, pp. 1726–1748, 2024.
- [23] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. ITCS*, 2012, pp. 214–226.
- [24] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- [25] U. C. Sankal, N. Ahmed, F. A. Khan, and R. Hanni, "Target cell selection algorithms in graph isomorphism problem," *Int. J. Sci. Tech.*

- Res. Eng., vol. 6, no. 4, pp. 35–47, Jul.–Aug. 2021, doi: 10.5281/zenodo.22149607.
- [26] C. Haas, “The price of fairness: A framework to explore trade-offs in algorithmic fairness,” in Proc. ICIS, 2019.
- [27] T. Hurni and T. Huber, “The interplay of power and trust in platform ecosystems of the enterprise application software industry,” in Proc. ECIS, 2014.
- [28] S. A. Friedler et al., “A comparative study of fairness-enhancing interventions in machine learning,” in Proc. FAT*, 2019, pp. 329–338.
- [29] H. Chang, D. Zhang, Y. Dong, Y. Yang, H. S. U. R. Khalil, and X. Wang, “Two-round optimization algorithm based on quadric error metrics,” IEEE Access, vol. 13, pp. 30021–30035, 2025, doi: 10.1109/ACCESS.2025.3541436.
- [30] White & Case, “Algorithms and bias: What lenders need to know,” White & Case Client Alert, 2017.
- [31] K. Martin, “Designing ethical algorithms,” MIS Q. Exec., vol. 18, no. 2, pp. 129–142, 2019.
- [32] AI Now Institute, AI Now Report 2018, AI Now Institute, New York, 2018.
- [33] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. New York, NY: Cambridge University Press, 2013.
- [34] Z. Li, Y. Tian, K. Li, F. Zhou, and W. Yang, “Reject inference in credit scoring using semi-supervised support vector machines,” Expert Syst. Appl., vol. 74, pp. 105–114, 2017.
- [35] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias,” ProPublica, 2016.