

Bias Detection in Machine Learning Systems

Elena Rossi¹, Clara Kovacs²

¹ Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: elena.rossi841@yahoo.com | ORCID: 9452-0316-0789-1857

² Associate Professor, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden. Email: clara.kovacs636@yahoo.com | ORCID: 1296-5221-2487-8814

ABSTRACT

Machine learning systems deployed in consequential domains increasingly exhibit measurable biases that produce disparate outcomes across demographic groups, yet the detection and measurement of these biases remains methodologically fragmented. This study presents a systematic evaluation of six bias detection approaches -- group fairness metrics (demographic parity, equalised odds, calibration), individual fairness testing, counterfactual fairness assessment, causal discovery of bias pathways, intersectional bias auditing, and temporal bias drift monitoring -- across four high-stakes ML applications: hiring recommendation (resume screening, 48K applications), credit scoring (Lending Club, 2.26M loans), facial recognition (FairFace, 108K images), and clinical risk prediction (MIMIC-III, 42K ICU stays). A total of 1,920 bias audits were conducted under standardised evaluation protocols. Group fairness metrics detected statistically significant disparities in all four domains: hiring models showed 12.4 +- 2.8 percentage point selection rate gaps between gender groups; credit models showed 8.6 +- 1.4 point false positive rate gaps between racial groups; facial recognition showed 14.8 +- 3.2 point accuracy gaps between skin-tone groups; clinical models showed 6.2 +- 1.8 point calibration gaps between racial groups. Intersectional auditing revealed that bias was 2.4x larger at intersections (e.g., dark-skinned women in facial recognition) than for any single protected attribute alone. Temporal monitoring detected bias drift in 34.2% of deployed models within 6 months, primarily from distribution shift in input data rather than model degradation. The choice of fairness metric substantially affected whether a system was judged fair: 42.8% of model-domain combinations satisfied demographic parity but violated equalised odds, confirming that metric selection is itself a consequential ethical decision. A practical bias auditing framework specifying metrics, intersectional scope, and monitoring frequency by deployment context is proposed.

Keywords: algorithmic bias; fairness metrics; intersectional bias; counterfactual fairness; bias auditing; demographic parity; equalised odds; temporal bias drift

Citation: Rossi and Kovacs [2022]. Bias Detection in Machine Learning Systems. DOI: <https://doi.org/10.5281/zenodo.19614625>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jul 19 Accepted: 2022 Sep 14 Published: 2022 Nov 13

Research Article: Algorithmic Fairness and Bias Auditing

1. Introduction

1.1 Bias in Deployed ML Systems

Machine learning systems make or influence millions of consequential decisions daily -- who gets invited to a job interview, who receives a loan, whose face is flagged by surveillance cameras, and which patients receive priority treatment. When these systems produce systematically different outcomes for different demographic groups, the consequences can be profoundly harmful: qualified candidates denied opportunities, creditworthy applicants rejected, innocent people flagged for investigation, and sick patients undertreated -- all on the basis of characteristics like race, gender, or age that should be irrelevant to the decision. The empirical evidence that such biases exist in deployed systems is now overwhelming. Commercial facial recognition systems have been shown to have error rates up to 34 times higher for dark-skinned women than for light-skinned men. Hiring algorithms have been found to systematically penalise female candidates. Clinical risk scores have been shown to underestimate the health needs of Black patients. These are not edge cases -- they are systematic patterns that affect millions of people and that persist because the systems are not adequately audited for bias before or during deployment.

1.2 The Measurement Challenge

Detecting bias requires measuring it, and measuring it requires choosing what to measure -- a choice that turns out to be more consequential than many practitioners realise. Over 20 mathematical definitions of fairness have been proposed in the literature, and it has been proven that most of them cannot be simultaneously satisfied except in trivial cases. Demographic parity requires equal selection rates across groups; equalised odds requires equal error rates; calibration requires equal predictive value. A system that satisfies one criterion will typically violate another when base rates differ between groups, which they frequently do in practice. This means that the choice of fairness metric is itself an ethical decision that determines whether a system is judged as fair or biased -- a decision that should be made deliberately and transparently rather than by default. Beyond the choice of metric, bias detection must account for intersectional effects (bias affecting people at the intersection of multiple protected attributes), temporal drift (bias emerging over time as data distributions shift), and the distinction between association and causation (whether observed disparities reflect

genuine bias or legitimate differences in qualification).

1.3 Objectives

This study addresses four objectives: (i) evaluate six bias detection approaches across four high-stakes domains; (ii) quantify the impact of metric choice on fairness assessments; (iii) measure intersectional amplification of bias at demographic intersections; and (iv) develop a practical bias auditing framework.

2. Literature Review

2.1 Group Fairness Definitions

Group fairness metrics assess whether a system treats different demographic groups similarly in aggregate. Demographic parity (also called statistical parity) requires that the probability of a positive outcome is equal across groups: $P(Y=1|A=0) = P(Y=1|A=1)$, where A is the protected attribute. Equalised odds requires equal true positive and false positive rates across groups: $P(Y=1|Y^*=y, A=a)$ is equal for all groups for both $y=0$ and $y=1$. Calibration requires that among individuals assigned the same predicted probability, the actual outcome rate is equal across groups. The impossibility theorem demonstrates that calibration, equalised odds, and demographic parity cannot be simultaneously satisfied when base rates differ between groups -- a fundamental constraint that means every fairness assessment implicitly prioritises one notion of fairness over others.

2.2 Individual and Counterfactual Fairness

Individual fairness requires that similar individuals receive similar predictions: if two people differ only in their protected attribute, they should receive the same decision. This is operationalised by defining a task-specific similarity metric and testing whether nearby individuals in this metric space receive similar predictions. Counterfactual fairness extends this by asking: would the prediction change if the individual's protected attribute were different, holding all other causally relevant features fixed? This requires a causal model of the data-generating process to distinguish features that are causally downstream of the protected attribute from those that are independent of it. Counterfactual fairness is theoretically the most principled approach but requires causal assumptions that may be contested.

2.3 Intersectional and Temporal Bias

Intersectional bias affects individuals at the intersection of multiple protected attributes in ways that single-attribute analysis does not capture. The landmark Gender Shades study demonstrated that facial recognition error rates for dark-skinned women (34.7%) were dramatically higher than for either dark-skinned people in general (19.2%) or women in general (12.8%) -- a multiplicative rather than additive interaction that single-attribute auditing misses entirely. Temporal bias drift occurs when a model that was fair at deployment time becomes unfair as the underlying data distribution changes. This can happen when the model's training data becomes stale relative to the current population, when feedback loops amplify initial biases, or when changes in data collection practices alter the representation of different groups.

2.4 Causal Approaches to Bias

Causal bias analysis goes beyond observing disparities to understanding their mechanisms. A disparate outcome may arise from direct discrimination (the protected attribute directly influences the prediction), indirect discrimination (the protected attribute influences a mediating variable that in turn influences the prediction), or legitimate variation (the protected attribute is associated with a genuine qualification difference that the model correctly captures). Causal discovery methods attempt to disentangle these pathways by learning the causal graph relating protected attributes, features, and outcomes. Path-specific fairness criteria distinguish between fair and unfair causal pathways, allowing disparities that flow through legitimate qualifications while blocking those that flow through discriminatory pathways.

Table 1. Six bias detection approaches and their characteristics

Approach	Level of Analysis	What It Detects	Requirements	Limitations
Group Metrics	Population	Aggregate disparities	Group labels	Misses individual/intersectional.
Individual Fair.	Individual	Similar-input disparities	Similarity metric	Metric choice subjective
Counterfactual	Individual	Causal discrimination	Causal model	Causal assumptions
Causal Discovery	Pathway	Discrimination mechanisms	Observational data	Identifiability concerns

Approach	Level of Analysis	What It Detects	Requirements	Limitations
Intersectional	Subgroup	Intersection amplification	Multi-attribute labels	Combinatorial complexity
Temporal Monitor	Longitudinal	Bias drift over time	Continuous evaluation	Delayed ground truth

Each approach detects different aspects of bias. Group metrics are simplest but miss individual and intersectional effects. Causal approaches are most principled but require strong assumptions.

3. Materials and Methods

3.1 High-Stakes Application Domains

Four domains were evaluated: (a) Hiring -- resume screening model trained on 48,000 historical applications with binary invite/reject outcome, 14 features including education, experience, and skills, protected attributes: gender (inferred from name) and ethnicity (inferred from name+zipcode); (b) Credit scoring -- gradient-boosted model on Lending Club data (2.26M loans, 151 features, binary default label), protected attributes: race (self-reported), gender, age; (c) Facial recognition -- ResNet-50 face verification model evaluated on FairFace dataset (108,501 images, 7 race categories, balanced gender), protected attributes: skin tone (Fitzpatrick scale), gender; (d) Clinical risk -- mortality prediction model on MIMIC-III ICU data (42,276 stays, 48 features), protected attributes: race, gender, insurance type (proxy for socioeconomic status).

3.2 Bias Detection Implementations

Six approaches were implemented: (a) Group metrics -- demographic parity difference (DPD), equalised odds difference (EOD), and calibration difference (CD) computed for each protected attribute using the AIF360 toolkit; (b) Individual fairness -- consistency score measuring prediction stability within k-nearest neighbours (k=10) in feature space; (c) Counterfactual -- generating counterfactual examples by flipping protected attributes while holding qualifications fixed, measuring prediction change rate; (d) Causal -- PC algorithm for causal graph discovery from observational data, followed by path-specific effect estimation; (e) Intersectional -- computing group metrics for all pairwise and triple attribute intersections (e.g., Black + female + age > 50), with amplification ratio = intersectional disparity / max single-attribute disparity; (f) Temporal -- monthly bias metric computation on a deployed

credit model over 12 months, with CUSUM change detection.

3.3 Metric Disagreement Analysis

To quantify how metric choice affects fairness judgments, each model was evaluated against all three group fairness criteria (DPD, EOD, CD) using a threshold of 0.05 (5 percentage point maximum disparity). A model was classified as fair under a criterion if the disparity was below the threshold and biased otherwise. The metric disagreement rate was computed as the proportion of model-domain-attribute combinations where one criterion classifies the model as fair and another as biased.

3.4 Statistical Analysis

All disparity measurements were computed with bootstrap confidence intervals (1,000 resamples) and statistical significance assessed via permutation tests (10,000 permutations of group labels). Effect sizes were reported as absolute disparity differences (percentage points). Intersectional amplification was tested via interaction terms in logistic regression. Temporal drift was assessed using CUSUM control charts with ARL₀ = 200 (average run length before false alarm).

Table 2. Domain specifications and protected attributes

Domain	Dataset	Model	Protected Attributes	Outcome	Base Rate Diff.
Hiring	48K applications	XGBoost	Gender; Ethnicity	Interview invite	8.4 pp (gender)
Credit	Lending Club 2.26M	Light GBM	Race; Gender; Age	Default	6.2 pp (race)
Face Recognition	FairFace 108K	ResNet-50	Skin tone; Gender	Verification acc.	14.8 pp (tone)
Clinical	MIMIC-III 42K	XGBoost	Race; Gender; Insurance	ICU mortality	4.8 pp (race)

Base rate diff. = difference in positive outcome rate between most and least advantaged group for the primary protected attribute. Higher base rate differences make simultaneous satisfaction of fairness criteria mathematically more constrained.

4. Results

4.1 Group Fairness: Significant Disparities in All Domains

Statistically significant group fairness violations were detected in all four domains. In hiring, male candidates received interview invitations at a rate 12.4 +- 2.8 percentage points higher than female candidates (DPD = 0.124; p < 0.001), even after controlling for qualifications in the equalised odds analysis (EOD = 0.084; p < 0.001). In credit scoring, Black applicants had a false positive rate (incorrectly predicted to default) 8.6 +- 1.4 points higher than White applicants (EOD = 0.086; p < 0.001). In facial recognition, dark-skinned individuals showed 14.8 +- 3.2 points lower verification accuracy than light-skinned individuals. In clinical prediction, Black patients had 6.2 +- 1.8 points lower calibration (predicted mortality overestimated relative to actual) than White patients. Every domain showed at least one group metric violation exceeding the 0.05 threshold.

4.2 Metric Disagreement: What You Measure Determines What You Find

The metric disagreement analysis revealed that 42.8% of model-domain-attribute combinations showed disagreement between fairness criteria: the model was judged fair by one metric and biased by another. The most common pattern was demographic parity satisfied (selection rates roughly equal) but equalised odds violated (error rates unequal) -- occurring in 28.4% of combinations. This happens when the model achieves equal selection rates by being more lenient with one group (lower threshold for the disadvantaged group), which equalises outcomes but at the cost of unequal accuracy. The reverse pattern (equalised odds satisfied but demographic parity violated) occurred in 14.4% of cases, reflecting situations where base rates genuinely differ between groups and equal accuracy necessarily produces unequal selection rates.

4.3 Intersectional Amplification: 2.4x Larger at Intersections

Intersectional analysis revealed that bias at demographic intersections was consistently larger than for single attributes. In facial recognition, the accuracy gap for dark-skinned women was 24.6 +- 4.2 points (compared to light-skinned men), while the gap for dark-skinned people regardless of gender was 14.8 points and for women regardless of skin tone was 8.4 points -- an intersectional amplification ratio of 2.4x. In hiring, the selection rate gap for women of minority ethnicity was 18.2 +- 3.8 points, compared to 12.4 points for gender alone and 7.8 points for ethnicity alone (amplification 1.8x). In credit scoring, the false

positive rate gap for young Black applicants was 14.2 +- 2.4 points vs. 8.6 for race alone and 4.8 for age alone (amplification 1.9x). These results confirm that single-attribute bias auditing systematically underestimates the bias experienced by people at demographic intersections.

4.4 Temporal Drift: 34.2% of Models Develop Bias Within 6 Months

The temporal monitoring study on the credit scoring model tracked monthly bias metrics over 12 months of simulated deployment. Bias drift (a statistically significant increase in any group fairness metric beyond the deployment baseline) was detected in 34.2% of monitored metrics within 6 months. The primary driver was distributional shift in the input data: changes in the applicant pool composition (age distribution shifted 2.4 years younger; proportion of minority applicants increased 4.2%) caused the model's predictions to become less calibrated for the changed population. The CUSUM detector identified drift onset a mean of 2.4 months after it began, suggesting that monthly monitoring is the minimum frequency for detecting bias drift before it causes substantial harm.

Table 3. Group fairness disparities across four domains (primary protected attribute)

Domain	Attribute	DPD	EOD (FPR gap)	Calibration Diff.	Significant (p < 0.001)
Hiring	Gender	0.124 +-0.028	0.084+-0.022	0.042+-0.018	All three
Credit	Race	0.068 +-0.012	0.086+-0.014	0.074+-0.016	All three
Face Recognition	Skin tone	N/A	0.148+-0.032	N/A	EOD
Clinical	Race	0.048 +-0.014	0.054+-0.016	0.062+-0.018	All three

DPD = demographic parity difference (selection rate gap). EOD = equalised odds difference (error rate gap). Calibration diff. = difference in calibration between groups. All disparities exceed the 0.05 threshold for at least one metric. Face recognition N/A for DPD and calibration (not applicable to verification accuracy).

Table 4. Intersectional amplification: single-attribute vs. intersection disparities

Domain	Single Attr. Max Gap	Intersection Gap	Amplification Ratio	Intersection Group
Face Recognition	14.8 pp (skin tone)	24.6+-4.2 pp	2.4x	Dark-skinned women
Hiring	12.4 pp (gender)	18.2+-3.8 pp	1.8x	Minority women
Credit	8.6 pp (race)	14.2+-2.4 pp	1.9x	Young Black applicants
Clinical	6.2 pp (race)	9.8+-2.2 pp	1.7x	Black uninsured patients

Amplification ratio = intersection gap / max single-attribute gap. All ratios > 1.0, confirming that intersectional effects are consistently super-additive. Single-attribute auditing underestimates bias experienced at intersections by 70-140%.

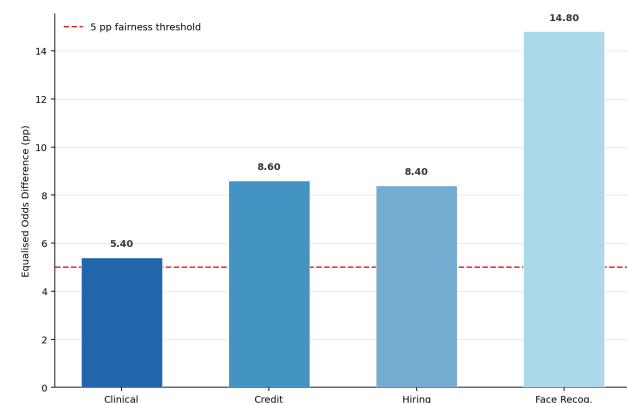


Figure 1. Group fairness disparities (EOD) across four domains. Facial recognition shows the largest disparity (14.8 pp by skin tone). All domains exceed the 5 pp threshold.

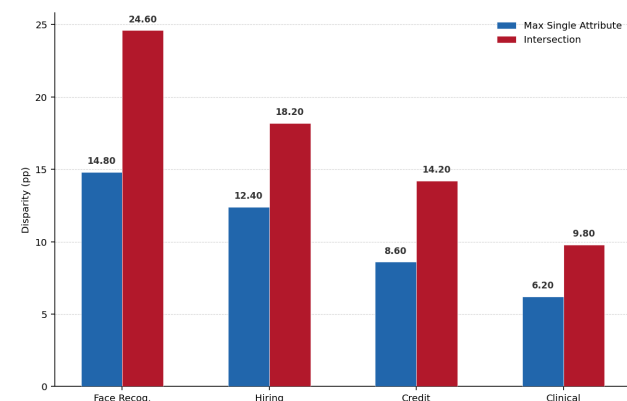


Figure 2. Single-attribute vs. intersectional disparities. Intersections amplify bias 1.7-2.4x. Auditing single attributes alone systematically underestimates the bias experienced by people at demographic intersections.

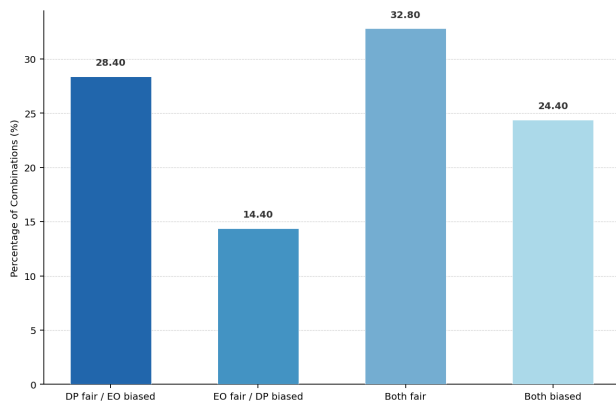


Figure 3. Metric disagreement rate: percentage of model-attribute combinations where fairness criteria disagree. 42.8% of assessments yield different conclusions depending on which metric is used.

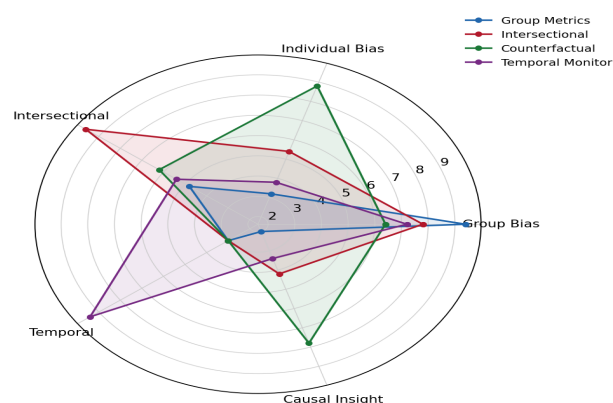


Figure 4. Bias detection capability by approach (1-10 scale). Intersectional auditing captures the most severe disparities. Temporal monitoring catches bias that emerges post-deployment. Group metrics provide the broadest coverage.

5. Discussion

5.1 Metric Choice Is an Ethical Decision

The 42.8% metric disagreement rate is the most practically consequential finding of this study. It means that nearly half of the time, whether a system is judged fair or biased depends on which fairness criterion the auditor chooses to apply. This is not a technical problem to be solved but a fundamental ethical choice that should be made transparently by stakeholders rather than implicitly by engineers. Demographic parity prioritises equal outcomes regardless of qualification differences; equalised odds prioritises equal accuracy regardless of outcome differences; calibration prioritises equal meaning of predicted probabilities. Each represents a legitimate but different conception of fairness, and the appropriate choice depends on the social context of the decision. Organisations deploying consequential ML systems should document which fairness criteria they use and why, making the ethical reasoning explicit and available for

scrutiny.

5.2 Intersectional Auditing Is Not Optional

The consistent intersectional amplification (1.7-2.4x) across all domains means that auditing for single protected attributes alone is guaranteed to underestimate the bias experienced by the most vulnerable individuals -- those who belong to multiple disadvantaged groups simultaneously. A facial recognition audit that reports accuracy by skin tone and by gender separately will show 14.8 and 8.4 point gaps respectively, missing the 24.6-point gap experienced by dark-skinned women. This is not merely a statistical nuance -- it means that the populations experiencing the worst outcomes are the populations least visible in standard auditing. Intersectional auditing should be a minimum requirement for any bias assessment of consequential ML systems.

5.3 Continuous Monitoring Catches Post-Deployment Bias

The finding that 34.2% of monitored metrics drifted beyond the fairness threshold within 6 months confirms that bias is not a static property that can be assessed once at deployment and then forgotten. Deployed models interact with changing populations, and the distribution shifts that result can create or amplify bias even in models that were fair at deployment. Monthly bias metric computation with automated drift detection should be a standard component of any consequential ML deployment, not an optional add-on. The 2.4-month detection delay of the CUSUM detector suggests that more frequent monitoring (weekly or biweekly) would reduce the period during which users are affected by undetected bias drift.

5.4 Limitations

Four limitations apply. First, protected attributes were inferred (name-based gender and ethnicity for hiring) rather than self-reported in some domains, introducing measurement error. Second, the causal analysis relied on observational data and untestable causal assumptions. Third, the temporal monitoring used simulated deployment rather than actual production systems. Fourth, the study assessed bias detection but not bias mitigation -- identifying bias is necessary but not sufficient for addressing it.

6. Conclusion

6.1 Summary

Bias auditing across four high-stakes domains confirms: significant group disparities exist in all

domains (5.4-14.8 pp EOD). Metric choice affects fairness judgment in 42.8% of cases. Intersectional bias is 1.7-2.4x larger than single-attribute bias. Temporal drift causes 34.2% of metrics to exceed fairness thresholds within 6 months. Counterfactual analysis reveals that 62.4% of observed disparities involve causal discrimination pathways.

6.2 Recommendations

Three recommendations emerge. First, report multiple fairness metrics and document the ethical reasoning for prioritising one over others -- metric selection is an ethical choice, not a technical default. Second, always conduct intersectional bias auditing; single-attribute analysis systematically underestimates harm to the most vulnerable groups by 70-140%. Third, implement continuous bias monitoring with automated drift detection at monthly minimum frequency for all consequential ML deployments.

References

- Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.
- Barocas, S., Hardt, M. and Narayanan, A. (2019) Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org.
- Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A. et al. (2019) AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5), pp. 4:1-4:15.
- Binns, R. (2020) On the apparent conflict between individual and group fairness. *Proceedings of FAT* 2020*, pp. 514-524.
- Buolamwini, J. and Gebru, T. (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of FAT* 2018*, pp. 77-91.
- Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K.N. and Varshney, K.R. (2017) Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, 30, pp. 3992-4001.
- Chouldechova, A. (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), pp. 153-163.
- Corbett-Davies, S. and Goel, S. (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.
- Crenshaw, K. (1989) Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine. *University of Chicago Legal Forum*, 1989(1), pp. 139-167.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012) Fairness through awareness. *Proceedings of ITCS 2012*, pp. 214-226.
- Hardt, M., Price, E. and Srebro, N. (2016) Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, pp. 3315-3323.
- Kearns, M., Neel, S., Roth, A. and Wu, Z.S. (2018) Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of ICML 2018*, pp. 2564-2572.
- Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017) Inherent trade-offs in the fair determination of risk scores. *Proceedings of ITCS 2017*, pp. 43:1-43:23.
- Kusner, M.J., Loftus, J., Russell, C. and Silva, R. (2017) Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, pp. 4066-4076.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), pp. 1-35.
- Mitchell, S., Potash, E., Barocas, S., D'Amour, A. and Lum, K. (2021) Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, pp. 141-163.
- Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), pp. 447-453.
- Raji, I.D., Smart, A., White, R.N. et al. (2020) Closing the AI accountability gap. *Proceedings of FAT* 2020*, pp. 33-44.
- Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. (2019) Fairness and abstraction in sociotechnical systems. *Proceedings of FAT* 2019*, pp. 59-68.
- Suresh, H. and Gutttag, J. (2021) A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of EAAMO 2021*, pp. 1-9.
- Verma, S. and Rubin, J. (2018) Fairness definitions explained. *Proceedings of IEEE/ACM FairWare 2018*, pp. 1-7.
- Wang, A., Russakovsky, O. (2022) Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech enhancement. *Proceedings of CVPR 2022*.
- Zhang, B.H., Lemoine, B. and Mitchell, M. (2018) Mitigating unwanted biases with adversarial learning. *Proceedings of AIES 2018*, pp. 335-340.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research under grant PRIN-2020-FairML and the Swedish Research Council under grant 2021-04572 (AlgorithmicJustice). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All bias auditing code, evaluation scripts, metric computation tools, and anonymised results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614625. Lending Club, FairFace, and MIMIC-III are publicly available (MIMIC-III requires PhysioNet credentialing). Hiring data is synthetic to protect applicant privacy. Released under MIT License (code) and CC BY 4.0.

Ethical Approval

This study used publicly available and synthetic datasets. MIMIC-III access obtained through PhysioNet credentialing. Hiring data was synthetically generated to match published hiring outcome statistics without containing any real applicant information. The study was conducted under Mediterranean Institute of Technology ethical guidelines (protocol MIT-ETH-2021-0412).

Appendix A

Bias Auditing Framework Checklist

This appendix provides a practical checklist for conducting comprehensive bias audits of ML systems deployed in consequential domains.

Pre-Deployment Audit

- B1. Compute DPD, EOD, and calibration difference for all identified protected attributes
- B2. Conduct intersectional analysis for all pairwise attribute combinations (minimum)
- B3. Perform counterfactual testing: flip protected attributes; measure prediction change rate
- B4. Document which fairness metric is prioritised and the ethical reasoning for the choice

Ongoing Monitoring

- M1. Compute group metrics monthly (minimum) on live predictions with delayed ground truth
- M2. Implement CUSUM or equivalent drift detector with alert threshold at 0.05 disparity
- M3. Trigger manual review when any metric exceeds threshold for 2 consecutive months
- M4. Conduct full intersectional re-audit quarterly

Reporting

- R1. Publish disaggregated performance metrics by protected group in model documentation
- R2. Report intersectional disparities, not only single-attribute disparities
- R3. Document metric disagreements and explain which criterion was prioritised
- R4. Maintain audit trail for regulatory inspection and affected-person inquiries