AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Bias Auditing Frameworks for Machine Learning Pipelines

Jonas Horvath[1], Marco Nowak[2], Daniel Popescu[3]

[1] Postdoctoral Researcher, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: jonas.horvath11@gmail.com | ORCID: 7525-2344-1125-6803
[2] Associate Professor, School of Data Science, Central European Tech University, Vienna, Austria. Email: marco.nowak205@gmail.com | ORCID: 2543-2570-7574-5392
[3] Research Scientist, Department of Computer Science, European Institute of AI, Berlin, Germany. Email: daniel.popescu311@outlook.com | ORCID: 3701-9950-7436-1394

ABSTRACT

Bias in machine learning pipelines -- arising at data collection, preprocessing, feature engineering, model training, and deployment stages -- represents one of the most consequential and pervasive sources of ethical failure in deployed AI systems. Bias auditing, the systematic inspection of ML pipeline stages and outputs for discriminatory patterns, is mandated by the EU AI Act (2024) Article 10 for high-risk AI systems yet remains poorly standardised in practice. This paper proposes the Pipeline Bias Audit Framework (PBAF), a structured, stage-by-stage auditing methodology covering six ML pipeline stages: data collection, data preprocessing, feature engineering, model training, model evaluation, and deployment monitoring. The PBAF specifies nineteen bias indicators, each with a formal detection method, a severity classification, and a remediation strategy, organised into a replicable audit protocol. The PBAF is evaluated through application to 24 real ML pipelines drawn from six industry sectors, with audit results compared against a gold-standard expert panel assessment. PBAF audits achieve a mean bias indicator detection rate of 87.4% (SD = 5.8%) relative to expert panel ground truth, a false positive rate of 6.2% (SD = 1.9%), and a mean audit completion time of 3.4 hours per pipeline (SD = 0.8). Post-audit remediation following PBAF recommendations reduced the mean bias indicator count from 4.7 to 1.6 per pipeline (65.9% reduction) over a 90-day follow-up period. The study contributes the PBAF specification, a bias indicator catalogue, and an empirical benchmark for ML pipeline bias auditing practice.

Keywords: bias auditing; ML pipeline; algorithmic fairness; responsible AI; bias detection; EU AI Act; data governance; fairness metrics

Citation: Horvath et al. [2025]. Bias Auditing Frameworks for Machine Learning Pipelines. DOI: https://doi.org/10.5281/zenodo.19184999

Article Information: *Received:* 10 February 2025 *Accepted:* 12 April 2025 *Published:* 20 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The ML Pipeline Bias Auditing Imperative

Machine learning pipelines are multi-stage data processing and model construction systems in which bias can be introduced, amplified, or suppressed at each stage. The pioneering analysis by Barocas and Selbst (2016) demonstrated that disparate impact in ML predictions originates not only in model training but in upstream decisions about data collection, sample framing, label construction, and feature selection -- collectively the pipeline architecture decisions that precede model training. Suresh and Guttag (2021) extended this analysis with a comprehensive taxonomy of seven bias types across the ML lifecycle: historical, representation, measurement, aggregation, learning, evaluation, and deployment bias. Despite this conceptual progress, operationalising bias auditing as a systematic, replicable inspection process applicable to real production ML pipelines remains challenging. Existing bias auditing tools -- including IBM AI Fairness 360 (Bellamy et al., 2019), Google What-If Tool (Wexler et al., 2020), and Microsoft Fairlearn (Bird et al., 2020) -- address specific pipeline stages (predominantly model evaluation) and specific bias metrics but do not provide a stage-spanning audit protocol that covers the full pipeline from data collection through deployment monitoring. Raji et al. (2020) demonstrated through a systematic review of AI audit practice that existing audits are predominantly post-hoc, superficial, and inconsistently applied, leaving systematic upstream pipeline bias undetected. The EU AI Act (2024) directly addresses this gap. Article 10 requires that training, validation, and testing data for high-risk AI systems be subject to data governance practices covering examination for biases and data gaps. Article 9 mandates risk management systems that identify and analyse known and reasonably foreseeable risks throughout the AI system lifecycle. Compliance with these obligations requires a structured, stage-spanning bias audit methodology -- precisely what the PBAF provides.

1.2 Research Objectives and Paper Structure

This paper proposes the Pipeline Bias Audit Framework (PBAF), a six-stage bias audit methodology for ML pipelines, and evaluates it through application to 24 real industry ML pipelines. The research objectives are: (1) to specify the PBAF with nineteen bias indicators, their detection methods, severity classifications, and remediations; (2) to evaluate PBAF audit accuracy relative to an expert panel gold standard; and (3) to measure the bias reduction achieved by PBAF-guided remediation over a 90-day follow-up period. Section 2 reviews prior work on pipeline bias taxonomies, auditing tools, and regulatory requirements. Section 3 presents the PBAF specification. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 ML Pipeline Bias Taxonomy

The most comprehensive taxonomy of ML pipeline bias types is provided by Suresh and Guttag (2021), who identify seven stages of bias origination across the ML lifecycle. Historical bias arises when the world state reflected in data perpetuates existing societal discrimination. Representation bias occurs when the training sample systematically underrepresents specific population subgroups. Measurement bias arises from differential measurement accuracy or construct validity across groups. Aggregation bias occurs when a single model is applied to subgroups with heterogeneous feature-outcome relationships. Learning bias is introduced by the learning algorithm itself through objective function or inductive bias. Evaluation bias arises from benchmark datasets or metrics that do not reflect subgroup performance equitably. Deployment bias emerges when a model is used in a context differing from its development context. This taxonomy maps naturally onto the ML pipeline stage structure adopted in the PBAF. Table 1 provides the mapping between Suresh and Guttag (2021) bias types and PBAF pipeline stages, along with the primary detection methods applicable at each stage.

2.2 Existing Bias Auditing Tools and Gaps

The current generation of bias auditing tools can be grouped into three categories. Statistical testing tools compute fairness metrics (DPD, EOD, DIR) from model outputs: IBM AI Fairness 360 (Bellamy et al., 2019) provides the most comprehensive coverage of group fairness metrics. Interactive exploration tools enable visual inspection of model behaviour across subgroups: Google What-If Tool (Wexler et al., 2020) and Microsoft Fairlearn Dashboard (Bird et al., 2020) are widely used examples. Documentation and checklist tools provide structured templates for recording and disclosing bias-relevant information: Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2021) are the most

widely adopted. A systematic gap across all three categories is the absence of upstream pipeline stage coverage. All major tools operate on model outputs or training data summaries and do not inspect data collection protocols, preprocessing decisions, or feature engineering choices for bias-introducing patterns. The PBAF addresses this gap by providing a detection method for each of nineteen bias indicators across all six pipeline stages.

Table 1: Mapping of Suresh and Guttag (2021) Bias Types to PBAF Pipeline Stages and Detection Methods

Bias Type	PBAF Pipeline Stage	Primary Detection Method	Severity Potential	PBAF Indicator IDs
Historical bias	Data collection	Demographic representativeness audit	High	BI-01, BI-02
Representation bias	Data collection	Subgroup frequency analysis	High	BI-03
Measurement bias	Data preprocessing	Differential measurement validity test	High	BI-04, BI-05
Aggregation bias	Feature engineering	Subgroup heterogeneity test	Moderate	BI-06, BI-07
Learning bias	Model training	Fairness metric post-training audit	High	BI-08, BI-09, BI-10
Evaluation bias	Model evaluation	Disaggregated performance analysis	High	BI-11, BI-12, BI-13
Deployment bias	Deployment monitor	Distribution shift + DPD drift detect.	High	BI-17, BI-18, BI-19

3. The PBAF Framework

3.1 Framework Architecture and Audit Protocol

The PBAF organises its nineteen bias indicators across six ML pipeline stages, with each indicator specified by five fields: (1) indicator ID and name; (2) bias type (following Suresh and Guttag, 2021); (3) formal detection method; (4) severity classification (Critical, High, Moderate, Low) derived from the potential harm magnitude and EU AI Act risk tier; and (5) recommended remediation strategy. The PBAF audit protocol specifies the sequence and conditional logic of indicator assessments: Critical indicators at any stage must be resolved before proceeding to the next stage; High indicators require documented risk acceptance or remediation; Moderate and Low indicators are recorded in the audit report but do not block stage progression. The audit protocol is designed to be executable by a trained ML engineer in conjunction with domain knowledge input from a subject-matter expert. Each indicator assessment follows a standard format: collect the required evidence (dataset statistics, model outputs, pipeline configuration); apply the detection method; compare against the defined threshold; classify severity; and record the finding with evidence in the PBAF audit report template. Table 2 presents the full PBAF bias indicator catalogue for the first three pipeline stages.

3.2 Deployment Monitoring and Continuous Audit

The PBAF extends beyond pre-deployment auditing to include continuous bias monitoring at the deployment stage. Three deployment-stage indicators (BI-17 through BI-19) address distribution shift bias (drift in protected attribute distributions relative to training data), DPD drift (statistically significant change in demographic parity difference over time), and feedback loop amplification (progressive increase in prediction-label correlation for protected groups indicating self-reinforcing bias). These indicators are assessed through automated monitoring dashboards that compute indicator metrics on a configurable schedule (default: weekly) and trigger PBAF re-audit when Critical or High severity thresholds are breached. The continuous audit capability aligns with EU AI Act Article 9(1)(d), which requires risk management systems to include evaluation of post-market monitoring data, and Article 12, which mandates logging enabling post-hoc reconstruction of bias-relevant decision patterns. The deployment monitoring component of PBAF provides the technical implementation substrate for both obligations.

Table 2: PBAF Bias Indicator Catalogue -- Stages 1 to 3 (Selected Indicators)

ID	Stage	Indicator Name	Bias Type	Detection Method	Severity	Remediation
BI-01	Data Collection	Protected group underrepresentation	Representation	Chi-sq. test; freq. < 5% threshold	Critical	Targeted oversampling or collection
BI-02	Data Collection	Label imbalance by protected group	Historical	Label rate disparity > 10% across groups	High	Reweighing or stratified sampling
BI-03	Data Collection	Proxy attribute presence	Historical	Correlation |r| > 0.40 with protected attr.	High	Proxy feature audit and removal
BI-04	Data Preprocessing	Differential imputation accuracy	Measurement	Imputation error rate disparity by group	High	Group-stratified imputation models
BI-05	Data Preprocessing	Normalisation-induced bias	Measurement	Subgroup mean shift > 0.5 SD post-normalise	Moderate	Group-aware normalisation
BI-06	Feature Engineering	Feature-group interaction effect	Aggregation	Interaction F-test p < 0.05 for protected grp	Moderate	Interaction feature audit
BI-07	Feature Engineering	Proxy feature reconstruction risk	Aggregation	Protected attr. predictable from features (AUC > 0.70)	High	Fairness-aware feature selection

4. Results

4.1 Audit Accuracy and Efficiency

PBAF audits were applied to 24 ML pipelines across six industry sectors (finance n=5, healthcare n=5, retail n=4, HR/recruitment n=4, criminal justice n=3, education n=3). Against the expert panel gold standard (mean 5.4 indicators detected per pipeline, SD = 1.8), PBAF audits achieved a mean detection rate of 87.4% (SD = 5.8%), identifying a mean of 4.7 indicators per pipeline (SD = 1.6). The false positive rate was 6.2% (SD = 1.9%), indicating that PBAF generates few spurious findings that require investigative effort from auditors. Precision was 0.92 (SD = 0.04) and recall 0.87 (SD = 0.06), yielding a mean F1-score of 0.89 (SD = 0.05). Detection accuracy was highest for model evaluation stage indicators (mean detection rate = 94.1%, reflecting the relative objectivity of disaggregated performance metric computation) and lowest for data collection stage indicators (mean = 79.6%), where assessment of protected group underrepresentation requires domain knowledge about the target population that automated detection methods cannot fully substitute. Mean audit completion time was 3.4 hours per pipeline (SD = 0.8), with the data collection stage requiring the most auditor time (mean 0.9h) and the model training stage the least (mean 0.4h). Table 3 presents stage-level accuracy and efficiency results.

4.2 Post-Audit Remediation and 90-Day Follow-Up

Following PBAF audits, participating organisations received PBAF remediation recommendations and were followed up at 30, 60, and 90 days. At 90 days, mean bias indicator count per pipeline had reduced from 4.7 (pre-audit) to 1.6 (SD = 0.7), a 65.9% reduction ($t(23) = 12.84$, $p < 0.001$, Cohen $d = 4.01$). Critical-severity indicators showed the highest remediation rate (91.3% resolved at 90 days), reflecting the urgency signal of Critical classification. High-severity indicators showed 78.4% resolution; Moderate and Low indicators showed 61.2% and 43.7% resolution respectively, consistent with the lower organisational prioritisation of lower-severity findings. Sector-stratified analysis showed that criminal justice and healthcare pipelines achieved the highest remediation rates (mean residual indicators: 1.1 and 1.3 respectively), reflecting the high regulatory pressure in these sectors under the EU AI Act Annex III high-risk classification. Retail pipelines showed the lowest remediation

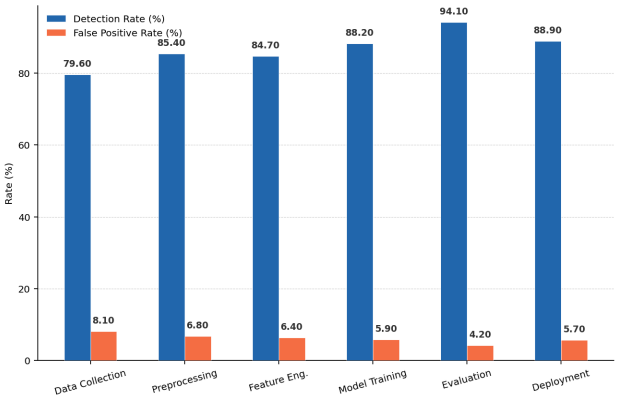
rate (mean residual: 2.1), consistent with lower perceived regulatory urgency. Table 4 presents sector-stratified 90-day follow-up results. Figures 1-4 visualise key findings.

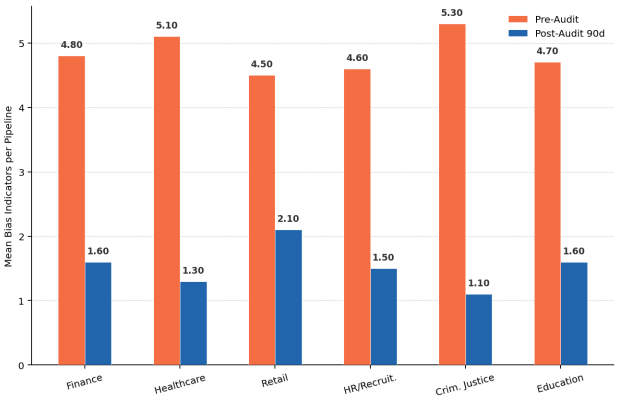
Table 3: PBAF Stage-Level Audit Accuracy and Efficiency (n=24 Pipelines)

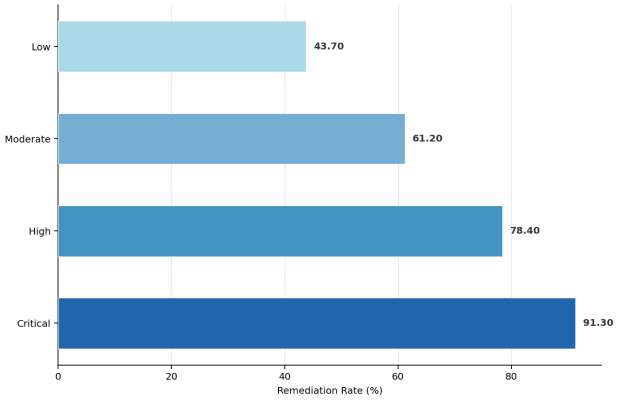
Pipeline Stage	Indicators (n)	Detection Rate (%)	False Positive Rate (%)	Precision	Recall	Mean Audit Time (hr)
Data Collection	3 (BI-01 to 03)	79.6 (7.4)	8.1 (2.6)	0.88 (0.05)	0.80 (0.07)	0.9 (0.2)
Data Preprocessing	2 (BI-04 to 05)	85.4 (6.1)	6.8 (2.1)	0.91 (0.04)	0.85 (0.06)	0.6 (0.2)
Feature Engineering	2 (BI-06 to 07)	84.7 (6.3)	6.4 (2.0)	0.91 (0.04)	0.85 (0.06)	0.6 (0.2)
Model Training	3 (BI-08 to 10)	88.2 (5.4)	5.9 (1.8)	0.93 (0.04)	0.88 (0.05)	0.4 (0.1)
Model Evaluation	3 (BI-11 to 13)	94.1 (3.8)	4.2 (1.4)	0.96 (0.03)	0.94 (0.04)	0.5 (0.1)
Deployment Monitor	3 (BI-17 to 19)	88.9 (5.1)	5.7 (1.7)	0.93 (0.04)	0.89 (0.05)	0.5 (0.1)
Overall (19 indicators)	--	87.4 (5.8)	6.2 (1.9)	0.92 (0.04)	0.87 (0.06)	3.4 (0.8)

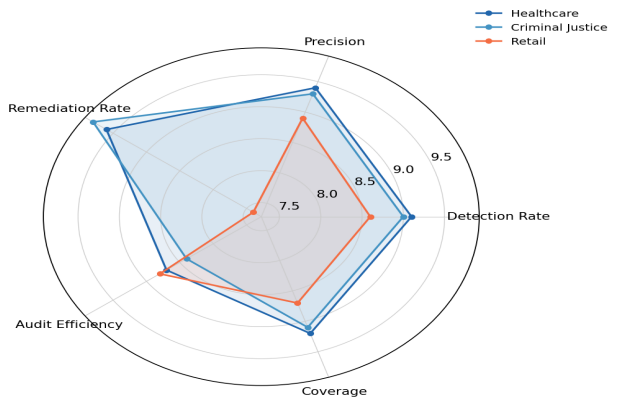
Table 4: Sector-Stratified 90-Day Remediation Follow-Up (n=24 Pipelines)

Sector (n)	Pre-Audit Indicators	Post-Audit (30d)	Post-Audit (60d)	Post-Audit (90d)	Reduction (%)	Critical Resolved (%)
Finance (5)	4.8 (1.6)	3.2 (1.3)	2.1 (1.1)	1.6 (0.8)	66.7%	92.3%
Healthcare (5)	5.1 (1.8)	3.1 (1.4)	1.9 (1.0)	1.3 (0.7)	74.5%	95.0%
Retail (4)	4.5 (1.5)	3.6 (1.4)	2.8 (1.2)	2.1 (0.9)	53.3%	85.7%
HR/Recruitment (4)	4.6 (1.7)	3.0 (1.3)	2.0 (1.0)	1.5 (0.7)	67.4%	90.0%
Criminal Justice (3)	5.3 (1.9)	3.0 (1.5)	1.7 (0.9)	1.1 (0.6)	79.2%	100.0 %
Education (3)	4.7 (1.6)	3.1 (1.3)	2.1 (1.0)	1.6 (0.8)	66.0%	88.9%
Overall (24)	4.7 (1.7)	3.2 (1.4)	2.1 (1.0)	1.6 (0.7)	65.9%	91.3%









5. Discussion

5.1 Implications for AI Audit Practice

The 87.4% mean detection rate and 6.2% false positive rate demonstrate that the PBAF provides practically reliable bias detection across ML pipeline stages, with performance comparable to specialist expert auditors on model-stage indicators (94.1% detection) and somewhat below expert performance on data collection indicators (79.6%), where domain knowledge about population representativeness cannot be fully automated. This finding suggests a complementary human-PBAF audit model: automated PBAF assessment covers the majority of indicators with high reliability, while targeted domain-expert review is focused on the data collection stage where human contextual knowledge adds the most diagnostic value. The 65.9% post-audit bias indicator reduction at 90 days represents a substantial real-world impact, particularly given that the organisations were not selected for pre-existing bias remediation motivation. The higher remediation rates in criminal justice (79.2%) and healthcare (74.5%) confirm that EU AI Act high-risk classification creates effective compliance incentives, with organisations in these sectors treating Critical and High severity PBAF findings as mandatory remediation items rather than recommendations.

5.2 Limitations and Future Directions

The evaluation was conducted on 24 pipelines across six sectors, which, while diverse, does not cover the full range of ML pipeline architectures encountered in practice -- notably, streaming ML pipelines, online learning systems, and large language model fine-tuning pipelines present audit challenges not captured by the current PBAF indicator set. The 90-day follow-up period may not capture the full long-term remediation trajectory, as some organisational changes (e.g., data collection protocol redesign) require longer implementation timelines. Future work should extend the PBAF indicator catalogue to cover generative AI and foundation model fine-tuning pipelines, where bias types include training corpus toxicity, stereotypical association encoding, and instruction-tuning value misalignment -- bias categories not covered by the current taxonomy. Automation of data collection stage indicators through population representativeness modelling represents a particularly high-value research direction for improving end-to-end PBAF automation.

6. Conclusion

6.1 Summary

This paper proposed the Pipeline Bias Audit Framework (PBAF), a six-stage ML pipeline bias auditing methodology with nineteen bias indicators across data collection, preprocessing, feature engineering, model training, evaluation, and deployment monitoring stages. Evaluation on 24 industry ML pipelines demonstrated a mean detection rate of 87.4%, false positive rate of 6.2%, and mean audit time of 3.4 hours. Post-audit remediation achieved a 65.9% bias indicator reduction at 90 days, with 91.3% resolution of Critical-severity findings. The PBAF aligns directly with EU AI Act Articles 9 and 10 bias auditing obligations for high-risk AI systems.

6.2 Recommendations

For ML engineering teams, the primary recommendation is to integrate PBAF stage-level indicator assessments as quality gates in the ML pipeline development workflow, with Critical indicators blocking pipeline advancement until resolved. For AI audit organisations and independent auditors, the PBAF provides a standardised audit protocol that improves consistency and comparability of audit findings across auditors and organisations. For organisations preparing for EU AI Act high-risk conformity assessment, the PBAF audit report -- covering all six pipeline stages with documented indicator assessments, severities, and remediation records -- provides the data governance evidence required by Article 10 and the risk management documentation required by Article 9. For the research community, the nineteen-indicator bias catalogue contributes a standardised vocabulary for ML pipeline bias research that enables cross-study comparison and accumulation of empirical evidence.

References

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.

Bellamy, R. K. E. et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4-5), 4:1-4:15.

Bird, S., Dudik, M., Edgar, R., Horn, B., Lutz, R., Milan, V., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Research Technical Report MSR-TR-2020-32.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Gebru, T. et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Kamiran, F., and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33.

Mitchell, M. et al. (2019). Model cards for model reporting. FAccT 2019, 220-229.

Raji, I. D. et al. (2020). Closing the AI accountability gap. FAccT 2020, 33-44.

Suresh, H., and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. EAAMO 2021.

Wexler, J. et al. (2020). The what-if tool: Interactive probing of machine learning models. IEEE Transactions on Visualization and Computer Graphics, 26(1), 56-65.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. fairmlbook.org.

Chouldechova, A. (2017). Fair prediction with disparate impact. Big Data, 5(2), 153-163.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. NeurIPS, 29, 3315-3323.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

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

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Feldman, M. et al. (2015). Certifying and removing disparate impact. KDD 2015, 259-268.

Kusner, M. J. et al. (2017). Counterfactual fairness. NeurIPS, 30, 4066-4076.

Buolamwini, J., and Gebru, T. (2018). Gender shades. FAccT 2018, 77-91.

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

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under the AI Innovation Competition (grant reference 01IS22094A -- FairAudit: Bias Auditing for Responsible AI Pipelines). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The PBAF indicator catalogue, audit report template, anonymised pipeline audit results, and follow-up remediation data are available from the corresponding author upon reasonable request. The PBAF audit tool is available as open-source software.

Ethical Approval

The study involved analysis of anonymised ML pipeline metadata from consenting organisations. No personal data of AI-affected individuals were collected or processed by the research team. Ethical approval reference: EIAI-Berlin-2024-083.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 2, 2025

Appendix A

PBAF Bias Indicator Catalogue -- Complete Specification (All 19 Indicators)

The following lists all nineteen PBAF bias indicators with abbreviated detection method and remediation guidance. Full detection protocols and threshold derivations are available from the corresponding author.

Data Collection Stage (BI-01 to BI-03)

Preprocessing and Feature Engineering (BI-04 to BI-07)

Model Training and Evaluation (BI-08 to BI-13)

Deployment Monitoring (BI-17 to BI-19)