Detecting Hidden Bias in Machine Learning Models: A Systematic Approach to Fairness Evaluation and Mitigation

Debabrata Pruseth
AI Architect & Applied AI Researcher
Singapore

Author Note

This article is a research-style companion version of author's blog post "How to Detect Hidden Bias in Your ML Model — A Step-by-Step Tutorial" and associated GitHub code experimentation "FairHire-Detecting-and-Fixing-Bias-in-AI-Hiring-Models".

This research presents an applied AI framework for detecting, interpreting, and mitigating hidden bias in machine learning systems. The study focuses on practical fairness evaluation in supervised decision models, particularly models used in domains such as hiring, lending, admissions, risk scoring, and customer eligibility assessment. The central motivation is that predictive accuracy alone is insufficient when model outcomes affect people differently across demographic or protected groups.

The manuscript is framed as an implementation-oriented research study. It develops a systematic workflow that connects data inspection, sensitive-group analysis, group-level performance evaluation, disparate impact measurement, fairness-aware mitigation, and responsible governance. The approach is designed for AI practitioners, data scientists, product teams, auditors, and decision-makers who need to evaluate whether a machine learning model is not only performant, but also equitable, explainable, and operationally accountable.

Abstract

Machine learning models increasingly influence high-stakes decisions in employment, lending, education, healthcare, insurance, and public services. Although these systems are often evaluated using aggregate accuracy, precision, recall, or area-under-curve metrics, such measurements can conceal unequal performance across demographic groups. A model may appear statistically effective overall while producing systematically unfavorable outcomes for a protected or underrepresented group. This study presents a systematic applied framework for detecting hidden bias in machine learning models and translating fairness evaluation into practical mitigation steps. The methodology begins with identification of decision context, target outcome, protected attributes, and possible proxy variables. It then evaluates model outcomes using group-level selection rates, error rates, false positive rates, false negative rates, and disparate impact ratios. The framework emphasizes that fairness cannot be reduced

to a single metric; rather, fairness evaluation requires contextual interpretation, stakeholder review, domain-specific thresholds, and continuous monitoring. Mitigation strategies are organized into pre-processing, in-processing, and post-processing interventions, including dataset rebalancing, feature review, fairness-constrained learning, threshold adjustment, and human oversight. The study also discusses responsible AI governance, including transparency, auditability, documentation, privacy protection, and risk of overclaiming fairness. The contribution of this work is a practical, structured fairness-evaluation workflow that supports both technical analysis and governance-oriented decision-making. The study is conceptual and implementation-oriented; it does not claim benchmark superiority or empirical validation across datasets.

Keywords

Machine learning fairness; hidden bias; responsible AI; algorithmic bias; disparate impact ratio; demographic parity; equalized odds; fairness evaluation; bias mitigation; model governance; AI audit; Fairlearn; AI Fairness 360; explainable AI; trustworthy AI

---

1. Introduction

Machine learning systems are increasingly used to support decisions that affect access to opportunity, credit, employment, education, healthcare, insurance, public resources, and digital services. In these settings, model performance is not only a technical question but also a social, ethical, and governance concern. A model that performs well on average may still produce systematic disadvantage for a subgroup if the training data, labels, feature engineering process, or decision threshold reflects historical inequity or statistical imbalance.

The central problem addressed in this study is hidden bias: bias that is not immediately visible through aggregate performance metrics. A classifier may achieve high accuracy while producing lower selection rates for one demographic group, higher false rejection rates for another, or higher false positive rates for a group that is exposed to greater downstream harm. Such patterns can remain undetected when evaluation is limited to overall accuracy, global loss, or business conversion metrics.

Bias can enter machine learning systems through multiple pathways. These include historically biased labels, underrepresentation in training data, measurement error, proxy variables correlated with sensitive attributes, imbalanced class distributions, feature leakage, biased sampling, and deployment environments that differ from the training context. Removing a protected attribute from the dataset is not sufficient when other variables indirectly encode similar information.

This research examines the following question:

How can machine learning practitioners systematically detect, interpret, and mitigate hidden bias in predictive models while preserving responsible governance and avoiding unsupported fairness claims?

The proposed answer is a structured fairness-evaluation workflow that integrates technical metrics with contextual interpretation. The framework is not limited to one fairness definition. Instead, it treats fairness as a multi-dimensional assessment involving selection rates, group error rates, model utility, legal or policy constraints, domain harms, and human accountability.

---

2. Research Objective

The objective of this study is to develop a practical and systematic approach for fairness evaluation and mitigation in machine learning models.

The specific objectives are:

1. To define hidden bias as a model-evaluation problem that may not appear in aggregate performance metrics.
2. To identify the main stages where bias can enter the machine learning lifecycle.
3. To formalize a step-by-step workflow for evaluating group-level model outcomes.
4. To explain practical fairness metrics such as selection rate, disparate impact ratio, false positive rate parity, false negative rate parity, demographic parity, and equalized odds.
5. To distinguish between technical disparity signals and final ethical, legal, or operational conclusions.
6. To organize mitigation strategies into pre-processing, in-processing, and post-processing interventions.
7. To position fairness evaluation as a continuous governance activity rather than a one-time model check.
8. To define limitations and responsible-use principles for fairness-oriented machine learning evaluation.

---

3. Background and Conceptual Foundation

3.1 Bias in Machine Learning Systems

In machine learning, bias refers to systematic patterns that cause a model to produce unequal or unfavorable outcomes for particular groups. This definition is broader than statistical bias in estimation. In applied AI governance, bias includes social, organizational, representational, historical, and measurement-related sources of unfairness.

Bias may appear at several stages of the machine learning lifecycle:

Lifecycle Stage	Potential Bias Source	Example
Problem definition	Objective misalignment	Optimizing for speed of hiring instead of quality and equity
Data collection	Underrepresentation	Fewer samples from a demographic group
Label generation	Historical decisions	Past loan approvals reflecting biased human judgment
Feature engineering	Proxy variables	Location, school, income, or employment gaps encoding sensitive-group patterns
Model training	Optimization imbalance	Model prioritizing majority-group accuracy
Threshold selection	Unequal decision effects	Same cutoff producing different rejection rates across groups
Deployment	Dataset shift	Model used in a population different from training data
Monitoring	Lack of subgroup tracking	Aggregate performance monitored but group-level harm ignored

This lifecycle view is important because hidden bias is not only a property of the trained model. It can arise from the full sociotechnical system in which the model is developed and deployed.

3.2 Protected Attributes and Proxy Variables

Protected attributes are demographic or sensitive variables such as gender, race, ethnicity, age, disability status, religion, nationality, or other legally or ethically significant categories. In practice, fairness evaluation often requires access to these attributes for auditing, even when they are excluded from model training.

A common misconception is that removing sensitive attributes eliminates bias. This is often incorrect. Other variables may act as proxies. For example, postal code, school attended, employment gaps, income band, browser language, device type, or neighborhood-level indicators may correlate with demographic attributes. A model can therefore reproduce group disparities without directly observing protected attributes.

3.3 Fairness as a Multi-Metric Problem

Fairness is not a single universal quantity. Fairness metrics often encode different ethical assumptions and may conflict with one another. For example, demographic parity focuses on

equal selection rates across groups, while equalized odds focuses on equal error behavior conditional on the true label. Fairlearn describes demographic parity as a condition where predictions are independent of sensitive group membership, often operationalized as equal selection rates in binary classification.

The practical implication is that fairness evaluation should not rely on a single number. Instead, it should compare multiple metrics and interpret them in relation to the decision domain, harm profile, and institutional obligations.

3.4 Disparate Impact Ratio

The disparate impact ratio compares the selection rate of a group with the selection rate of a reference group. In employment contexts, the four-fifths rule has historically been used as a rule of thumb: a selection rate below 80% of the highest group selection rate may indicate adverse impact and trigger further investigation, but it is not a final legal conclusion.

For a binary decision model:

$$Selection\ Rate_g = \frac{Number\ of\ favorable\ predictions\ for\ group\ g}{Total\ number\ of\ individuals\ in\ group\ g}$$

$$Disparate\ Impact\ Ratio = \frac{Selection\ Rate_{unprivileged}}{Selection\ Rate_{privileged}}$$

A low ratio may indicate that one group receives favorable outcomes at a substantially lower rate than another group. However, this metric must be interpreted carefully because it does not explain why the disparity exists or whether the model is legally, ethically, or operationally acceptable.

---

4. Proposed Framework for Hidden Bias Detection

The proposed framework consists of eight stages.

Stage	Purpose	Output
1. Define decision context	Clarify model use and harm profile	Fairness assessment scope
2. Identify protected and proxy attributes	Determine groups and potential indirect signals	Sensitive-feature inventory
3. Establish baseline performance	Evaluate aggregate utility	Accuracy, precision, recall, AUC, calibration

Stage	Purpose	Output
4. Perform group-level outcome analysis	Detect selection-rate disparities	Group selection table
5. Perform group-level error analysis	Detect unequal model mistakes	FPR, FNR, TPR, TNR by group
6. Compute fairness metrics	Quantify disparity	DIR, demographic parity difference, equalized odds difference
7. Apply mitigation strategy	Reduce identified disparity	Revised data, model, or threshold
8. Govern and monitor	Sustain fairness over time	Audit record and monitoring plan

4.1 Stage 1: Define the Decision Context

The first stage is to define the model's decision context. Fairness evaluation is meaningful only when the practitioner understands what the model predicts, how the prediction is used, who is affected, and what harm may occur from false positives and false negatives.

For example, in a hiring-screening model, a false negative may incorrectly reject a qualified candidate. In a credit model, a false negative may deny access to a loan. In a medical triage model, a false negative may delay treatment. The relative harm of different errors should influence which fairness metrics receive priority.

4.2 Stage 2: Identify Protected Attributes and Proxy Variables

The second stage is to identify protected attributes and plausible proxy variables. This step should include data scientists, domain experts, compliance teams, and stakeholders familiar with the affected population.

A practical sensitive-feature inventory should include:

Category	Examples	Fairness Relevance
Direct protected attributes	Gender, race, age, disability	Used for group-level audit
Quasi-sensitive variables	Nationality, language, location	May correlate with protected groups
Socioeconomic indicators	Income, education, employment history	May encode structural inequality

Category	Examples	Fairness Relevance
Behavioral variables	Device type, response time, platform usage	May reflect access differences
Historical decision variables	Previous approvals, prior ratings	May encode past discrimination

The purpose is not necessarily to remove all such variables. The purpose is to understand their role and test whether they contribute to unfair outcome patterns.

4.3 Stage 3: Establish Baseline Model Performance

Before fairness mitigation, the model's baseline utility should be measured. This includes standard metrics such as accuracy, precision, recall, F1-score, area under the ROC curve, calibration, and confusion-matrix behavior.

However, baseline performance should be treated as incomplete. A single global score can hide subgroup failure. Therefore, the same metrics should be recomputed by group.

4.4 Stage 4: Analyze Group-Level Selection Rates

The next stage is to compare favorable prediction rates across groups. In a hiring example, the favorable outcome may be "shortlisted." In a lending example, it may be "approved." In a healthcare eligibility model, it may be "recommended for intervention."

A simplified fairness-audit table is shown below.

Group	Total Applicants	Favorable Predictions	Selection Rate
Group A	1,000	600	0.60
Group B	800	360	0.45
Group C	500	150	0.30

If Group A has the highest selection rate, the disparate impact ratios are:

Group	Selection Rate	Disparate Impact Ratio
Group A	0.60	1.00
Group B	0.45	0.75
Group C	0.30	0.50

This table does not prove discrimination, but it identifies a disparity pattern that requires further analysis.

4.5 Stage 5: Analyze Group-Level Error Rates

Selection-rate analysis alone is insufficient. The model may select groups at similar rates but still make different types of errors across groups. Therefore, fairness evaluation should include group-level confusion-matrix analysis.

Metric	Definition	Fairness Interpretation
True Positive Rate	Qualified individuals correctly accepted	Measures opportunity recognition
False Negative Rate	Qualified individuals incorrectly rejected	Measures missed opportunity
False Positive Rate	Unqualified individuals incorrectly accepted	Measures risk allocation
False Discovery Rate	Accepted individuals who are actually negative	Measures decision reliability
Calibration	Whether predicted probabilities match observed outcomes	Measures probability trustworthiness

Equalized odds focuses on equalizing error behavior across groups, particularly true positive and false positive rates. Fairlearn supports fairness assessment and mitigation for classification and regression systems, including disparity-based evaluation across affected populations.

4.6 Stage 6: Compute Fairness Metrics

The proposed framework recommends using a metric set rather than a single fairness score.

Fairness Metric	Question Answered	Practical Use
Demographic parity difference	Are selection rates similar across groups?	Useful for access and allocation decisions
Disparate impact ratio	Is one group selected much less often?	Useful as an initial disparity screen
Equal opportunity difference	Are qualified individuals recognized equally?	Useful when false negatives are harmful

Fairness Metric	Question Answered	Practical Use
Equalized odds difference	Are error rates similar across groups?	Useful in high-stakes classification
Group calibration	Do predicted probabilities mean the same thing across groups?	Useful for risk scoring
Worst-group accuracy	Which group receives the lowest performance?	Useful for deployment safety
Error concentration	Where are errors clustered?	Useful for remediation planning

The output of this stage should be a fairness profile that shows both overall utility and subgroup behavior.

4.7 Stage 7: Interpret the Disparity

A disparity metric is a signal, not a conclusion. Interpretation should examine:

1. Whether the disparity is statistically and practically significant.
2. Whether the disparity appears in outcomes, errors, calibration, or all three.
3. Whether the disparity is caused by data imbalance, label bias, feature proxies, model choice, thresholding, or deployment shift.
4. Whether the disparity creates meaningful harm.
5. Whether mitigation introduces unacceptable trade-offs.
6. Whether human review or policy redesign is required.

This step prevents fairness evaluation from becoming mechanical. A fairness metric can identify risk, but responsible decision-making requires domain interpretation.

4.8 Stage 8: Monitor Fairness After Deployment

Fairness is not static. A model that appears fair during validation may become unfair after deployment due to population drift, behavioral adaptation, policy changes, economic changes, or feedback loops.

Post-deployment monitoring should track:

Monitoring Area	Example Question
Population drift	Has the demographic composition changed?
Outcome drift	Are approval or rejection rates shifting by group?
Error drift	Are false negatives increasing for one group?

Monitoring Area	Example Question
Proxy drift	Are new variables acting as indirect protected-attribute signals?
Human override patterns	Are human reviewers overriding one group more often?
Complaint signals	Are affected users reporting unequal treatment?

NIST's AI Risk Management Framework emphasizes ongoing measurement, monitoring, validity, reliability, and management of harmful bias as part of trustworthy AI practice.

---

5. Methodology for Practical Fairness Evaluation

The methodology can be expressed as a repeatable audit workflow.

5.1 Inputs

The workflow requires:

Input	Description
Model predictions	Predicted labels or scores
Ground-truth labels	Actual outcomes where available
Sensitive attributes	Protected-group membership for audit
Feature metadata	Description of model inputs
Decision threshold	Cutoff used to convert scores into decisions
Domain context	Meaning of favorable and unfavorable outcomes
Risk criteria	Harm profile and governance requirements

5.2 Process

The fairness-evaluation process consists of the following steps:

1. Define the favorable outcome.
2. Segment the validation or test dataset by protected group.
3. Compute group selection rates.
4. Compute disparate impact ratios.
5. Compute confusion matrices for each group.

6. Compare true positive, false positive, false negative, and false discovery rates.
7. Evaluate calibration by group if the model produces probabilities.
8. Identify features that may act as proxies.
9. Test alternative thresholds or mitigation strategies.
10. Document findings, assumptions, limitations, and decisions.

5.3 Outputs

The workflow produces:

Output	Purpose
Fairness metric table	Quantifies disparity
Group confusion matrix	Identifies unequal error patterns
Proxy-variable assessment	Explains possible indirect bias
Mitigation recommendation	Defines next technical intervention
Governance record	Supports auditability
Monitoring plan	Supports post-deployment fairness management

---

6. Technical Analysis and Practical Findings

This study does not report new benchmark experiments. Instead, it formalizes practical findings derived from applied fairness evaluation.

6.1 Aggregate Accuracy Can Hide Subgroup Harm

A model may achieve high global accuracy while underperforming for a minority group. This occurs when the majority group dominates the dataset or when the model learns patterns that work well for common cases but fail for underrepresented populations.

For example, if 90% of the dataset belongs to Group A and 10% belongs to Group B, a model may optimize primarily for Group A performance. Without subgroup evaluation, Group B errors may remain hidden.

6.2 Removing Protected Attributes Is Not Sufficient

Excluding sensitive attributes from model training does not guarantee fairness. Proxy variables can preserve similar information. For example, geography may correlate with race or income; employment gaps may correlate with caregiving responsibilities; school names may correlate with socioeconomic status.

Therefore, fairness evaluation should test outputs, not merely inspect inputs.

6.3 Disparate Impact Ratio Is Useful but Incomplete

The disparate impact ratio provides an accessible way to identify selection-rate differences. However, it does not reveal whether the model has unequal false negative rates, unequal false positive rates, or calibration differences. It also does not determine legal liability or ethical acceptability by itself. The EEOC's guidance treats the four-fifths rule as a rule of thumb and notes that it does not resolve the ultimate question of unlawful discrimination.

6.4 Fairness Metrics May Conflict

A model may satisfy one fairness criterion while violating another. Equal selection rates may not imply equal error rates. Equal error rates may not imply equal calibration. This is why fairness evaluation should be framed as a decision process rather than a single optimization target.

6.5 Mitigation Requires Context-Specific Trade-Offs

Bias mitigation can affect model accuracy, group-specific utility, operational cost, and decision consistency. A fairness intervention should therefore be evaluated using both fairness and performance metrics. The aim is not to manipulate the model until a metric looks acceptable, but to improve decision quality under responsible constraints.

---

7. Bias Mitigation Strategies

Bias mitigation can occur at three levels: before model training, during model training, and after model training.

7.1 Pre-Processing Mitigation

Pre-processing methods modify the data before training.

Method	Description	Use Case
Rebalancing	Adjust group representation	Underrepresented groups
Reweighting	Assign weights to samples	Imbalanced labels by group
Feature review	Remove or transform risky proxy variables	Proxy leakage
Label audit	Examine whether labels encode historical bias	Hiring or lending history

Method	Description	Use Case
Data augmentation	Add representative examples	Sparse subgroup data

Pre-processing is often useful when bias originates from the dataset rather than the model architecture.

7.2 In-Processing Mitigation

In-processing methods modify the learning process itself.

Method	Description	Use Case
Fairness-constrained optimization	Add fairness constraint to objective	Explicit fairness target
Adversarial debiasing	Reduce protected-attribute predictability	Proxy-sensitive models
Regularization	Penalize disparity-related behavior	Controlled trade-off
Cost-sensitive learning	Adjust cost of group-specific errors	Unequal harm profiles

Fairlearn includes mitigation algorithms and fairness constraints that support practical fairness-aware modeling.

7.3 Post-Processing Mitigation

Post-processing methods modify predictions after the model is trained.

Method	Description	Use Case
Threshold adjustment	Use different cutoffs to reduce disparity	Score-based classifiers
Reject-option classification	Review uncertain decisions	Borderline cases
Human-in-the-loop review	Add expert oversight	High-impact decisions
Calibrated post-processing	Adjust probabilities or labels	Risk-score applications

Post-processing can be useful when the model cannot be retrained or when governance requires a controlled decision layer.

7.4 Tooling Support

Several fairness toolkits support practical bias detection and mitigation. AI Fairness 360 is an open-source toolkit containing metrics and mitigation algorithms for datasets and models. Fairlearn provides assessment and mitigation capabilities for AI fairness and supports evaluation across affected populations.

---

8. Discussion

The proposed framework emphasizes that fairness evaluation is both technical and sociotechnical. Technical metrics are necessary because they make disparity measurable. However, they are not sufficient because fairness depends on context, harm, stakeholder expectations, legal constraints, organizational accountability, and deployment conditions.

A major challenge is that machine learning teams often treat fairness as a late-stage compliance check. This is inadequate. Bias should be considered during problem formulation, data collection, model development, validation, deployment, and monitoring. The earlier fairness risks are identified, the easier they are to address.

Another challenge is fairness-performance trade-off framing. Organizations may assume that fairness always reduces accuracy. This is an oversimplification. In some cases, fairness interventions can improve model robustness by correcting underrepresentation, label noise, or proxy leakage. In other cases, trade-offs may exist and must be documented transparently.

The framework also highlights the importance of explainability. If a model shows disparity across groups, practitioners must investigate why. This may involve feature importance analysis, counterfactual testing, subgroup error review, label-quality analysis, and stakeholder consultation.

Finally, fairness should be treated as an ongoing operational responsibility. Models deployed in real environments interact with changing populations and institutional processes. Without monitoring, even a carefully evaluated model can become unfair over time.

---

9. Research Contribution

This study contributes a structured applied AI framework for hidden-bias detection and fairness mitigation in machine learning systems.

The main contributions are:

1. **A systematic fairness-evaluation workflow**
   The study organizes fairness assessment into decision-context definition, sensitive-

feature inventory, group-level outcome analysis, error-rate analysis, mitigation, and monitoring.

2. **A practical distinction between performance and fairness**
   The framework shows why aggregate accuracy is insufficient for high-impact model evaluation.
3. **A multi-metric fairness approach**
   The study integrates selection-rate metrics, disparate impact ratio, group error rates, demographic parity, equalized odds, and calibration.
4. **A mitigation taxonomy**
   Bias mitigation is organized into pre-processing, in-processing, and post-processing interventions.
5. **A governance-oriented fairness model**
   The framework connects technical audit results with documentation, human oversight, monitoring, and responsible AI controls.
6. **A cautionary interpretation model**
   The study avoids treating any single fairness metric as proof of fairness or unfairness and instead frames metrics as evidence requiring contextual review.

---

10. Limitations

This study has several limitations.

First, the framework is conceptual and implementation-oriented. It does not present new empirical experiments, benchmark comparisons, or statistical validation across multiple datasets.

Second, fairness metrics are domain-dependent. A metric that is appropriate for one use case may be inappropriate for another. For example, demographic parity may be relevant for access allocation, while equalized odds may be more relevant where error harms are central.

Third, protected-attribute data may not always be available due to privacy, legal, or organizational constraints. Without such data, fairness auditing becomes more difficult and may require privacy-preserving or proxy-based approaches.

Fourth, fairness mitigation can introduce trade-offs between model utility, group-level outcomes, operational consistency, and interpretability.

Fifth, the framework does not provide legal advice. Metrics such as disparate impact ratio can flag potential concern, but legal interpretation requires jurisdiction-specific review by qualified experts.

Sixth, fairness is not solved through technical intervention alone. Organizational policy, human decision-making, appeal mechanisms, and accountability structures remain essential.

Finally, hidden bias may emerge after deployment due to feedback loops, dataset drift, changing user behavior, and institutional adaptation. Continuous monitoring is therefore necessary.

11. Governance and Responsible Use

Fairness evaluation should be embedded into broader responsible AI governance. A responsible workflow should include:

Governance Principle	Practical Requirement
Transparency	Document model purpose, data sources, metrics, and limitations
Accountability	Assign ownership for fairness review and remediation
Privacy	Protect sensitive attributes used for auditing
Auditability	Maintain records of model versions, thresholds, and evaluations
Human oversight	Review high-impact or uncertain decisions
Explainability	Provide interpretable reasons for model behavior where feasible
Monitoring	Track subgroup outcomes after deployment
Contestability	Allow affected individuals to challenge decisions
Non-overclaiming	Avoid declaring a model “fair” based on one metric

NIST’s AI RMF frames trustworthy AI in relation to validity, reliability, safety, accountability, transparency, explainability, privacy, and fairness, including management of harmful bias. This aligns with the proposed view that fairness evaluation should be part of a broader risk-management process.

12. Future Work

Future work can extend this framework in several directions.

First, the framework should be validated across real-world datasets from hiring, lending, healthcare, education, and public-sector decision systems.

Second, future research should compare the effectiveness of different mitigation strategies under varying levels of data imbalance, proxy leakage, and label bias.

Third, statistical testing should be integrated more formally into the workflow to distinguish random variation from practically meaningful disparity.

Fourth, privacy-preserving fairness auditing should be explored for environments where sensitive attributes cannot be freely stored or processed.

Fifth, fairness monitoring dashboards should be developed to support post-deployment governance.

Sixth, stakeholder-centered fairness evaluation should be incorporated so that affected communities can help define relevant harms and acceptable trade-offs.

Seventh, fairness evaluation should be integrated with explainable AI techniques to identify the drivers of group-level disparities.

Finally, future work should examine fairness in generative AI and large language model evaluators, where bias may appear through ranking, summarization, scoring, and recommendation behavior rather than conventional classification alone.

---

13. Conclusion

This research presented a systematic approach to detecting hidden bias in machine learning models. The study argued that aggregate accuracy and global performance metrics are insufficient when model decisions affect people across different demographic groups. Hidden bias may appear through unequal selection rates, unequal error rates, proxy variables, historical labels, threshold effects, and deployment drift.

The proposed framework organizes fairness evaluation into a practical workflow: define the decision context, identify protected and proxy variables, measure baseline performance, compute group-level selection and error metrics, interpret disparity signals, apply mitigation strategies, and monitor outcomes after deployment. The framework emphasizes that fairness cannot be reduced to a single number. Metrics such as disparate impact ratio, demographic parity, and equalized odds provide useful evidence, but they must be interpreted in context.

The central contribution of this work is an applied fairness-evaluation methodology that connects machine learning practice with responsible AI governance. It supports practitioners who need to move beyond performance optimization toward transparent, auditable, and accountable model evaluation. The study remains conceptual and requires empirical validation, but it provides a structured foundation for detecting and mitigating hidden bias in real-world AI systems.

References

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D.,

Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. arXiv.

Fairlearn. *Common fairness metrics: Demographic parity*. Fairlearn documentation.

Fairlearn. *Fairlearn: A Python package to assess and improve fairness of AI systems*. GitHub documentation.

IBM Research. (2018). *Introducing AI Fairness 360*. IBM Research.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework: AI RMF 1.0*. NIST.

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023). *Fairlearn: Assessing and improving fairness of AI systems*. arXiv.

Pruseth, D. (2025). *How to Detect Hidden Bias in Your ML Model — A Step-by-Step Tutorial*. Debabrata Pruseth AI blog.

Suggested Citation

Pruseth, D. (2026). *Detecting Hidden Bias in Machine Learning Models: A Systematic Approach to Fairness Evaluation and Mitigation*. Debabrata Pruseth AI blog.