*International Journal of Social Science and Economic Research*

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

# Fairness by Design: Addressing Gender Bias in AI-Driven Credit Scoring through the REACT Framework and Gender Fairness Index

Samriddhi Shah

The International School Bangalore

Guided by Dr. Sukanya Kemp

DOI: 10.46609/IJSSER.2025.v10i11.011 URL: https://doi.org/10.46609/IJSSER.2025.v10i11.011

Received: 2 November 2025 / Accepted: 18 November 2025 / Published: 30 November 2025

## ABSTRACT

*Credit scoring systems are key to financial decisions. They affect access to loans and terms like interest rates. AI models improve efficiency but can embed gender bias. These biases disadvantage women. This paper explores how biases enter models, their effects, detection challenges, and solutions for fairness and profitability. We propose the REACT framework and Gender Fairness Index (GFI) as original tools. Using literature review, toy data simulation, and framework proposals, we show that fair models boost inclusion without harming accuracy. Findings apply to FinTech firms in emerging economies and school simulations.*

**Keywords:** Accountability, Gender bias, Proxy variables, System gender disparities

## 1. Introduction

Credit scoring systems are foundational for modern financial decision-making, influencing access to loans, credit cards, mortgages, and borrowing terms like interest rates and credit limits. While perceived as objective and data-driven, these systems can embed gender biases arising from algorithmic design and deployment.

AI-based credit scoring algorithms predict default or delinquency probability, providing Credit Service Providers (CSPs) with efficient creditworthiness assessments. The growing use of Artificial Intelligence and Machine Learning has automated decision-making, enabling quick processing of large datasets, real-time decisions, reduced operational costs, and improved accuracy, scalability, and access for underserved populations (The World Bank Group et al., 2019).

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

However, AI-driven credit scoring carries embedded gender bias—a form of algorithmic bias that reproduces or amplifies gender inequalities, systematically disadvantaging women in credit assessments. AI models exhibit these biases in two ways. Firstly, through training on datasets containing historical gender biases or disparities, which their conclusions then reflect; secondly, through embedded developer assumptions favoring certain outcomes (SAP, 2024).

Substantial evidence demonstrates gender bias in algorithms. Women often receive lower credit limits or worse terms despite comparable creditworthiness. The 2019 Apple Card controversy illustrates this—Goldman Sachs was accused of assigning women lower limits than men in shared households (Gibson, 2019; Nedlund, 2019).

Similar disparities exist globally. In the European Union, female entrepreneurs face lower approval rates and tighter collateral demands relative to men with similar credentials (Popov et al., 2015). In the Global South, women in microfinance receive high-risk scores despite strong repayments, restricting access to growth capital in India and Sub-Saharan Africa (Women's World Banking, 2021). These cases highlight gender bias as a global challenge embedded in both traditional and AI-driven systems.

This research investigates gender bias in AI credit scoring across four areas: processes through which biases integrate into models, consequences for lending decisions, complexities of detecting and addressing biases, and actionable frameworks for fairness and profitability.

This paper targets a critical research gap by focusing on how AI-powered, proxy-laden systems reproduce and scale gender bias while offering practical remediation. Most studies note bias; few detail mechanisms or tested interventions at operational levels.

This study argues that addressing gender bias achieves both algorithmic fairness and inclusive access. Mitigating bias is both a moral imperative and a business opportunity—fairer algorithms unlock markets, improve trust, and drive inclusion. Inclusive digital credit empowers women-owned SMEs—the "missing middle"—which fuel growth and innovation (Women's World Banking, 2021).

## 2. Analytic Approach (Methods)

This paper uses a literature review to understand gender bias mechanisms. Sources include policy reports from the World Bank, Brookings, and SME Finance Forum. It also includes academic papers from arXiv and journals. This ensures policy relevance and technical rigor.

We simulate the Gender Fairness Index (GFI) with toy data. This illustrates disparities in true positive rates.

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

We propose the REACT framework and GFI as original contributions. These synthesize literature into actionable tools. This approach distinguishes original work from reviewed literature. It focuses on fairness metrics like true positive rates (TPR) and false negative rates (FNR).

## 3. Mechanisms of Gender Bias

### 3.1 Historical Data and Systemic Gender Disparities

Credit-scoring models train historical financial data. This data often reflects a world where men had greater access to formal employment, property titles, and credit limits. This translated into higher approval odds for men in training data (Brookings Institution, 2021).

Employment patterns show disparities between men and women. These span the 19th and 20th centuries, extending into the mid to late 20th century. Past employment data reveal men in higher-paying formal jobs like banking, law, and medicine. Women were in lower-paid or informal work like teaching, nursing, and domestic services (Hartmann et al., 2024).

Savings account records show women with smaller balances. This is due to lower wages and interrupted career patterns, which creates a compounding effect. This reinforces perceived financial instability in algorithms, where the disparities become predictive norms—which are replicated. As models maximize predictive performance, they recognize such patterns, reward features common among men, and penalize those common among women.

### 3.2 Data Sparsity and Model Performance Issues

Due to structural barriers, women are underrepresented in financial data. This contributes to data sparsity (limited data points). It diminishes model performance for women's financial activities, reducing predictive accuracy. Consequently, models exhibit reduced effectiveness for women, resulting in biased outcomes.

Structural barriers include limited access to financial services. They include lack of continuous usage and a lower adoption of digital products. This leads to fewer data points for training. As models rely on rich, balanced data, they suffer from sample bias and underrepresentation—perpetuating gender disparities in financial inclusion (Kelly et al., 2021).

### 3.3 Prohibition of Explicit Gender Features and Proxy Variables

Rules forbid using gender as a variable, preventing overt bias. However, factors like income, job type, marital status, and caregiving duties reflect gender—enabling models to inadvertently sustain bias.

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

This is an example of indirect discrimination or proxy bias. Proxy features are variables that appear neutral, but act as protected characteristics—like using occupation as a stand-in for gender. Even without gender, models unfairly evaluate individuals due to correlated variables (Algorithmic Bias in Women-MSMEs Credit Scoring: Insights & Solutions Community of Practice (CoP) | SME Finance Forum, n.d.).

### 3.4 Algorithmic Amplification of Biases

Machine predictive models learn to maximize prediction accuracy, learning patterns from past instances. If data includes systematic discrimination, models incorporate these as patterns.

As such, decisions by the model perpetuate past injustices. They consider biased patterns as standard behavior, generating a feedback loop. Biased predictions affect decisions, like hiring or lending, which discriminate against marginalized groups. This then biases subsequent data, reinforcing the initial discrimination and making it more difficult for women to bridge barriers (Garvan, 2024; Vidal & Caire, 2024).

### 3.5 Sampling, Labeling, and Outcome Pathways

When women are underrepresented in approved loan data, credit models struggle to learn their repayment behavior. As the data shows historical prejudice toward women, this leads to fewer approvals. Labels for creditworthiness (good or bad) are then based on biased results, which include systemic discrimination from uneven credit conditions for women.

Sampling bias is a concern, as underrepresentation limits views of women's repayment. Label bias worsens this, as labels mirror discriminatory practices. While models assume results are unbiased, unequal access distorts data, reducing fairness and accuracy for women. (Vidal et al., 2019).

Several recurring patterns contribute to gender bias in AI credit models. Clarifying these is essential as it helps direct mitigation across specific factors: pipeline layers, like data, modeling, and decision-making.

**Table 1. Mechanisms of Gender Bias in Credit Scoring Models and Their Risks to Women**

| Description | Example Signal | Risk to Women |
|---|---|---|
| Women underrepresented in training, especially among prior approvals | Sparse or missing features for women segments | Conservative scores and higher denials |

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

| | | |
|---|---|---|
| Past biased denials embedded in “good/bad” labels | Label correlates with reviewer discretion | Learned replication of past discrimination |
| Outcomes reflect worse terms or product fit | Higher observed delinquency from smaller limits or stricter terms | Misattribution of structural effects to individual risk |
| Non-gender features correlate with gender | Occupation, geography, phone usage patterns | Reintroduction of gender via correlated inputs |
| Feature relationships differ by gender | Income–tenure–sector interactions | Mis-specified risk in subgroup-specific regimes |

Note. Adapted from multiple sources on gender bias in credit scoring models (CGAP, 2024; SME Finance Forum, n.d.; Women’s World Banking, 2021; European Central Bank, 2022).

## 4. Mitigation Methods and Frameworks

### 4.1 Original Contribution: REACT Framework and Gender Fairness Index (GFI)

This paper introduces the REACT framework and GFI as original contributions. REACT integrates regulatory oversight, explainable decisions, accountability, collaboration, and transparency; GFI measures fairness via true positive rates.

Persistent gender bias in credit scoring arises from systemic inequalities. It comes from data, modeling, post-decision contexts, transparency, and governance. Fairer outcomes need multi-level interventions that span the credit scoring cycle, combining technical solutions with reforms. Mitigation occurs at every one of the lifecycle stages as mentioned above. Each stage offers levers to reduce disparities, embedding fairness in systems (Financial Conduct Authority et al., 2024).

### 4.2 Data Collection and Preparation

Representative sampling ensures equitable representation in AI training datasets. It includes women and intersectional subgroups (groups defined by overlapping identities, such as gender and race, e.g., Black women). This mitigates biases in machine learning models.

Wang et al. (2022) emphasize intersectionality in ML pipelines. Select demographic attributes via empirical validation and domain knowledge. Such practices handle underrepresentation by inferring from related groups and use ROC AUC metrics (receiver operating characteristic area

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

under the curve, a measure of model performance) to evaluate fairness algorithms in intersectional settings.

These include RWT (reweighting scheme), RDC (cost-sensitive classification), LOS (fairness regularizer), GRP (probabilistic logistic regression combinations), and GRY (cost-sensitive from zero-sum games). Tests on US Census datasets like ACSIncome promote targeted data collection. They advocate metrics like ranking correlations for subgroup fairness, and avoid traditional techniques like reweighting, resampling, or synthetic data (e.g., SMOTE or GANs) due to normative concerns.

Rejected inference addresses selection bias in credit-scoring AI models. It estimates hypothetical outcomes, like default probabilities, for rejected applicants. Their absence skews predictions toward lower-risk profiles.

Key methods include parceling, which segments rejected cases into risk bands and assigns probabilistic labels based on prudence factors. The augmentation uses nearest neighbors and infers labels by matching to accepted cases. Semi-supervised learning uses unlabeled reject data with labeled accepts. Algorithms include Gaussian mixtures or SVMs.

Validation needs prospective pilots, using control groups where all applicants are accepted—creating unbiased test sets that ensure robustness before deployment (Ehrhardt et al., 2021).

Feature sanitation identifies and constrains proxy features in credit-scoring AI models. Proxy features are variables like ZIP code or occupation that encode structural inequities, like gender or racial biases, instead of intrinsic risk.

Using causal analysis techniques—including propensity score matching or structural equation modeling—features are detected and transformed to ensure fairer predictions. This mitigates bias amplification, ensuring model integrity (Hardt et al., 2016).

### 4.3 Modeling and Training

Fairness-aware objectives incorporate constraints or penalties in credit-scoring AI models. They minimize disparities in true positive rates (TPR) and false negative rates (FNR) across groups, preserving calibration.

Techniques include adversarial debiasing, which trains a classifier with an adversary to remove sensitive information. Constrained optimization integrates fairness metrics into the loss function (Yang et al., 2023).

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

Hardt's framework emphasizes statistical parity and equalized odds, ensuring similar error rates across groups by focusing on measurable group fairness in credit scoring (Hardt et al., 2016).

Wang's intersectional approach broadens this. It incorporates overlapping identities—like gender and race—and addresses underrepresentation via subgroup analyses and tailored metrics (Wang et al., 2022).

While Hardt's methods are clear and implementable, with its simplicity favouring regulators, Wang's provides nuance for diverse populations. For credit providers, a hybrid model balances Hardt's constraints with Wang's insights, yielding equitable models.

Interpretable baselines develop scorecards in credit-scoring AI models that assign points to features for transparency using monotonic gradient-boosted trees. This enforces directional constraints based on domain knowledge.

These reveal feature-outcome relationships. Black-box models obscure them. They act as fairness and reliability guardrails (Laugel et al., 2017).

Regularization and monotonicity enforce monotone relationships in credit-scoring AI. For economically meaningful features, like higher income improving scores, penalized constraints or simplified architectures are used.

This reduces spurious interactions. They disproportionately affect women. It ensures regulatory compliance with fairness (Chen & Ye, 2022).

**4.4 Post-Processing and Decisioning**

Setting decision thresholds equalizes true positive rates (TPR) across genders. This balances fairness with portfolio risk.

Adjust cutoffs separately for men and women. Address disparities without sacrificing profitability. Monitor false positive rates and calibration as guardrails (CGAP, 2024).

Approval bands involve structured manual reviews in credit-scoring AI. Use bias-aware checklists for applicants near thresholds. This minimizes false negatives.

Double-blind reviews in sensitive cases reduce evaluator bias. This ensures fairer outcomes.

The hybrid approach integrates human judgment with AI. It enhances fairness and accuracy (Hardt et al., 2016).

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

Product Tailoring or offering credit products tailored to women's cash flow patterns. Examples include starter credit limits, smooth line increases, or flexible repayment schedules.

These align access with real needs. They avoid penalizing women for differing risk profiles. They support financial inclusion (CGAP, 2024).

### 4.5 Transparency and Consumer Recourse

Adverse action notices explain loan denials or pricing in credit-scoring AI. Link model factors, like high debt, to simple reasons.

This transparency helps applicants understand decisions. It supports fair recourse (Chodorow-Reich & Coglianese, 2020).

Counterfactual explanations offer guidance. For example, "reduce credit utilization by 10% to improve approval odds."

They help applicants understand and act on denials. This enhances fairness and transparency (Wachter et al., 2017b).

### 4.6 Governance and Culture

Cross-functional oversight engages risk, legal, and diversity teams in credit-scoring AI. They design and monitor models.

This ensures fairness metrics balance with financial goals. It leads to equitable lending (Tan et al., 2019).

Model documentation uses clear reports in credit-scoring AI. Outline data, model limits, and fairness results across groups.

This helps identify and address biases transparently (Gebru et al., 2018).

Ongoing monitoring conducts regular fairness audits in credit-scoring AI. Update models to detect bias or performance shifts.

This ensures equitable treatment over time (Bellamy et al., 2018).

### 4.7 REACT Framework for Accountability and Regulation

The REACT framework is a novel approach for fair credit-scoring AI. It integrates five components: Regulatory oversight, Explainable algorithmic decisions, Accountability through inclusion KPIs, Collaboration across industry, and Transparency in data and pathways.

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

Each draws from fairness literature. It addresses biases in lending models. Regulatory oversight ensures compliance. Explainability clarifies decisions. Accountability tracks equitable outcomes. Collaboration leverages shared resources. Transparency elucidates data use.

This framework synthesizes principles into a cohesive structure. It enhances fairness and governance.

It is necessary due to pervasive biases in AI lending. They perpetuate discrimination against underrepresented groups. They amplify inequalities. They undermine trust.

Fragmented approaches fail to mitigate risks comprehensively. Unified frameworks ensure ethical deployment, compliance, and equitable access without compromising accuracy.

Regulatory oversight in REACT ensures fairness. It enforces standards. It requires pre-deployment bias checks, third-party validation, and subgroup reports. This promotes equitable lending.

Explaining decisions uses clear reason codes and simple explanations. For example, why a loan was denied. This helps applicants understand and appeal fairly (Goodman & Flaxman, 2017).

Accountability tracks fairness metrics, like equal error rates across groups. Set as key goals. Ensure models reduce unfair gaps without harming assessment (Dwork et al., 2011).

Industry collaboration shares privacy-safe datasets and best practices via consortia. This enables lenders to adopt fairness solutions efficiently (Mehrabi et al., 2019).

Transparency clarifies data sources and decision impacts. Publish fairness metrics and bias reduction plans. Provide regular updates to ensure equitable models (Gebru et al., 2018b).

**4.8 Novel Proposals**

Aggregating behavioral data from platforms like social media and e-commerce into credit-scoring AI models risks amplifying gender bias. It happens through proxy variables, such as gendered online activity patterns. These disproportionately disadvantage women.

This hypothesis extends REACT’s transparency focus. It emphasizes scrutiny of multi-source data integration.

Systemic barriers limit women's financial data. Examples include informal employment. This exacerbates proxy bias. Algorithms may prioritize male-dominated patterns.

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

Empirical testing could use fairness metrics like demographic parity. Assess platform-specific feature impacts.

Research highlights risks of multi-source data encoding societal biases. It necessitates causal analysis to detect hidden correlations (Bellamy et al., 2018; Dastin et al., 2018).

REACT's collaboration supports shared datasets to study effects. This ensures equitable lending.

The Gender Fairness Index (GFI) audits AI credit-scoring models. It measures variable impacts across genders to detect proxy biases, like spending patterns. This ensures equitable outcomes.

Aligned with REACT, GFI enhances accountability. It addresses underrepresentation for women, including intersectional subgroups.

GFI integrates into monitoring, per REACT's governance. It balances fairness and profitability.

Research supports metrics for bias detection (Hardt et al., 2016). GFI fosters trust and compliance in lending.

A critical contribution is the proposal and demonstration of GFI for credit scoring algorithms.

GFI quantifies disparities in model performance across genders. It compares true positive rates (TPR) between women and men. This metric allows assessment of fairness. It provides a goal for improvement.

### 4.9 Gender Fairness Index (GFI)

A higher GFI (closer to 1) indicates greater fairness. 1 represents a perfectly equal opportunity. Lower values signal disparity.

## 5. Empirical Evidence and Case Studies

**Table 2. Toy model results for credit approval prediction:**

| Gender | True Positives | False Negatives | True Positive Rate (TPR) |
|---|---|---|---|
| Women | 80 | 20 | 80 / (80 + 20) = 0.80 |
| Men | 90 | 10 | 90 / (90 + 10) = 0.90 |

Average TPR = (0.80 + 0.90) / 2 = 0.85
GFI = 1 - |0.80 - 0.90| / 0.85 = 1 - 0.10 / 0.85 = 1 - 0.1176 ≈ 0.88

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

A GFI of 0.88 signals moderate gender disparity. The TPR gap is about 12%. For an operational model, improve to push GFI closer to 1. Refine data, constraints, or thresholding.

Researchers can use GFI in routine audits. Compare mitigation impacts. Set regulatory benchmarks in disparate contexts.

## 6. Integration of the Gender Fairness Index (GFI) into Fairness Toolkits

To promote adoption and reproducibility, the proposed Gender Fairness Index (GFI) can be integrated into existing open-source fairness auditing frameworks such as **IBM's AI Fairness 360 (AIF360)** or **Google's What-If Tool**. Both toolkits are widely used for evaluating algorithmic fairness across multiple datasets and fairness definitions. However, they currently emphasize generic parity-based metrics (e.g., statistical parity difference, equal opportunity difference) and lack a *gender-specific interpretive fairness score* that captures disparities across both statistical and behavioral dimensions.

### 6.1 Architectural Integration

In AIF360, fairness metrics are implemented as subclasses within the Metric class, which operates on a BinaryLabelDataset. The GFI can be added as a custom metric class—GenderFairnessIndex—that computes the ratio of true positive rates (TPR) or acceptance rates between female and male groups. The class would inherit AIF360's base metric structure, ensuring compatibility with other evaluation modules and visualization functions.

This module would integrate seamlessly with AIF360's pipeline:

1. **Data ingestion** using BinaryLabelDataset with gender as the protected attribute.
2. **Model training and prediction** using any classifier (e.g., logistic regression, random forest).
3. **Fairness evaluation** using GenderFairnessIndex.compute_gfi() alongside standard metrics.
4. **Visualization** through the existing FairnessDashboard module, plotting GFI alongside accuracy and parity metrics.

### 6.2 Expected Outcomes

Integrating the GFI metric within AIF360 or What-If Tool enables organizations and regulators to:

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

- Quantitatively monitor **gender-specific fairness** over time.
- Compare **bias mitigation algorithms** (e.g., reweighing, adversarial debiasing) in terms of their impact on gender balance.
- Combine **GFI** with existing fairness measures to form a *multi-metric fairness dashboard*.
- Facilitate **auditable model governance** by producing standardized fairness reports that include gender disparity ratios.

The GFI addition thus extends the functionality of existing fairness toolkits toward more nuanced socio-economic contexts, particularly in credit scoring and financial inclusion. By bridging academic theory with operational tooling, this integration supports the practical realization of fairness-by-design principles in financial AI systems.

### 6.3 Case Example of Practice in India

In India, Lendingkart is a leading digital lender for MSMEs. It exemplifies fairness-oriented practices in AI credit scoring. It uses regular bias audits for equitable treatment across genders.

A comprehensive audit by Women's World Banking partnered with the University of Zurich. It assessed approval likelihood, loan terms, and repayment rates. It revealed parity in outcomes for women and men in comparable risk bands.

This shows accuracy-driven modeling achieves fairness without trade-offs. It blends technical controls like fairness metrics in validation, periodic reviews, and proxy-limiting refinements. It includes organizational commitments to inclusivity.

Such approaches align with REACT's transparency and accountability. India's ecosystem has digital rails like account aggregators and AI regulations. It facilitates representative data for women entrepreneurs with thin files.

Banks and NBFCs can adopt similar pipelines. Target women-led MSMEs. Foster inclusion and diversification. Address historical disparities (Women's World Banking, 2023).

This case highlights replicable pathways. Inclusive models advance gender equity. They maintain predictive performance and compliance.

### 6.4 Practical Techniques in Emerging Economies

Pre-processing strategies reweight training data. This corrects sampling biases. It enhances women's representation in credit models. This is key in markets where digital access favors men.

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

In-processing integrates fairness penalties during training. This minimizes disparate impacts.

Post-processing adjusts outputs for equal acceptance rates across genders. It avoids severe accuracy compromise.

Operational mitigations include regular audits and multidisciplinary teams. They reduce biases by involving stakeholders and reviews.

These methods work in emerging economies. They support financial inclusion. They open credit for women while maintaining viability (Kelly et al., 2021).

### 6.5 Decomposing Gender Disparities in Credit Allocation

Decomposition analyses reveal differences in observable factors like income and scores account for small gender gaps in bankcard limits. The majority stems from varying effects of characteristics. Men receive higher credit returns than women with similar profiles.

This unexplained disparity grows at higher quantiles and over time. It may reflect socioeconomic inequalities or subtle biases in automated systems.

Policy interventions focus on transparency in decisions. Address differential treatment. Challenges include isolating causal mechanisms and complying with laws.

Integrating insights into risk management mitigates inequities. It preserves profitability (Blascak et al., 2021).

## 7. Implementation Roadmap for Institutions

Gender bias in credit scoring demands a multi-layered response. It spans data, modeling, governance, and regulation. Integrate lifecycle strategies, GFI, and India evidence. This synthesizes pathways showing fairness advances of equity, profitability, and trust.

Phase 1: Diagnose and Align

- Compute disparities in approval rates, pricing, credit limits, and error rates across gender and intersectional segments.
- Validate calibration within groups for predictive accuracy.
- Quantify lost revenue from false negatives; model uplift from fair thresholding.
- Establish a cross-functional fairness committee, define KPIs, and set audit schedules.

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

Phase 2: Fix the Data

- Augment data via women's business networks and alternative data (e.g., digital payment histories).
- Train uplift and semi-supervised models for rejected inference pilots.
- Identify and constrain proxy variables (e.g., occupation) via causal analysis.
- Document justifications for retained features.

Phase 3: Retrofit the Model

- Integrate equal opportunity metrics into model training.
- Set group-aware thresholds to balance approval rates.
- Trigger human reviews for near-threshold cases.
- Stress test models under economic shocks for fairness.

Phase 4: Embed Transparency and Learning

- Deploy adverse action notices with counterfactuals (e.g., "reduce debt by $1,000").
- Streamline appeals to workflows to correct errors.
- Automated dashboards tracking TPR/FNR gaps and calibration drift.
- Rotate challenger models and update data to reduce bias accumulation.

Note. Adapted from multiple sources on algorithmic fairness and transparency in machine learning (Binns, 2018; Hardt et al., 2016; Wachter et al., 2018).

An Indian FinTech company could implement GFI in bias audits. In a school-run credit scoring simulation, students can apply REACT to analyze toy datasets and propose mitigations.

## 8. Conclusion

Gender bias in AI-driven credit scoring systems is a tangible issue. It stems from biased historical data and features laden with proxies for gender. This leads to measurable disparities, including higher false negative rates for women, stricter credit limits, costlier loan terms, and a

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

persistent financing gap. This gap hinders entrepreneurial opportunities and household financial stability.

However, these outcomes are not inevitable. Comprehensive audits, inclusive and representative datasets, and fairness-conscious model training can aid in carefully calibrating decision thresholds to success. Transparent model explanations support. Robust governance frameworks enabled. Lenders can mitigate unjust disparities. They can maintain or enhance risk-adjusted returns.

Addressing this challenge requires collaboration. Financial institutions should prioritize investments in fairness-enhancing tools. They need comprehensive model documentation. They require accessible customer recourse mechanisms.

Regulators must establish clear, outcome-focused standards. This ensures accountability.

Industry coalitions can share best practices and fairness benchmarks.

Technologists should develop models relying on causally robust signals to avoid structural inequities.

Real-world examples like Lendingkart show equitable scoring is achievable.

By designing AI systems to account for historical biases and promote inclusivity, the financial sector can transform credit scoring. It becomes a tool for financial inclusion, not perpetuating exclusions.

Our GFI and REACT framework enable this. For an Indian FinTech, use GFI to audit models and REACT for governance. In school simulations, apply them to toy data for hands-on learning.

**References**

1. Basel Committee on Banking Supervision, European Data Protection Board, European Securities and Markets Authority, International Committee on Credit Reporting (ICCR), Monetary Authority of Singapore, Roberts, T., Francisco, L., Ahmed, U., Trinh, T. D., Siddiqi, N., SAS Institute, U.S. Federal Reserve System, Data Science team, Wiley & Sons, & World Bank Group. (2019). *Credit scoring approaches guidelines*. World Bank Group. https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf

2. 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

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv*. https://arxiv.org/abs/1810.01943

3. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Journal of Machine Learning Research*, 18(81), 1–11. https://jmlr.org/papers/v18/17-635.html

4. Blascak, N., Cheney, J. S., Hunt, R. M., Mikhed, V., Ritter, D., & Varghese, M. (2021). Gender disparities in credit allocation. *Federal Reserve Bank of Philadelphia Working Paper*. (Adjusted for APA; original citation as Blascak et al., 2021)

5. Brookings Institution. (2021, January). The history of women's work and wages and how it has created success for us all. *Brookings*. https://www.brookings.edu/articles/the-history-of-womens-work-and-wages-and-how-it-has-created-success-for-us-all/

6. Chen, D., & Ye, W. (2022). Monotonic neural additive models: Pursuing regulated machine learning models for credit scoring. *arXiv*. https://arxiv.org/abs/2209.10070

7. Chodorow-Reich, G., & Coglianese, J. (2020). Adverse action notices in credit-scoring AI. (Adjusted; original citation as Chodorow-Reich & Coglianese, 2020)

8. Consultative Group to Assist the Poor (CGAP). (2024). *Gender-intentional credit scoring: Technical guide*. https://www.cgap.org/sites/default/files/publications/Tech%20Guide_Gender%20Intentional%20Credit%20Scoring%202024%20(1).pdf

9. Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2018). The measure and mismeasure of fairness. *arXiv*. https://arxiv.org/abs/1808.00023

10. Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. (Added based on text; Dastin et al., 2018 likely typo for Dastin, 2018)

11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255 (Adjusted for full citation)

12. Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., & Beben, S. (2021). Reject inference methods in credit scoring. *Journal of Applied Statistics*, 48(13–15), 2734–2754. https://doi.org/10.1080/02664763.2021.1929090

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

13. European Central Bank. (2022). Gender bias and credit access (Working Paper No. 1822). https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1822.en.pdf

14. Financial Conduct Authority, Bogiatzis-Gibbons, D., Charles, L., Dewing, H., Gretschel, C., Jomy, M., Reid, A., & Slack, R. (2024). A literature review on bias in supervised machine learning (Research Note). https://www.fca.org.uk/publication/research-notes/literature-review-bias-in-supervised-machine-learning.pdf

15. Garvan. (2024, October 11). The future of credit: AI or human judgment? *EMILDAI*. https://emildai.eu/the-future-of-credit-ai-or-human-judgment/

16. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv*. https://arxiv.org/abs/1803.09010 (For Gebru et al., 2018)

17. Gibson, K. (2019, November 12). Is Apple Card sexist? Goldman Sachs offers to review gender-bias claims. *CBS News*. https://www.cbsnews.com/news/apple-credit-card-goldman-sachs-disputes-claims-that-apple-card-is-sexist/

18. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

19. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems: 29* (pp. 3315–3323). https://papers.nips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

20. Hartmann, H., Hayes, J., Huber, E., Morrissey, T., & Rasheed, D. (2024). The long way to gender equality: Gender pay differences in Germany 1871–2021 (WID.world Working Paper). *World Inequality Database*. https://wid.world/document/the-long-way-to-gender-equality-gender-pay-differences-in-germany-1871-2021/

21. Kelly, S., Mirpourian, M., & Women's World Banking. (2021). *Algorithmic bias, financial inclusion, and gender*. Women's World Banking. https://www.womensworldbanking.org/wp-content/uploads/2021/02/2021_Algorithmic_Bias_Report.pdf

22. Konorski, J., & Rydzewski, K. (2017). Nodal cooperation equilibrium analysis in multi-hop wireless ad hoc networks with a reputation system. In *Lecture notes in computer science* (pp. 131–142). Springer. https://doi.org/10.1007/978-3-319-65127-9_11

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

23. Laugel, T., Lesot, M., Marsala, C., Renard, X., & Detyniecki, M. (2017). Inverse classification for comparison-based interpretability in machine learning. *arXiv*. https://arxiv.org/abs/1712.08443

24. Liu, Z., & Liang, H. (n.d.). Gender discrepancies in creditworthiness: Evidence from alternative credit data. *Gies College of Business, University of Illinois Urbana-Champaign*. https://ssrn.com/abstract=4897497

25. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

26. Nedlund, C. (2019, November 12). Apple Card is accused of gender bias. Here's how that can happen. *CNN Business*. https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html

27. Ongena, S., Popov, A., & European Central Bank. (2015). Gender bias and credit access (Working Paper No. 1822). *European Central Bank*. https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1822.en.pdf

28. SAP. (n.d.). What is AI bias? Causes, effects, and mitigation strategies. *SAP*. https://www.sap.com/resources/what-is-ai-bias

29. SME Finance Forum. (n.d.). Algorithmic bias in Women-MSMEs credit scoring: Insights & solutions (Community of Practice). https://www.smefinanceforum.org/post/algorithmic-bias-in-women-msmes-credit-scoring-insights-solutions

30. Song, Z., Rehman, S. U., PingNg, C., Zhou, Y., Washington, P., & Verschueren, R. (2024). Do FinTech algorithms reduce gender inequality in banks loans? A quantitative study from the USA. *Journal of Applied Economics*, 27(1). https://doi.org/10.1080/15140326.2024.2324247

31. Tan, S., Caruana, R., Hooker, G., Koch, P., & Gordo, A. (2019). Learning global additive explanations for neural nets using model distillation. *arXiv*. (Adjusted for Tan et al., 2019)

32. The World Bank Group et al. (2019). See Basel Committee above.

  

ISSN: 2455-8834

Volume:10, Issue:11 "November 2025"

33. Vidal, M. F., & Caire, D. (2024). Gender-intentional credit scoring. *CGAP*. https://www.cgap.org/research/publication/gender-intentional-credit-scoring

34. Vidal, M. F., Barbon, F., Consultative Group to Assist the Poor, & CGAP/World Bank. (2019). Credit scoring in financial inclusion. *CGAP/World Bank*. https://www.cgap.org/sites/default/files/publications/2019_07_Technical_Guide_CreditScore.pdf

35. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-Wachter-Mittelstadt-Russell.pdf

36. Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 336–349. https://doi.org/10.1145/3531146.3533101

37. Women's World Banking. (2023, June 21). In a world of gender bias, Lendingkart's AI-based credit model stands apart. *Women's World Banking*. https://www.womensworldbanking.org/insights/in-a-world-of-gender-bias-lendingkarts-ai-based-credit-model-stands-apart/

38. Women's World Banking. (2021). See Kelly et al., 2021.

39. Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y., & Clifton, D. A. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Medicine*, 6(1). https://doi.org/10.1038/s41746-023-00805-y

  