Journal of Economics Intelligence and Technology

ISSN(Online): 3082-3994
ISSN (Print): 3082-3986

Volume- 2| Issue- 01| 2026

Review Article

Received: 12-12-2025 | Accepted: 07-01-2026 | Published: 22-01-2026

# Bias, Fairness, and Explainability in AI-Driven Credit Scoring: A Critical Review of Algorithmic Governance in Financial Risk Assessment

***Clement Abugri*[1] *and Valerie Colley*[2]**

[1]*Denver, Colorado, USA*
[2]*Department of Statistics and Actuarial Science, KNUST*

**Abstract:** The application of artificial intelligence and machine learning to credit scoring has transformed traditional financial risk assessment but has raised serious concerns about algorithmic bias, fairness, and transparency. As machine-learning models increasingly determine credit decisions, the ethical considerations and governance structures that underpin them are critical to fintech inclusivity and regulatory compliance. This narrative review condenses recent research on bias detection, fairness methods, and explainability techniques for credit scoring models, while considering the underlying challenges to algorithmic decision-making in the U.S. financial system. Several key findings are highlighted: Recent research demonstrates significant advances in fairness-enhancing interventions, such as pre-processing bias mitigation, in-processing fairness constraints, and post-processing calibration methods. However, persistent challenges remain, including ongoing trade-offs between predictive performance and fairness metrics, providing meaningful explainability for complex ensemble models, and addressing intersectional discrimination. The integration of alternative data sources can increase inclusion, but risks introducing new forms of bias. While technological improvements have advanced bias detection and mitigation in credit scoring, tensions persist among fairness definitions, model performance, and regulatory demands. The review concludes by noting the importance of accounting for the context-dependency of fairness, improving evaluation schemes for real-world use, and establishing governance frameworks for algorithm accountability.

**Keywords:** Credit scoring; algorithmic bias; fairness in machine learning; explainable AI; financial inclusion.

## INTRODUCTION

Advances in applying machine learning-based models (under artificial intelligence and machine learning) to financial areas, especially in credit scoring, have changed risk assessment and decision-making by introducing cost-effectiveness and data-processing capabilities. This integration also creates challenges regarding transparency, fairness, and regulatory compliance, particularly due to the opacity of advanced AI systems. Credit scores are central to individual determinations of creditworthiness and access to financial services and opportunities (Kozodoi *et al*., 2022). As the financial industry increasingly adopts AI-based credit scoring systems, it is essential to understand how these automated models may entrench or introduce new biases, leading to unfair outcomes. Concerns are further heightened by the 'black box' nature of advanced machine learning algorithms in many cases, which may diminish transparency about decision-making and thereby make detection and correction of unfair practices challenging (Shokrzade *et al*., 2021).

The reshaping of credit score systems by AI and ML is a prominent use of algorithmic decision-making in financial services. Credit score models are used to determine access to financial products such as mortgages, personal loans, credit cards, and small business lending, as well as broader economic opportunities. As these systems increasingly depend on complex ML algorithms that process large datasets, concerns about bias, fairness, and transparency have become the main challenges for financial institutions, regulators, and civil society (Mehrabi *et al*., 2021).

Classic credit scoring systems, such as FICO scores, have used relatively small datasets from credit bureaus and applied transparent statistical techniques. With the rise of alternative data sources (e.g., utility payment and rental history, mobile phone usage, educational records, and social media activity) paired with ML frameworks such as gradient boosting machines, deep neural networks, and ensemble methods, the credit scoring landscape has changed (Berg *et al*., 2020). These developments offer more accurate predictions and broader financial inclusion but also introduce greater opacity and the risk that historical bias can be replicated or amplified.

The U.S. regulatory regime for credit scoring is centered on the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA). These laws prohibit discrimination based on protected attributes such as race, color, religion, national origin, sex, marital status, and age. Although these laws were enacted before modern AI, applying them to complex, non-linear models





***Corresponding Author:** Valerie Colley

DOI-https://doi.org/10.5281/zenodo.18334779

remain difficult. In these models, protected attributes may be inferred from variables that seem neutral (Hurlin *et al*., 2024). Recently, the Consumer Financial Protection Bureau (CFPB) has shown more interest in algorithmic fairness. However, policy guidance is still developing, and technical requirements can be vague.

Recent cases highlight the practical implications of biased credit scoring. Racial minorities, women, and low-income groups have all faced systematic disadvantages in algorithmic credit assessments, even when protected characteristics are not included in the model inputs (Fuster *et al*., 2022). Disparate impacts arise from several causes: biased training data from historical discrimination, proxy variables associated with protected features, feedback loops that perpetuate inequities, and fairness-accuracy trade-offs. These factors can worsen performance for subgroups when models are optimized for the entire population.

The ML community has produced significant work on fair-aware algorithms, leading to various mathematical definitions of fairness (demographic parity, equalized odds, calibration, individual fairness) and corresponding interventions (Chouldechova, 2020). Interpretable models, such as Shapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and attention mechanisms, have been developed by researchers in the explainable AI (XAI) community (Lundberg *et al*., 2020). However, theoretical research remains in the early stages for practical application in the production of credit scoring systems.

## RESEARCH GAP AND OBJECTIVES

Although research on algorithmic fairness and explainability is increasing, key gaps remain in mitigating bias in credit scoring. First, most fairness literature focuses on theory and benchmark datasets rather than real financial data or deployment (Chouldechova, 2020; Mehrabi *et al*., 2021; Agboola, 2025; Hurlin *et al*., 2024). Second, few studies examine the intersection of biases (e.g., racial, gender, socioeconomic) and patterned inequality (Foulds *et al*., 2020; Kozodoi *et al*., 2022; Berg *et al*., 2020; Osuyi, 2025). Third, little is known about trade-offs between different fairness measures and between fairness and accuracy in credit scoring. In these contexts, financial risk and social equity must be balanced (Chouldechova, 2020; Mehrabi *et al*., 2021; Kozodoi *et al*., 2022; Fuster *et al*., 2022). Fourth, regulatory compliance frameworks lack clear technical guidelines for AI-based credit unwillingness models. This leaves financial institutions unsure how to build fairness-aware systems (Sargeant, 2023; Waziri & Hassan, 2025; Valdrighi *et al*., 2025; Agboola, 2025).

This narrative review addresses these gaps by critically synthesizing recent literature from 2020 to 2025. We focus on bias, fairness, and explainability problems in credit scoring models. Our specific objectives are to:

1. Review the current state of fairness-oriented machine learning algorithms and their corresponding credit-scoring applications, including preprocessing, in-processing, and post-processing solutions.
2. Assess explainability techniques to make complex credit models understandable to customers, lenders, and regulators.
3. Evaluate the consequences of alternative data for the expansion of FIs and the emergence of new bias risk.
4. Analyze the regulatory and governance frameworks for algorithmic accountability in credit scoring.
5. Identify key research gaps and propose next steps to create even-handed, open, defensible, and anti-discriminatory credit-scoring programs.

Instead of providing systematic coverage, critically interpreting the literature to highlight themes and debates that have informed recent discussions on algorithmic fairness in consumer finance is done. Special attention is paid to the applications in the U.S. financial system, where certain regulatory and demographic contexts introduce distinct challenges and opportunities. By analyzing the field of both technological solution and sociotechnical consequences, this review aims to assist researchers, practitioners, and policy makers interested in discussing the intricate landscape created by AI-driven credit assessment.

## EMERGING TRENDS AND THEMATIC ANALYSIS

### Fairness Definitions and Measurement Frameworks in Credit Scoring

The operationalization of fairness in credit scoring is a central challenge that sits at the crossroads of mathematical precision, legal compliance, and ethical considerations. Recent literature has also heightened attention to the diversity of fairness definitions, and the competing demands they place





**Publisher: SARC Publisher**

on model design and evaluation (Mehrabi *et al*., 2021).

**Group Fairness Metrics**

Demographic parity (or statistical parity) requires that positive predictions (approved credits) be made at the same rate across protected groups. In credit scoring applications, this would require that the approval rates for diverse racial groups be equal, conditional on all other factors. Yet researchers caution that there might be tension between demographic parity and risk assessment if the actual default risks of groups differ because, in past generations, some faced economic discrimination (Chouldechova, 2020). Recent work by Hurlin *et al*. (2024) shows that demographic parity may even hurt the groups it is designed to help when there are differences in base rates, thereby lowering or raising standards for dominant groups or underprivileged ones.

Equalized odds, in contrast, require equality of true positive rates (TPR) and false positive rates (FPR) across groups. For credit scores, this means that if one considers only qualified borrowers (borrowers who would always repay), approval rates should be the same across protected groups, and similarly for unqualified borrowers (who will default). New empirical research demonstrates that equalized odds may be closer to anti-discrimination law than demographic parity, as it is a function of the target outcome rather than ignoring it altogether (Kozodoi *et al*., 2022).

In the context of group fairness, another essential requirement is well-calibrated scores across groups at each level, meaning that the predicted probability of default should equal the realized default rate for each protected group. If a model predicts that 20% of applicants from various demographic groups will default, then roughly 20% of individuals in each group should default. Mathematical impossibility results show that, subject to degenerate or finite measure, it is not possible to achieve calibration, equalized odds, and equal base rates across groups at the same time, a natural trade-off with far-reaching implications for fair credit scoring (Hurlin *et al*., 2024).

**Fairness Through Awareness vs. Fairness Through Unawareness**

There is a trade-off at play between "fairness by awareness" approaches, where we explicitly consider protected attributes to ensure treatment is fair, and 'fairness by unawareness' (blindness), too, where we just purposefully exclude the variables that give rise to them from our models. Pursuant to fair lending law in the U.S., protected classes cannot be considered in making a credit decision, and thus implicitly requires that the unawareness approach be taken. Yet, multiple studies have shown that merely removing protected variables does not safeguard against biased impact, as even proxy variables (zip code, ethnicity derived from names, educational institutions) can encode protected information (Fuster *et al*., 2022).

Several recent works have proposed "fairness through awareness" interventions that utilize protected attributes during model training to enforce fairness constraints, but not in the scoring algorithms deployed at test time. Valdrighi *et al*. (2025) proposed a best practice for responsible ML in credit scoring, demonstrating that fairness constraints can be imposed during training while preserving competitive accuracy. Their framework has been particularly influential in credit scoring, providing practical tools for implementing various fairness definitions.

**Contextual Fairness in Financial Services**

Recent work stresses the importance of domain-specific fairness definitions for credit scoring. The stage of the credit lifecycle may require different fairness criteria, from marketing and pre-screening that aim to ensure equal opportunity across demographic lines, to final lending decisions that prioritize calibration for financial sustainability (Kozodoi *et al*., 2022). The purpose of the credit (a mortgage for a primary residence versus a discretionary credit card) may also give rise to different fairness considerations, as the stakes and social policy objectives differ.

Recent work by Fuster *et al*. (2022) studying fintech lending data showed that algorithms incorporating alternative data can mitigate racial disparities in approval rates while simultaneously increasing or maintaining predictive accuracy (contrary to simplistic narratives of fairness-accuracy trade-offs being inevitable). But they caution that these advantages depend on the context and may not apply to all loan types or demographic groups.

**Measurement Challenges and Evaluation Frameworks**

However, the use of fairness metrics in production credit systems poses practical challenges. First, protected group membership is usually absent from training data for legal reasons (it cannot be





**Publisher: SARC Publisher**

collected or used), and researchers must rely on proxies derived from names or geolocation, or conduct separate bias audits using external demographic databases. Second, intersectional fairness: fairness across intersections of protected attributes (e.g., Black women or Asian elderly) becomes far more computationally intractable as the number of protected groups increases exponentially (Foulds *et al*., 2020). Third, the temporal stability of fairness metrics has been relatively underexplored; a fair model at deployment might learn to become biased over time as population distributions change or feedback loops develop.

**Bias Sources and Mitigation Techniques**

Bias can arise at different stages of the ML pipeline, including data collection, feature engineering, model training, and deployment (Mehrabi *et al*., 2021), thereby necessitating a holistic approach to debiasing.

**Preprocessing: Data Acquisition and Feature Extraction**

Historical discrimination in labeled data also poses a key challenge for fair credit scoring. Traditional credit data is inherently biased toward excluding populations without traditional credit histories (e.g., immigrants, young people, and those in lower-income brackets), leading to sample selection bias. Models trained on data biased towards advantaged populations tend not to perform well or to make equitable predictions for underrepresented groups (Valdrighi *et al*., 2025).

Several pre-processing interventions have been investigated in recent work. Reweighing methods weight the training samples to ensure balanced representation across protected groups, reducing disparate impact without disregarding any observations. Data augmentation methods enhance minority groups by generating additional samples; however, simplistic synthetic data generation can yield unrealistic feature combinations, thereby compromising classifier performance (Sargeant, 2023).

Feature engineering is another important pre-processing phase. Proxy features or characteristics that are statistically associated with protected identities can propagate bias even when protected attributes are not explicitly included. For instance, zip code is tightly linked to race due to residential segregation; educational facilities have a strong correlation with ethnicity and socioeconomic status (Hurlin *et al*., 2024).

Alternate data integration is a promising but potentially dangerous direction in pre-processing. Adding payment history for utilities, rent, and telecom will increase access for credit-invisible populations. But alternative data sources can also encode present-day inequalities – rental payment data, for example, disadvantages of people in discriminatory housing markets (Berg *et al*., 2020). Previous studies have been very careful about feature selection and bias auditing when incorporating new external data.

**In-Processing: Fairness-Aware Model Training**

In-processing methods encode notions of fairness directly into the objective function used for model training, with an explicit trade-off between reductions in predictive accuracy and fairness. This approach has received significant attention in recent years, as, in addition to its theoretical elegance, it should be highly effective in practice (Kozodoi *et al*., 2022).

Multi-objective optimization methods treat fairness and accuracy as separate objectives, yielding Pareto-optimal solutions that reflect trade-offs between the two. More recently, the power of multi-objective approaches to mortgage lending was demonstrated by showing that fairness gains could be achieved at very low accuracy costs in some parts of the trade-off surface (Valdrighi *et al*., 2025).

**Post-Processing: Calibration and Threshold Optimization**

Pre-processing methods transform input data instances before training the model to satisfy the fairness criterion. These methods are especially appealing for those who work with legacy models or third-party scoring systems, for which retraining stakeholders' models to accommodate the change is not an option (Dat *et al*., 2025).

Calibration methods aim to produce fair score distributions or probability estimates. Past research has shown that individual-level calibration can be achieved by learning distinct calibration functions for several population subgroups, so that the modelled probabilities match observed outcomes in each subgroup (Hurlin *et al*., 2024).

## LIMITATIONS AND ONGOING CHALLENGES

Although comprehensive work has been done, some severe defects still exist in bias reduction methods. First, as noted previously, most methods assume that sensitive attributes are known and accurately measured; this is not the case in real





**Publisher: SARC Publisher**

credit data. Second, once interventions are optimized for one fairness metric, their performance on other metrics may degrade due to mathematical infeasibilities (Chouldechova, 2020). Third, methods for addressing bias can reduce the model's overall predictive power, and it is not clear what the optimal trade-off between fairness and accuracy should be across applications (Agboola, 2025).

More importantly, mitigating technical bias does not address structural inequities in the underlying economic and social systems. Such a focus on the narrow technical intervention of algorithms can lead to ignoring wider systemic elements that produce differential outcomes - discriminatory housing markets, unfair education systems, employment discrimination, wealth disparities (Ford 2025; Osuyi, 2025).

**Explainable and Interpretable in Credit Scoring Models**

The opacity on the human-interpretable end of the spectrum makes it problematic to use the most advanced ML models for credit scoring, where transparency is necessary for consumer protection, regulatory risk, and lender risk management. Recent developments in explainable AI (XAI) offer new ways to interpret complex credit models; however, ongoing challenges persist between technical explainability and human interpretability (Lundberg *et al*., 2020).

**Global Interpretability: Understanding Model Behavior**

Global interpretability techniques describe the general model behavior and significance patterns across all predictions. SHAP (Shapley Additive exPlanations) values are becoming a popular technique for global interpretability. SHAP uses game-theoretic concepts to assign model predictions back to input features in a way that has many desirable properties, including local accuracy, robustness to missing values, and consistency. Lundberg *et al*. (2020) showed that SHAP can offer both global and local interpretability to practitioners, providing insight into the features that drive model decisions across populations and explanations of individual predictions.

In recent credit default prediction applications, it has been observed that payment history, debt-to-availability ratio, and account age were drivers of model decisions, providing positive empirical support for traditional credit theories (Gramegna & Giudici, 2021). Nonetheless, it was shown that some sensitive characteristics, such as the use of proxy variables (e.g., specific merchant categories), may contain demographic information that was abused in prior works on consumption data.

In a comprehensive study by Alqahtani *et al*. (2025), 150 peer-reviewed papers on model-agnostic explainable AI in finance were reviewed, and it was concluded that SHAP and LIME are the most used methods for credit scoring. They observed that credit risk models made loan approval more transparent and fairer, though trade-offs between interpretability and predictive performance persisted.

**Local Interpretability: Explaining Individual Predictions**

These local interpretability methods interpret individual applicant predictions, which are the predictions of interest in the context of regulatory requirements for adverse action notices and consumer "right to explanation" requests. The Fair Credit Reporting Act requires that lenders give reasons for credit denials, so explaining the samples is necessary.

LIME (Local Interpretable Model-agnostic Explanations) approximates complex models locally using simple, interpretable linear models. Similarly, LIME determines which features influenced the score of a given credit applicant by perturbing the inputs and observing changes in the prediction (Gramegna & Giudici, 2021). Several fintech lenders have used LIME to generate consumer-facing explanations, but they remain cautious about their stability; slight changes in applicant's features can lead to very different explanations.

Gramegna and Giudici (2021) conducted an extensive comparison of SHAP and LIME in credit risk applications and tested their discrimination power on real data sets for SMEs. They observed that, with both techniques yielding useful information, the SHAP method obtained more consistent results and better identified true feature importance, compared to LIME’s exclusively local approximations, which may sometimes mislead about the model's global behavior.

Recent efforts have introduced five dimensions for assessing explainability in credit risk, including interpretability, global explanations, local explanations, consistency, and complexity (Shokrzade *et al*., 2023). This framework provides





**Publisher: SARC Publisher**

systematic guidance on whether the explanation satisfies regulatory requirements and user demand.

**Regulatory Compliance and Legal Standards**

The need for explainability in credit scoring comes from different regulations. The FCRA requires that adverse action notices specify the "principal reasons" for credit denials. The ECOA generally prohibits discrimination and mandates that credit standards be based on rationality. These regulations were drafted before modern ML, so it is unclear what compliance with complex models would entail.

The most recent direction has been that explainability techniques should provide "accurate, meaningful and useful" explanations rather than surface or misleading ones (Hurlin *et al*., 2024). But it is unclear what the specific technical standards are, and research suggests that current XAI techniques may not meet legal requirements' explanations; they currently yield statistical correlations rather than a causal account.

**Trade-offs Between Accuracy and Interpretability**

There is always a trade-off between model performance and interpretability. Linear models and trees are inherently transparent but often less accurate than complex ensembles of trees, forests, or other non-linear methods in the prediction task; as a result, lenders must trade off their risk management and profitability objectives against transparency.

New studies question simplistic "accuracy vs. interpretability." One prevailing view is that for high-stakes tasks such as lending, it is preferable to use models with built-in interpretability rather than highly complex ones with only post-hoc interpretations, which may be unfaithful or misleading (De Bock *et al*., 2024). A well-constructed scoring system can be highly accurate while maintaining full transparency.

**Limitations of Current Explainability Approaches**

However, in the credit domain, XAI algorithms have had limited success. First, explanation fidelity is still under debate; it's unclear whether explanations provide access to a model's true reasoning, especially when models are highly nonlinear. Second, explanations are often non-robust: small distortions can yield vastly different explanations, even when the predicted class probabilities are similar (Alqahtani *et al*., 2025). And third, we have cognitive science work which shows that humans are bad at understanding probabilistic reasoning and feature interactions (so having technically correct but complex explanations give you limited value) (Khan *et al*., 2025).

But, most fundamentally, transparency does not guarantee fairness or accountability. Even with a fully explainable model, it can represent unjust social patterns and be optimized for an unfair goal. Explainability is required but not sufficient for responsible credit scoring; it must be combined with fairness guarantees, robust validation, and meaningful stakeholder engagement (Sargeant, 2023).

**Alternative Data Access: Opportunities and Risks**

The trend towards the use of non-traditional data sources in credit underwriting is one of the most profound in consumer finance, with extremely deep implications for both financial inclusion and bias risks. Alternative data refers to information other than traditional credit bureau reports, and may include utility payments, rent payments, bank transaction data (checking account or savings account), academic degrees and employment history, mobile phone usage, and, more recently, digital footprints on e-commerce platforms (Berg *et al*., 2020; Salami *et al*., 2025).

**Financial Inclusion Potential**

Roughly tens of millions of U.S. adults are "credit invisible", meaning they lack credit histories with the three major bureaus, and another 19 million have "unscorable" files due to insufficient histories, which contain insufficient data to gauge a score traditionally (Consumer Financial Protection Bureau, 2023; Board of Governors of the Federal Reserve System, 2022). They are disproportionately racial minorities, immigrants, young adults, and lower-income people who have been systemically excluded from the financial mainstream. Alternative data offers potential avenues for inclusion by enabling the assessment of creditworthiness for individuals without traditional credit histories.

Recent empirical evidence points to positive effects on inclusion. Based on fintech lending data, Fuster *et al*. (2022) find that ML models using alternative data sources increased approval rates for minorities by only 13 percentage points, while default rates remained unchanged, suggesting that credit was extended without increased risk. Berg *et al*. (2020) examined digital footprints in e-





**Publisher: SARC Publisher**

commerce data and found that transaction history, device details, and shopping activities could be used to predict defaults among thin-file borrowers, with performance like traditional credit scores for prime applicants.

Yet in a 2024 survey by Nova Credit, 90% of lenders said having more alternative data points would enable them to say yes to more creditworthy applicants, though only 43% currently include alternative data in their risk assessments alongside traditional scores. Adoption barriers include regulatory concerns, data reliability issues, and integration challenges.

TransUnion found that when rent payments were added to consumers' credit reports, the scores of those with low scores rose by an average of almost 60 points. But only 10% of renters have their on-time rental payments included in their scores, demonstrating that they fall quite short of collecting comprehensive alternative data.

**Bias Risks and Discrimination Concerns**

However, when using data commonly used in AI and ML algorithms, introducing bias from alternative sources can exacerbate or propagate existing biases. These risks arise at different levels and should be addressed through meticulous investigation.

**Historical Discrimination Embedded in the Data:** Secondary data frequently mirrors the outputs of discriminatory systems. Rental payment histories disadvantage those subject to housing discrimination; educational qualifications are linked to family wealth and encode racial disparities of access to educational opportunities; employment history mirrors discriminatory hiring practices. When such data is injected into credit models, it may introduce historical biases into automated systems (Sargeant, 2023).

**Digital Divides and Inequalities of Access:** Relying on alternative feeds may, ironically, be biased against the populations it seeks to serve. Digital trace-based scoring disadvantages of people who do not have smartphones, internet access, or tech literacy, who tend to be older, rural, and in low-income communities. In writing about risk, research shows that alternative data can create "poverty penalties," in which failure to be online is seen as risky, and low-income customers are excluded because they cannot afford digital participation.

**Inferential Discrimination:** ML models can infer protected attributes from seemingly neutral proxy information with fairly high accuracy. Even credit models based on behavioral data can engage in proxy discrimination in the absence of explicit protected attributes (Fuster *et al*., 2022).

**Lack of Standardization and Validation:** Unlike traditional credit information, which is subject to FCRA rules requiring data to be accurate, complete, and allowing for dispute/verification, there are no standards for collecting (e.g., scraping), verifying, or correcting alternative data sets. Mistakes on utility bills, telecom statements, or e-commerce transactions feed credit models that lack meaningful protections for consumers.

**Regulatory Landscape and Policy Debates**

Guidance on alternative data in credit scoring remains piecemeal and evolving. Alternative data use is applicable to traditional consumer protection statutes (FCRA, ECOA), but implementation standards are unspecified. Recent Congressional Research Service reports insist that alternative data should exhibit basic levels of accuracy, predictive power, and non-discrimination, even if the specifics are fuzzy.

The CFPB Section 1033 "Personal Financial Data Right" rule, proposed by the CFPB, could ease the adoption of alternative data by making it easier to share it, and 75% of respondents believe it will encourage its use. But policymakers face a dilemma when it comes to the push-and-pull between encouraging competition and protecting consumers.

**Emerging Best Practices**

Despite the ongoing debate, there is growing convergence on some of the best practices for responsible integration of alternative data:

1. **Purpose Limitation:** Alternative data should be used only if it can be shown to demonstrably improve predictive accuracy or fairness for underserved populations.
2. **Clear Validation:** Lenders should be able to log locale data, validation method, and bias audit.
3. **Consumer Control:** People should have visibility into the alternative data used to calculate their scores and a process to correct any errors.
4. **Continuous Monitoring:** Models utilizing other data need to be continuously monitored for fairness as data changes over time





**Publisher: SARC Publisher**

5. **Impact assessment:** Deployment should be preceded by an impact assessment of the algorithm on the protected groups.

## APPLICATIONS ACROSS CREDIT PRODUCT CATEGORIES

### Consumer and Personal Credit

Consumer lending applications (such as personal loans, credit cards, and auto financing) are the most common use-case of AI-powered credit scoring systems. Large financial institutions have built machine learning models with hundreds of features based on a variety of other sources beyond credit score, such as transaction history, account dynamics and customer interactions to establish sophisticated decision support for credit line management and default predictions.

But recent scandals demonstrate bias risks. This led to allegations that some fintech credit products were using an automated algorithm which set a lower credit limit for women than men with the same financial history, which is now being investigated by regulators. Although institutions defended their use of protected characteristics as not influencing credit decisions, the incidents made evident that proxy variables and biased training data could generate discriminatory outcomes even without them.

Research by Bartlett *et al*. (2022), based on mortgage lending data, fintech algorithms discriminated against the minority by 40% less than traditional human underwriters, indicating that ML can attenuate human bias under some circumstances. But they caution that this result is limited to the context of mortgages, where regulations are particularly strict and may not generalize to other consumer credit products which face less regulation.

### Mortgage Underwriting

Mortgage credit is the most significant source of applications for households and has major effects on wealth accumulation as well as intergenerational economic upward mobility. The use of machine learning (ML) algorithms to power automated underwriting systems (AUS) in the mortgage industry has become prevalent; however, questions about algorithmic fairness in lending have been raised.

Fannie Mae's Desktop Underwriter and Freddie Mac's Loan Product Advisor receive millions of mortgage applications each year, generating risk assessments based on an algorithm that takes into account creditworthiness, collateral risk and repayment ability. Analyses of algorithmic mortgage underwriting in recent studies do show nuanced results. Bartlett *et al*. (2022) found that fintech marketplace lenders using algorithms charged racially-minority borrowers' lower interest rates than banks did when the two sets of borrowers had similar loan characteristics, indicating that ML may have lowered discriminatory pricing. However, Fuster *et al*. (2022) explains that those gains were concentrated among higher-credit-score minorities, and that subprime minority borrowers continued to experience disadvantaged outcomes.

### Small Business Lending

Commercial small business risk scoring is a different application domain which poses specific data challenges and fairness related issues. The ascendancy of fintech players like PayPal Working Capital, Square Capital, and Kabbage is the result of ML models digesting alternative data sources to reinvent small business lending. These platforms use transaction data from payment processing, e-commerce sales trends, and customer reviews to evaluate a business's creditworthiness.

Preliminary evidence indicates that these approaches widened access for minority-owned and women-owned businesses that have had limited access to business funding from mainstream banks. But in small business settings, alternative data is subject to bias. Transaction-based lending helps those that do a lot of business, which can disadvantage service providers and seasonal businesses often favored by entrepreneurs who are immigrants.

## BENEFITS, LIMITATIONS, AND ETHICAL CONSIDERATIONS

### Advantages of the AI -Based Credit Scoring Model

#### Enhanced Predictive Accuracy

ML models continue to significantly outperform traditional credit scoring technologies in various lending scenarios (Ojo & Adeyemi, 2025). Research demonstrates that ML methods increase the accuracy of classification by 6-8% over logistic regression in consumer credit, with significant improvements in default losses or risk-equivalent extensions of lending at those levels (Kozodoi *et al*., 2022).

#### Expanded Financial Inclusion

Using non-traditional (i.e. alternative) data sources allow for a credit assessment of "invisible" people that might in turn, open access to credit





**Publisher: SARC Publisher**

lines for millions of underserved consumers. Fuster *et al*. (2022) showed ML models to have improved minority approval rates by 13 percentage points without changing default rates, suggesting that inclusion and risk management are not at odds.

**Reduced Human Bias**

Machine decision-making may help to minimize arbitrary human bias in credit scoring, including stereotyping, implicit bias, and discriminatory predilections. Bartlett *et al*. (2022) observed that in mortgage lending algorithmic decisioning was 40% less biased against a protected class than human underwriters, leading them to conclude that machine learning models could obviate bias.

**Operational Efficiency**

Credit decisions and processing times are cut significantly with AI-based credit systems. Processes previously taking days or weeks of human underwriting are now approved immediately, enhancing customer experience and lowering costs for lenders.

**Limitations and Challenges**

**Fairness-Accuracy Trade-offs**

There are basic mathematical impossibilities in trying to optimize all our fairness objectives along with predictive accuracy. It has been shown empirically that it is not the case; on average, accuracy deteriorates when we enforce fairness constraints and the extent of this damage may differ from context to context (Kozodoi *et al*., 2022).

**Data Quality and Representativeness**

Bias in training data is propagated and aggravated by ML models (Mehrabi *et al*., 2021; Chouldechova, 2020). Credit histories are the legacy of centuries of discriminatory lending, employment discrimination and wealth disparity that get built into algorithms trained on such data (Bartlett *et al*., 2022; Fuster *et al*., 2022; Sargeant, 2023). Minority populations are even more impacted by quality and fairness problems of the data; erroneous entries in credit reports, low coverage and more biased representations existing in alternative data tables, unbalanced nature of training samples all stand behind biased predictions (Berg *et al*., 2020; Agboola, 2025; Osuyi, 2025).

**Opacity and Black-Box Decision-Making**

Despite strides in XAI, complex ML models are central to most stakeholders, such as consumers and regulators, and even developers. The current deep learning revolution has led to systems with millions of parameters working on hundreds of features that are difficult for humans to comprehend leading to a lack of accountability (and reduced controllability) and transparency (Alqahtani *et al*., 2025).

**Feedback Loops and Amplification**

Credit scoring systems generate feedback loops that may reinforce initial biases over time. Unsuccessful applicants are unable to provide evidence of creditworthiness from successful repayment as this exclusion persists. Algorithmic credit decisions also affect downstream financial outcomes, resulting in self-fulfilling prophecies whereby initial algorithmic categorizations determine future economic reality (Ford 2025).

**Ethics and Stakeholders Views**

**Competing Values and Ethical Frameworks**

Fairness in credit scoring is a matter of weighing conflicting values that can never be entirely harmonized. What equality requires is treatment of like cases as like, and what counts as a like case can be a matter of contestable value judgment. Lenders stress risk management, consumers care about fair treatment, civil rights advocates focus on structural inequality, and regulators have to maintain both financial stability and consumer protection.

**Algorithmic Accountability and Governance**

Establishing liability for the algorithmic determination of credit is a basic problem. When complicated ML systems create discriminative decisions, responsibility can be attributed to multiple parties such as providers of data, models, deploying institutions and the algorithmic system itself. Classical mechanisms of accountability fail in the context of opacity and technical complexity.

## FUTURE DIRECTIONS AND RESEARCH GAPS

**Technical Research Frontiers**

**Intersectional Fairness**

Most of the fairness literature focuses on isolated demographics while discrimination is commonly at the intersections. They note that black women, elderly immigrants and disabled LGBTQ+ people are doubly wronged in ways not fully captured by considering race, age, disability or sexual orientation in isolation. The generation of fair metrics and mitigations for intersectional groups is an important area of research (Foulds *et al*., 2020).





**Publisher: SARC Publisher**

### Causal Fairness Frameworks

Existing fairness metrics are largely correlational and only compare outcomes between groups without considering the causal mechanisms. Yet separating real causal connections from discriminatory pathways demands causal thinking. The development of useful (practical) causal fairness methods also faces several challenges, including how to go about finding suitable causal graphs for describing the complex socio-economic systems that evolved over history.

### Fairness Under Distribution Shift

Most of the fairness works adopt static data distributions, but in practical credit systems, continuous distribution shift is present due to economic environment movement, regulatory reform and demographic evolution. Under what conditions do fairness properties weaken due to shift in the distribution? Can we provide fairness guarantees against distributional shifts? These are questions that demand both theoretical and empirical inquiry.

## NEEDS OF EMPIRICAL AND APPLIED RESEARCH

### Real-World Deployment Studies

Existing fairness work is based on either public datasets or simulations and provides few insights about real deployment impacts. Research is needed to monitor various measures of fairness, financial outcome, and institutional practices within production credit systems on a longitudinal basis (Valdrighi *et al*., 2025).

### Comparative Effectiveness Research

Although many bias-reducing methods have been suggested, it is not yet clear which of these are effective and whether their performance varies across credit contexts. What approaches are most effective for various lending products, types of borrowers, and equity goals? Objective comparative analysis in standardized credit cases would assist clinicians in selecting their method.

### Stakeholder-Centered Design Research

The technical literature involves little stakeholder involvement, and different communities may care about different definitions of fairness. Participatory design research with impacted communities to specify fairness objectives and assess systems could lead to fairness criteria that are more contextually relevant.

## REGULATORY AND POLICY RESEARCH

### Optimal Regulatory Frameworks

What "rules of the game" will foster fair AI credit scoring? If the rules must stipulate which fairness measure is acceptable, or define performance criteria measurable via auditing? How should regulators weigh the desire to stimulate innovation against that of safeguarding consumers? To address these questions, interdisciplinary investigations that draw upon legal analysis, economic modeling, and technical expertise are needed.

### Algorithmic Auditing Standards

In comparison, third-party auditing for algorithmic credit systems is a ubiquitous proposal without sufficient norms and methods of auditing yet in development. What should auditors check; training data, model architecture, measures of fairness, explanations, documentation? Designing standardized review procedures like financial logging norms could allow control (Hurlin *et al*., 2024).

### Adverse Action Explanation Requirements

The existing requirements for adverse action notice were designed with a simple scoring system in mind, not complex ML models. How best to reform these requirements considering modern algorithms, yet maintaining sensible transparency, is a big policy question. If explanations highlight high-level features, changes in terms of counterfactuals, or the profile of similar approved applicants? Experimental studies that compare the impact of various formats of explanations on consumer understanding could inform regulatory standards.

## INTERDISCIPLINARY INTEGRATION

### Behavioral Economics and Algorithm Design

Behavioral economics has discovered certain regularities in human processing of financial information (present-bias, mental-accounting, framing effects), but these insights are not readily incorporated into credit scoring algorithms. Can behaviorally motivated scoring even better predict default while remaining attuned to human psychology? What's the relation between algorithmic credit scoring and behavioral biases? Blending the Two This integration necessitates partnership between ML specialists and behavioral economists (Sargeant, 2023).

### Social Science Approaches to Fair Algorithm Design

Sociology, anthropology, and political science contribute important lessons about power





**Publisher: SARC Publisher**

dynamics, inequality, and institutionalized discrimination that are largely neglected in technical work on fairness. How do credit algorithms fit into larger systems of economic extraction? How does the credit mechanism help create social classes? More interdisciplinary work merging social science views with technical approaches could offer a more complete picture of challenges to fairness in algorithms.

**Legal Theory and Technical Implementation**

Improved collaboration between legal scholars and ML experts is essential to create algorithms that meet the standard dictated by law, not only a technical standard of fairness. What is the legal notion of non-discrimination, and how does it correspond to technical definitions of fairness? What are justifiable uses of protected attributes in fairness interventions? Research that would convert those legal standards into more technical implementable requirements would promote the development of such compliant systems.

## CONCLUSION

AI-powered credit scoring is rewiring access to financial opportunities, bringing technical tools to the task of dismantling bias and uncovering deeper structural problems. Recent advancements in fairness-aware machine learning, transparency techniques like SHAP, LIME and counterfactual explanations, and the use of alternative data illustrate that with conscious model design we can address bias and promote inclusion. But competing criteria of fairness, trade-offs between accuracy and fairness and non-causal explainability limitations bound to what technical methods can do. Such methods alone cannot solve the socioeconomic inequities behind disparate credit outcomes. In some cases, they are promoted mainly to avoid addressing these fundamental issues.

Responsible deployment demands the consideration of fairness as a primary design principle. Such remedies include engaging in methodologically rigorous bias audits; committing to fairness metrics which are sensitive to legal and stakeholder concerns; monitoring systems regularly (or in real time); privileging interpretable models where practical and communicating clearly about the strong limitations of the model. Regulators need to modernize oversight by establishing clear fairness and auditing standards, amping explanation requirements, allowing for responsible innovation approaches, and tackling the structural inequities shaping creditworthiness. Insights from Credit Scoring are relevant to more general settings in which AI is used to enable access to critical life opportunities, such as work, education, and healthcare. To truly ensure fair, transparent, and accountable credit algorithms, we must prioritize interdisciplinary research, engage stakeholders directly in the design process, and demand coordinated institutional support. Stakeholders and policymakers must commit to embedding algorithmic fairness within broader initiatives that build more equitable economic systems. We must act now to ensure AI expands opportunity rather than perpetuates existing inequalities.

## REFERENCES


1. Abdel-Basset, M., Mohamed, R., Abouhawwash, M., Chakrabortty, R. K., & Ryan, M. J. "EA-MSCA: An effective energy-aware multi-objective modified sine-cosine algorithm for real-time task scheduling in multiprocessor systems: Methods and analysis." *Expert systems with applications* 173 (2021): 114699.
2. Agboola, O. K. "Auditing bias in AI and machine learning-based credit algorithms: A data science perspective on fairness and ethics in FinTech." *International Journal of Technology, Management and Humanities* 11.02 (2025): 1-11.
3. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. "Consumer-lending discrimination in the FinTech Era." *Journal of Financial Economics* 143.1 (2022): 30-56.
4. Berg, T., Burg, V., Gombović, A., & Puri, M. "On the rise of fintechs: Credit scoring using digital footprints." *The Review of Financial Studies* 33.7 (2020): 2845-2897.
5. Chouldechova, A., & Roth, A. "A snapshot of the frontiers of fairness in machine learning." *Communications of the ACM* 63.5 (2020): 82-89.
6. Coraglia, G., Genco, F. A., Piantadosi, P., Bagli, E., Giuffrida, P., Posillipo, D., & Primiero, G. "Evaluating ai fairness in credit scoring with the brio tool." *arXiv preprint arXiv:2406.03292* (2024).
7. Kambara, M., & Luce, C. "Technical Correction and Update To The CFPB's Credit Invisibles Estimate." *Consumer Financial Protection Bureau Office of Research Reports Series* 25-5 (2025).
8. Dasgupta, K., Larrimore, J., Lloro, A., Merchant, Z., Merry, E. A., Shaalan, F., & Tranfaglia, A. “Economic Well-Being of US






**Publisher: SARC Publisher**

Households in 2023." No. 5133. Board of Governors of the Federal Reserve System (US), (2024).
9. Das, S., Huang, X., Adeshina, S., Yang, P., & Bachega, L. "Credit risk modeling with graph machine learning." *INFORMS Journal on Data Science* 2.2 (2023): 197-217.
10. Dat, N. Q., Dai Phong, N., & Ban, D. T. "Mitigating Algorithmic Bias in Credit Scoring: A CNN-SMOTE Framework." *Asian Journal of Mathematics and Computer Research* 32.3 (2025): 77-83.
11. De Bock, K. W., Coussement, K., De Caigny, A., Słowiński, R., Baesens, B., Boute, R. N., ... & Weber, R. "Explainable AI for operational research: A defining framework, methods, applications, and a research agenda." *European Journal of Operational Research* 317.2 (2024): 249-272.
12. Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. "An intersectional definition of fairness." *2020 IEEE 36th international conference on data engineering (ICDE)*. IEEE, 2020.
13. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. "Predictably unequal? The effects of machine learning on credit markets." *The Journal of Finance* 77.1 (2022): 5-47.
14. Gramegna, A., & Giudici, P. "SHAP and LIME: an evaluation of discriminative power in credit risk." *Frontiers in Artificial Intelligence* 4 (2021): 752558.
15. Hurlin, C., Pérignon, C., & Saurin, S. "The fairness of credit scoring models." *Management Science* 72.1 (2026): 406-425.
16. Khan, F. S., Mazhar, S. S., Mazhar, K., A. AlSaleh, D., & Mazhar, A. "Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions." *Artificial Intelligence Review* 58.8 (2025): 232.
17. Kozodoi, N., Jacob, J., & Lessmann, S. "Fairness in credit scoring: Assessment, implementation and profit implications." *European Journal of Operational Research* 297.3 (2022): 1083-1094.
18. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. "From local explanations to global understanding with explainable AI for trees." *Nature machine intelligence* 2.1 (2020): 56-67.
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54.6 (2021): 1-35.
20. Ojo, O. E., & Adeyemi, R. A. "Machine learning for credit scoring and loan default prediction: A comprehensive review." *World Journal of Advanced Research and Reviews*, 26.3 (2025): 884–904.
21. Osuyi, E. "Bias and fairness in AI-based credit scoring: A comparative study across demographics." *International Journal of Advanced Research in Computer Science*, 16.2 (2025): 45–68.
22. Salami, I. A., Popoola, A. D., Gbadebo, M. O., Kolo, F. H. O., & Adesokan-Imran, T. O. "AI-powered behavioural biometrics for fraud detection in digital banking: A next-generation approach to financial cybersecurity." *Asian Journal of Research in Computer Science* 18.4 (2025): 473-494.
23. Sargeant, M. H. "Economic and Normative Implications of Algorithmic Credit Scoring." (2023).
24. Shokrzade, A., Ramezani, M., Tab, F. A., & Mohammad, M. A. "A novel extreme learning machine based kNN classification method for dealing with big data." *Expert Systems with Applications* 183 (2021): 115293.
25. Valdrighi, G., M. Ribeiro, A., SB Pereira, J., Guardieiro, V., Hendricks, A., Miranda Filho, D., ... & Medeiros Raimundo, M. "Best practices for responsible machine learning in credit scoring." *Neural Computing and Applications* 37.25 (2025): 20781-20821.
26. Lui, A. T., Lamb, G., & Durodola, L. "A right to explanation for algorithmic credit decisions in the UK." *Law, Innovation and Technology* 17.1 (2025): 289-317.

**Source of support:** Nil; **Conflict of interest:** Nil.

**Cite this article as:**
Abugri, C. & Colley, V. "Bias, Fairness, and Explainability in AI-Driven Credit Scoring: A Critical Review of Algorithmic Governance in Financial Risk Assessment." *Journal of Economics Intelligence and Technology* 2.1 (2026): pp 1-12.





**Publisher: SARC Publisher**