International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

# Bias Mitigation in AI-Driven Banking: Legal, Ethical, and Regulatory Perspectives

OYEYEMI AKINRELE *

*Stetson-Hatcher School of Business, Mercer University, USA.*

International Journal of Science and Research Archive, 2025, 17(02), 373-379

Publication history: Received on 30 September 2025; revised on 08 November 2025; accepted on 12 November 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.17.2.3013

## Abstract

AI-based decision-making in banking, particularly in lending, credit scoring, and fraud detection, raises concerns about systemic bias, fairness, and discrimination. This paper critically examines the intersection of algorithmic bias and U.S. anti-discrimination laws (Equal Credit Opportunity Act, Fair Housing Act). It proposes a hybrid framework that combines technical bias mitigation strategies with legal and ethical oversight. Through qualitative analysis of existing litigation and quantitative fairness audits of selected models, the paper identifies practical solutions to enhance equity in banking AI systems.

**Keywords:** Artificial Intelligent; Financial; US; Banking; Bias

## 1. Introduction

Currently, AI implementation in the financial sector is limited to specific applications such as customer service and process automation, but significant expansion is anticipated(Lawrence Damilare Oyeniyi et al., 2024). Respondents reported ongoing monitoring of market trends to identify emerging risks. The classification of risks is expected to evolve, as AI may amplify existing risks and create spillover effects across business areas, potentially generating new risks. The most frequently cited risk categories were cybersecurity, market manipulation, and bias and discrimination. Cybersecurity risks are linked to the vulnerability of AI systems to cyber-attacks, including adversarial attacks that target decision-making processes (Roylance, 2001). For example, generative AI may facilitate advanced phishing emails or deepfakes, increasing incidents of identity theft and fraud. Market manipulation risks are heightened by AI-driven herding effects, uninformed investment recommendations, hallucinations, and deepfakes(B. Singh et al., 2024). Market manipulation, discrimination, and privacy risks all share challenges related to data quality. The complexity and volume of financial data present difficulties in ensuring data quality, relevance, security, and confidentiality, which can lead to data poisoning and misappropriation of personal information. Poor governance was also frequently reported as a risk. Governance systems are complex, and the relationships between variables in AI models are often not transparent. The opaque or 'black box' nature of AI increases the risk of limited explainability and interpretability. While high autonomy in AI systems can improve operational efficiency, it also complicates the evaluation and critique of outcomes. AI algorithms often have sparse architectures with numerous parameters and may consist of ensembles of interacting models, making input signals difficult to specify or unknown. This complexity poses challenges for banks at both operational and reporting levels, hindering the detection of flaws.

A lack of explainability can prevent assessment of whether an AI approach is conceptually sound, as it is difficult to present decision-making processes to regulators, customers, and other stakeholders (de Bruijn et al., 2022). As AI models are increasingly used across business lines, insufficient understanding of their techniques and limitations can result in significant model risk, a concern frequently cited by survey respondents. Additional risks arise from financial institutions' reliance on third parties for AI services, which increases the likelihood of fraud and introduces new operational risks. Governments noted that greater transparency in banks' IT infrastructures and increased reliance on

[*] Corresponding author: OYEYEMI AKINRELE

third-party suppliers heighten third-party dependency risks (de Bruijn et al., 2022). These include limited knowledge and control over third parties, vendor lock-in, concentration risk, and challenges in model maintenance. Technical arrangements may also exacerbate risks of money laundering and fraud. In some cases, financial regulators lack supervisory authority over third parties, leading to untested risks, such as those associated with cloud service providers. Respondents also identified operational risks such as reliability issues, technical failures, system errors, and model drift, all of which can disrupt financial processes and contribute to financial loss and instability. Data privacy, discrimination, and model risk can occur at various stages of AI system interaction (Shahriar et al., 2023). Data quality issues at the input level can compromise system performance, while even high-quality data may be distorted during modeling, leading to model-related risks. The deployment of AI systems is often contingent on governance structures, as illustrated by the use of AI in credit decision processes, which can introduce additional risks. Low-quality data may result in biased datasets and models, increasing the risk of consumer discrimination, financial exclusion, and concentration in credit and investment. The use of chatbots and virtual assistants may exclude customers with limited technological access or digital literacy. Interviewees also noted that AI models may contain defects that deny service access without clear justification, a challenge exacerbated by the black box nature of these systems. Each of these risk domains can result in significant reputational, legal, and regulatory consequences for financial institutions. The deployment of flawed or biased AI systems may lead to data leaks, cyber-attacks, or fraud, undermining consumer trust. Other risks identified by respondents include auditability, competition, and copyright infringement. Some jurisdictions have highlighted risks stemming from the rapid development of AI, such as difficulties in recruiting highly qualified personnel. Approaches to addressing AI-specific financial risks vary among respondents, with some treating AI risk as a subset of broader risk categories rather than as a distinct category within the financial sector.

## 2. Regulatory Background, Anti-discrimination and Financial Compliance

A starting challenge for US compliance professionals is that there is no uniform definition of NFM (Non-Financial Misconduct) across the SEC (Securities and Exchange Commission), CFTC (Commodity Futures Trading Commission), and FINRA (Financial Industry Regulatory Authority). Unlike the UK FCA, which has explicitly defined NFM to encompass behavior like bullying, discrimination, and sexual harassment or misconduct, US regulators address the underlying behavior through other existing frameworks.

In the US, "workplace misconduct," "unethical conduct," "conduct not consistent with just and equitable principles of trade" (FINRA Rule 2010), "failure to supervise" (FINRA Rule 3110, CFTC Rule 166.3), or conduct impacting "fitness and propriety" (CFTC Part 3) are generally the functional concepts under which NFM-related issues are addressed.

Consequently, US regulators primarily tackle NFM by applying existing rules related to:

- Disclosure & Internal Controls (SEC): Establishing whether companies make disclosures regarding risks associated with their people and culture, and possess adequate controls to evaluate these risks.
- Supervision (FINRA, CFTC): Determining if companies adequately supervise staff activity to prevent rule violations and misconduct.
- General Conduct Standards (FINRA): Regarding whether conduct is up to "high standards of commercial honor and just and equitable principles of trade".
- Registrant Fitness (CFTC): Whether misconduct affects the fitness of an individual for registration.
- Whistleblower Protections (SEC, CFTC): Prohibiting companies from impeding potential violations reporting, e.g., misconduct.

One of the foundations of the SEC's plans is Exchange Act Rule 13a-15(a), where public companies are made to maintain disclosure controls and procedures (DCP). An example was the case of Activision Blizzard (an online gaming firm) in 2023: the firm was penalized $35 million by the SEC for failing to maintain adequate disclosure controls to collect and analyze worker complaints of workplace misconduct. Because the company did not have controls for measuring the severity and number of reported harassments, management did not have a way to know whether there were material issues that needed to be disclosed to investors. A SEC official cautioned that without such controls, firms "lack the means to determine whether larger issues existed that needed to be disclosed to investors.". Individually, the SEC charged the company with violating whistleblower protection laws by using separation agreements that could silence employees. The clear message is that in America or otherwise, significant cultural problems (and how they're handled within) can become securities law violations - e.g., if they reveal management's failure to offer proper internal controls or if investors are misled as to the causes of significant executives' resignations.

**Table 1** USA Financial AI Approach

| Feature | US Approach (SEC/CFTC/FINRA Composite) |
|---|---|
| Regulatory Framework | Indirect, fragmented, leverages existing rules. SEC: Exchange Act Rules 13a-15(a), 21F-17 |
| Key Rules Leveraged | CFTC: Rule 166.3, Part 3. FINRA: Rules 3110, 2010; Form U5. |
| Definition of NFM | Implicit; varies by regulator/context (workplace misconduct, failure to supervise, unethical conduct, impact on fitness) |
| Primary Focus | Impact of NFM on disclosure, supervision, market integrity, investor protection, and registrant fitness |
| Enforcement Mechanism | Actions for related rule breaches (e.g., disclosure control failure, supervisory lapse, impeding whistleblowers) |

## 3. Sources of Bias in Financial AI Models

Biases in financial AI systems are a consequence of several sources, which recursively affect one another in a complex way (Ferrara, 2024). Data-based biases represent a primary concern, where historical bias arises when training data reflects discriminatory past practices, such as redlining or differential treatment by demographic groups. Empirical studies have indicated that historical bias is a large problem in lending models, where several decades of well-documented discriminatory practices result in the encoding of issue-ridden patterns inherently within datasets (Bansal et al., 2023). Representation bias occurs when certain demographic groups are underrepresented in training sets, so models for these groups are inaccurate. Selection bias occurs from non-random sampling procedures that fail to capture the diversity of financial behaviors, particularly across underbanked individuals (Banking on Cooperation: Testing Evolutionary Theories of Human Cooperation via Microfinance Loans, 2023). Measurement bias occurs when methods of collecting data vary across population segments and therefore lead to inconsistent quality or depth of financial information (Liu, 2020). Model-induced biases add an additional layer to the complexity of fairness in finance AI. Algorithmic bias occurs when model structures subtly favor predominant trends in the data, essentially optimizing for majority groups at the expense of others (Cowgill et al., 2019). Research has shown that even with highly balanced training sets, some model structures have elevated error rates for minority groups. Proxy bias is a challenging issue in which models learn correlations between protected traits and seemingly neutral variables, such as zip code, education providers, or purchasing behaviors. Money models create intricate proxy mechanisms that can effectively reconstruct safeguarded features based on combinations of permitted variables (Essifi, n.d.). Feedback loop bias produces reinforcing loops when model predictions influence future data collection, gradually developing early bias through repeated use. Deployment biases occur when AI systems are implemented with users and environments. User interface bias occurs when financial applications are not equally accessible or usable to different demographics (Sarkar et al., 2016), which can restrict access to useful features or information. Interpretation bias occurs when human decision-makers in hybrid decision systems interpret model outputs differently based on customer profiles, normally because it continues to reinforce existing stereotypes or prejudices. System integration bias is a growing risk where banks and other financial organizations are deploying multiple AI systems in sequence to create interaction effects that have the potential to prolong customer journey biases.

## 4. Fairness audits and case studies of Unintentional Bias

Mortgage approval algorithms provide a glaring illustration of algorithmic bias in practice (Pappil Kothandapani, 2025). A 2022 investigation of mortgage approval algorithms used by five major U.S. banks discovered that Black and Hispanic applicants were 40-80% (BHUTTA et al., 2025), more likely to be denied loans than white applicants with similar qualifications. Detailed analysis revealed that the algorithms had been trained on historical lending data that reflected decades of discriminatory practices, with the impact of a self-reinforcing cycle. After controlling for all legitimate risk factors, the disparity in unexplained approvals persisted across geographic areas and income levels. The distribution of small business loans during the COVID-19 pandemic showed how algorithmic bias can be exacerbated by a crisis. AI systems used to triage Paycheck Protection Program (PPP) applications showed significant bias toward well-established firms in wealthier neighborhoods (Howell et al., 2022), at the expense of minority-owned businesses and those in poorer neighborhoods. The algorithms privileged long credit histories and extended banking relationships, dimensions that historically vary by socioeconomic status and race. Analysis of Public-Private Partnership (PPP) distribution patterns discovered that AI-driven.

Observation
Induction
EMPIRICAL RESEARCH METHODOLOGY CYCLE
Evaluation
Deduction
Testing

**Figure 1** Empirical Research Methodology Cycle

Loan origination systems approved loans to firms in white-majority zip codes at nearly twice the rate of firms in majority-minority neighborhoods with identical economic profiles (Schweitzer & Guo, 2024). Automated wealth management systems have also been found to exhibit concerning gender and age discrimination patterns. Analysis of robo-advisor recommendations revealed that investment advice provided varied significantly by perceived age and gender, with more conservative investment advice being provided to younger women despite the same risk tolerance and investment objectives. When researchers created test profiles with identical financial characteristics but differing demographic indicators, they found consistent differences in asset allocation advice that would result in significantly lower predicted returns for women investors over extended time periods. These trends were observed on numerous automated advisory sites with consistency.

## 5. Adversarial Testing for Bias Detection

Adversarial testing has also been a successful technique for revealing hidden biases in money-related AI systems (Gowda & Gowda, 2024) . The method involves testing models in a systematic way with specifically designed inputs that aim to produce differential treatment based on demographics. The model supports three levels of adversarial testing (L. K. Singh & Khanna, 2023): Counterfactual Input Testing Developing (CITD) matched inputs identical in every aspect except in protected features (e.g., the same mortgage application with varying applicant names suggesting different ethnicities) allows for direct comparison of model output. Research indicates that doing so can reveal distinguishable patterns of bias hidden in aggregate metrics. Measuring output divergence provides quantitative evidence of disparate treatment, whereas identifying decision thresholds at which bias increases facilitates triaging remediation efforts. Fairness Attack Vectors Designing (FAVD) targeted prompts requesting biased assumptions in LLMs is an improvement over traditional fairness testing.

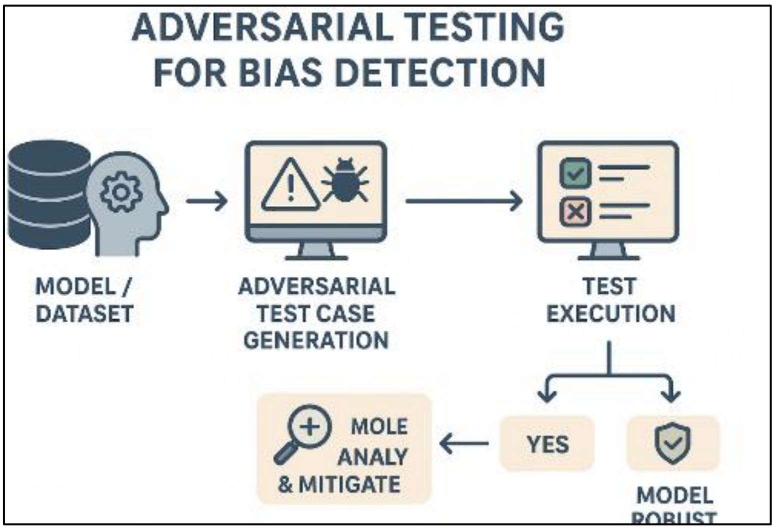
ADVERSARIAL TESTING FOR BIAS DETECTION
MODEL / DATASET
ADVERSARIAL TEST CASE GENERATION
TEST EXECUTION
MOLE ANALY & MITIGATE
YES
MODEL ROBUST

**Figure 2** Adversarial Testing for Bias Detection

Science shows that carefully designed contexts can show deep-seated bias even in models that pass as fair under standard testing, as shown in Figure 2. Testing against ambiguous financial situations and evaluating responses to intersectional identities helps identify complex expressions of bias that would otherwise go undetected. Temporal Stability Testing is all about asking questions about models with historical patterns known to be correlated with discriminatory actions, which shows how well fairness interventions are maintained across time. Finance-specific research demonstrates that temporal testing can identify models with "fairness decay" as data distributions evolve. This approach helps organizations to predict and prevent bias recurrence rather than responding only after biased events have occurred. In one example, the adversarial test suite for a major credit-scoring Large Language Model (LLM) showed that gender bias was strongest in applicants with credit scores from 680 to 720—precisely the "borderline" decision zone where human judgment long had the largest impact.

## 6. Implications for U.S. banks, regulators, and customers

The United States (U.S.) employs a dual banking system in which banks may be chartered at either the state or federal level (Sykes, 2018), and which facilitates a pluralistic environment of large, global banks and small, regional banks.

As U.S. banks continue to navigate through a complex and evolving regulatory landscape necessitated by economic pressures (Joshi, 2025), shifting enforcement priorities, and emerging financial technology, a recent but now waning climate of rising interest rates has made bank profitability more robust. Banks currently also have greater funding costs, credit exposures, and liquidity pressures. Notable bank failures in 2023 pinned stricter regulatory focus on capital adequacy, risk, and liquidity planning, fueling current controversy over whether regulators should tighten prudential requirements for mid-sized banks.

Additionally, the new U.S. presidential government's change will bring regulatory priority changes. Policymakers signaled a rollback of some consumer protection policies and more pro-business financial regulation (Nofsinger, 2012), in particular, on capital requirements, fintech partnerships, and digital assets. At the same time, there will be continued monitoring of systemic risk, anti-money laundering (AML) compliance, and future technology risks. In particular, with the world pressure on more financial crime enforcement from the world regulatory bodies, banks and financial institutions with international operations will keep analyzing such regulatory changes.

## 7. Conclusion

Governance requirements for banks in the U.S. are shaped by a combination of regulatory mandates, supervisory expectations, and best practices issued by federal financial regulators, including the OCC (Office of the Comptroller of the Currency), FDIC (Federal Deposit Insurance Corporation), and FRB (Federal Reserve Board). The regulatory framework emphasizes board oversight, risk governance, internal controls, and accountability mechanisms to ensure financial stability and resilience. The OCC has issued heightened standards for bank governance, primarily through its OCC Guidelines, which apply to insured national banks, federal savings associations, and federal branches of foreign banks with $50 billion or more in average total consolidated assets. These guidelines require institutions to maintain a written risk governance framework that delegates authority from the board of directors to management committees and executives, and is updated routinely. The OCC also requires independent risk management to oversee risk-taking activities and to design a risk governance framework.

Under the OCC Guidelines, banks are required to have front-line units responsible for assessing and managing risk associated with their activities. The guidelines also require that at least two members of a covered bank’s board of directors be independent, and the board must receive ongoing training covering risks that could impact the bank.

The FDIC has introduced, but not yet finalized, similar requirements (the FDIC Guidelines) applicable to FDIC-supervised institutions with $10 billion or more in total consolidated assets. If finalized, the FDIC Guidelines would require a majority of independent directors on the board, with specific responsibilities for risk oversight, governance, and compliance. These guidelines would also mandate the formation of a dedicated risk committee chaired by an independent director to oversee risk management.

The role of bank boards has come under heightened scrutiny following a wave of bank failures in 2023. Regulatory investigations concluded that board-level failures in risk management oversight contributed to these banks’ vulnerabilities, particularly regarding liquidity risk and capital management. For example, the FRB’s post-mortem analysis of SVB highlighted the board’s failure to appreciate risks associated with the bank’s high level of uninsured deposits and the absence of risk management metrics in executive compensation structures as contributing factors to its collapse. These developments have led to increased regulatory emphasis on board accountability, particularly in relation to risk governance, internal controls, and executive compensation.

## References

[1] BANKING ON COOPERATION: TESTING EVOLUTIONARY THEORIES OF HUMAN COOPERATION VIA MICROFINANCE LOANS. (2023).

[2] Bansal, C., Pandey, K. K., Goel, R., Sharma, A., & Jangirala, S. (2023). Artificial intelligence (AI) bias impacts: classification framework for effective mitigation. Issues in Information Systems, 24(4), 367–389. https://doi.org/10.48009/4_iis_2023_128

[3] BHUTTA, N., HIZMO, A., & RINGO, D. (2025). How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions. The Journal of Finance, 80(3), 1463–1496. https://doi.org/10.1111/jofi.13444

[4] Cowgill, B., Tucker, C., Blei, D., Brown, C., Corbett-Davies, S., Dell'acqua, F., Impink, M., Horton, R., Kahneman, D., Kogut, B., Lipton, Z., Morgan, J., Miller, A., Seamans, R., Shelef, O., Sibony, O., Stevenson, M., Thompson, N., Yeomans, M., & Zhou, A. (2019). Economics, Fairness and Algorithmic Bias *. http://www.aies-conference.com/

[5] de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, 39(2), 101666. https://doi.org/10.1016/j.giq.2021.101666

[6] Essifi, Y. (n.d.). Historical Reconstruction of Credit Spreads and Cross-Currency Basis : A Comparative Analysis of Machine Learning Techniques and Ensemble Models; Historical Reconstruction of Credit Spreads and Cross-Currency Basis : A Comparative Analysis of Machine Learning Techniques and Ensemble Models; Historisk rekonstruktion av kreditspreadar och cross-currency basis – En jämförande analys av tekniker för maskininlärning och ensemblemodeller: Vol. TRITA – EECS-EX.

[7] Ferrara, E. (2024). The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. Machine Learning with Applications, 15, 100525. https://doi.org/10.1016/j.mlwa.2024.100525

[8] Gowda, P., & Gowda, A. N. (n.d.). Benefits and Risks of Generative AI in FinTech. Journal of Scientific and Engineering Research, 2024(5), 267–275. www.jsaer.com

[9] Howell, S. T., Kuchler, T., Snitkof, D., Stroebel, J., Wong, J., Anbil, R., Arora, V., Bala, B., Blake, K., Chandler, C., Dobridge, K., Mack, K., Mills, D., Musto, H., Oudghiri, M., Ross, K., & Ruppel, S. T. (2022). Lender Automation and Racial Disparities in Credit Access.

[10] Joshi, V. . C. (2025). The Journey Ahead. In Changing Dimensions of Financial Services and Banking Regulation (pp. 185–193). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-5443-7_11

[11] Lawrence Damilare Oyeniyi, Chinonye Esther Ugochukwu, & Noluthando Zamanjomane Mhlongo. (2024). Implementing AI in banking customer service: A review of current trends and future applications. International Journal of Science and Research Archive, 11(2), 1492–1509. https://doi.org/10.30574/ijsra.2024.11.2.0639

[12] Liu, G. (2020). Data quality problems troubling business and financial researchers: A literature review and synthetic analysis. Journal of Business & Finance Librarianship, 25(3–4), 315–371. https://doi.org/10.1080/08963568.2020.1847555

[13] Nofsinger, J. R. (2012). Household behavior and boom/bust cycles. Journal of Financial Stability, 8(3), 161–173. https://doi.org/10.1016/j.jfs.2011.05.004

[14] Pappil Kothandapani, H. (2025). Social Implications of Algorithmic Decision-Making in Housing Finance: Examining the broader social impacts of deploying machine learning in lending decisions, including potential disparities and community effects. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (Online), 4(1), 78–97. https://doi.org/10.60087/jklst.v4.n1.009

[15] Roylance, D. (2001). PRESSURE VESSELS.

[16] Sarkar, U., Gourley, G. I., Lyles, C. R., Tieu, L., Clarity, C., Newmark, L., Singh, K., & Bates, D. W. (2016). Usability of Commercially Available Mobile Applications for Diverse Patients. Journal of General Internal Medicine, 31(12), 1417–1426. https://doi.org/10.1007/s11606-016-3771-6

[17] Schweitzer, M., & Guo, A. (2024). Basic facts on the coverage of the paycheck protection program. Business Economics, 59(1), 10–30. https://doi.org/10.1057/s11369-023-00345-z

[18] Shahriar, S., Allana, S., Hazratifard, S. M., & Dara, R. (2023). A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. IEEE Access, 11, 61829–61854. https://doi.org/10.1109/ACCESS.2023.3287195

[19] Singh, B., Kaunert, C., & Gautam, R. (2024). Artificial Intelligence in Detecting Herding and Market Overreaction (pp. 1–22). https://doi.org/10.4018/979-8-3693-7827-4.ch001

[20] Singh, L. K., & Khanna, M. (2023). Introduction to artificial intelligence and current trends. In Innovations in Artificial Intelligence and Human-Computer Interaction in the Digital Era (pp. 31–66). Elsevier. https://doi.org/10.1016/B978-0-323-99891-8.00001-2

[21] Sykes, J. B. (2018). Banking Law: An Overview of Federal Preemption in the Dual Banking System. www.crs.gov