Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

Fairness, Accountability, and Transparency in Financial AI: Addressing Bias through Responsible Regulation and Auditable Design

Mohammad Talha Siddiqui[1], Sumit Kumar[2], Shashi Bharti[3], Yusairah Ahmad[4], Mohd Suhail[5]

[1]University of Lucknow, Lucknow-226031, India.

[2]University of Lucknow, Lucknow-226031, India.

[3]University of Lucknow, Lucknow-226031, India.

[4]Khwaja Moinuddin Chisti Language University, Lucknow, India.

[5]BIM Manager, Egis, Tabuk, Kingdom of Saudi Arabia

Abstract

AI systems increasingly shape decision-making in financial services, particularly in credit scoring, insurance underwriting, and fraud detection. While these systems improve efficiency and predictive accuracy, they may also reproduce and amplify existing social and economic inequalities. This paper examines how fairness, accountability, and transparency (FAT) principles can be systematically integrated into financial AI systems. Using a conceptual and critical approach, the study synthesises literature on algorithmic bias, analyses the regulatory frameworks of the EU AI Act, GDPR, and US fair lending law, and incorporates fairness impossibility results from Chouldechova (2017) and Kleinberg et al. (2017). The analysis identifies key sources of bias, including data imbalance, proxy variables, measurement error, label contamination, and feedback loops. It further highlights that fairness metrics such as demographic parity, predictive parity, and equalised odds cannot be satisfied simultaneously in many contexts, creating unavoidable trade-offs for regulators and practitioners. To address these challenges, the paper proposes a FAT Lifecycle Framework covering six stages of AI development and governance. The framework offers practical guidance for organisations and regulators seeking to operationalise FAT principles across the full AI lifecycle while supporting equitable access to financial services.

Keywords: Financial Artificial Intelligence; Algorithmic Fairness; Accountability in AI Systems; Transparency and Explainability; Ethical AI Governance; Bias Mitigation in FinTech; Fairness Impossibility; Regulatory Compliance.

1. Introduction

Financial institutions have reshaped how they assess risk, approve loans, and deliver products, and AI is the engine behind much of that transformation. These systems can process far more data than human decision-makers and generate outputs faster. The catch is that this growing reliance on AI has also raised serious concerns about fairness, accountability, and transparency in automated decision-making. Together, these three principles, often abbreviated as FAT, provide a framework for evaluating whether AI systems treat individuals equitably, operate under clear responsibility structures, and remain open to scrutiny (Oyasiji et al., 2023).

Algorithmic bias in financial AI shows up across credit scoring, mortgage approval, insurance underwriting, and fraud detection. Studies have shown that these systems can reproduce patterns of discrimination already baked into historical data, often disadvantaging groups protected under anti-discrimination law (Bahangulu & Owusu-Berko, 2025). Individuals from marginalised backgrounds may face lower credit limits, higher premiums, or outright denials of service, not because of their financial behaviour but because the training data reflects past social inequalities (Thiruma Valavan, 2023). The bias is inherited, not invented.

Scale changes the stakes. A biased algorithm applies discriminatory patterns across millions of consumers, producing systemic effects far more widespread than traditional human bias (Mensah, 2023). As AI becomes more deeply embedded in financial decision-making, these biases gain permanence. Addressing them has therefore become both a regulatory and an ethical priority.

Legal and policy frameworks have begun to respond. The EU's General Data Protection Regulation introduced early provisions for transparency and accountability in automated decision-making, including the 'right to explanation' under Article 22, though scholars debate the practical scope of this right (Wachter et al., 2018; Metikoš & Ausloos, 2025). More recently, the EU AI Act has established detailed obligations for high-risk AI systems, which cover most financial applications (Hacker & Passoth, 2022; Juliussen, 2025). In the United States, the Equal Credit Opportunity Act and the Fair Housing Act are increasingly being applied to algorithmic credit decisions (Kumar et al., 2022; Wu, 2024).

A challenge that receives far less attention than it deserves is that fairness itself is mathematically constrained. Chouldechova (2017) demonstrated that when base rates differ across groups, it is impossible to simultaneously satisfy calibration, predictive parity, and equal false-positive rates, a result extended by Kleinberg et al. (2017). These impossibility theorems mean that every fairness intervention involves trade-offs, and every regulatory compliance framework must implicitly choose which conception of fairness to prioritise. We treat these constraints as central to the discussion, not peripheral.

Our aim in this paper is to contribute to responsible financial AI by connecting three perspectives: technical approaches to bias reduction, legal and regulatory frameworks, and institutional governance. We argue that achieving fairness requires an integrated approach, one that combines technical design, ethical reflection, and policy oversight, and we propose an original FAT Lifecycle Framework to operationalise this integration.

2. Algorithmic Bias In Financial Ai Systems

Bias in financial AI springs from both technical limitations and the social inequalities already embedded in data. Understanding how these biases develop is a prerequisite for effective mitigation. Research shows that bias enters financial algorithms through several pathways, including data imbalance, feature selection, label contamination, and feedback effects (Trinh & Zhang, 2024).

2.1 Mechanisms of Bias

Data imbalance is among the most common sources of bias. Datasets used to train AI models often carry uneven demographic representation. Algorithms thus perform better for majority populations and less accurately for underrepresented ones, producing higher error rates for certain applicants even when sensitive variables are formally excluded from the model (Garcia et al., 2024). The disparity is structural, not incidental.

Feature selection creates additional risk. Some variables appear neutral but act as proxies for protected characteristics. ZIP codes, schools attended, or employment history can correlate strongly with race, gender, or socioeconomic status. Because these variables often appear predictive, their discriminatory effects can evade standard fairness tests (Bhutta et al., 2025; Datta et al., 2017). The proxy problem is difficult to detect and even harder to eliminate without sacrificing model accuracy.

Feedback loops transform temporary bias into structural inequality. When algorithms deny credit to certain groups, those individuals lose opportunities to build positive credit histories. Over time, the data reinforces statistical patterns that suggest higher risk, producing continued exclusion. Hurlin et al. (2026) and Wyllie et al. (2024) have shown that these dynamics are self-amplifying, particularly when models are retrained on outputs that already reflect prior biased decisions. The loop closes on itself.

Label bias develops when outcome variables reflect historically discriminatory human decisions. If past loan approvals were influenced by biased judgment, algorithms trained on these outcomes learn to replicate the same discriminatory patterns (Rizzi et al., 2021). Measurement bias further distorts outcomes: credit utilisation patterns driven by limited financial resources may be misclassified as indicators of poor credit management,

penalising constrained but responsible borrowers (Wyllie et al., 2024). What looks like a neutral data point is anything but.

2.2 The Fairness Impossibility Problem

Any serious treatment of algorithmic fairness must confront an uncomfortable fact: multiple intuitively reasonable fairness criteria are mathematically incompatible. Chouldechova (2017) proved that when a risk score is well-calibrated and base rates differ across demographic groups, it is impossible to simultaneously achieve equal false-positive rates and equal false-negative rates between groups. Kleinberg et al. (2017) extended this analysis to show that calibration and balance conditions cannot be jointly satisfied except in degenerate cases. The arithmetic is unforgiving.

These impossibility results have direct practical implications for financial AI. A credit model cannot simultaneously achieve demographic parity, predictive parity, and equalised odds when the underlying default rates differ, which they typically do, owing to structural economic inequalities. Practitioners and regulators must therefore make explicit choices about which fairness criterion to prioritise, and must acknowledge the trade-offs those choices entail. A regulatory framework that simply mandates 'fairness' without specifying the operative conception risks producing compliance theatre rather than substantive equity. We return to this point repeatedly because it shapes every subsequent argument in the paper.

2.3 Consequences for Marginalised Groups

The Apple Card controversy of 2019 laid bare how algorithmic systems can reproduce long-standing patterns of discrimination while appearing objective. Multiple customers reported that credit limits assigned by the Goldman Sachs-issued Apple Card were substantially lower for women than for men with comparable financial profiles. Although gender was not an explicit model input, indirect variables such as employment history and spending patterns appear to have functioned as proxies (Kelley et al., 2022). The New York Department of Financial Services subsequently investigated and did not find a violation of existing law, illustrating a critical gap: current legal frameworks may be insufficient to address indirect algorithmic discrimination even when disparate outcomes are observable. The case exposes both the mechanism of proxy discrimination and the limitations of enforcement tools designed for intentional, direct discrimination.

The consequences extend well beyond individual transactions. Systematic bias in credit scoring, lending, or insurance decisions contributes to broader economic inequality, limiting the ability of marginalised groups to build wealth and achieve financial stability (Kelley et al., 2022). West et al. (2019) describe these dynamics as a modern infrastructure of economic exclusion, noting that AI-mediated discrimination operates at a scale and speed that human bias simply cannot match. What begins as a modelling decision ends as a structural disadvantage.

Fraud detection systems produce their own unfair outcomes. Differences in spending or saving behaviour between demographic groups may be misinterpreted as suspicious activity, resulting in elevated false-fraud-alert rates for certain customers even when their behaviour is legitimate (Pagan et al., 2023). Prince and Schwarcz (2020) document how proxy discrimination in insurance and credit scoring produces systematically adverse outcomes for minority communities across multiple financial product categories. The pattern is consistent, and it is troubling.

3. Regulatory And Legal Context

3.1 European Union Frameworks

The EU AI Act, which entered into force on 1 August 2024, introduces a risk-based regulatory architecture. It classifies AI systems according to potential harm and imposes strict requirements on high-risk applications. Financial systems used for credit scoring, loan approval, and insurance underwriting fall within this high-risk category and must comply with requirements for transparency, human oversight, and bias monitoring (Hacker & Passoth, 2022). Applicability is phased: prohibitions on unacceptable-risk AI became operative in February 2025,

while obligations for high-risk systems apply from August 2026. This transition period creates compliance uncertainty that institutions must manage proactively.

The EU AI Act's requirement for conformity assessments and continuous post-market monitoring represents a meaningful advance over prior sector-specific rules. Several boundary cases remain contested, though. Fraud detection tools, credit pre-screening algorithms, and risk-scoring models used for marketing segmentation may or may not meet the threshold for 'high-risk' classification depending on implementation context, a question that regulators have yet to resolve definitively (Nisevic et al., 2024). The boundaries are porous, and the stakes are high.

The GDPR complements the AI Act through Article 22, which grants individuals the right to meaningful information about the logic behind automated decisions that significantly affect them. There is substantial scholarly disagreement about the scope of this right. Wachter et al. (2018) argued that Article 22 does not create a legally enforceable right to a full algorithmic explanation, but rather a right to information about the existence of automated decision-making. Metikoš and Ausloos (2025) and Juliussen (2025) analyse how emerging case law under both the GDPR and the AI Act is gradually clarifying these obligations, though significant ambiguity persists. The law is catching up with the technology; it just has not caught up yet.

3.2 United States Frameworks

In the United States, fair lending laws, principally the Equal Credit Opportunity Act and the Fair Housing Act, prohibit discrimination in lending based on protected characteristics, and recognise both disparate treatment, that is, intentional discrimination, and disparate impact, meaning neutral policies producing discriminatory outcomes. As AI takes on more decision-making functions, these legal frameworks are being applied to evaluate algorithmic systems (Kumar et al., 2022; Wu, 2024).

The US regulatory approach differs from the European model in an important respect: it is primarily enforcement-oriented rather than ex ante design-oriented. Institutions are not required to obtain pre-deployment approval for credit algorithms, and regulatory intervention typically follows identified harm rather than anticipating it. Wu (2024) argues that this creates asymmetric incentives that may allow biased systems to persist until litigation or regulatory action compels correction. The Consumer Financial Protection Bureau has issued guidance on adverse action notices for AI-driven credit decisions, but sweeping legislation comparable to the EU AI Act has not been enacted. The system waits for harm to surface before it acts.

3.3 Cross-Jurisdictional Compliance Challenges

Financial institutions operating across jurisdictions face substantial compliance complexity. European regulation emphasises proactive bias prevention and design-stage transparency, while US approaches focus on enforcement after discrimination has occurred. The coexistence of these models creates conflicting obligations for multinational firms (Juliussen, 2025). The EU AI Act's requirement for detailed documentation of model design and bias testing, for instance, may conflict with US intellectual property protections that treat model architecture as proprietary. Navigating both regimes simultaneously is no small task.

Regulatory Technology solutions have emerged to help manage compliance obligations. These systems automate monitoring, generate audit trails, and assist with regulatory reporting (Akinwumi et al., 2021; Kothandapani, 2024). Yet they raise a reflexive question that the literature has not fully addressed: how are the compliance systems themselves made transparent and accountable? A RegTech tool that monitors bias in a credit model is itself an algorithmic decision-support system subject to the same fairness concerns it is designed to address (Olaiya et al., 2024; Abikoye et al., 2024). The watchdog needs watching.

3.4 Existing Professional Standards

Beyond legislation, professional standards and guidelines shape organisational practice. The EU AI Act requires institutions to follow ISO/IEC 23053:2022 for managing bias across the AI lifecycle. The IEEE and ACM have issued ethical guidelines emphasising FAT principles in AI development, providing an operational translation of abstract ethical commitments (Metikoš & Ausloos, 2025). The Brookings Institution's AI fair

lending policy agenda (Akinwumi et al., 2021) offers a sector-specific treatment of how regulatory instruments can be calibrated to address algorithmic discrimination in consumer credit markets. While non-binding, these frameworks shape organisational norms and inform regulatory interpretation, particularly where legislation is ambiguous. They fill gaps, but they do not close them.

3.5 Compliance Challenges

Technical interpretability remains a primary obstacle to regulatory compliance. Many advanced AI systems, particularly those based on deep learning or ensemble models, operate as black boxes whose internal logic resists human interpretation (Nisevic et al., 2024). Regulations increasingly call for explainable AI, but meaningful explanation of complex models is technically difficult and contextually dependent: what constitutes a satisfactory explanation varies between regulators, developers, and affected consumers. One size does not fit all.

Proprietary protections create additional friction. Financial institutions frequently rely on third-party algorithms protected by intellectual property agreements, which can prevent full regulatory disclosure even when transparency is legally required (Hacker & Passoth, 2022). Smaller institutions face resource constraints in developing and maintaining fairness auditing infrastructure, creating competitive asymmetries that may reduce market diversity (Metikoš & Ausloos, 2025). The burden of compliance falls unevenly, and the smallest players bear the greatest relative cost.

4. An Integrated Fat Lifecycle Framework For Financial Ai

The literature reviewed above, covering technical bias mechanisms, fairness impossibility results, and regulatory obligations, points toward a clear conclusion: piecemeal interventions are insufficient. Bias can enter at any stage of a system's lifecycle, and no single technical fix or regulatory requirement addresses all its manifestations. We propose an integrated FAT Lifecycle Framework that maps fairness, accountability, and transparency obligations onto six sequential stages of AI system development. Table 1 below present this framework.

The framework's central premise, informed by the impossibility theorems discussed in Section 2.2, is that institutions must make explicit choices about which fairness criteria to prioritise at each stage, document those choices transparently, and subject them to governance review. Our framework operationalises this requirement by specifying, for each lifecycle stage, which FAT principle is primary, what actions are required, what tools are available, and which regulatory provisions apply.

Table 1. FAT Lifecycle Framework for Financial AI

Stage	Principle	Key Actions	Tools / Methods	Regulatory Alignment
Data Collection	Fairness	Representational sampling; remove discriminatory historical labels	Reweighting, synthetic data generation, data cards	GDPR Art. 5 (data quality); EU AI Act Annex IV
Model Development	Fairness + Transparency	Embed fairness metrics alongside accuracy; adversarial training; bias testing throughout	Regularisation, adversarial debiasing, SHAP feature analysis	EU AI Act Art. 10–15 (high-risk AI requirements)
Validation & Testing	Fairness + Accountability	Multi-metric fairness evaluation; demographic parity	Chouldechova (2017) impossibility tests; counterfactual fairness checks	ECOA adverse action notice; ISO/IEC 23053:2022

		vs. equal opportunity trade-off analysis		
Deployment	Transparency	Model Cards; System Cards; explainable outputs for affected individuals	LIME, SHAP, counterfactual explanations	GDPR Art. 22 (right to explanation); EU AI Act Art. 13–14
Post-Deployment Monitoring	Accountability + Fairness	Continuous bias monitoring; drift detection; independent audits; consumer feedback loops	Automated monitoring dashboards, audit trails, RegTech tools	EU AI Act Art. 9 (risk management); CFPB enforcement guidance
Governance	Accountability	Assign institutional fairness ownership; escalation procedures; stakeholder engagement	Algorithmic impact assessments, Model Cards, bias incident reporting	EU AI Act Art. 16–29 (obligations for deployers); national fair lending laws

4.1 Data Collection: The Foundation of Fairness

Data quality and representational balance are preconditions for any subsequent fairness intervention. Data used to train financial AI must reflect the diversity of the populations the system will affect. We must identify historical patterns of discrimination that distort training data, such as the underrepresentation of women and minority groups in historical credit datasets resulting from prior exclusionary lending practices (Rizzi et al., 2021; Prince & Schwarcz, 2020). Datasheets for datasets (Pushkarna et al., 2022) provide a standardised mechanism for documenting collection practices, representational limitations, and known biases, enabling downstream developers and auditors to assess data suitability before model training begins. Without clean foundations, everything built on top is suspect.

4.2 Model Development: Embedding Fairness by Design

During model development, fairness considerations must be woven into design choices and performance objectives from the outset. The fairness-by-design approach (Oguntibeju, 2024) treats ethical and social considerations as core components of the AI lifecycle rather than afterthoughts. We argue that selecting fairness metrics which reflect the institution's ethical and regulatory obligations, and embedding bias testing throughout the training process rather than applying it only at validation, is not optional. It is the bare minimum.

Pre-processing techniques correct bias before training begins, including statistical reweighting, demographic resampling, and the removal of proxy variables (Petersen et al., 2021). In-processing methods embed fairness directly into the learning algorithm through regularisation, adversarial training, or constrained optimisation (Zeng et al., 2024). Post-processing approaches adjust outputs after training through threshold calibration and output modification (Lohia et al., 2019). In practice, hybrid strategies combining all three are most effective, as bias can arise at multiple points and no single method is sufficient (Petersen et al., 2021). One technique alone will not do the job.

The choice of fairness metric cannot be made purely on technical grounds. The impossibility results of Chouldechova (2017) and Kleinberg et al. (2017) require institutions to make explicit, documented decisions about which fairness criterion, whether demographic parity, predictive parity, or equalised odds, is most appropriate for the application context and the affected populations. These decisions should be reviewed by both technical and non-technical stakeholders and recorded in model documentation. Leaving the choice implicit is itself a decision, and a consequential one.

4.3 Validation and Testing

Fairness validation requires evaluating models against multiple metrics simultaneously and documenting the trade-offs. Because different metrics can produce conflicting results, validation must consider both the statistical properties of the model and the values of affected stakeholders (Zeng et al., 2024). Counterfactual fairness tests, which assess whether the model's predictions would change if a protected attribute were altered, complement distributional fairness metrics and provide more granular insight into individual-level discrimination (Karimi et al., 2021).

ISO/IEC 23053:2022 provides structured methods for bias identification and mitigation across the AI lifecycle, aligning technical validation with regulatory expectations. Institutions should document validation procedures, fairness metrics used, trade-offs acknowledged, and decisions made, in compliance with EU AI Act Annex IV requirements for high-risk systems. The paper trail matters, both for accountability and for legal defensibility.

4.4 Deployment: Transparency and Explainability

Transparency at the deployment stage requires that affected individuals, regulators, and internal oversight functions can understand how automated decisions are made. Local Interpretable Model-agnostic Explanations provide case-by-case insights into specific decisions by approximating the local behaviour of complex models. In loan approval systems, LIME can identify which factors most strongly influenced an individual decision, supporting both consumer transparency and compliance officer review (Černevičienė & Kabašinskas, 2024).

SHapley Additive exPlanations assign importance values to each input feature, showing how much each contributed to a prediction. Comparing SHAP values across demographic groups enables systematic detection of patterns suggesting unequal treatment (Rane et al., 2023). Counterfactual explanations describe what changes would have led to a different outcome, informing an applicant, for instance, that reducing their debt-to-income ratio would improve loan eligibility. Such explanations support transparency and consumer agency; they must be designed carefully, though, to ensure the suggested changes are realistic and do not reinforce biased expectations (Karimi et al., 2021; Verma et al., 2024). A counterfactual that is technically correct but practically impossible serves no one.

Model Cards (Pushkarna et al., 2022) provide standardised documentation of a model's purpose, data sources, performance metrics, and known limitations. System Cards (Golpayegani et al., 2024) extend this to the full AI system, documenting architecture, governance processes, and operational conditions. Both tools support communication with regulators and the public while meeting EU AI Act transparency requirements. Different audiences require different types of explanation: regulators need audit documentation; developers need technical metrics; consumers need clear, actionable explanations of decisions that affect them (Verma et al., 2024). One document cannot serve all three audiences equally well.

4.5 Post-Deployment Monitoring

Deployment is not the end of the fairness lifecycle. Economic conditions, demographic shifts, and changes in user behaviour can alter model performance over time, introducing new forms of bias even when initial validation showed fair results (Rane et al., 2023). Continuous monitoring systems must track fairness metrics alongside accuracy metrics, with automated alerts when disparate impact thresholds are breached. Wyllie et al. (2024) demonstrate that models retrained on synthetic or biased outputs can amplify existing inequalities, a risk that post-deployment monitoring must specifically address. Vigilance does not expire.

Detailed audit trails document how data was collected, processed, and used; they capture system inputs, model versions, parameter updates, and deployment changes, enabling reconstruction of decisions when disputes arise (Raji et al., 2020). The EU AI Act's Art. 9 requirements for continuous risk management, and the CFPB's adverse action guidance, both presuppose monitoring infrastructure of this kind. Without it, accountability collapses.

4.6 Governance: Accountability Structures

Effective integration of FAT principles depends on institutional governance that assigns clear responsibility for fairness outcomes. Raji et al. (2020) argue that accountability gaps persist not because organisations lack ethical intent but because responsibility is diffuse, with no individual or team owning fairness outcomes across the full lifecycle. Their end-to-end framework for internal algorithmic auditing proposes assigning explicit governance ownership at each lifecycle stage, with escalation procedures and documented handoffs between technical, legal, and business functions.

Engaging stakeholders and affected communities is equally important. Pushkarna et al. (2022) and Huang et al. (2021) argue that technical audits alone are insufficient: consumer and advocacy group perspectives can surface forms of bias that standard fairness metrics miss. Institutions should create structured mechanisms for community engagement, with findings feeding back into model governance. Building a culture of continuous improvement, learning from past errors and adapting to new research, regulation, and public expectations, requires leadership commitment and should be embedded in institutional governance rather than delegated to compliance functions (Kumar et al., 2022; Wu, 2024). The job of fairness is never done.

5. Toward A Responsible Financial Ai Ecosystem

Building a responsible AI ecosystem in the financial sector requires collaboration among financial institutions, regulators, technology providers, and civil society (Kothandapani, 2024). Each actor plays a distinct role, and the ecosystem fails if any one of them defaults on its responsibilities. No single party can carry the weight alone.

Financial institutions wield the most direct influence because they deploy AI systems in real-world decision-making. They must integrate fairness goals into business strategy, risk management, and operational processes, and treat algorithmic bias risk with the same seriousness as traditional credit, market, and operational risks (Abikoye et al., 2024; Olaiya et al., 2024). Fairness investments may not yield immediate returns. They do, however, strengthen credibility, reduce regulatory risk, and contribute to long-term sustainability.

Regulators must balance innovation promotion with consumer protection. Oversight frameworks should encourage experimentation while setting clear limits. Sandbox environments that allow supervised testing of new systems can promote learning while managing risk (Abikoye et al., 2024). Civil society and consumer advocacy organisations, exemplified by Butler and O'Brien's (2019) analysis of RegTech governance, monitor social effects and represent affected communities, ensuring that fairness definitions reflect lived experience rather than purely technical interpretations. The voice of those affected must be part of the conversation.

Technology providers share responsibility for embedding fairness in their tools. Designing products with fairness considerations from the outset helps downstream financial institutions meet ethical and legal standards more effectively. A key market-level challenge is that fairness investments yield broader social benefits not fully captured by individual firms; consistent regulatory enforcement across competitors is therefore essential to prevent a race to the bottom in which cost-cutting on fairness becomes a competitive strategy. The economics of fairness are collective, and the rules must reflect that.

5.1 Balancing Innovation and Oversight

The relationship between innovation and regulatory oversight in financial AI is frequently framed as a tension. We think that framing is misleading. When guided by sound regulation and ethical design, innovation can support greater fairness rather than threaten it. The challenge is ensuring that flexibility does not become a loophole.

Regulators must avoid rigid, prescriptive frameworks that stifle experimentation or entrench incumbents. Principle-based regulation, specifying outcomes rather than methods, allows institutions to adapt technical approaches while maintaining consistent fairness standards (Olaiya et al., 2024). Institutions must integrate fairness risk into existing enterprise risk frameworks, recognising that reputational, legal, and operational risks

are interdependent with fairness performance (Kothandapani, 2024). The absence of international regulatory co-ordination is a significant structural gap. Multinational financial institutions must contend with divergent frameworks across jurisdictions, creating compliance overhead and, in some cases, incentives to locate AI development in less-regulated environments (O'Neil et al., 2024). Greater international alignment, potentially modelled on data protection co-ordination under the GDPR, would reduce these arbitrage opportunities and raise the global floor for fairness standards.

6. Future Research And Policy Directions

Several research and policy gaps deserve attention. First, while the impossibility theorems of Chouldechova (2017) and Kleinberg et al. (2017) establish that fairness trade-offs are unavoidable, the literature lacks empirical documentation of how specific institutions manage these trade-offs in practice. Qualitative and quantitative research on actual fairness decision-making processes in financial institutions would substantially advance understanding. We simply do not know enough about what happens inside organisations when the theory meets the practice.

Second, standardised fairness metrics and testing procedures are lacking. No universally accepted measurement framework exists, and different metrics produce conflicting results. Developing agreed-upon standards through collaboration between technical experts, regulators, and community stakeholders would enable more consistent assessment across institutions and jurisdictions (Raji et al., 2020; O'Neil et al., 2024). The field needs shared benchmarks, not just individual good intentions.

Third, current auditing practices remain limited. Future frameworks should emphasise ongoing, independent auditing with transparent reporting, evaluating not only point-in-time compliance but how systems evolve under deployment conditions (Pushkarna et al., 2022). Research on the governance of RegTech compliance tools themselves, including how to ensure their own transparency and accountability, is particularly underexplored. Who audits the auditors is a question that still needs a proper answer.

Fourth, the interaction between fairness interventions and financial inclusion deserves sustained attention. Rizzi et al. (2021) document how bias reduction can sometimes conflict with efficiency or profitability goals. Long-term empirical studies of the economic effects of fair lending and inclusive AI design are needed to provide evidence supporting policies that promote equity without sacrificing systemic stability (Olaiya et al., 2024). The tension between fairness and profit is real, and pretending otherwise does not help anyone.

Finally, interdisciplinary education is essential. Future professionals in finance, law, and technology require training that combines technical competence with ethical and legal reasoning. Universities, regulators, and industry bodies should develop programmes that prepare practitioners to design, evaluate, and govern fair AI systems (Hacker & Passoth, 2022). The next generation of practitioners needs more than technical skill; they need the judgment to know when and how to apply it.

7. Conclusion

The growing use of artificial intelligence in financial services has created both opportunities and risks. AI systems can improve efficiency and expand access to credit. They also have the capacity to reproduce and amplify social inequalities at a scale and speed that traditional human bias cannot match. Addressing these challenges requires a thorough approach grounded in fairness, accountability, and transparency.

We have made three contributions in this paper. We have synthesised the technical mechanisms through which bias enters financial AI systems, including the under-discussed problem of feedback loops that transform temporary bias into structural inequality. We have critically evaluated the emerging regulatory architecture, the EU AI Act, GDPR, and US fair lending law, identifying both their advances and their limitations, particularly the gap between the EU AI Act's phased applicability timeline and the complexity of real-world compliance. And we have proposed an integrated FAT Lifecycle Framework that maps specific FAT obligations, tools, and regulatory

requirements onto each stage of AI system development, from data collection through to post-deployment governance.

Our central analytical contribution is the engagement with formal fairness impossibility results. The theorems of Chouldechova (2017) and Kleinberg et al. (2017) demonstrate that fairness is not a single property that can be achieved through a combination of techniques, but a family of properties between which institutions must make explicit, documented, and governable trade-offs. Regulatory frameworks that ignore this constraint risk producing compliance exercises that satisfy formal requirements while leaving substantive inequities unaddressed. We should be honest about that.

Building a responsible financial AI ecosystem requires genuine collaboration among financial institutions, regulators, technology providers, and civil society. Technical expertise must be combined with ethical reflection and policy oversight. The success of financial AI will ultimately be measured not only by predictive accuracy or profitability but by its contribution to equitable access and public trust, a standard that demands ongoing commitment, not a single point of compliance.

References

[1] Abikoye, B. E., Umeorah, S. C., Adelaja, A. O., Ayodele, O., & Ogunsuji, Y. M. (2024). Regulatory compliance and efficiency in financial technologies: Challenges and innovations. World Journal of Advanced Research and Reviews, 23(2), 1045–1058.

[2] Akinwumi, M., Merrill, J., Rice, L., Saleh, K., & Yap, M. (2021). An AI fair lending policy agenda for the federal financial regulators. Brookings Policy Brief.

[3] Bahangulu, J. K., & Owusu-Berko, L. (2025). Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications. World Journal of Advanced Research and Reviews, 25(2), 1746–1763.

[4] 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.

[5] Butler, T., & O'Brien, L. (2019). Understanding RegTech for digital regulatory compliance. In Disrupting Finance (pp. 85–102). Palgrave Macmillan.

[6] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. Artificial Intelligence Review, 57(8), Article 216.

[7] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

[8] Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy non-discrimination in data-driven systems. arXiv preprint arXiv:1707.08120.

[9] Farinu, U. (2025). Fairness, accountability, and transparency in AI: Ethical challenges in data-driven decision-making. Available at SSRN 5128174.

[10] Garcia, A. C. B., Garcia, M. G. P., & Rigobon, R. (2024). Algorithmic discrimination in the credit domain: What do we know about it? AI & Society, 39(4), 2059–2098.

[11] Golpayegani, D., Hupont, I., Panigutti, C., Pandit, H. J., Schade, S., O'Sullivan, D., & Lewis, D. (2024). AI cards: Towards an applied framework for machine-readable AI and risk documentation inspired by the EU AI Act. In Annual Privacy Forum (pp. 48–72). Springer.

[12] Hacker, P., & Passoth, J. H. (2022). Varieties of AI explanations under the law: From the GDPR to the AIA, and beyond. In Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 343–373). Springer.

[13] Huang, C., Nourian, A., & Griest, K. (2021). Hidden technical debts for fair machine learning in financial services. arXiv preprint arXiv:2103.10510.

[14] Hurlin, C., Pérignon, C., & Saurin, S. (2026). The fairness of credit scoring models. Management Science, 72(1), 406–425.

[15] Juliussen, B. A. (2025). The right to an explanation under the GDPR and the AI Act. In International Conference on Multimedia Modeling (pp. 184–197). Springer.

[16] Karimi, A. H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 353–362).

[17] Kelley, S., Ovchinnikov, A., Hardoon, D., & Heinrich, K. (2022). Antidiscrimination laws, artificial intelligence, and gender bias: A case study in nonmortgage fintech lending. Manufacturing & Service Operations Management, 24(6), 3039–3059.

[18] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS), Article 43.

[19] Kothandapani, H. P. (2024). Automating financial compliance with AI: A new era in regulatory technology (RegTech). International Journal of Science and Research Archive, 11(1), 2646–2659. https://doi.org/10.30574/ijsra.2024.11.1.0040

[20] Kumar, I. E., Hines, K. E., & Dickerson, J. P. (2022). Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with US fair lending regulation. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 357–368).

[21] Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In ICASSP 2019 – IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 2847–2851).

[22] Mensah, G. B. (2023). Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in AI systems. Africa Journal for Regulatory Affairs (AJFRA), 2(1).

[23] Metikoš, L., & Ausloos, J. (2025). The right to an explanation in practice: Insights from case law for the GDPR and the AI Act. Law, Innovation and Technology, 17(1), 205–240.

[24] Nisevic, M., Cuypers, A., & De Bruyne, J. (2024). Explainable AI: Can the AI Act and the GDPR go out for a date? In 2024 International Joint Conference on Neural Networks (pp. 1–8). IEEE.

[25] Oguntibeju, O. O. (2024). Mitigating artificial intelligence bias in financial systems: A comparative analysis of debiasing techniques. Asian Journal of Research in Computer Science, 17(12), 165–178.

[26] Olaiya, O. P., Adesoga, T. O., & Pieterson, K. (2024). RegTech solutions: Enhancing compliance and risk management in the financial industry. GSC Advanced Research and Reviews, 19(2), 234–256.

[27] O'Neil, C., Sargeant, H., & Appel, J. (2024). Explainable fairness in regulatory algorithmic auditing. West Virginia Law Review, 127(1), 79–133.

[28] Oyasiji, O., Okesiji, A., Imediegwu, C. C., Adeyemi, B., & Oluwole, T. (2023). Ethical AI in financial decision-making: Transparency, bias, and regulation. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 9(5), 453–471.

[29] Pagan, N., Baumann, J., Elokda, E., De Pasquale, G., Bolognani, S., & Hannák, A. (2023). A classification of feedback loops and their relation to biases in automated decision-making systems. Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–14.

[30] Petersen, F., Mukherjee, D., Sun, Y., & Yurochkin, M. (2021). Post-processing for individual fairness. In Advances in Neural Information Processing Systems (Vol. 34, pp. 25480–25492).

[31] Prince, A. E. R., & Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Review, 105(3), 1257–1318.

[32] Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1776–1826).

[33] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

[34] Rane, N., Choudhary, S., & Rane, J. (2023). Explainable artificial intelligence (XAI) approaches for transparency and accountability in financial decision-making. Available at SSRN 4640316.

[35] Rizzi, A., Kessler, A., & Menajovsky, J. (2021). The stories algorithms tell: Bias and financial inclusion at the data margins. Center for Financial Inclusion, Accion.

[36] Thiruma Valavan, A. (2023). AI ethics and bias: Exploratory study on the ethical considerations and potential biases in AI and data-driven decision-making in banking, with a focus on fairness, transparency, and accountability. World Journal of Advanced Research and Reviews, 20(3), 456–478.

[37] Trinh, T. K., & Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems, 4(2), 36–49.

[38] Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J. P., & Shah, C. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review. ACM Computing Surveys, 56(12), Article 312.

[39] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law and Technology, 31(2), 841–887.

[40] West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute.

[41] Wu, J. J. X. (2024). Algorithmic fairness in consumer credit underwriting: Towards a harm-based framework for AI fair lending. Berkeley Business Law Journal, 21(1), 65–122.

[42] Wyllie, S., Shumailov, I., & Papernot, N. (2024). Fairness feedback loops: Training on synthetic data amplifies bias. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 567–582).

[43] Zeng, X., Jiang, K., Cheng, G., & Dobriban, E. (2024). Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing. arXiv preprint arXiv:2402.02817.