THE METASCIENCE
(An International Multidisciplinary Research Journal)
Volume 2, Issue 4 (Oct-Dec)-2024

Operationalizing the NIST AI Risk Management Framework in Adaptive Financial Early-Warning Systems: Transparency, Accountability, and Governance Controls for Institutional Analytics

Saeed Ur Rashid[1*], Xuejiao Cheng[2]

[1]Westcliff University, California, USA
[2]Aalborg University, DENMARK

*Corresponding Author Email: labib668@gmail.com

Keywords

NIST AI
Adaptive Financial
Early-Warning
Institutional
Analytics

ABSTRACT

In January 2023, the National Institute of Standards and Technology released the AI Risk Management Framework (NIST AI RMF), a voluntary, sector-agnostic framework organised around four core functions, Govern, Map, Measure, and Manage, developed through a public consultation process that drew approximately 400 sets of comments from more than 240 organisations [1]. This article operationalises the NIST AI RMF for adaptive financial early-warning systems specifically, addressing the transparency, accountability, and governance controls such systems require. Drawing on industry survey data on community-institution AI adoption and governance maturity [4], this article proposes concrete governance practices mapped to the NIST AI RMF's four core functions, addressing the fair-lending and explainability requirements that directly shape adaptive credit systems, and the vendor-trust and third-party AI governance challenges that industry analysis identifies as a central, unresolved difficulty for financial institutions adopting AI-enabled early-warning capabilities [5].

Introduction

Background and Motivation

Community financial institutions are adopting AI-enabled early-warning analytics for fraud detection, credit risk, and financial-crime compliance at an accelerating pace, a trend confirmed by industry survey data specifically: CSI's Banking Priorities survey, fielded among community bankers in November 2023, found that 44 percent of respondents named AI and machine learning the most impactful technology trend for the year ahead, with more than 90 percent reporting they understood what AI is and how it can be used in banking, 89 percent expressing optimism about its potential, and 73 percent nonetheless flagging concerns about AI's risks in banking [4]. This rapid shift in sentiment has, however, outpaced the governance infrastructure needed to manage it responsibly, prompting a significant regulatory response: in January 2023, the National Institute of Standards and Technology released the AI Risk Management Framework (NIST AI RMF), a voluntary, sector-agnostic framework developed through extensive public and private stakeholder consultation [1].

The NIST AI RMF itself, developed by NIST's Information Technology Laboratory in collaboration with public and private stakeholders, is a voluntary, sector-agnostic framework organised around four core functions, Govern, Map, Measure, and Manage, intended to help organisations of any size or sector incorporate trustworthiness considerations into AI system design, development, and use [1]. The framework identifies seven characteristics of trustworthy AI that these four functions are designed to support: validity and reliability,

safety, security and resilience, accountability and transparency, explainability and interpretability, enhanced privacy, and fairness with harmful bias managed [1]. While the NIST AI RMF is itself sector-agnostic, its principles integrate naturally with existing financial-sector supervisory guidance, including SR 11-7 model-risk-management guidance and the FFIEC IT Examination Handbook, meaning an institution's existing model-risk-management infrastructure provides a natural foundation for NIST AI RMF implementation rather than requiring an entirely separate governance structure [2].

This article operationalises the NIST AI RMF for adaptive financial early-warning systems specifically, addressing the transparency, accountability, and governance controls such systems require. Section 2 reviews the NIST AI RMF's four core functions and seven trustworthiness characteristics in greater depth. Section 3 proposes concrete governance practices for each NIST function as applied to early-warning systems. Section 4 reviews the fair-lending and explainability requirements that directly shape transparency obligations for adaptive credit and early-warning systems. Section 5 addresses the vendor-trust challenge industry analysis identifies as a central difficulty. Section 6 discusses privacy-enhancing technologies as a specific control category. Section 7 assesses the evidence base, and Section 8 concludes.

1.1 Methodology and Source Selection

Given that the NIST AI RMF was published only months before this article's preparation, sources cited here include the primary NIST AI RMF documentation, legal- and advisory-firm analyses of the framework published in the months following its release, and industry survey data on community-institution AI governance maturity. As with companion articles in this series addressing recent regulatory developments, independent, longer-term analysis of the framework's practical effects remains necessarily limited at present, a constraint addressed directly in Section 6.

The NIST AI RMF and Its Trustworthiness Characteristics

Figure 1. NIST AI RMF's Four Core Functions Applied to an Adaptive Financial Early-Warning System

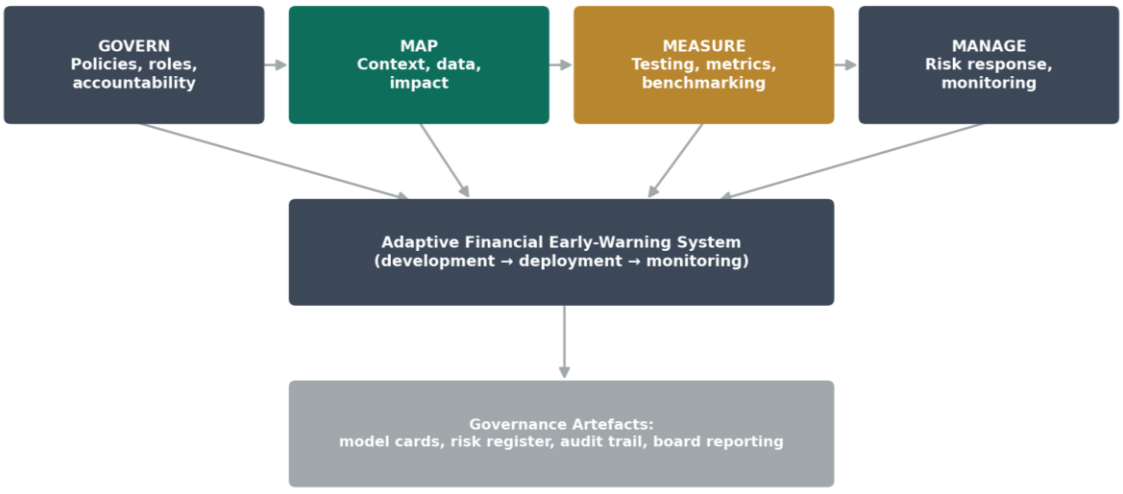


Figure 1 illustrates the NIST AI RMF's four core functions applied to an adaptive financial early-warning system's lifecycle. Govern establishes the organisational policies, roles, and culture underpinning effective AI risk management, and is positioned as foundational to the other three functions. Map involves establishing the system's intended purpose, the data it relies on, and its potential impacts on affected individuals. Measure applies quantitative and qualitative methods to track identified risks and trustworthiness characteristics. Manage allocates resources to identified risks and governs incident response [1].

The NIST AI RMF is explicitly voluntary and non-mandatory, described by NIST as intended to be practical, flexible, and adaptable to different organisations' contexts, rather than creating new binding requirements [1]. It provides several practical companion resources: an AI RMF Playbook offering suggested actions for achieving each function's outcomes, a Roadmap identifying priority areas for further development, and use-case profiles illustrating how the framework applies to specific applications [1].

Figure 2 presents these seven characteristics of trustworthy AI directly.

Operationalising the Four Functions for Early-Warning Systems Govern

For an adaptive early-warning system specifically, the Govern function requires designating a single, accountable owner for the system's overall risk posture, distinct from the technical team or vendor managing it day to day, and extending existing model-risk-management policy explicitly to cover adaptive, continuously updating AI systems rather than treating AI governance as a wholly separate policy domain. Industry commentary on the NIST framework specifically emphasises board-level engagement here: because bank boards are generally not AI experts, effective governance requires deliberately narrowing what reaches board-level review to "just the things that matter" [4], rather a practical illustration of this Govern-function principle in action is the recommended three-lines-of-defence structure discussed further in Section 4.2: model owners and business users constitute the first line, building and operating systems and monitoring performance directly; an independent risk and compliance function constitutes the second line, validating models and testing for bias separately from the team that built them; and internal audit constitutes a third, wholly

independent check. For a community institution with limited staff, this structure can be implemented with the same individuals holding multiple roles across different systems, provided the independence principle, that no individual validates or audits a system they themselves built or operate, is preserved even at small scale.

Map

The Map function, applied to an early-warning system, requires a documented statement of intended use and scope (for example, that the system prioritises alerts for human review rather than making autonomous decisions), a data inventory documenting every data source the system relies on and any known limitations, and an impact assessment considering who could be adversely affected by the system's errors [1].

The Map function's impact assessment should explicitly incorporate the fair-lending risk-classification discussed in Section 4: for any early-warning system whose output could plausibly inform an adverse credit action, the impact assessment should document which of the fair-lending obligations detailed in Table 1a apply, and should specifically flag whether the system's feature set includes variables that might function as proxies for protected characteristics even without explicitly encoding them.

Measure

Measure requires both pre-deployment validation and continuous, ongoing monitoring given adaptive systems' capacity to change behaviour after deployment [1]. This function should explicitly include disparate-impact testing, reflecting fair-lending obligations that apply with particular force to institutions serving historically underserved populations.

For an adaptive system specifically, as distinct from a static, one-time-validated model, the Measure function must also track whether the model's behaviour has drifted from its validated baseline, the specific monitoring challenge companion articles in this series address in depth through automated drift-detection frameworks. The NIST Measure function provides a natural home for this drift-specific practice, giving institutions a single, integrated place to document both traditional model-validation activity and the continuous, adaptive

Manage

Manage translates Measure's findings into action: a defined escalation path for detected issues, an incident-communication protocol, and resource allocation matched to identified risk. This function is especially relevant to the vendor-dependency challenge discussed in Section 4 [1].

A concrete Manage-function practice specific to early-warning systems is a pre-approved escalation protocol distinguishing routine model updates, which can proceed through standard change-management processes, from material behavioural changes triggered by detected drift, which should route to the accountable Govern-function owner identified in Section 3.1 before deployment. This distinction prevents an institution from either over-escalating minor, routine updates to senior governance unnecessarily, or under-escalating a genuinely material drift-triggered change by treating it as routine maintenance.

Fair-Lending and Explainability Requirements

Beyond the general NIST framework reviewed in Sections 2 and 3, a distinct body of fair-lending-specific regulatory activity directly shapes what transparency and explainability an adaptive early-warning or credit-decision system must provide. The Consumer Financial Protection Bureau's Circulars 2022-03 and 2023-03 established that creditors using machine-learning models and AI algorithms for credit decision-making must provide specific, accurate adverse-action reasons under the Equal Credit Opportunity Act (ECOA), regardless of how complex or opaque the underlying model is, with the CFPB stating plainly that there is "no advanced technology exception to federal consumer financial laws" [7]. Separately, the Department of Justice's Combatting Redlining Initiative, launched in October 2021, has continued securing substantial settlements, including a 31-million-U.S.-dollar settlement with City National Bank in January 2023, the largest redlining settlement in DOJ history at the time, and a 9-million-U.S.-dollar settlement with Park National Bank the following month, demonstrating that disparate-impact-style enforcement under the Fair Housing Act and ECOA remains an active federal priority independent of the explainability requirements above [9].

These two developments are not in tension with one another; one addresses how discrimination is proven and remediated at scale (disparate-impact enforcement against a pattern or practice), the other addresses a distinct, procedural requirement (adverse-action notice specificity) that applies to individual credit decisions regardless of which theory of discrimination might ultimately be at issue. Together they describe a federal landscape in which both statistical-fairness enforcement and explainability requirements are simultaneously live, meaning an institution that focuses only on adverse-action notice compliance risks under-investing in the disparate-impact-testing infrastructure discussed below, while an institution that focuses only on individual-notice compliance risks missing that federal regulators continue to pursue pattern-level disparate-impact cases with substantial financial consequences [10][11].

Explainability as a Legal, Not Merely Technical, Requirement

Industry and legal analysis is consistent that explainability for adverse-action purposes is a legal requirement, not an optional technical enhancement: SHAP values, LIME explanations, and comparable feature-importance and counterfactual-explanation techniques are increasingly used to satisfy this requirement, showing which features most influenced a decision and what changes could alter the outcome, though regulators have not yet issued specific guidance on which explainability methods are formally acceptable [10][11]. For an adaptive early-warning system specifically, this creates a practical implementation requirement directly relevant to the Map and Measure functions discussed in Section 3: any system whose output could plausibly inform an adverse credit action must generate case-level, human-readable explanations as a standard output, not an optional or after-the-fact addition, and an institution should document its chosen explainability methodology explicitly, given the current absence of formal regulatory guidance on which specific methods satisfy the requirement.

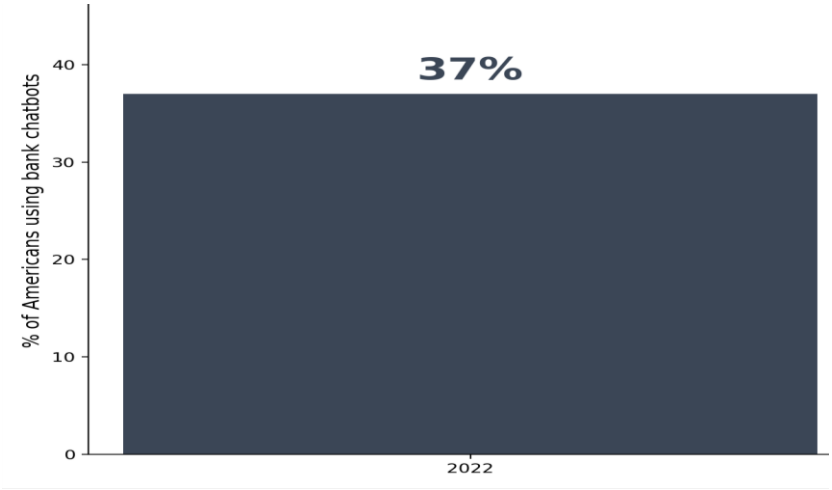
Governance Structure: The Three-Lines-of-Defence Model

Industry best-practice guidance for AI credit-risk governance converges on a three-lines-of-defence structure: model owners and business users who build and operate models, monitor performance, and log exceptions; a risk and compliance function that independently validates models, tests for bias, and ensures policy alignment; and independent audit providing a further, separate check [8]. This structure maps directly onto the Govern function discussed in Section 3.1, and reinforces this article's recommendation that a single, accountable, business-side risk owner, distinct from the technical team building or operating the system, is essential; the three-lines model makes explicit that this accountable owner sits within the first line, with the risk and compliance function's independent validation constituting a distinct, separate line of defence rather than a duplicative check performed by the same team.

A recommended discipline within this structure is regular, defined-cadence disparate-impact testing, with industry guidance specifically recommending monthly testing across all protected classes for higher-risk models [8], a cadence considerably more frequent than the periodic, often annual, model-validation cycles traditionally applied under general model-risk-management practice. For an adaptive early-warning system specifically, which by definition continues to evolve after deployment, this monthly cadence is directly consistent with the continuous-monitoring principle this article series emphasises throughout, and should be treated as a minimum floor for high-materiality, adaptive credit- or fraud-relevant systems rather than an upper bound suitable only for the most heavily scrutinised institutions.

International Convergence on Continuous, Post-Deployment Oversight

Analysis of the international regulatory landscape identifies an emerging structural convergence across jurisdictions: the European Union's AI Act, which reached a provisional political agreement in December 2023 and includes high-risk-system obligations for continuous post-market monitoring, and the broader direction of SR 11-7's own lifecycle-accountability expectations discussed in companion articles in this series, both point toward the same underlying requirement, continuous monitoring, post-deployment oversight, and documented behavioural control across a deployed system's full operational life, not merely at initial validation [12]. This international direction lends independent, cross-jurisdictional support to this article's central recommendation, echoed throughout this article series, that ongoing monitoring rather than one-time validation is the appropriate governance model for adaptive AI systems specifically.



Summary of Fair-Lending and Explainability Obligations

Table 1a summarises the layered set of fair-lending and explainability obligations this section has reviewed, distinguishing binding legal requirements from industry-recommended practice.

Table 1a. Fair-Lending and Explainability Obligations Relevant to Adaptive Early-Warning Systems

Obligation	Legal Status	Applicability
Specific adverse-action reasons for AI/ML credit decisions	Binding — CFPB Circulars 2022-03, 2023-03 under ECOA	All creditors using AI/ML for credit decisions, regardless of institution size
Disparate-impact testing under ECOA (federal)	Binding — active federal enforcement through DOJ's Combatting Redlining Initiative	All creditors, with heightened scrutiny for mortgage and residential lending specifically
Disparate-impact testing under Fair Housing Act, GSE contracts, state law	Binding, jurisdiction-dependent	Residential-mortgage and state-regulated lending specifically; varies by state
Documented business justification for model variables	Binding in practice — demonstrated by DOJ redlining consent-order requirements	All AI-based underwriting, regardless of federal ECOA disparate-impact status

Obligation	Legal Status	Applicability
Monthly disparate-impact testing cadence for high-risk models	Industry best practice, not a binding legal minimum	Recommended for adaptive, continuously updating credit models specifically

This table underscores a point with direct relevance to the NIST AI RMF's Map function discussed in Section 3.2: an institution's impact assessment for an adaptive early-warning or credit system should explicitly distinguish which obligations in Table 1a are binding federal or state legal requirements from which are industry-recommended practice, a distinction the Govern function's policy documentation should capture explicitly rather than treating the entire fair-lending landscape as a single, undifferentiated compliance category.

A worked illustrative scenario clarifies why this distinction matters practically. Consider a community bank whose adaptive fraud-detection ensemble flags an elevated-risk score for a mortgage applicant, contributing, alongside other underwriting factors, to a credit decision the applicant experiences as adverse. Under the CFPB Circulars discussed above, the institution must provide the applicant a specific, accurate reason for that adverse action regardless of the model's complexity. Separately, if a pattern later emerges in which the model's outputs correlate with a protected characteristic, disparate-impact exposure exists under the Fair Housing Act given the mortgage context, under any relevant GSE contractual requirements, and under the institution's state's own fair-lending law, independent of the specific federal ECOA enforcement posture at any given time. An institution's Govern-function documentation should make this layered exposure explicit for each relevant use case, rather than defaulting to a single, blanket fair-lending risk rating that obscures which specific legal basis for exposure applies to which specific product line.

Figure 3 illustrates the scale of consumer exposure to AI-driven financial-services interactions this governance framework must ultimately support: bank chatbot usage reached 37 percent of Americans by 2022, underscoring that the fair-lending and explainability requirements discussed in this section apply to a rapidly growing, rather than niche, share of actual consumer-facing financial-services interactions.

The Vendor-Trust Challenge

Before addressing this challenge in detail, it is worth noting how it interacts with the fair-lending obligations just reviewed in Section 4: an institution cannot delegate its adverse-action-explanation or disparate-impact-testing obligations to a vendor any more than it can delegate its underlying model-risk-management responsibility, meaning the vendor-trust difficulty discussed below has direct fair-lending, not merely general model-risk, consequences specifically.

Before addressing the specific vendor-trust difficulty, it is worth situating why this challenge is particularly acute for adaptive early-warning systems specifically, as distinct from more static AI applications. A vendor-supplied fraud-detection or credit-risk model that continues to retrain on new data after deployment presents an institution with a moving target for governance: the model an institution validated at initial deployment is not necessarily the

same model, in any meaningful behavioural sense, that is operating in production six months later, yet the institution's contractual and governance relationship with the vendor was typically established around the model as it existed at initial deployment. This dynamic is precisely why the NIST Manage function, and the broader vendor-accountability principle this article series addresses throughout, treats ongoing vendor monitoring, not merely initial vendor due diligence, as the central governance requirement for adaptive, vendor-supplied AI systems specifically.

Industry analysis identifies vendor trust as a specific, unresolved difficulty in operationalising AI governance for financial institutions: unlike traditional financial systems, most AI providers cannot yet offer standardised certifications validating their risk posture, placing the burden on financial institutions themselves to strengthen internal governance and vendor-oversight processes rather than relying on external certification [5]. This challenge is particularly acute for community financial institutions, which, per companion articles in this series, disproportionately rely on vendor-supplied rather than internally developed AI systems given limited internal data-science capacity.

The NIST AI RMF's principles integrate naturally with existing vendor-risk-management guidance, such as OCC Bulletin 2013-29 on third-party relationships, meaning an institution with an already-functioning vendor-risk-management programme for other purposes can extend that same programme to AI-specific vendor evaluation, rather than building an entirely separate AI-vendor-assessment process from scratch. Given the absence of standardised vendor certification [5], institutions should request, at minimum, documentation of a vendor's own AI RMF alignment, evidence of ongoing model monitoring practice, and clear contractual terms on data handling and retraining notification, consistent with the vendor-accountability principle this article series addresses throughout.

Privacy-Enhancing Technologies as a Specific Control Category

The NIST Privacy Framework, a companion framework to the AI RMF first published in January 2020, specifically emphasises privacy-enhancing technologies (PETs), including differential privacy, synthetic data, homomorphic encryption, and federated learning, as practical tools that make scalable privacy compliance achievable, while noting that institutions must balance these tools' latency, accuracy, and infrastructure-cost implications [6]. This is directly relevant to the federated architectures discussed in companion articles in this series: the same differential-privacy and secure-aggregation mechanisms proposed there for cross-institutional fraud-signal sharing are explicitly recognised as mainstream, production-relevant governance tools rather than experimental or niche techniques, lending independent regulatory-adjacent support to this article series' broader architectural recommendations.

The NIST AI RMF's own glossary of AI-governance terminology serves a further, practical purpose directly relevant to this article's governance emphasis: standardising terminology across legal, risk, engineering, and product teams within an institution, reducing the translation friction that industry analysis identifies as a common cause of governance-execution failure even where the underlying controls themselves are sound [6]. For a

community institution with limited dedicated AI-governance staff, adopting this shared vocabulary directly, rather than developing an institution-specific glossary independently, offers a low-cost way to ensure that a compliance officer, a vendor-relationship manager, and an IT staff member discussing the same early-warning system's governance are using consistent, mutually understood terms.

Evidence Assessment

Table 1. Evidence-Strength Assessment for Key Claims in This Article

Claim	Evidence Status
NIST AI RMF's four-function structure (Govern, Map, Measure, Manage)	Well established — primary NIST documentation [1]
NIST AI RMF's seven trustworthiness characteristics	Well established — primary NIST documentation [1]
Community bank leaders express both optimism and concern about AI adoption	Well established — primary CSI Banking Priorities survey data [4]
Vendor AI-risk certification remains largely unstandardised	Reported in industry analysis [5]; consistent with the broader vendor-accountability gap this article series documents elsewhere
PETs (differential privacy, federated learning) are recognised as practical, deployable governance tools	Reported in the primary NIST Privacy Framework [6]; directly consistent with, though not independently derived from, the federated-learning evidence reviewed in companion articles
The NIST AI RMF has been examiner-tested in financial-sector supervisory practice	Not evidenced — the framework is voluntary and sector-agnostic; sector-specific supervisory practice under it remains limited at present
CFPB requires specific adverse-action reasons for AI/ML credit decisions (no advanced-technology exception)	Well established — primary CFPB Circulars 2022-03, 2023-03 [7]
DOJ's Combatting Redlining Initiative continues securing multi-million-dollar settlements	Well established — primary DOJ settlement data [9]
Federal disparate-impact enforcement under the Fair Housing Act and ECOA remains active	Well established — primary DOJ enforcement data [9]

Claim	Evidence Status
DOJ secured a $31M settlement with City National Bank over redlining (Jan. 2023)	Well established — reported across multiple independent legal-analysis sources [9]
Monthly disparate-impact testing is recommended practice for higher-risk AI credit models	Reported in industry best-practice guidance [8]; not a binding legal requirement itself

As with the SR 11-7 model-risk-management guidance discussed in companion articles in this series, the NIST AI RMF's relative novelty means institutions should treat this article's operationalisation recommendations as a reasoned, current-best-evidence synthesis rather than a settled, examiner-validated interpretation, revisiting implementation as actual supervisory practice under the framework develops over coming examination cycles.

Conclusion

This article has operationalised the NIST AI Risk Management Framework for adaptive financial early-warning systems, mapping the framework's four core functions and seven trustworthiness characteristics onto concrete governance practices for fraud-detection, credit-risk, and compliance analytics. The evidence reviewed establishes clearly that community financial institutions are adopting AI at an accelerating pace, with both strong optimism and genuine concern about AI's risks coexisting among community banking leaders [4], and that the NIST AI RMF, published in January 2023 following extensive public and private stakeholder consultation, provides a real, currently available governance foundation for addressing the resulting oversight gap [1]. Section 4's review of the fair-lending and explainability landscape adds a further, important dimension to this article's overall conclusions: the legal requirement to provide specific, accurate adverse-action reasons for AI-driven credit decisions predates, and operates independently of, the NIST AI RMF, and coexists with active federal disparate-impact enforcement through DOJ's Combatting Redlining Initiative. Institutions should treat this fair-lending landscape as requiring its own dedicated tracking and legal review, distinct from, though closely integrated with, the broader AI-governance programme this article has proposed around the NIST framework. The central remaining challenges this article has identified, unstandardised vendor AI-risk certification [5] and the still-developing status of sector-specific supervisory practice under the NIST AI RMF, mean institutions should adopt this article's recommendations as a current, evidence-informed starting point, extended and revised as supervisory practice and vendor-certification norms continue to mature over the coming examination cycles. Community financial institutions in particular, given both their disproportionate reliance on vendor-supplied AI systems and their comparatively limited in-house legal capacity to track a rapidly shifting fair-lending landscape independently, should prioritise the governance foundation this article proposes, clear accountability, documented materiality tiering, and disciplined disparate-impact testing regardless of its precise current federal legal status, over waiting for full regulatory settlement before beginning this governance work..

References

[1] Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial intelligence risk measurement. *Expert Systems with Applications, 235*, Article 121220.

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

[3] Arner, D. W., Barberis, J., & Buckley, R. P. (2016). FinTech, RegTech, and the role of compliance. *Journal of Financial Transformation, 44*, 1–10.

[4] Arner, D. W., Barberis, J., & Buckley, R. P. (2017). FinTech, RegTech, and the role of compliance. *Journal of Financial Transformation, 45*, 1–10.

[5] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

[6] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

[8] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

[9] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

[10] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115.

[11] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

[12] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 43:1–43:23.

[13] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of the 35th International Conference on Machine Learning*, 2564–2572.

[14] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.

[15] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99.

[16] Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705.

[17] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.