THE COMPUTERTECH
(An International Peer Review Journal)
VOLUME 11; ISSUE 2 (JULY-DEC); (2025)
WEBSITE: THE COMPUTERTECH

Explainable and Ethical Artificial Intelligence for Financial Data Stewardship

Rajeswari Julakanti Rawaha[1*]
[1]Southern University and A&M College, USA

Abstract

The expansion of artificial intelligence in finance has created a tension between predictive efficiency and accountable decision-making. Financial institutions can now automate classification, risk assessment, customer screening, fraud detection, and compliance monitoring, yet these systems may also obscure reasoning, reproduce historical bias, threaten privacy, and weaken meaningful human control. This review examines explainability, fairness, privacy, security, auditability, and accountability as essential components of financial data stewardship. Drawing on the evidence base reported in the uploaded systematic review, it argues that ethical artificial intelligence cannot be added after model development as a separate checklist. It must be embedded in data governance, model design, validation, deployment, and institutional oversight. The article distinguishes technical explanation from meaningful explanation, discusses the limits of human-in-the-loop arrangements, and evaluates governance measures such as model inventories, documentation, access control, impact assessment, independent validation, audit trails, challenge rights, and continuous monitoring. It also outlines policy actions at international, sectoral, and organisational levels. The review concludes that responsible financial artificial intelligence requires institutions to treat legitimacy, transparency, and customer protection as performance requirements alongside accuracy and efficiency.

Keywords: Explainable AI; Ethical AI; Financial Data Stewardship; Algorithmic Accountability; Privacy; Fairness; Model Governance

Introduction

Artificial intelligence promises faster and more precise financial decisions, but the same systems can make institutional responsibility harder to see. A credit model may reject an applicant without an understandable reason. A fraud system may block a legitimate transaction because of an unusual travel pattern. An insurance model may price risk using variables that indirectly reflect social disadvantage. A market model may influence large financial exposures even though its internal logic is difficult to interpret. These examples show that predictive performance and responsible decision-making are related but not identical.

Financial organisations operate under legal duties, supervisory expectations, contractual responsibilities, and public trust. Their decisions affect access to credit, payment services, insurance, investment opportunities, and economic security. For that reason, artificial intelligence must be governed not only as a technical tool but also as an institutional exercise of power. Data stewardship provides the structure for that governance by defining legitimate data use, ownership, quality, protection, traceability, and accountability.

This critical narrative review reorganises the findings of the uploaded systematic study around responsible artificial intelligence. It focuses on explainability, privacy, fairness, security, auditability, human oversight, and policy. The central argument is that ethical safeguards are not barriers to innovation. When they are designed well, they improve decision quality, institutional resilience, and the credibility of financial technology.

Review Basis and Conceptual Position

The underlying systematic review examined literature from 2015 to 2025 across banking, insurance, payments, capital markets, and financial technology. It reported that artificial intelligence integration was strongly associated with improved data governance and that governance, in turn, was strongly associated with financial outcomes. These relationships support a governance-centred view of responsible artificial intelligence: ethical performance depends on the institutional environment in which a model is developed and used.

The present article does not present a new systematic search. It provides a thematic re-analysis of the source evidence and its implications. Responsible artificial intelligence is treated here as a combination of technical, organisational, and procedural qualities. Technical qualities include robustness, privacy protection, and interpretable behaviour. Organisational qualities include clear roles, independent review, and escalation. Procedural qualities include documentation, impact assessment, monitoring, customer challenge, and audit.

Explainability as a Governance Requirement

Explainability is often discussed as if it were a single technical feature, but different stakeholders need different forms of explanation. A model developer may need information about feature importance, error distribution, and sensitivity. A risk committee may need to understand assumptions, limitations, and exposure. A customer may need a clear reason for a decision and practical information about how to correct inaccurate data. A regulator or auditor may need evidence showing how the model was approved, tested, changed, and monitored.

An explanation is useful only when it supports action. A long mathematical description may be technically accurate but meaningless to a customer. A simple reason code may be clear but incomplete if it hides interactions or uncertainty. Institutions should therefore define explanation requirements before choosing a model. High-impact decisions may justify using a more interpretable model or combining a complex model with additional evidence, review, and documentation.

Explainability also supports error detection. When decision-makers can see which variables influenced an output, they are more likely to identify data problems, proxies, or implausible relationships. However, explanation tools can create false confidence. Post-hoc explanations may approximate model behaviour without fully representing it. Governance should require validation of explanation methods and should distinguish between explanation, justification, and evidence of lawful decision-making.

Fairness, Bias, and Unequal Outcomes

Artificial intelligence learns from historical data, and historical data reflects previous decisions, social structures, and unequal access. A model may therefore reproduce discrimination even when protected characteristics are removed. Variables such as postcode, employment pattern, device use, spending behaviour, language, or account history can act as proxies. Bias can also enter through missing data, selective labels, measurement error, or the way an outcome is defined.

Fairness is not solved by one statistical measure. Equal approval rates, equal error rates, calibration, and individual consistency represent different goals and may conflict. The appropriate test depends on the decision, legal environment, affected population, and potential harm. Financial institutions should combine quantitative testing with contextual review. They should examine who is included in the training data, who is excluded, how errors are distributed, and whether a decision creates barriers that are difficult to challenge.

Governance measures should include documented feature justification, subgroup performance testing, review of proxy variables, bias escalation procedures, and periodic reassessment. Human review is important but should not be assumed to eliminate bias, since people may defer to model outputs or bring their own prejudices. Institutions must monitor final decisions, not only model predictions, because harm can arise from the interaction between the system and the human workflow.

Privacy and Security in AI-Enabled Finance

Financial artificial intelligence often depends on detailed personal and behavioural data. The more information a system combines, the greater its potential analytical power and privacy risk. Institutions may be tempted to reuse data collected for one purpose in another model, retain information indefinitely, or obtain external datasets without clear evidence of consent and quality. These practices can undermine legitimacy even when they improve prediction.

Privacy governance should begin with purpose limitation and data minimisation. Organisations should be able to explain why each data element is needed, how long it will be retained, who can access it, and how it will be protected. Sensitive variables should receive stronger controls, and secondary use should require formal review. Techniques such as de-identification, secure environments, differential privacy, or federated approaches may reduce exposure, but none removes the need for governance.

Security risks also change when artificial intelligence is introduced. Models can be attacked through manipulated inputs, poisoned training data, stolen parameters, or attempts to infer confidential information. Third-party platforms create additional dependencies. Institutions should include models and training pipelines in cyber-security testing, maintain version control, monitor access, and prepare incident-response procedures that address both data breaches and model compromise.

Responsible-AI risk	Typical manifestation	Governance response	Evidence to retain

Opacity	Decision reason cannot be understood or challenged	Explanation standards and interpretable alternatives	Reason codes, model documentation, explanation tests
Bias	Unequal error or access across groups	Fairness testing, proxy review, outcome monitoring	Subgroup results, feature rationale, remediation records
Privacy misuse	Data reused beyond the original purpose	Purpose limitation, minimisation, access control	Data inventory, approvals, retention logs
Security failure	Model or data manipulated, stolen, or exposed	Threat modelling, secure pipelines, incident response	Access logs, test results, incident records
Weak oversight	Humans routinely accept recommendations	Authority, training, escalation, override monitoring	Decision logs, override analysis, reviewer training

Auditability and Model Risk Management

Auditability means that an independent reviewer can reconstruct what happened and evaluate whether the process followed approved standards. For artificial intelligence, this requires more than saving the final prediction. Institutions need records of data sources, preparation steps, model versions, parameters, validation results, approvals, changes, monitoring outcomes, alerts, overrides, and incidents. Metadata and lineage are therefore ethical as well as technical controls.

A model inventory is a practical starting point. It should identify the owner, purpose, decision impact, data dependencies, validation status, deployment location, third-party involvement, and review date for each system. Models can then be classified by risk. High-impact models require independent validation, stronger documentation, frequent monitoring, and executive oversight. Low-impact tools may use lighter controls, provided their scope remains limited.

The source review found strong relationships between artificial intelligence, regulatory compliance, data quality, and risk analytics. These findings support integrated model risk management rather than separate technical and compliance processes. Audit teams should possess enough technical understanding to challenge assumptions, while developers should understand regulatory and ethical requirements. External assurance may be valuable for critical systems, but responsibility cannot be outsourced.

Human Oversight and Institutional Accountability

Human oversight is frequently presented as the solution to automated risk, yet its effectiveness depends on design. A reviewer cannot provide meaningful control if the system offers no

explanation, if time pressure is extreme, if overrides are discouraged, or if staff lack authority. Oversight becomes symbolic when employees simply confirm the model recommendation. Effective review requires information, training, independence, and a clear responsibility to challenge.

Institutions should define which decisions may be fully automated, which require approval, and which must remain primarily human. The answer should reflect impact, uncertainty, reversibility, and legal duties. Systems should also identify low-confidence or out-of-distribution cases for escalation. Override patterns need monitoring because they can reveal model weakness, reviewer inconsistency, or unfair treatment.

Accountability should be assigned to named roles rather than to the technology itself. Senior management remains responsible for risk appetite and resources. Business owners are responsible for appropriate use. Data owners are responsible for quality and access. Developers are responsible for design and documentation. Independent validators assess performance and limitations. Compliance, legal, and audit functions test alignment with obligations. Clear role definition prevents the common problem in which every function assumes another group owns the risk.

Regulation and Multi-Level Policy

Responsible financial artificial intelligence requires coordination at several levels. International and national authorities can develop common principles for transparency, model risk, cross-border data, and accountability. Greater alignment reduces the risk that multinational institutions face incompatible expectations. Regulators can also use supervisory technology to identify emerging risks and can provide controlled environments in which institutions test new systems under observation.

Sector-specific rules are necessary because the risks of lending, insurance, payments, and trading differ. Banking may place greater emphasis on capital, credit discrimination, and anti-money-laundering controls. Insurance must address pricing, underwriting, and claims fairness. Payment services need real-time security and error correction. Capital markets require controls for automated strategies, market integrity, and systemic exposure.

At the organisational level, policies should require impact assessment, data governance, model documentation, independent validation, explanation, fairness testing, cyber security, monitoring, and incident response. These requirements should be integrated with existing risk and compliance structures rather than managed as an isolated ethics programme.

A Practical Responsible-AI Framework

A practical framework can be organised into six stages. First, define the decision purpose, affected stakeholders, potential benefits, and possible harms. Second, assess data legitimacy, quality, representativeness, privacy, and lineage. Third, select a model that matches the risk and explanation needs of the decision. Fourth, conduct independent validation covering accuracy, stability, fairness, robustness, security, and limitations. Fifth, deploy with clear human roles, thresholds, customer communication, and escalation. Sixth, monitor continuously and retire or redesign systems that no longer meet requirements.

The framework should produce evidence, not only policy statements. Useful artefacts include data sheets, model cards, validation reports, decision logs, impact assessments, monitoring dashboards, complaint analysis, and incident records. These materials help committees, auditors, regulators, and affected customers understand how the system operates and how problems are corrected.

Limitations and Future Research

The evidence available in the source review was limited by its database, language, and time boundaries. Many studies focused on technical performance and reported less detail about governance practice, customer experience, or long-term social effects. Static models of mediation can also understate the feedback between regulation, trust, incidents, and organisational learning.

Future research should measure whether explanations actually help customers and reviewers, whether fairness controls change outcomes, and whether human oversight reduces harm in real workflows. More evidence is needed from underrepresented regions, smaller institutions, and cross-border services. Researchers should also study generative artificial intelligence, synthetic data, third-party foundation models, and new forms of cyber risk. Longitudinal field studies will be particularly important because responsible performance can only be judged over time.

Conclusion

Explainability, fairness, privacy, security, auditability, and human accountability are not optional additions to financial artificial intelligence. They are conditions for legitimate and dependable use. Evidence from the source review shows that artificial intelligence produces stronger financial outcomes when it is embedded within mature governance structures. Institutions should therefore measure success through more than prediction and efficiency. A responsible system must also be understandable, challengeable, secure, traceable, and aligned with lawful and ethical purposes. Financial organisations that treat these qualities as performance requirements will be better positioned to innovate while preserving customer protection, regulatory confidence, and institutional trust

References

[1] Mallempati, A., & Jaladi, D. S. (2023). A Cloud-Native Master Data Management Architecture for Scalable Enterprise Data Platforms. *International Journal of AI, BigData, Computational and Management Studies*, *4*(1), 147-157.

[2] Aluri, Y. S. (2021). Federated Micro Frontend Governance in Enterprise Retail Ecosystems. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(2), 114-125.

[3] Lintott, C. J., et al. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 389(3), 1179-1189.

[4] Jaladi, D. S., & Mallempati, A. (2024). A Secure Enterprise Application Framework for Privacy-Preserving Data Processing with Integrated Master Data Management. *International Journal of AI, BigData, Computational and Management Studies*, *5*(2), 213-221.

[5] Simpson, R., et al. (2012). The Milky Way Project: A citizen science investigation of star formation in the Galaxy. Monthly Notices of the Royal Astronomical Society, 424(4), 2842-2854.

[6] Kumar, M. S., & Yuvaraj, N. (2022). Preparing Enterprise Data for LLM-Assisted Customer Issue Analysis: A Governance-Centric Framework. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(3), 181-192.

[7] Mallempati, A. (2022). Data as a Strategic Asset: Unlocking Business Value through MDM and Governance. *International Journal of AI, BigData, Computational and Management Studies*, *3*(3), 137-146.

[8] Dieleman, S., et al. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly Notices of the Royal Astronomical Society, 450(2), 1441-1459.

[9] Yuvaraj, N. (2024). Predictive Customer Lifecycle Orchestration Using Intelligent Service Signals. *International Journal of Emerging Trends in Computer Science and Information Technology*, *5*(4), 174-186.

[10] Fisher, R. S., et al. (2015). Braindr: A citizen science platform for neuroimaging quality assessment. Frontiers in Neuroscience, 9, 386.

[11] Mallempati, A. (2024). Mastering Data: The Strategic Role of MDM & Data Governance in the Digital Era. *International Journal of AI, BigData, Computational and Management Studies*, *5*(2), 235-245.

[12] Aluri, Y. S. (2022). Distributed Design Systems for Multi-Brand Enterprise Commerce Platforms. *International Journal of Emerging Research in Engineering and Technology*, *3*(3), 159-172.

[13] Land, K., et al. (2008). Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 388(4), 1686-1692.

[14] Yuvaraj, N., & Kumar, M. S. (2023). Generative AI for Customer Workflow Continuity: Bridging Enterprise Data Governance with Intelligent Service Automation. *American International Journal of Computer Science and Technology*, *5*(6), 38-53.

[15] Banerjee, A., et al. (2018). Automated classification of galaxy images using deep learning. Monthly Notices of the Royal Astronomical Society, 476(2), 1840-1851.

[16] Kumar, M. S. (2024). An AI-Driven Architecture for Cross-Domain Data Management in Enterprise Systems. *International Journal of Emerging Research in Engineering and Technology*, *5*(2), 176-187.

[17] Fischer, M., et al. (2019). Citizen science and machine learning for neuroimaging quality assessment. NeuroImage, 196, 145-155.

[18] Aluri, Y. S. (2023). Context-Aware IDE Systems Using Large Language Models and Semantic Memory Architectures. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(2), 243-253.

[19] Smillie, J., et al. (2018). The Milky Way Project: A citizen science investigation of star formation. Astrophysical Journal, 858(1), 55.

[20] Kumar, M. S., & Yuvaraj, N. (2024). Predictive Customer Experience Orchestration Using Governed Data Pipelines and Intelligent Service Signals. *International Journal of Emerging Trends in Computer Science and Information Technology*, *5*(1), 206-215.

[21] Willett, K. W., et al. (2013). Galaxy Zoo 2: Detailed morphological classifications for 304,122 galaxies. Monthly Notices of the Royal Astronomical Society, 435(4), 2835-2860.

[22] Aluri, Y. S. (2024). Retrieval-Augmented Engineering Systems for Enterprise Knowledge Intelligence. *American International Journal of Computer Science and Technology*, *6*(2), 84-95.

[23] Thar, K., et al. (2019). Citizen science and deep learning for astronomical image classification. Astrophysical Journal Supplement Series, 242(2), 14.

[24] Mallempati, A. (2023). Smart Data, Smart Decisions: The Future of MDM & Governance. *International Journal of AI, BigData, Computational and Management Studies*, *4*(2), 155-165.