ShodhAI: Journal of Artificial Intelligence
January-June 2026 3(1)
https://shodhai.org

Original Article

AUDIT-READY EXPLAINABLE AI FOR FRAUD OPERATIONS: PERSISTED AND REPLAYABLE DECISION ARTEFACTS FOR MODEL GOVERNANCE AND INVESTIGATOR TRUST

Rajeew Vishvakarma [1*]
[1] B.Sc. M.C.A., Project Manager, Infosys Bengaluru, India

ABSTRACT

Fraud detection operations increasingly depend on machine-learning systems to prioritise suspicious events for investigator review. In regulated environments, however, operational defensibility requires more than predictive accuracy and post-hoc explanation. A review, challenge, or audit must be able to reconstruct which model, feature contract, threshold or alert-budget policy, explainer configuration, and workflow actions produced a given alert at decision time. This paper argues that explainability is not the same as audit readiness. It proposes an audit-ready design pattern for fraud operations in which each alert is treated as a governed decision artefact with persisted score and explanation snapshots, version and threshold lineage, investigator disposition records, and monitoring evidence for drift and replayability. The manuscript contributes four core outputs: a six-dimension audit-readiness rubric, a minimum alert-artefact schema, an architecture pattern for persisted and replayable explanations, and an evaluation blueprint that separates predictive quality, explanation quality, workflow utility, and audit readiness. The paper also analyses privacy, security, and retention risks introduced by persisted artefacts and proposes practical controls for role-based disclosure, minimisation, immutable access logging, and evidence-preserving storage. The result is a publication-ready framework for converting explainable fraud models into traceable operational systems.

Keywords: Fraud Detection, Explainable AI, Auditability, Model Governance, Decision Provenance, Drift Monitoring

INTRODUCTION

Machine-learning fraud detection is often discussed as a classification problem: rank events, set a decision threshold, and optimize metrics such as AUROC, PR-AUC, or recall at a given alert budget. In real operations, however, a fraud alert is not merely a score. It can trigger manual review, customer friction, delayed payments, escalations to compliance teams, suspicious-activity workflows, and retrospective audit. That operational setting changes the design problem. Institutions need to know not only whether a model is accurate, but also whether an alert can be reconstructed and defended after model versions, feature logic, or threshold policies evolve.

This manuscript takes the position that a fraud alert should be treated as a governed decision artefact. The decisive question is not simply 'why did the model produce this score?' but 'what exactly was known, shown, applied, stored, and actioned at the moment

*Corresponding Author:
Email address: Rajeew Vishvakarma (rajeew.vishvakarma@gmail.com)
Received: 25 March 2026; **Accepted:** 09 April 2026; **Published** 30 May 2026
DOI: 10.29121/ShodhAI.v3.i1.2026.77
Page Number: 71-78
Journal Title: ShodhAI: Journal of Artificial Intelligence
Journal Abbreviation: ShodhAI J. Artif. Intell.
Online ISSN: 3048-9245, **Print ISSN:** 3108-1940
Publisher: Granthaalayah Publications and Printers, India
Conflict of Interests: The authors declare that they have no competing interests.
Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Authors' Contributions: Each author made an equal contribution to the conception and design of the study. All authors have reviewed and approved the final version of the manuscript for publication.
Transparency: The authors affirm that this manuscript presents an honest, accurate, and transparent account of the study. All essential aspects have been included, and any deviations from the original study plan have been clearly explained. The writing process strictly adhered to established ethical standards.

the alert was created?' That distinction matters because fraud systems are high-volume, adversarial, and non-stationary. A model can be explainable in the narrow sense of supporting feature attributions, yet still be operationally fragile if explanation snapshots are not persisted, policy changes are not versioned, or review outcomes cannot be linked back to the original decision context.

The paper therefore distinguishes explainability from audit readiness. Explainability concerns interpretive access to model behaviour; audit readiness concerns whether a system can reproduce, justify, challenge, and govern decisions over time using durable evidence. This framing aligns with established model risk management expectations that emphasise documentation, validation, change control, oversight, and effective challenge, as well as more recent AI governance frameworks that make lifecycle traceability and logging explicit design requirements.

The contributions of the paper are fivefold. First, it defines audit readiness as a measurable operational capability rather than a vague governance aspiration. Second, it formalises a six-dimension rubric for assessing audit readiness in fraud alerting systems. Third, it specifies a minimum alert-artefact schema that preserves model lineage, policy lineage, explanation state, and workflow state. Fourth, it provides an evaluation blueprint that separates predictive quality, explanation quality, workflow utility, and audit readiness rather than conflating them. Fifth, it analyses the privacy and security implications of persisting explanations and decision traces, and it proposes practical controls suitable for regulated fraud operations.

The manuscript is intentionally framed as a framework and design paper rather than an empirical benchmark study. Public fraud datasets are useful for future validation, but they do not by themselves resolve governance questions such as replayability, evidence persistence, or challenge logging. The goal here is to provide a publication-ready conceptual and technical foundation that can later be extended through a demonstration harness, an enterprise case study, or a human-grounded workflow evaluation.

RELATED WORK AND MOTIVATION

EXPLAINABILITY IN FRAUD AND FINANCIAL-RISK MODELLING

Post-hoc explanation methods such as SHAP and LIME are now common in tabular fraud detection because they provide local attribution summaries for individual alerts and, in some settings, aggregated global views of model behaviour. Counterfactual explanations extend that toolkit by describing minimal changes that would have produced a different decision, thereby supporting contestability and actionability. In graph-based anti-money-laundering contexts, explainers such as GNNExplainer help analysts identify influential nodes, neighbourhoods, and features behind graph predictions.

These methods are valuable, but they solve only part of the operational problem. A local explanation displayed on screen at review time is not, by itself, a durable governance artefact. Unless the exact explanation snapshot, explainer version, feature contract, and decision policy are persisted, the institution may later be unable to reconstruct what the investigator saw or why a case moved down a particular workflow branch. Recent fraud-XAI literature reinforces this concern. User-centred work in finance demonstrates the practical importance of explanation design for regulators, analysts, and auditors, while newer review literature highlights persistent weaknesses in evaluation practice, especially around faithfulness, stability, and workflow relevance in imbalanced and adversarial settings.

SHAP is especially influential because it provides an additive attribution framework with a clear cooperative-game-theoretic interpretation Lundberg and Lee (2017). LIME remains a widely used local surrogate method for explaining individual predictions Ribeiro et al. (2016). Counterfactual explanations are particularly relevant where decisions may be contested because they foreground minimal changes associated with a different outcome Wachter et al. (2018). For graph-based fraud and AML, GNNExplainer provides a natural point of reference for explanatory subgraph extraction Ying et al. (2019). At the same time, explanation reliability cannot be taken for granted; sanity-check work shows that explanation methods can appear convincing while being weakly tied to learned model parameters, which is why stability and faithfulness testing belong in any future empirical extension Adebayo et al. (2018). Applied fraud literature further underscores the need for stakeholder-aware explanation design Zhou et al. (2023) and for stronger evaluation discipline in fraud-specific XAI research Zafar and Wu (2026).

GOVERNANCE, PROVENANCE, AND LOGGING

Model governance literature has long treated documentation, validation, and effective challenge as core lifecycle controls rather than administrative afterthoughts. In financial services, supervisory guidance on model risk management emphasises that robust development must be accompanied by change control, tracking, oversight, and transparent documentation. Cross-sector AI risk management frameworks similarly push organisations to map, measure, and manage risks throughout the lifecycle rather than focus exclusively on model construction.

This governance emphasis converges with two especially relevant strands of recent work. The first is decision provenance: the idea that accountable systems require reconstructable records of how information moved through a decision pipeline. The second is logging for continuous auditing of ML applications: emerging research argues that responsible AI cannot be audited reliably unless logging is designed to capture the information needed to evaluate performance, fairness, transparency, security, and change over time.

In finance, the core governance anchor remains SR 11-7 on model risk management, which emphasises robust development, validation, governance, and effective challenge Board of Governors of the Federal Reserve System, and Office of the Comptroller of the Currency (2011). The OCC's model risk handbook translates similar expectations into examiner-oriented operational controls Office of the Comptroller of the Currency (2021). At a broader AI-governance level, NIST AI RMF 1.0 frames lifecycle risk management around mapping, measuring, managing, and governing AI risks (NIST, 2023), while the EU AI Act makes logging and record-keeping explicit requirements for certain high-risk AI uses European Union (2024). Decision provenance supplies the conceptual basis for reconstructable decision pipelines Singh et al. (2019), and recent logging research argues that ML systems cannot be continuously audited without purpose-built logging practice and tooling Foalem et al. (2025).

For fraud operations, the implication is straightforward. A model explanation is necessary for transparency, but logging, lineage, and replay are necessary for auditability. An operationally mature fraud system therefore needs to preserve not only the output of an explainer but also the policy and workflow context in which that explanation acquired institutional meaning.

NOVELTY RELATIVE TO ADJACENT LITERATURE

The novelty of this paper is not the use of SHAP, LIME, counterfactual explanations, or graph explainers in fraud. Those are established tools. The novelty lies in changing the unit of analysis from 'prediction plus explanation' to 'alert as a versioned, replayable, governable decision artefact'. User-centred fraud-XAI studies focus on explanation design for stakeholder understanding; graph-based AML studies focus on predictive performance and relational structure; provenance and logging studies provide powerful accountability concepts but remain domain-agnostic. This paper brings those strands together for fraud operations and makes them operational through explicit evidence dimensions, a minimum schema, and a validation blueprint.

AUDIT READINESS AS A MEASURABLE CONSTRUCT

Audit readiness is defined here as the measurable capability of a fraud alerting system to reproduce, justify, challenge, and govern each alert decision over time using persisted artefacts. A system is audit-ready when it can answer, at minimum, the following operational questions: Which model and preprocessing version produced this alert? Which threshold or alert-budget policy was active? What explanation was shown at decision time? What action did the investigator take, and can that action be linked to a rationale? What evidence exists that the system was monitored for score drift, explanation drift, and policy change?

This definition deliberately separates audit readiness from both model accuracy and explanation quality. A system may be accurate but not replayable. It may produce concise explanations but fail to preserve them. It may have strong validation at deployment time but weak change control thereafter. By treating audit readiness as a distinct construct, organisations can evaluate governance capability without pretending that traditional predictive metrics capture it.

This distinction is consistent with both governance and standards literature. Documentation, tracking, and challenge are first-class concerns in financial model governance Board of Governors of the Federal Reserve System, and Office of the Comptroller of the Currency (2011), while ISO/IEC 42001 and ISO/IEC 23894 treat AI governance and risk management as organisational capabilities rather than model-only properties International Organization for Standardization (2023a), International Organization for Standardization (2023b).

Six design principles follow from this definition. First, anything that changes the meaning of a decision must be versioned. Second, the explanation shown at decision time must be storable in a form that can be replayed or at least compared to a tolerance-bounded reconstruction. Third, model behaviour and policy behaviour must be recorded separately, because thresholds and alert budgets can change without model retraining. Fourth, review activity must be linked to the original decision artefact instead of being stored as an unrelated case note. Fifth, explanation disclosure must be role-sensitive because full transparency can increase adversarial risk. Sixth, drift monitoring must include both predictive outputs and explanatory patterns when explanations are operationally relied upon.

AUDIT-READINESS RUBRIC

Table 1

Table 1 Six Dimensions of Audit Readiness and Their Minimum Evidence Requirements

Dimension	What it requires	Illustrative measure	Minimum evidence
Explanation persistence	Persist the explanation snapshot that was actually displayed at decision time.	Persistence rate	Alert ID linked to explanation snapshot, explainer version, and timestamp.

Version and threshold traceability	Recover model version, feature contract, preprocessing hash, and active policy for any alert.	Recovery rate	Immutable version references and policy ledger.
Replayability	Recompute or rehydrate the alert package to a defined tolerance.	Replay success rate	Replay harness, equivalence rules, and failure logs.
Investigator usability	Provide concise, role-appropriate reasons that fit review-time constraints.	Median review time; ambiguity rate	Role-based presentation templates and reviewer feedback loop.
Challengeability	Record overrides, reviewer disagreement, and rationale for challenge or escalation.	Challenge-log completeness	Structured override fields and audit trail.
Drift evidence	Track changes in score distribution, explanations, and policy over time.	Drift-evidence completeness	Monitoring records linked to model and policy changes.

The rubric is intentionally evidence-oriented. Each dimension can be audited through concrete artefacts and not merely through policy statements. In practice, organisations may combine the individual measures into a composite score, but the component-level evidence should remain visible because a strong aggregate score can otherwise obscure serious weaknesses in replay, challenge logging, or policy traceability.

AUDIT-READY EXPLANATION PIPELINE

An audit-ready fraud pipeline should produce a compact but durable alert package at the point of decision. The architecture proposed here has six operational stages: event ingestion, versioned feature generation, model scoring, threshold or alert-budget policy application, explanation snapshot generation, and investigator review. What differentiates this pipeline from a conventional fraud stack is the persisted decision artefact store that captures the state needed to replay and govern the alert later.

The persisted artefact store is not synonymous with a feature store, case-management system, or model registry, although it may integrate with all three. Its purpose is narrower and more exacting: preserve the smallest set of evidence needed to reconstruct what the institution knew and showed when it acted on the alert. That evidence must include both machine state and workflow state. If the store omits the threshold policy, a score may be reproducible yet the decision outcome may not. If it omits the explanation snapshot, model state may be reproducible yet reviewer context may be lost. If it omits review actions, the institution may know how the alert was created but not how it was challenged or resolved.

Figure 1

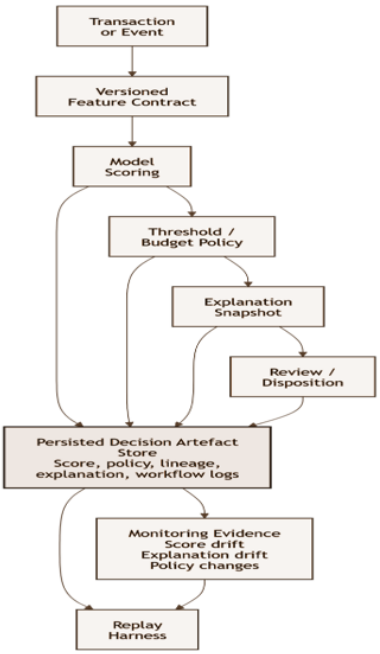

Figure 1 Audit-Ready Fraud-Alert Lifecycle and Persisted Decision Artefact Store

MINIMUM ALERT-ARTEFACT SCHEMA
Table 2

Table 2 Minimum fields for a persisted alert artefact

Field	Layer	Purpose	Sensitivity
alert_id	Workflow	Primary key linking score, explanation, review, and audit logs.	Low
event_timestamp	Event	Anchors temporal ordering, replay windows, and change analysis.	Low
model_version	Model	Identifies scoring model binary or registry entry used at decision time.	Medium
feature_schema_version	Data	Captures the feature contract and column semantics.	Medium
preprocessing_config_hash	Data	Pins the exact transformation and encoding logic.	Medium
threshold_policy_id	Policy	Separates the score from the business rule that converted it into an alert.	Low
score	Model	Stores the raw or calibrated risk output used by downstream policy logic.	Medium
decision	Policy	Records the alert outcome after threshold or budget logic.	Low
explanation_snapshot_id	Explainability	Links to the explanation that was actually presented to reviewers.	Medium
explainer_version_and_params	Explainability	Supports interpretation and bounded replay of explanation behaviour.	Medium
investigator_action	Workflow	Captures disposition, escalation, or override applied during review.	Medium
audit_log_pointer	Governance	Links to immutable access and modification records.	High
retention_policy_id	Governance	Connects the artefact to storage, access, and deletion rules.	Low

This minimum schema is deliberately compact. It does not require storing raw personal data inside every artefact record. Where possible, sensitive identifiers can be tokenised or linked through secure references, while hashes and registry pointers preserve replay and audit utility.

REPLAY SEMANTICS AND CHANGE CONTROL

Replayability should be specified explicitly rather than assumed. Two forms are useful. Deterministic replay requires the same model, feature contract, and explainer configuration to reproduce the original output exactly. Tolerance-bounded replay accepts small numerical differences but requires them to remain within documented bounds. Fraud systems often depend on external services, model calibrators, or updated libraries, so a tolerance-bounded approach may be more practical, but its limits must be published in validation documentation.

Change control is equally important. Thresholds, investigator alert budgets, escalation rules, and customer-treatment policies can materially alter operational outcomes even when the model remains fixed. For that reason, policy objects should be versioned separately from model objects. The decision artefact must preserve both, otherwise post-hoc reviews may wrongly attribute an outcome to the model when the proximate cause was a policy change.

EVALUATION BLUEPRINT

A central weakness in much fraud-XAI work is the conflation of distinct evaluation goals. Predictive quality, explanation quality, workflow utility, and governance capability answer different questions and should therefore be measured separately. The blueprint below is designed to support a future empirical follow-on study without forcing governance claims to rest on non-comparable benchmark results.

For public-data validation, suitable datasets include IEEE-CIS for large-scale tabular fraud, PaySim for simulated financial transactions, and Elliptic for graph-based illicit transaction detection. A credible experimental protocol should use temporal or blocked splits rather than naïve random partitioning, report uncertainty for imbalanced metrics, and disclose the full feature contract, preprocessing configuration, and policy definition used in scoring.

IEEE-CIS is a widely used open benchmark distributed through Kaggle. PaySim was created specifically to address the scarcity of shareable financial transaction data and remains useful for controlled fraud-system experiments Lopez-Rojas et al. (2016). Elliptic extends the evaluation space to illicit-transaction graphs and motivates graph-aware modelling as well as explanation challenges at realistic scale Weber et al. (2019).

Table 3

Table 3 Multi-Layer Evaluation Blueprint

Layer	Primary question	Representative metrics	Evidence type
Predictive quality	How well does the model rank or classify suspicious events?	AUROC, PR-AUC, recall@K, false-positive burden, calibration	Benchmark experiments
Explanation quality	Do explanations reflect model behaviour and remain stable enough to use?	Faithfulness, stability, sparsity, sanity checks	Explainer tests
Workflow utility	Do explanations help investigators review alerts efficiently and consistently?	Median review time, override rate, ambiguity rate	User study or validated simulation
Audit readiness	Can alerts be reconstructed, challenged, and defended over time?	Persistence rate, replay success rate, version recovery rate, challenge-log completeness, drift-evidence completeness	System and audit tests

AUDIT QUESTIONS AND EVIDENCE MAPPING

Table 4

Table 4 Typical Audit Questions and the Evidence Required to Answer Them

Audit question	Required evidence	Failure mode if absent
Which model created this alert?	model_version, feature_schema_version, preprocessing_config_hash	Cannot attribute responsibility or reproduce score
Why did the alert fire at that time?	score, threshold_policy_id, decision, explanation_snapshot_id	Score may be reproducible, but outcome is not
What did the investigator see?	Persisted explanation snapshot and presentation layer	Later rationalisation replaces decision-time context
Was the alert challenged or overridden?	investigator_action, challenge fields, escalation notes, audit logs	No evidence of effective challenge
Did drift or policy change affect this class of alerts?	Monitoring outputs, change-control records, policy history	Root cause analysis becomes speculative

RECOMMENDED EMPIRICAL FOLLOW-ON PROTOCOL

A compact empirical extension of this framework could proceed as follows. Train a baseline fraud model using a temporal split on a public dataset. Generate decision-time explanations with a fixed explainer configuration. Persist all fields in the minimum artefact schema. Introduce controlled changes to policy thresholds, preprocessing versions, or library environments, then run a replay harness to measure reconstruction success, mismatch rates, and failure causes. Finally, assess explanation stability across adjacent time windows and evaluate whether a small group of reviewers can use the explanation views to triage cases faster or with lower ambiguity.

Such a protocol would not need to solve every operational challenge in order to be informative. Even a modest demonstration would materially strengthen the claim that audit readiness is an engineering capability that can be designed, measured, and validated independently of headline benchmark performance.

SECURITY, PRIVACY, AND OPERATIONAL CONTROLS

Persisted decision artefacts create real governance value, but they also increase the sensitivity of the system. Explanation records, feature references, device signals, and investigator notes may reveal behavioural patterns, thresholds, or internal heuristics that should not be broadly disclosed. In adversarial fraud settings, careless transparency can become a route to threshold probing, evasion, or social engineering.

Three design controls are therefore essential. First, explanation disclosure should be role-based. Internal investigators may need rich feature-level reason sets, validators may need configuration-level detail, and external communications may need only controlled reason codes or counterfactual summaries. Second, storage should follow data-minimisation principles. The persisted artefact should carry only the information needed for replay, challenge, and audit. Where possible, direct identifiers should be replaced by secure references, hashes, or tokenised surrogates. Third, access to artefacts should itself be auditable. Read access, export events, policy changes, and replay operations should generate immutable logs with role, timestamp, and purpose fields.

These controls are not merely prudent engineering. They are aligned with external record-keeping and traceability expectations. The EU AI Act requires high-risk AI systems to support automatic logging over the lifetime of the system for traceability and post-market oversight European Union (2024). In anti-money-laundering practice, FATF Recommendation 11 likewise expects record retention sufficient to reconstruct transactions and customer-due-diligence history (FATF, 2025). Taken together, those obligations support an architecture in which persistence is deliberate, access is controlled, and retention is explicitly governed.

Retention policy must be treated as part of system design rather than post-hoc compliance paperwork. Fraud operations often face overlapping obligations: sufficient retention for reconstruction, operational replay, and internal audit, but controlled deletion or archiving when business or legal need expires. A clean design separates operational replay stores from long-horizon regulatory record stores and binds both to explicit retention-policy identifiers inside the artifact schema.

DISCUSSION

The practical advantage of the audit-readiness framing is that it converts governance from a narrative claim into an engineering target. Organisations often describe systems as 'explainable', 'responsible', or 'traceable', but those labels are hard to challenge unless they are tied to evidence. The rubric and schema proposed here provide a way to ask more precise questions: can an alert be replayed, can a policy change be recovered, can an override be explained, and can explanation behaviour itself be monitored over time?

This framing also sharpens the distinction between model governance and operations governance. A fraud model may be well validated at deployment, yet its operational environment may still degrade auditability through undocumented threshold changes, opaque escalation logic, or missing explanation persistence. Conversely, a system may have rigorous lineage and replay controls even while its predictive model still needs improvement. Treating these as separate assessment layers helps reviewers, validators, and engineering teams avoid misleading trade-offs.

For practitioners, the immediate implication is architectural: decision-time evidence must be captured when the alert is generated, not reconstructed ad hoc months later. For researchers, the implication is methodological: governance claims in fraud-XAI papers should increasingly be backed by artefact designs, replay tests, and workflow evidence rather than by explanation screenshots alone.

LIMITATIONS AND FUTURE WORK

This manuscript does not claim empirical performance gains, causal interpretability, or validated human-factors improvements. It is a framework paper that specifies what should be measured and stored if fraud operations are to become genuinely auditable. Future work should therefore extend the design in three directions.

First, a demonstration study should instantiate the artefact schema on at least one open fraud benchmark using a fully disclosed temporal protocol. Second, explanation robustness should be tested through stability checks, perturbation-based analyses, and randomisation-style sanity checks where appropriate. Third, workflow studies should evaluate whether persisted and layered explanation views actually reduce ambiguity, rework, or review time for investigators and validators.

There is also room for deeper domain-specific analysis. Graph-based anti-money-laundering systems raise distinct challenges around neighbourhood explanations, subgraph storage, and privacy-preserving replay. Customer-facing fraud prevention systems raise additional questions about adverse-action communication, contestability, and cross-jurisdictional retention rules. Those extensions are important, but they reinforce rather than weaken the central claim: operational auditability has to be designed as a first-class property of the system.

CONCLUSION

Explainable fraud models are not automatically audit-ready fraud systems. In operational settings, defensibility depends on whether an institution can persist, reconstruct, challenge, and monitor the full decision context of an alert as models, policies, and workflows evolve. This paper has argued for a shift in emphasis from transient explanations to governed decision artefacts and has provided the technical building blocks for that shift: a measurable rubric, a minimum schema, a lifecycle architecture, and an evaluation blueprint.

The framework is deliberately pragmatic. It does not require organisations to abandon existing fraud models or explanation methods. Instead, it identifies the evidentiary and architectural layers that must be added if those models are to support trustworthy

investigation, model governance, and audit over time. That makes audit readiness both a research agenda and an implementable systems requirement.

ACKNOWLEDGMENTS

None.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity Checks for Saliency Maps. Advances in Neural Information Processing Systems, 31.

Board of Governors of the Federal Reserve System, and Office of the Comptroller of the Currency. (2011). Supervisory Guidance on Model Risk Management (SR 11-7). Author.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Financial Action Task Force. (2025). International Standards on Combating Money Laundering and the Financing of Terrorism and Proliferation: Recommendation 11 on Record keeping. FATF.

Foalem, P. L., Da Silva, L., Khomh, F., Li, H., and Merlo, E. (2025). Logging Requirement for Continuous Auditing of Responsible Machine Learning-Based Applications. Empirical Software Engineering. Advance online publication. https://doi.org/10.1007/s10664-025-10656-8

International Organization for Standardization. (2023a). ISO/IEC 42001:2023 Information Technology—Artificial Intelligence—Management System. ISO.

International Organization for Standardization. (2023b). ISO/IEC 23894:2023 Information Technology—Artificial Intelligence—Guidance on Risk Management. ISO.

López-Rojas, E. A., Elmir, A., and Axelsson, S. (2016). PaySim: A Financial Mobile Money Simulator for Fraud Detection. In Proceedings of the 28th European Modeling and Simulation Symposium (EMSS).

Lundberg, S. M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Office of the Comptroller of the Currency. (2021). Model Risk Management: Comptroller's handbook. Author.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Singh, J., Cobbe, J., and Norval, C. (2019). Decision Provenance: Harnessing Data Flow for Accountable Systems. IEEE Access, 7, 6562–6574. https://doi.org/10.1109/ACCESS.2018.2887201

Wachter, S., Mittelstadt, B., and 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. https://doi.org/10.2139/ssrn.3063289

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., and Leiserson, C. E. (2019). Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics (arXiv No. 1908.02591). arXiv. https://doi.org/10.48550/arXiv.1908.02591

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. Advances in Neural Information Processing Systems, 32.

Zafar, U., and Wu, F. (2026). Methodological Challenges in Explainable AI for Fraud Detection: A Systematic Literature Review. Artificial Intelligence Review. Advance online publication. https://doi.org/10.1007/s10462-026-11516-7

Zhou, Y., Li, H., Xiao, Z., and Qiu, J. (2023). A User-Centered Explainable Artificial Intelligence Approach for Financial Fraud Detection. Finance Research Letters, 58(Part A), Article 104309. https://doi.org/10.1016/j.frl.2023.104309