Lee Chiaw Boon (2026). Operationalising Responsible AI Governance in Financial Crime Compliance: An Evidence-Based Control Framework for Financial Institutions. *Academos Journal, Vol. 2, No. 3,* 2025, 33-45

Original scientific paper

UDC: [004.8:343.851]:[343.9.024:336.7]
DOI: https://doi.org/10.65921/wg5rcn92

Received:24.07.2026.
Revised: 25.08.2026.
Accepted: 30.08.2026.

Operationalising Responsible AI Governance in Financial Crime Compliance: An Evidence-Based Control Framework for Financial Institutions

Lee Chiaw Boon[1]
[1]Golden Gate University, San Francisco, California, USA; mail: sitasia@hotmail.com

Abstract: Purpose: The deployment of artificial intelligence in financial institutions now focuses more on transaction-monitoring alerts and investigative support, but responsible-AI principles at a higher level do not address the evidentiary requirements for reconstructing decisions. This paper provides an auditable governance solution to this challenge. **Methodology**: Conceptual design science research applies framework synthesis and thematic analysis to the Monetary Authority of Singapore FEAT principles and Veritas methodology, the NIST AI Risk Management Framework and Generative AI Profile, Financial Action Task Force technology guidance, the European Union Artificial Intelligence Act, model risk guidance, and explainability and human factors literature. **Results**: The Governance-to-Evidence Responsible AI Framework maps the mandate and ownership, data and fairness, model validation, decision orchestration, human accountability, and continuous assurance control domains to decision rights, controls, and retained evidence. An illustrative transaction-monitoring use case demonstrates how controls depend on consequence. **Conclusions**: The responsible adoption of AI requires the governance of the entire decision-making process, rather than only the accuracy of models and human-in-the-loop. **Recommendations**: Institutions should consider use case classification, independent validation, user-oriented explanation, decision logging, third-party controls, sampling for closure, and suspension controls. **Additional data**: The framework is conceptual; there were no human participants, no confidential data, and no production results.
Keywords: financial crime compliance; anti-money laundering; FATF risk-based assessment; AI governance; human oversight; Governance-to-Evidence Responsible AI Framework.

Introduction

Problem context and transaction-monitoring scope

Detecting complex forms of financial crime across more customers, transactions, devices, and counterparties requires more sophisticated alerting technology. Transaction monitoring on a rules-based engine is still necessary but generally results in big populations of alerts that need to be reviewed manually. Machine learning can prioritize alerts, recognize non-linear patterns, and show relationships that cannot be recognized by static rules. Artificial intelligence can find procedures, synthesize case documents, and draft narratives. The application of interest in this paper is alert prioritization: Alerts are generated by existing rules, the AI model estimates their investigative priority, and human operators decide on their further action.

There are risks on both sides of a decision. A false positive will slow down or harm a legitimate customer; a false negative will let suspicious transactions go unnoticed. AI can affect the order in which alerts are investigated, which investigations will get less attention, and what evidence will come into play for a decision maker. Accuracy will not be sufficient to control these effects. Purpose, data, threshold, human authority, explanation, record-keeping, change control, and challenge should all be governed as part of the decision process.

Transaction monitoring is a good case study for responsible-AI governance since the algorithm works in combination with other mechanisms in a regulated process of customer-risk assessment, sanctions controls, investigation decision-making, quality assurance, suspicious transaction reporting, document retention, and oversight. A prioritization score may be just advisory in theory, but its place in the process may determine which alerts are to be considered

in case of scarce investigative resources. Responsible governance then requires paying attention not only to the technical characteristics of AI but to how it influences the process, including such parameters as function-specific influence, data availability and quality, threshold, human operator's authority, transparency of the AI process, record-keeping, and change control.

The same issue arises regarding efficiency claims. Reducing the number of alerts, decreasing the time of investigations, or reducing the backlog can be important, but each of these measures is insufficient without understanding of residual risk and the quality of decisions made. The system might become more efficient on average while accumulating more risk within one customer segment or a new typology. Alternatively, it could increase efficiency by delegating verification tasks to quality assurance team members. Therefore, responsible operating model must define efficiency as sustainable improvement in detection, investigation quality, impact on customers, resilience, and cost.

Conceptual foundations and research gap

Responsible-AI guidance centres on the principles of fairness, transparency, accountability, privacy, robustness, and human oversight (Jobin et al., 2019). In finance, the Monetary Authority of Singapore (MAS, 2018) captures the FEAT principles through fairness, ethics, accountability, and transparency, while the Veritas method helps with the actual assessment (MAS, 2022a, 2022b). Risk management at the National Institute of Standards and Technology (NIST, 2023) consists of Govern, Map, Measure, and Manage steps; the Generative AI Profile covers the risks such as confabulation, information integrity, privacy, security, and human–AI interaction (Autio et al., 2024). Financial Action Task Force (FATF, 2021) encourages technology that enhances AML/CFT effectiveness without disrupting the risk-based approach. These guiding documents give direction but do not explicitly describe how a financial institution will show that certain AI-assisted result was approved, reviewed, and monitored. Thus, the operational gap is the gap between the principle and the proof. While the policy states that the AI should be fair and accountable, the auditor still needs to know which model and version of data was used, which threshold, whose review was performed, why there was an escalation (or not), and what happened when the limitations were identified.

The EU Artificial Intelligence Act strengthens risk management, documentation, record-keeping, human oversight, accuracy, robustness, and cybersecurity of covered systems (European Parliament & Council of the European Union, 2024). Guidance on model-risk and operational resilience further stresses the importance of challenge process, change control, governance, and recoverability (Basel Committee on Banking Supervision, 2021; Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011). These documents are not considered as equally binding; they provide different control needs for the synthesis.

The socio-technical systems theory states that the outcome is a product of joint design of the technology, people, task, and organization (Bostrom & Heinen, 1977). The institutional theory shows how the regulatory and professional pressures shape the governance structure (DiMaggio & Powell, 1983). The dynamic-capabilities theory concentrates on sensing and reconfiguration of the resources (Teece et al., 1997). Taken together, the above-mentioned theories suggest that the responsible adoption should be based on technical assurance, institutional accountability, and operating capability that adapt to changing crime typology, data, and models.

The second gap is related to the unit of governance. Many responsible-AI publications focus on the model, whereas the regulated outcome is an artefact of the combined design of the data pipelines, business rules, model scores, thresholds, interfaces, people, processes, vendors, and fallback arrangements. While the model remains unchanged, threshold changes, data source

switch, interface redesign, or staff changes significantly alter the result. At the same time, minor changes in the model might require increased review if the population routed to the investigation changes. Thus, GERAI considers the decision pathway as the object of governance and models as components of that pathway.

The third gap is related to the symmetry of evidence. Most of the governance programs keep more information about the escalated cases since escalation leads to investigation and reporting. However, AI prioritization leads to one of the greatest blind spots in the form of alerts that do not require increased attention. When an institution cannot recreate the non-escalation pathway, it is impossible to understand whether an adverse outcome is caused by the model score, the threshold, the interaction of the rules, missing data, interface design, reviewer workload, or authorized override. The symmetric evidence does not mean that every case needs to be independently reviewed; it means that there should be enough version-controlled information and risk-based sampling of the pathways.

Research question and contribution

The main research question is: how may financial organizations apply responsible AI principles as auditable controls for AI-supported financial crime compliance? Some subordinate research questions will address special transaction monitoring risks, proportional deployment of AI and evidence needed to reproduce escalation and non-escalation decisions. The paper provides contributions in the form of GERAI Framework, standards mapping, illustrative case study, and testable propositions. GERAI is not a substitute for AML/CFT programs or model risk management but rather links them through decision rights and evidence retained.

Key terms and abbreviations

AI stands for artificial intelligence; AML/CFT for anti-money laundering and countering the financing of terrorism; FATF for Financial Action Task Force; FEAT for fairness, ethics, accountability, and transparency; GERAI for Governance-to-Evidence Responsible AI Framework; MLRO for money laundering reporting officer; NIST AI RMF for NIST Artificial Intelligence Risk Management Framework; and XAI for explainable artificial intelligence. In the current research, evidence object refers to a retained artifact that provides evidence of a control or decision, whereas decision orchestration refers to the thresholds, routing rules, overrides, and fallbacks associated with the model output.

Materials and methods

Research design

This is a conceptual, non-empirical study using design science methodology. It is fitting in design science methodology because it involves creating and assessing artefacts to solve known organisational problems (Hevner et al., 2004; Peffers et al., 2007). In this case, it is not software but a governance control framework. It has the following design goals: proportionality, lifecycle consideration, traceability, accountable human intervention, challenge, compliance governance integration, and applicability to internally created and third party systems.

The design science process involved six connected activities based on a proven methodology: problem identification, definition of design goals, artefact construction, demonstration through illustrative use case, conceptual evaluation against source requirements and limitations and propositions communication. The problem to be solved is the existence of the gap between the principles and their reviewable evidence. Design goals were obtained from the recurring requirements within the selected governance tools. Then, the construction was done in iterations: candidates domains were combined when they referred to the same decision right

and divided when there was different owners, controls and evidence required. Next, the transaction monitoring example was used for revealing gaps in controls and accountability.
The artefact was evaluated through traceability, completeness, internal consistency, proportionality and implementation. Traceability questioned whether all domains were linked to a recognised governance requirement. Completeness asked whether the framework included design, deployment, operation, change, incident and retirement considerations. Internal consistency questioned whether the decision rights corresponded with the control ownership and evidence. Proportionality questioned whether the approach was able to differentiate insignificant assistance from material automation. Implementation questioned whether the evidence suggested was feasible by ordinary governance, model risk, compliance, technology and audit procedures. This is an analytical evaluation and not empirical proof of effectiveness.

Search strategy and selection criteria

A literature search for relevant academic articles was undertaken between February and August 2026 and updated on 10 August 2026. Academic sources were found via Google Scholar and publishers' archives; regulatory and standards sources were obtained from the MAS, NIST, FATF, EU, BIS, and United States banking supervisor archives. Search combinations used were: ("responsible AI" AND finance), ("AI governance" AND AML), ("transaction monitoring" AND explainability), ("human oversight" AND automation bias), fairness without demographics, counterfactual explanations, and (generative OR agentic AI) AND governance. Inclusion criteria required sources to be published/accepted in English, relevant to a specific governance requirement/control design, and either authoritative regulation or standards, or peer-reviewed research that served as foundational or application-specific literature. Theory predating the search period was included if necessary. Duplicate sources, commentaries without methodology, vendor marketing, and irrelevant sources to the artefact were removed. There remained 23 sources after synthesis. As the objective was artefact development rather than estimating an intervention effect, the review is purposive and not systematic in nature, with no meta-analysis undertaken.

Framework synthesis and conceptual evaluation

The material selected was examined through thematic content analysis and framework synthesis. Requirements in regard to governance, fairness, transparency, validation, human supervision, documentation, third party involvement, and monitoring were coded and aggregated. Each of these themes was checked against four questions: What risk is mitigated? To whom the decision rights belong? What kind of control enables the requirement to work? What kind of evidence allows future reconstruction and validation? The resulting areas were compared to FEAT and the NIST AI RMF in terms of coverage and were then used to check coherence and usability in case of transaction monitoring. The evaluation is conceptual; neither any claims about production effectiveness nor any claims of statistical validation are made.
Coding was done at the level of actionable requirements and not just keywords. Thus, accountability references were separated into ownership, approval authority, review competency, escalation, and evidence preservation since each of these means something different in terms of control. Similarly, transparency was divided into data lineage, model documentation, local interpretation, user communication, and management reporting. Human supervision was separated not by its presence or absence but rather by reviewer's access to information, time, competence, authority, and ability to influence the outcome. That allowed avoiding counting one policy as control for several different requirements.
The crosswalk has been designed in order to find a design tool and not to prove that all frameworks cover the same area and have the same legal status. FEAT provided finance-

specific principles; Veritas provided guidance in terms of assessment; the NIST AI RMF provided lifecycle risk functions; FATF placed the technology in the context of risk-based AML/CFT; the EU AI Act supported the ideas of documentation and oversight; and banking guidelines provided guidance in terms of validation, change control, and resilience. Necessary differences were preserved. The synthesis, therefore, helps designing institutional control but does not give any jurisdictional legal opinion or determine if an application belongs to the statutory risk category.

Results

The GERAI artefact

GERAI operates on a basic design rule: a principle is auditable if it has an accountable actor, a control, and an evidence object to retain. Figure 1 illustrates the chain of logic from external principles to governance outcomes. The framework comprises six control domains, which are not phases in an ad hoc project.

Mandate and ownership define permitted purpose, prohibition of uses, autonomy, risk classification, approval and suspension authority. Data and fairness manage provenance, quality, relevance of features, proxy risk, missingness, and differential effects. Validation verifies discrimination, calibration, stability, sensitivity, explainability, security, resilience, and user interface. Decision orchestration controls thresholds, routing, override conditions, mandatory escalation, segregation of duties, and fallback. Human accountability includes reviewers competent in information and process, having enough time and authority to disagree. Continuous assurance monitors drift, data quality, performance, overrides, complaints, quality-assurance results, incidents, and typology evolution.

Table 1 correlates each domain with the evidence objects that management, validation, internal audit and supervisors can sample. GERAI enables prospective accountability prior to deployment, concurrent accountability during deployment, and retrospective accountability in case an outcome is disputed.

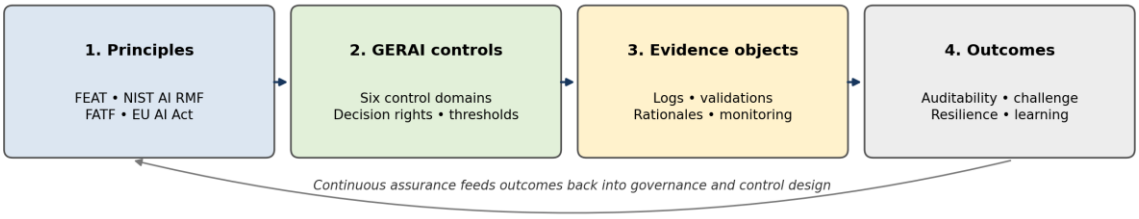
The six domains operate as a system of control. Mandate defines what the use case is allowed to do; data and fairness specify what information may influence the decision; validation specifies under what conditions trust is warranted; orchestration transforms output into workflow; human accountability controls judgment and challenge; and continuous assurance decides whether the original reasons for approval are still valid. Weakness in one domain cannot be made up for by strength in another. For instance, a well-calibrated model does not cover the problem of an unauthorized use, and a human approval step does not cover the problems of an uncertain interface and of workload making the review perfunctory.

Evidence objects need to be designed in tandem with controls. A threshold approval control is incomplete if the organization cannot find out about the approved value, its validity period, reasons, covered population, approving entity, dependencies of validation, and rollback condition. Similarly, a human review control is incomplete if the record does not include information presented, relevant warnings issued by the system, reviewer's decision and rationale where needed, and any escalation and override actions. The design of evidence objects specifies the retention, access, integrity, and correlation needed to preserve usability of the record without generating uncontrolled repositories of sensitive information.

GERAI uses three assurance horizons. Prospective assurance focuses on purpose, risk classification, data suitability, testing, approval, training and fallback prior to deployment. Concurrent assurance documents versions, inputs, scores, decision rules, human actions and exceptions during the decision-making process. Retrospective assurance samples outcomes, investigates incidents, assesses drift and complaints, and provides feedback into thresholds,

training, validation or suspension. The horizons make the governance dynamic: evidence is not only stored for audit but also used to check whether the underlying assumptions of the approval are still valid.

Figure 1: Governance-to-evidence logic of GERAI



Source: Author's synthesis.

Table 1: GERAI control domains and core evidence

Domain	Control objective	Core evidence
Mandate and ownership	Define purpose, autonomy, owners, approvals, and suspension	Use-case charter; risk classification; RACI; approval
Data and fairness	Control provenance, quality, feature rationale, proxies, and outcomes	Lineage; data tests; feature register; segment analysis
Model validation	Establish fitness, limitations, robustness, and explanation quality	Model card; validation report; limitations log
Decision orchestration	Govern thresholds, routing, overrides, escalation, and fallback	Decision log; threshold approval; fallback test
Human accountability	Enable competent, authorised, meaningful challenge	Rationales; overrides; QA results; training records
Continuous assurance	Detect drift, incidents, and control deterioration	Dashboard; incidents; revalidation; board reporting

Source: Author's synthesis based on MAS (2018, 2022a, 2022b), NIST (2023), and FATF (2021).

Proportional controls and explainability

Control Intensity relies on consequence, autonomy, uncertainty, data sensitivity and customer or regulatory effect. An alert retrieval assistant which identifies approved procedures but cannot make changes to a case involves less consequence than an automatic alert closure or customer limitation. This is demonstrated in Table 2. Less consequence never indicates ungovernance: access, source integrity, testing, logging, and ownership are all still required.

Interpretability is context-dependent, depending on audience and decision being made. LIME explains predictions locally (Ribeiro et al., 2016) whereas SHapley Additive exPlanations (SHAP) explain how the prediction was attributed (Lundberg & Lee, 2017). Counterfactual explanations provide changes that would be associated with a different prediction (Wachter et al., 2017). All of these are tools and do not provide justification or evidence of causation and correctness. The validator will require stability and fidelity tests; the investigator requires concise risk indicators and source evidence; the board requires limitations, trends, and incidents. User-centred design should start with the question that is to be asked and then test if explanations help with a proper challenge to the system, not passive acceptance (Liao et al., 2020). In consequential decisions, an interpretable model may be better than the post hoc explanation of the black box (Rudin, 2019).

Fairness in AML/CFT does not involve statistical parity irrespective of risk. The differentiation in terms of risk might be justified when explained and proportionate. When protected attributes cannot be legally or practically used by the institution, they should not presume that fairness

cannot be evaluated. They can perform quality and error rate checks across available operational segments, do worst-case sub-group analysis and proxy analysis, check complaints and overrides, and employ techniques developed for evaluating fairness without using demographic variables as additional diagnostic techniques (Lahoti et al., 2020). Such techniques must be validated properly.

Table 2: Proportional governance by use-case consequence

Use case	Consequence	Minimum control response
Retrieve approved procedure	Lower	Approved sources; access control; citation; answer logging; owner
Summarise case evidence	Moderate	Source traceability; factuality check; human edit; no unsupported claims
Prioritise alert queue	High	Independent validation; segment tests; overrides; closure sampling; fallback
Close alert or restrict customer	Very high	Named human approval; enhanced validation; complete decision record; appeal/escalation

Source: Author's synthesis.

Illustrative transaction-monitoring application

Take a model that maximises detection of alerts produced by current rules. Mandated prohibitions preclude automatic reporting of suspicious transactions, closing accounts and reusing model predictions. Compliance responsibility belongs to the MLRO, not the owner of the model who cannot independently tune a threshold affecting the scope of the investigation. There are data lineages from source transactions and customer attributes to approved features with a rationale for their relationship with financial crime. Historical dispositions are assumed noisy labels, not an objective truth.

The validation tests the model against the current process and any alternative process. They include recall of quality-assured escalations, precision at proposed thresholds, calibration, stability, sensitivity to missing data, segment errors, scenario tests, service failure, and investigator usability. Volume reduction is only counted as success when residual risk and quality of investigations are acceptable. Low-priority alerting option comes into play only after governance accepted the residual risk; sanctions escalation, law-enforcement requests, specified high-risk typologies, and missing critical data always take precedence over the score. In case of a model failure, rules-based backup process is used after testing.

Decision log keeps alert identifier, time stamp, referenced inputs, version of the model and feature pipeline, score, threshold, explanation, override trigger, assigned reviewer, disposition, rationale and approvals. Quality assurance purposely uses low-priority closures because governance needs to explain why the alert was not escalated, not only why another alert was. Human oversight is calibrated by consequences, and user interface presents uncertainty and evidence rather than score as the answer. This approach is suggested as bounded automation architecture and will have to prove itself empirically.

The implementation starts with establishing baseline for the current process. Institution describes rule coverage, alert volumes, age of queues, investigator resources, quality assurance findings, gaps in typologies knowledge, distribution of outcomes. Baseline makes comparison possible and protects new model from taking credit for improvements resulting from other factors, like tuning rules or hiring more people. At the same time, it helps to establish controls that need to operate independently of the model, including sanctions escalation, law-enforcement requests, mandatory scenarios and manual referral routes. Approval criteria must be defined before the pilot to avoid motivated interpretation of the results.

A controlled pilot can utilise shadow operation prior to live application of the model. Scores and explanations are calculated and logged, but the current process still controls the routing of

alerts. The reviewers evaluate recommendations against actual cases, register any disagreements, and point out flaws in explanations or user interface. When live usage starts, its scope can be limited by customer population, typology, geography or queue. Expanding into new territories happens based on pre-defined criteria, not just feeling of success. Rollback conditions include material drift, data failure, unexplained segment degradation, excessive overrides, missed mandatory scenarios, service instability, and evidence gaps that make reconstruction impossible.

Population for low-priority alerts requires explicit quality assurance because outcomes are likely to appear after some time. Samples can be taken at random, risk-based, near-threshold cases, new customer/new product segments, unusual missing data patterns, and alerts that were impacted by recent changes. Reviewers evaluate whether there is enough evidence to close the alert and whether the model or workflow prevented escalation of this particular case. Findings should be traced to root cause and corrective actions, not just presented as aggregate error rate. Information that arrives later, such as escalation or law-enforcement response, should be considered when possible.

Third-party, generative, and agentic AI governance

Third-party procurement does not shift responsibility. Due diligence and contractual arrangements should cover data rights, confidentiality, security, documentation of training-data and model where available, validation access, performance assurances, incident reporting, subcontracting, change notification, continuity, audit privileges, and exit provisions. Institutions should document components they cannot independently evaluate and assess whether compensating controls mitigate residual risk to an acceptable level. Vendor model changes should not occur without internal change control approval.

For generative AI, evidence should include the use approval, sources retrieved, prompt and model version where proportional, citations, automated and human reviews, edits and approval. Unsubstantiated language should not enter a regulatory report just because it sounds plausible. Agentic multi-step workflows need controls over permissions, action limitations, memory, delegation, intermediate decision-making, and termination of process. Logs should cover tool invocation and hand-offs; high-impact actions require human oversight; and independent testing should validate prompt injection, unauthorized tool invocation, cascading error, and inability to terminate. These extensions extend the NIST governance principles to higher levels of system autonomy and not assume that autonomy is safe (Autio et al., 2024; NIST, 2023).

Vendor governance record should specify what the institution knows, what vendor guarantees, what an independent third party has tested and what is opaque. Opaque components increase rather than eliminate the necessity for outcome testing, contractual notification, interface management, monitoring, and exit strategy. Significant changes made by vendor should be traced to internal validation requirements, such as changes in training data, model architecture, safety controls, hosting provider, subprocessor, data location, or service agreement terms. Concentration and continuity analysis should cover availability of evidence, configurations, prompts, and decision history in case of service suspension or replacement.

Generative systems also complicate the distinction between retrieval and creation. A response approved through the process might still omit a qualification, combine sources inappropriately or express a tentative judgment with undue certainty. As a result, controls should confirm that the source is authoritative, up-to-date, aligned with the citation, consistent with the facts, avoids prohibited data, and the distinction between summary and decision-making. In case where generated content serves as basis for a regulatory report or customer correspondence, the human approving the document should have access to the relevant source materials and edits. The documentation kept for the purpose should be proportional to consequence and avoid retention of sensitive prompts any longer than necessary.

Agentic systems also introduce additional risks related to state and delegation. Agent may choose the tool to use, create a subtask, invoke another agent, modify plan, or take action based on intermediate results. Governance framework covers the whole execution graph: who initiates, objective, permissible tools, actors delegating, data accessed, intermediate decisions, approvals, termination criteria, and the final action. Least privilege, transaction limits, sandboxing, timeouts, and human review limit impact. Evaluation includes adverse and failure scenarios like conflicting instructions, tool unavailability, corrupt memory, cyclic delegation, repeated action, and partial execution. The final answer is not enough evidence if the process can affect external records or case status.

Metrics and operating model

Third-party procurement does not relieve anyone from their responsibilities. Due diligence and contracts should address data rights, confidentiality, security, training data and model documentation when applicable, validation access, performance guarantees, incident reporting, subcontracting, change notification, continuity, audit rights, and contract termination terms. The organizations should document the components that are outside their capability to evaluate and determine whether mitigating controls reduce the remaining risk to an acceptable level. Changes to vendor models should not happen without internal change control approval.

When dealing with generative AI, the evidence base should include the use approval, sources used, prompt and model version when proportional, citations, automated and human reviews, edit and approval. Unsupported language should not be included into a regulatory report only because it sounds like a plausible statement. Agentic multi-step workflows will require controls around permission, action restriction, memory, delegation, intermediate decision making, and process termination. Log should capture tool invocation and hand-offs; high impact actions need to have human approval; and independent testing should validate prompt injection, tool invocation without authorization, cascading errors, and inability to stop the process. These additions extend the NIST governance principles to the higher levels of the system autonomy and do not assume that autonomy is safe (Autio et al., 2024; NIST, 2023).

The vendor governance record should include what is known by the institution, what the vendor guaranteed, what was tested by the independent third party and what is opaque. Opaque components increase rather than decrease the necessity for outcome testing, contract notification, interface management, monitoring and exit strategy. Major changes made by the vendor should be tracked to the internal validation requirement, such as changes in training data, model architecture, safety controls, host, subprocessor, data location, and service agreement terms. Concentration and continuity analysis should be done in relation to the availability of evidence, configurations, prompts and decision history in the event of suspension or replacement of the service.

Generative systems make the difference between retrieval and creation more ambiguous. An output that was approved through the governance process could still omit a qualification, combine the sources improperly, or state a tentative opinion with unwarranted confidence. As a result, the controls should confirm that the source is authoritative, current, aligns with the citation, is factually correct, does not use prohibited data, and the distinction between summarization and decision-making. When the output becomes the basis for a regulatory report or customer communication, the person approving it should have access to the source material and be able to make the necessary corrections. Documentation that needs to be maintained for this purpose should be proportional to the consequence and should not retain sensitive prompts longer than needed.

State and delegation are new risks associated with agentic systems. The agent could decide on the tool usage, initiate a subtask, call another agent, modify the plan, or act based on the intermediate result. Governance framework should cover the entire execution graph, including

who started the process, goal, permitted tools, actors delegating, data used, intermediate decisions, approvals, termination criteria, and the final action. Least privilege, transaction limits, sandboxing, timeouts, and human review reduce the impact. Evaluation should account for adverse and failure scenarios like inconsistent instructions, unavailable tools, corrupted memory, cyclic delegation, repeated actions, and incomplete execution. The answer itself is not enough evidence if the process has an impact on the external records or case status.

Table 3: Balanced governance scorecard

Dimension	Illustrative indicators	Governance question
Effectiveness	Recall; escalation quality; typology coverage; review time	Is detection improving within residual-risk appetite?
Fairness	Segment errors; delays; overrides; complaints	Is differential treatment risk-relevant and justified?
Reliability	Drift; missingness; outages; open findings	Does the system remain within validated limits?
Human oversight	Override patterns; QA defects; review time; competence	Can reviewers understand and challenge outputs?
Operational value	Queue age; hours redeployed; case quality	Does value remain consistent with control quality?

Source: Author's synthesis.

Discussions

Contribution and theoretical propositions

The GERAI framework contributes in three ways. First, it reconceptualizes responsible AI as an evidence architecture. Documentation is not merely administrative detritus but the means by which to challenge, assure, and learn from decisions. Second, it takes non-escalation as a governance outcome. Prioritization places risks in the less-attended population, making closure sampling and delayed-outcome analysis critical. Third, it couples responsibility with bounded efficiency. Explicit constraints and evidence enable limited forms of automation rather than an either/or between innovation and control.

In turn, the framework suggests four propositions for future research. P1: There is a positive association between greater governance-to-evidence traceability and auditability. P2: Data governance and independent validation moderate the association between AI capability and compliance effectiveness. P3: Meaningful human oversight mediates the association between model recommendation and accountable outcomes. P4: Continuous assurance contributes to operational resilience through early detection and treatment of drift, data failure, and control breakdown. The four propositions bridge socio-technical joint design, institutional pressure, and dynamic reconfiguration.

The evidence-architecture approach explains why governance maturity should not be measured on the basis of the existence of policies and committees alone. Two institutions could implement identical principles yet vary considerably in the extent to which they are able to reconstruct decisions, locate accountable owners, prove that constraints have been followed, and respond to degradation. This framework proposes a maturity path from the development of principles to controls to linked evidence and analysis, and even beyond to continuous analysis and reconfiguration. The highest level is not maximal automation. It is the consistent ability to change or terminate automation when evidence no longer supports it.

In addition, the framework implies some testable constructs. Traceability may be measured as the percentage of decisions for which purpose, version, threshold, owner, rationale, and assurance status are retrievable. Meaningful human oversight may be measured as the extent to which reviewers comprehend explanations, disagree appropriately, override recommendations effectively, escalate problems, and the time and data that are available to

them. Continuous assurance may be measured as the time required to detect, investigate, and remedy problems of drift and control breakdown. Further research may test the association between these constructs and audit findings, decision quality, and incident severity.

Human factors, practical tensions, and proportional implementation

A nominal human-in-the-loop approach is inadequate where throughput goals, persuasive interfaces, or absence of authority pressures agreement. Automation could be abused through over-reliance or disrespected through under-reliance (Parasuraman & Riley, 1997). Supervision should therefore be tested via disagreements, quality outcomes, speed, and the ability to get evidence and elevate limits. Explanation fidelity, rather than visual appeal, should be evaluated using appropriate standards (Doshi-Velez & Kim, 2017).

Retention of evidence can clash with issues of privacy, minimization, security, and retention duties. The segment analysis may be limited by legal access and sample size. Explanations may reveal sensitive detection processes or promote gaming. The above contradictions need documentation of balanced decisions, access controls, retention schedule, aggregation, and risk tolerance—rather than the assumption that transparency equals unlimited disclosure.

The basic standard for smaller financial institutions and fintech businesses would include: AI inventory, accountable owner, statement on use allowed, data records, proportionate validation, logging of decisions with material impact, human escalation, incident management, oversight and monitoring, and third party exit plans. Independent expert knowledge can be secured from external parties; however, the management is responsible for accountability. Additional activities by bigger organizations can include specialized validation, automated evidence pipeline, cross-use-case analytics, and assurance. Proportionality impacts the depth and frequency of controls, not the existence of essential controls.

Limitations and future research

GERAI itself is a theoretical construct, emerging from a purposeful review of the literature and not yet validated through empirical studies in production environments. The example application is a simplification of numerous models, pipelines, vendors, investigation teams, and reporting regimes. Inadequate ground truth limits assessment of AML/CFT systems and the preservation of evidence introduces additional privacy and security risks. The framework does not answer jurisdictionally specific legal questions.

Further research will rely on expert interviews and comparative case studies, followed by controlled pilots that evaluate investigator challenge, quality of decisions, closure certainty, and auditability. Survey research could be conducted to examine the hypotheses about relationships between the maturity of the governance, the pressure from regulators, adoption and operational resilience. Additional work is required on explanation fidelity, fairness without the presence of protected attributes, delayed outcomes, privacy-preserving collaborative analytics, and governance of agentic systems involving delegation and tools.

Empirical evaluation should not equate the deployment to one intervention. Comparative designs would allow investigating how variations in threshold values, explanations, workload, training of reviewers, and evidence design impact outcomes while keeping the underlying model constant. The mixed-methods approach would be especially useful: the quantitative analysis can help to measure error rates, delays, overrides and drift, and interviews and observations can identify whether reviewers understand explanations, perceive challenges to recommendations, and develop workaround strategies. The research designs should include populations of low-priority and non-escalated cases whenever possible because otherwise the bias of exclusive focus on confirmed cases would replicate the very asymmetry that GERAI aims to address.

The cross-institution research can investigate proportionality in community banks, fintech firms, multinational banks and outsourced compliance. Among other things, such research should evaluate the adequacy of external validation to ensure independence of small institutions, what evidence objects can be standardized across vendors and how the expectations of supervisory authorities vary depending on advisory or autonomous function of an institution. For the generative and agentic applications, the longitudinal design is required to separate the effect of pilot performance from the effects of typology change, staff turnover, vendor releases and production pressure.

Conclusions

What matters is not whether an organization utilizes AI, but whether it can control its integrated system of data, models, rules, people, vendors, and decisions. GERAI takes well-established guidelines on responsible AI and applies them across six domains of control linked to responsible actors, controls, and evidence objects. It includes the focus on decision orchestration, meaningful human challenge, third party governance, and non-escalation path, expanding the scope of governance beyond model performance.

Practitioners have a proportional way of implementing principles through GERAI. For supervisory bodies, it serves as the source of challenge. For researchers, it serves as the testable artifact and propositions. The responsible implementation must be evaluated according to the ability of the organization to understand what happened, why it was possible, who was responsible, what controls were used, and what lessons were learned from it. Empirical validation is the logical step forward.

Acknowledgements

The author appreciates the application of generative artificial intelligence technologies in structuring, initial writing, language correction, and formatting consistency. The author has personally edited and revised the paper and ensured that all references cited are accurate.

Conflicts of Interest

The author declares no conflict of interest.

References

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., & Hall, P. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Basel Committee on Banking Supervision. (2021). Principles for operational resilience. Bank for International Settlements. https://www.bis.org/bcbs/publ/d516.htm

Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR Letter 11-7). https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: A socio-technical perspective, part II: The application of socio-technical theory. MIS Quarterly, 1(4), 11–28. https://doi.org/10.2307/249019

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review, 48(2), 147–160. https://doi.org/10.2307/2095101

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Financial Action Task Force. (2021). Opportunities and challenges of new technologies for AML/CFT. https://www.fatf-gafi.org/content/dam/fatf-gafi/guidance/Opportunities-Challenges-of-New-Technologies-for-AML-CFT.pdf

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., & Chi, E. H. (2020). Fairness without demographics through adversarially reweighted learning. Advances in Neural Information Processing Systems, 33, 728–740. https://proceedings.neurips.cc/paper/2020/hash/07fc15c9d169ee48573edd749d25945d-Abstract.html

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3313831.3376590

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Monetary Authority of Singapore. (2018). Principles to promote fairness, ethics, accountability and transparency in the use of artificial intelligence and data analytics in Singapore's financial sector. https://www.mas.gov.sg/-/media/MAS/News-and-Publications/Monographs-and-Information-Papers/FEAT-Principles-Final.pdf

Monetary Authority of Singapore. (2022a). FEAT principles assessment methodology. https://www.mas.gov.sg/-/media/mas-media-library/news/media-releases/2022/veritas-document-3---feat-principles-assessment-methodology.pdf

Monetary Authority of Singapore. (2022b). FEAT principles assessment case studies. https://www.mas.gov.sg/-/media/mas/news/media-releases/veritas-document-6---feat-principles-assessment-case-studies.pdf

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

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. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), 509–533. https://doi.org/10.1002/(SICI)1097-0266(199708)18:7%3C509::AID-SMJ882%3E3.0.CO;2-Z

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841–887. https://arxiv.org/abs/1711.00399