International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 3, 140-149, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I3P116

Original Article

Trusted Enterprise AI: Governance, Explainability, and Compliance for Mission-Critical Systems

Amirtha Saminathan
Independent Researcher USA.

Received On: 14/07/2026 ***Revised On: 21/08/2026*** ***Accepted On: 30/08/2026*** ***Published On: 05/09/2026***

Abstract *- Deploying AI in audited, business-critical enterprise systems requires more than model accuracy. Decisions must be traceable, explainable, governed and defensible to business stakeholders and auditors, because an incorrect decision does not stay inside the application: it propagates into accounting records, regulatory reports and operational actions. This paper proposes a Trusted Enterprise AI reference architecture that embeds governance, explainability, human oversight, continuous monitoring and compliance controls across the AI lifecycle rather than attaching them after deployment. The contribution is not a new explainability algorithm or compliance standard; it is an architectural pattern for making AI decisions governable by design. Each consequential decision generates a traceable evidence chain connecting model and version, input context, explanation, confidence, applicable policies, control checks, human intervention where required, and the resulting business action, with continuous monitoring extending those controls afterward to detect drift, abnormal behavior and policy violations. Trust is decomposed into governance, explainability and compliance, each with a concrete control, a named owner, and evidence of execution rather than evidence of intent. The framework has not been implemented end to end as a single production system, and no governance outcomes are claimed as measured.*

Keywords *- Trusted Enterprise AI, AI Governance, Explainable AI, Responsible AI, AI Compliance, AI Auditability, Decision Traceability, Human-in-the-loop, Model Governance, Model Risk Management, AI Observability, Enterprise Architecture, Mission-Critical Systems, Continuous Compliance.*

1. Introduction

A system is mission-critical when its outputs directly affect financial accuracy, regulatory reporting, customer obligations, operational continuity, or the movement and reconciliation of money. The defining property is not importance but propagation: in these environments an incorrect decision does not remain isolated within the application. It flows into downstream systems, accounting records, reports and operational actions, and each hop makes it harder to find and more expensive to reverse.

Introducing AI into such a system is therefore a different act from adding AI to an ordinary application. The ordinary case optimizes for model accuracy and user experience. This case has to satisfy an existing control environment that predates the model, was designed around deterministic processing, and will be tested by an auditor who was not consulted about the model's introduction.

Enterprises commonly add AI through APIs, embedded models, decision services, copilots or analytics platforms. Governance is then handled through model documentation and approval processes, explainability through separate model-specific tooling, and compliance through policies, audit logs and periodic review. Each of those controls can be effective individually, and they are frequently disconnected from the actual execution of an AI-driven business decision.

That separation is particularly costly in a system already governed and audited, because of what such a system is required to reconstruct. An auditor or business owner must be able to establish not only what the AI predicted, but which model and version produced the decision, what data and policy context applied, why the result was accepted, whether required controls executed, who approved or overrode it, and what downstream action followed. Assembling that after the fact from three disconnected control systems is expensive when it is possible and impossible when it is not.

The proposal is to treat governance, explainability and compliance as one architectural control layer surrounding the AI decision lifecycle, so that each consequential decision generates a traceable evidence chain as a by-product of being made rather than as a reconstruction exercise afterward.

The contribution is architectural rather than algorithmic. It is not a new explainability method [1], [2], [3] or a new compliance standard. It is a pattern for making AI decisions governable by design, and the sections that follow specify where each control sits, what evidence each produces, and where the approach reaches its limits.

2. Background

Explainability has a substantial and maturing literature. Surveys establish the taxonomy of methods and the distinction between interpretable models and post-hoc

explanation [1], [2], [3], and model-agnostic techniques are widely available [4], [5]. A sustained critique holds that post-hoc explanation of an opaque model is not equivalent to interpretability and should not be treated as such for high-stakes decisions [6], and separate work has examined what an explanation is actually for, and to whom [7], [8].

AI governance has developed largely at the level of principle. Surveys of published ethics guidelines find broad convergence on high-level values and considerable divergence on implementation [9], [10]. Work on accountability has moved toward auditable process, particularly internal audit frameworks intended to be executed before deployment rather than after harm [11]. Organizational treatments frame AI management as a distinct capability rather than an extension of software management [12].

The engineering side supplies the complementary account. Machine learning systems accumulate maintenance burden in ways conventional software does not, and the practices that keep them healthy differ correspondingly [13]. Production readiness has been formalized as a testable rubric covering data, model, infrastructure and monitoring [14], and studies of deployed explainability find a persistent gap between what the tooling provides and what practitioners need from it [15].

What is comparatively thin is the connection between these bodies of work at the level of a single business decision inside a system of record. The explainability literature largely addresses the model; the governance literature largely addresses the organization; the engineering literature largely addresses the pipeline. An auditor's question concerns one transaction, and answering it requires all three at once.

3. What Breaks without Controls

Consider a large retailer reconciling cash and other store transactions across thousands of locations. An AI capability might be introduced to classify reconciliation exceptions, identify suspicious variances, or recommend whether discrepancies can be cleared automatically.

The two failure directions are asymmetric and both are costly. If the model incorrectly classifies a legitimate discrepancy as resolved, that decision flows into reconciliation records and ultimately affects financial reporting. If it repeatedly flags valid transactions, it generates large volumes of unnecessary investigation and delays financial close. The first is a correctness failure that hides; the second is an efficiency failure that announces itself.

The problem becomes materially worse when the AI cannot explain the decision. Finance teams and auditors need to reconstruct what data was used, which model and version produced the result, what rule or confidence threshold applied, and whether a human reviewed the exception. Without those controls the organization faces incorrect ledger entries, delayed reconciliation and close, costly manual rework, audit findings, and loss of confidence in automated decisions.

At enterprise scale small error rates become operationally significant. As an arithmetic illustration rather than a measurement: a one percent error rate across one hundred thousand automated financial decisions implies a thousand decisions requiring investigation or correction. The cost is not the value of one incorrect prediction. It is downstream financial exposure, investigation effort, reporting delay, compliance risk, and the operational cost of reconstructing decisions that were never adequately recorded.

That last item is the one the architecture addresses directly. Investigation cost is a function of how much evidence was retained at the moment of decision, and it is the only one of these costs that is fully determined by design choices made before anything goes wrong.

4. Decomposing Trust

Figure 1 sets out the decomposition.

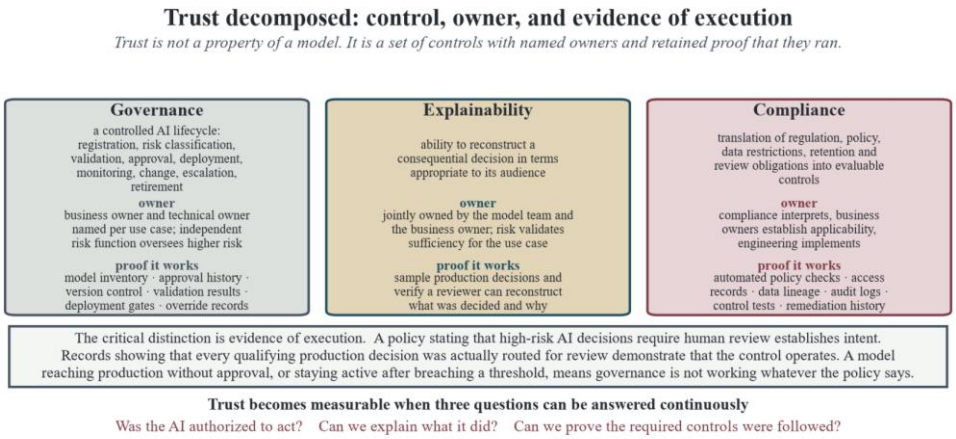

Fig 1: Trust Decomposed into Control, Owner and Evidence of Execution

Trust in a regulated enterprise system should not be treated as an abstract quality or a policy statement. It decomposes into three parts, each requiring an operational control, an accountable owner, and verifiable evidence.

Table 1: The Three Components of Trust

Component	The Concrete Requirement	Who Owns it	How it is Proven
Governance	A controlled AI lifecycle covering model registration, risk classification, validation, approval, deployment, monitoring, change management, escalation and retirement	A named business owner and a named technical or model owner per use case; an independent risk or governance function oversees higher-risk use cases	Model inventory records, approval histories, version-controlled changes, validation results, deployment gates, override records, monitoring alerts
Explainability	The ability to reconstruct a consequential AI-assisted decision in terms appropriate to its audience	Jointly owned by the model team and the business owner, with risk or compliance validating whether explanations are sufficient for the use case	Sampling production decisions and verifying that reviewers can reconstruct what the system decided, why, what influenced it, and how it affected downstream processing
Compliance	Translation of applicable regulations, internal policies, data restrictions, retention requirements, approval rules and human-review obligations into controls evaluable across the lifecycle	Compliance or legal interprets the obligations, business owners establish applicability, engineering implements enforceable technical controls	Automated policy checks, access records, data lineage, audit logs, retention records, exception handling, control-test results, remediation histories

4.1. Evidence of Execution, not Evidence of Intent

This is the distinction the whole framework rests on. A policy stating that high-risk AI decisions require human review establishes intent. Records showing that every qualifying production decision was actually routed for review demonstrate that the control operates. Only the second is evidence.

The negative test is equally concrete. A model reaching production without the required approval, or remaining active after breaching a defined threshold, indicates that governance is not operating effectively, whatever the documented process says.

Trust therefore becomes measurable when an enterprise can continuously answer three questions: was the AI authorized to act, can we explain what it did, and can we prove the required controls were followed?

5. The Reference Architecture

Figure 2 gives the end-to-end path.

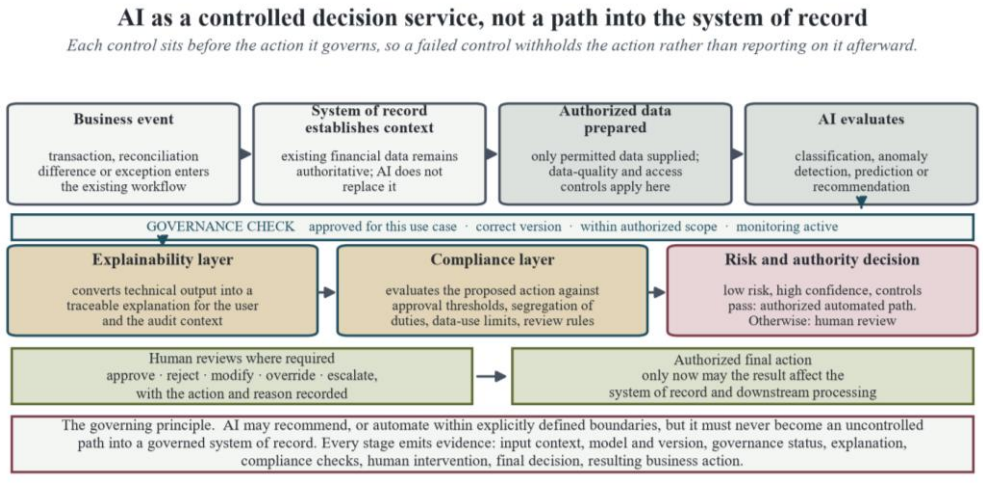

Fig 2: AI as a Controlled Decision Service, not a Path into the System of Record

The architecture places AI inside an existing mission-critical workflow as a controlled decision service rather than allowing it to act directly on financial or operational records.

A business transaction enters through the existing system of record: a reconciliation platform, payment workflow, operational ledger or exception-management

process. The workflow sends only the required context to the AI component.

Before the AI result can influence the business process, a governance layer verifies that the model is approved for this use case, the deployed version is valid, required monitoring is active, and the request falls within the model's authorized scope. The AI service then produces a result together with supporting metadata: model version, confidence, relevant inputs, decision rationale.

An explainability layer converts that technical output into a traceable explanation appropriate to the user and the audit context. For a financial exception, the system should be able to show why the transaction was flagged, which factors influenced the recommendation, and how confident the model was.

A compliance layer evaluates the proposed action against applicable policies and controls: approval thresholds, segregation of duties, data-use restrictions, human-review requirements, and whether the decision may proceed automatically at all.

The final action is determined by risk. Low-risk, high-confidence decisions satisfying all controls may proceed through an approved automated path. Decisions involving material financial impact, low confidence, unusual behavior or policy exceptions route to a human reviewer, who can approve, reject, modify or escalate, with that action recorded.

Only after the required controls and approvals are satisfied does the workflow update the system of record or trigger downstream processing.

Table 2: The Ordered Decision Path and where Each Control Sits

#	Step	Control at this Point
1	Business event enters the existing workflow	None; the workflow is unchanged
2	System of record establishes authoritative context	AI does not replace the system of record
3	Required data prepared for AI evaluation	Data-quality and access controls; only authorized data supplied
4	Pre-decision governance check	Model approved for use case, correct production version, within authorized scope, monitoring active
5	AI performs evaluation	Permitted task types only
6	Decision evidence captured	Model and version, input and source references, output, timestamp, confidence
7	Explanation produced	Reason codes, contributing factors, audience-appropriate depth
8	Compliance controls evaluate the proposed action	Policy requirements, approval thresholds, data restrictions, segregation of duties, human-review requirements
9	Risk and authority decision	Automated path only if low risk, permitted, sufficient confidence and all controls pass
10	Human review where required	Approve, reject, modify, override, escalate
11	Human action and reason recorded	
12	Authorized final action occurs	Only now may the result affect the governed workflow
13	System of record updated	
14	Complete decision lineage retained	Source event through to system-of-record impact
15	Post-decision monitoring	Drift, exception patterns, overrides, downstream corrections, reconciliation differences, control failures, business outcomes
16	Authority can be reduced	Automation restricted, routed to human review, rolled back or disabled

The governing principle is that AI may recommend, or automate within explicitly defined boundaries, but it must never become an uncontrolled path into a governed system of record.

Step 16 is the one most often absent from published governance designs, and it is what makes the arrangement dynamic rather than a one-time approval.

6. Explainability in Practice

Figure 3 sets out the audiences.

Explainability serves different users by translating an AI output into evidence they can understand, verify and act on. The objective is not to expose how a model works technically, but to answer a practical question: why did the system recommend or take this action?

For an operator, the explanation should be concise and actionable: the recommended action, the main reasons, the confidence, the exception condition, and whether approval or escalation is required. Enough context to investigate, accept, reject or escalate.

For a finance team, the explanation should connect the decision to its financial consequence: which transaction or exception was evaluated, what factors drove the recommendation, whether it was executed automatically or approved manually, and what downstream record was affected.

For an auditor, explainability becomes evidence. The organization must be able to reconstruct a historical

decision by showing source data reference, model and version, relevant input features, output, confidence where applicable, explanation, applicable controls, human approval or override, and final business action.

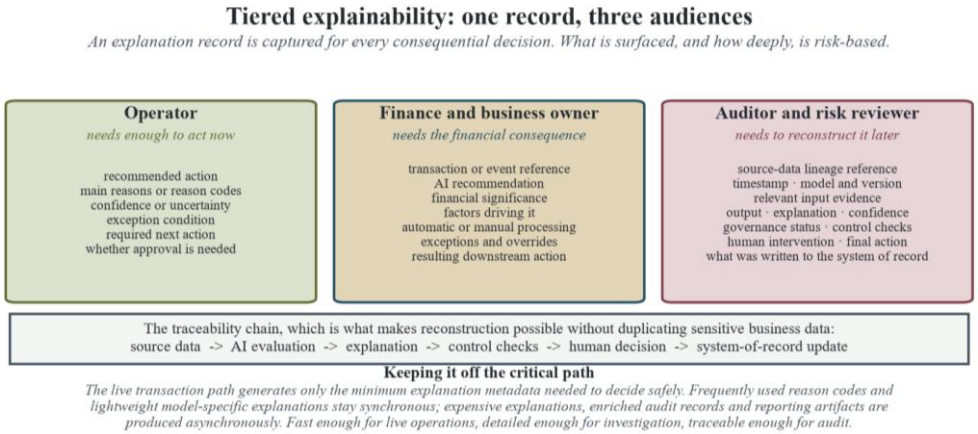

Fig 3: Tiered Explainability: One Record, Three Audiences

6.1. Coverage and Depth

An explanation record is captured for every consequential AI decision, even where the user does not need to see the full explanation at the time. Depth is then risk-based: routine low-risk decisions expose a concise reason code and confidence indicator during normal processing, while high-value transactions, exceptions, overrides and audit investigations retrieve a fuller explanation.

Traceability to source data is essential and does not require duplicating sensitive business data. The explanation record retains transaction identifiers, data lineage references, timestamps, model version, relevant feature values or references, and integrity metadata. That produces a chain from source data through AI evaluation, explanation, control checks and human decision to the system-of-record update.

6.2. Keeping Explainability off the Critical Path

Explainability must not become a performance bottleneck in a system running live operations. The live transaction path generates only the minimum explanation metadata required to make the decision safely. More computationally expensive explanations, enriched audit records and reporting artifacts are generated asynchronously. Frequently used reason codes and lightweight model-specific explanations remain synchronous, while deeper analysis is produced on demand.

The result is tiered explainability: fast enough for live operations, detailed enough for investigation, traceable enough for audit. The goal is not that every user understands the mathematics of the model. It is that every consequential decision can be understood at the appropriate level and reconstructed when accountability requires it [7], [8], [15].

7. Where the Framework Breaks Down

Explainability has a hard ceiling. Some complex models cannot provide a complete human-understandable account of their internal reasoning, and post-hoc explanations may indicate influential factors without faithfully representing how the model reached its conclusion [6]. For consequential financial decisions the framework therefore requires a minimum explainability threshold: if a model cannot produce evidence sufficient for the required level of audit or business review, it should not be permitted to execute that class of decision autonomously. It may operate as a recommendation system with human approval, or a more interpretable model may be required.

Regulations overlap and change. Financial and operational systems may be subject to multiple regulations, jurisdictions, internal policies, contractual requirements and retention rules that move over time. The framework cannot determine the correct legal interpretation of any requirement; legal, compliance and risk functions remain responsible for that. What an implementation can do is maintain versioned, machine-evaluable policy controls that record which requirements applied when each decision was made, so that a decision is assessed against the rules in force at the time rather than the rules in force at audit.

AI must not destabilize a system that cannot go down. A system responsible for reconciliation, reporting or money movement cannot be redesigned around an experimental component. AI should initially operate outside the authoritative transaction path through shadow processing, recommendation, or parallel evaluation, with outputs compared against the established process before controlled automation is progressively enabled. Feature flags, rollback mechanisms, circuit breakers, fallback rules and conventional deterministic processing must allow the underlying workflow to continue when the AI service is unavailable or behaving unexpectedly. The system of record must not depend on AI availability for basic operational continuity.

7.1. The Quietly Wrong Model

Figure 4 states the hardest failure mode.

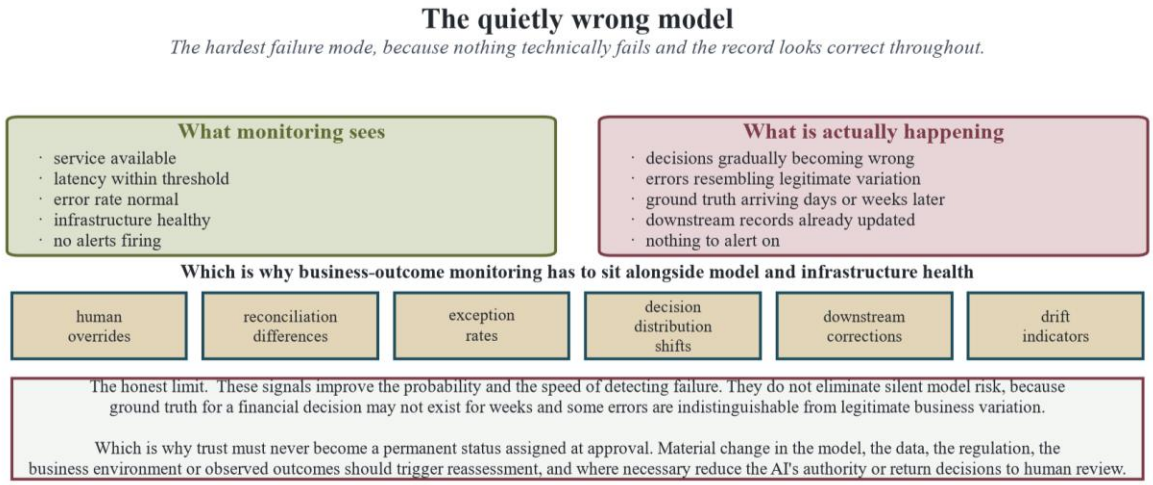

Fig 4: The Quietly Wrong Model

A model may remain technically healthy, available, fast and within infrastructure thresholds, while gradually making incorrect business decisions. Conventional system monitoring will not detect this, because nothing has technically failed.

The framework therefore requires monitoring of business outcomes alongside model and infrastructure health: human overrides, reconciliation differences, exception rates, changes in decision distribution, downstream corrections, drift indicators and financial outcomes.

These controls cannot guarantee timely detection, and the paper does not claim they can. Ground truth may arrive days or weeks after a decision, and some errors resemble legitimate business variation closely enough to be indistinguishable in aggregate. Observability improves the probability and the speed of detecting failure. It does not eliminate silent model risk.

For that reason trust should never become a permanent status conferred at approval. Material change in the model, the data, the regulations, the business environment or observed outcomes should trigger reassessment and, where warranted, reduce automation or return decisions to human review.

7.2. The Limit that Architecture cannot Cross

Architecture cannot remove accountability from people. The framework can determine whether required controls ran, preserve evidence, detect warning signals and prevent specific unauthorized actions. It cannot prove that any given AI decision is correct, and it cannot substitute for professional judgement in interpreting regulation or material financial risk.

Its purpose is therefore not to produce an AI system that must be trusted, but an enterprise environment in which AI is trusted only to the extent that current evidence justifies that trust, and its authority can be reduced when that evidence weakens.

8. Future Work: A Continuous Assurance Layer

The natural extension is a continuous AI assurance layer combining policy-as-code, versioned regulatory controls, automated evidence collection, decision lineage, outcome-based monitoring, drift detection, periodic control testing, independent model validation and dynamic risk thresholds.

The behavior that layer would enable is specific: a decision class moves automatically from autonomous processing to human review when confidence, data quality, model behavior or compliance conditions deteriorate, and moves back only when evidence supports it. That converts the authority reduction in step 16 of Table II from a manual intervention into a property of the system.

9. Operating Context

This section reports outcomes from enterprise modernization work that establish the context the framework was designed for. These figures are outcomes of that modernization initiative. They are not attributable to the governance framework proposed in this paper, and they are reported here separately from any claim about it.

In a large-scale cash-office modernization initiative, process improvements reduced daily cash-office processing time from approximately three hours to approximately one hour per store, representing roughly fourteen hours of weekly effort saved per location. The broader modernization programme was associated with approximately thirty-two million dollars in annual operating-cost reduction, a fifty-five percent reduction in dead or float cash, and approximately four point eight million dollars in annual interest benefit.

What these figures establish is scale and stakes, not framework effectiveness. They describe an operating environment in which reconciliation runs across thousands of locations, financial close depends on it, and small per-transaction inefficiencies aggregate into material amounts. That is the environment in which an AI decision affecting reconciliation would propagate, and it is the reason the architecture treats the system of record as something AI may inform but not directly write to.

10. What a Production Implementation would Measure

The framework has not been implemented end to end and measured. No figures are reported for AI workflows governed, AI decisions explained, exceptions caught, or audit outcomes. A production implementation would make its effectiveness quantitatively assessable, and the architecture is specified to capture what that assessment requires.

Table 3: Measurement Model for Future Validation

Measure	What it would Establish
Number and percentage of AI decisions under governance	Governance coverage, and whether ungoverned paths exist
Automated versus human-reviewed decision split	Whether the risk routing operates as designed
Policy violations detected	Whether compliance controls fire
Low-confidence decisions escalated	Whether confidence thresholds route correctly
Human override rate and reasons	Whether the model is trusted appropriately, and an early indicator of quiet wrongness
Model or data drift events	Detection sensitivity
Unexplained decisions	Explanation coverage failures
Control failures	Whether the control layer itself is reliable
Reconciliation exceptions	Business-outcome signal
Decision latency	Whether governance is affordable on the live path
Manual review time	Operational cost of the human-in-the-loop requirement
Audit evidence retrieval time	Whether the evidence chain is usable in practice, not merely present
Completeness of audit evidence	Whether reconstruction actually succeeds on sampling

The validation that would matter most is the audit reconstruction test: sample production decisions after the fact and measure what proportion can be fully reconstructed, and how long each takes. That test is executable on a single workflow, requires no comparison group, and directly measures the property the framework exists to produce.

11. Limitations

The complete framework has not been implemented or measured end to end. It is a reference architecture synthesized from modernization experience and established responsible AI principles, not a validated system.

The figures in Section IX belong to a different intervention. They describe the operating context and say nothing about whether this framework works.

The error-rate illustration in Section III is arithmetic, chosen to show how small rates aggregate at enterprise scale. It is not an observed error rate.

Silent model risk is reduced, not eliminated. Section VII-A states this directly, and it is the framework's most consequential residual risk rather than a formality.

Regulatory interpretation remains human. The architecture can record which policy version applied and enforce encoded controls. It cannot determine what a regulation requires.

Single sector and single organizational vantage. The experience informing this work is enterprise architecture and modernization in large-scale retail platforms. Whether the decomposition transfers to other regulated sectors is untested.

Explainability sufficiency is a judgement, not a measurement. The minimum explainability threshold in Section VII is a requirement the framework imposes without supplying an objective test for whether a given explanation meets it.

No review has occurred. A prospective reviewer with relevant vantage has been identified and not approached, and no external assessment of this draft exists.

12. Conclusion

Adding AI to an audited, business-critical system is not the same act as adding AI to an ordinary application, because the decision does not stay inside the application. It reaches the ledger, the report and the reconciliation, and it reaches them through a control environment that was designed for deterministic processing and will be tested by someone who was not consulted about the model.

This paper proposes treating governance, explainability and compliance as one control layer around the decision lifecycle rather than three separate practices attached afterward. Trust decomposes into a control, an owner and evidence of execution for each of the three, and the distinction that matters throughout is between a policy

stating that a control should run and a record showing that it did.

The architecture places AI as a controlled decision service: a governance check before the model may influence the process, an explanation produced at the point of decision, compliance evaluated against the proposed action, risk determining whether a human must approve, and the system of record updated only after the required controls are satisfied. Every stage emits evidence, and the AI's authority can be reduced when monitoring indicates deterioration.

Its limits are real and are stated rather than deferred. Observability improves the probability and speed of detecting a quietly wrong model without eliminating that risk, because ground truth for a financial decision may not exist for weeks. Architecture cannot interpret regulation. And it cannot remove accountability from people.

The framework has not been implemented end to end or measured, and Section X specifies what a production implementation would need to capture to establish whether it works. The purpose is not an AI system that must be trusted, but an environment in which AI is trusted only as far as current evidence justifies, and in which that trust can be withdrawn when the evidence weakens.

References

[1] A. Barredo Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020. doi: 10.1016/j.inffus.2019.12.012

[2] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018. doi: 10.1109/ACCESS.2018.2870052

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, pp. 1-42, 2018. doi: 10.1145/3236009

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144. doi: 10.1145/2939672.2939778

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," arXiv preprint, 2017. doi: 10.48550/arXiv.1705.07874

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019. doi: 10.1038/s42256-019-0048-x

[7] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," in Proc. Conference on Fairness, Accountability, and Transparency (FAT*), 2019, pp. 279-288. doi: 10.1145/3287560.3287574

[8] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv preprint, 2017. doi: 10.48550/arXiv.1702.08608

[9] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389-399, 2019. doi: 10.1038/s42256-019-0088-2

[10] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People, An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," Minds and Machines, vol. 28, no. 4, pp. 689-707, 2018. doi: 10.1007/s11023-018-9482-5

[11] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in Proc. Conference on Fairness, Accountability, and Transparency (FAT*), 2020, pp. 33-44. doi: 10.1145/3351095.3372873

[12] N. Berente, B. Gu, J. Recker, and R. Santhanam, "Managing Artificial Intelligence," MIS Quarterly, vol. 45, no. 3, pp. 1433-1450, 2021. doi: 10.25300/MISQ/2021/16274

[13] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software Engineering for Machine Learning: A Case Study," in Proc. IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 2019, pp. 291-300. doi: 10.1109/ICSE-SEIP.2019.00042

[14] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ML test score: A rubric for ML production readiness and technical debt reduction," in Proc. IEEE International Conference on Big Data, 2017, pp. 1123-1132. doi: 10.1109/BigData.2017.8258038

[15] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in Proc. Conference on Fairness, Accountability, and Transparency (FAT*), 2020, pp. 648-657. doi: 10.1145/3351095.3375624

Appendix A. Governance Requirements as Checkable Items

For every AI component used in a mission-critical workflow.

#	Requirement
1	The use case has a named business owner
2	The component has a named technical or model owner
3	Use case and intended business purpose are documented
4	Financial or operational impact is identified
5	Risk level is assigned before production use
6	Permitted and prohibited uses are defined
7	The model and version in production are registered
8	Training and validation information is documented where applicable
9	Required validation is complete before deployment
10	Production approval is recorded
11	Decision and confidence thresholds are defined
12	Conditions requiring human review are defined
13	Maximum authority given to the AI is explicitly defined
14	Human override capability exists for consequential decisions
15	Overrides and their reasons are recorded
16	Model, input-data and business-outcome monitoring are active
17	Drift and abnormal decision patterns are monitored
18	Escalation thresholds are defined
19	Model changes trigger appropriate revalidation
20	Rollback and fallback procedures exist
21	AI failure does not unnecessarily prevent the underlying workflow from operating
22	A model retirement and deactivation process exists
23	Evidence of approvals, changes, exceptions and monitoring is retained

Appendix B. Explainability Requirements as Checkable Items

For each consequential AI decision.

#	Requirement
1	A unique decision or transaction identifier is recorded
2	Timestamp is recorded
3	The AI use case is identifiable
4	Model and model version are recorded
5	Relevant source-data references are retained
6	Important inputs or features are identifiable where technically appropriate
7	The AI output or recommendation is recorded
8	Confidence or uncertainty is recorded where meaningful
9	Human-understandable reasons or reason codes are available
10	The applicable decision threshold is identifiable
11	Human review, where required, is recorded
12	Human approval, rejection or override and its reason are recorded
13	The final business action is linked to the AI recommendation
14	The downstream system-of-record update is traceable
15	Historical decisions can be reconstructed later

Appendix C. Compliance Requirements as Checkable Items

#	Requirement
1	Applicable regulations and internal policies are identified for the use case
2	Compliance requirements are mapped to specific controls
3	Required controls are versioned
4	Data permitted for AI processing is defined
5	Restricted and sensitive data handling requirements are enforced
6	Access is role-based and logged
7	Required segregation of duties is maintained
8	Financial and materiality thresholds are defined where applicable

9	Human-review requirements are encoded into the workflow
10	Required approvals cannot be bypassed by the AI
11	Retention requirements are defined
12	Decision evidence is retained for the required period
13	The policy or control version applicable at decision time is identifiable
14	Compliance exceptions are logged
15	Exceptions are routed to an accountable person
16	Remediation is tracked
17	Regulatory or policy changes trigger control review
18	Periodic control testing is performed