FSB FINANCIAL STABILITY BOARD

Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation report

Response to Consultation

GOLDAi

1. **Do you agree with the benefits and risks of AI adoption by financial institutions described in this report? Are there any substantive benefits or risks not covered?**

   The taxonomy of benefits and risks is accurate and unusually candid. The report's treatment of agentic risk is its strongest analytical contribution: it recognises that agents can take unauthorised or erroneous actions whose remediation may be infeasible, and that real-time human monitoring of agent decisions does not scale. One risk deserves explicit treatment as its own entry: the migration of institutional authority into model behaviour. As agents integrate with payment systems, customer records, and trading and compliance infrastructure, the institution's authority to act can come to reside, implicitly and unobserved, in the disposition of a probabilistic system. None of the listed risks names this directly, yet several of them, including unauthorised actions and weakened oversight, are its downstream symptoms. It is an architectural risk rather than a behavioural one, and it is preventable by design rather than detectable by monitoring.

2. **Are the sound practices sufficiently comprehensive and clear to enable financial institutions' responsible AI adoption?**

   The twelve practices are comprehensive for the governance of AI behaviour across the lifecycle. I recommend one addition, complementing rather than replacing Sound Practice 10 on human oversight: consequential actions initiated or materially shaped by AI should require an explicit institutional determination

before execution. Such a determination identifies the actor requesting the action, the delegated authority under which it is requested, the governing obligations and policies, the relevant institutional facts, the resulting authorisation, and any conditions attached. It exists independently of any natural-language explanation a model produces. This practice gives the report's proportionality principle a concrete anchor for the highest-autonomy systems: the greater the autonomy, the more important it is that authorisation be located outside the model.

3. **Do the sound practices strike an appropriate balance between managing risks relating to all forms of AI, and addressing some of the risks relating to emerging and new complex forms of AI, such as GenAI and agentic AI?**

   For advisory and analytical AI, the balance is appropriate. For agentic AI, the practices lean on behavioral controls whose limits the report itself acknowledges. Where the report observes that effective monitoring of agents may require augmentation with another AI, I would note that AI monitoring AI still governs behaviour; it does not answer the prior question of what authorizes a regulated action in the first place. The durable answer for agentic AI is a separation of concerns: probabilistic systems may interpret intent, retrieve information, propose plans, and explain outcomes; institutional determinations, independently governed, versioned, and reproducible, authorize regulated effects; and execution follows authorization. Each determination should produce a record sufficient for an independent reviewer to reconstruct, after the fact, the applicable obligations, the facts considered, the policy version in force, the authorization pathway, the outcome, and any human intervention. Behavioral evidence characterises what a system tended to do. Determination records establish why a specific regulated action was or was not permitted. Supervision of agentic AI will need the second kind of evidence.

4. **Are the sound practices sufficiently flexible to accommodate and address newer types of AI and responsible adoption over time?**

Practices indexed to the behaviour of particular AI technologies will require continuous revision as capabilities evolve. Practices indexed to institutional outcomes will not. I encourage the final report to frame the agentic-AI expectations in outcome terms that are technology-neutral by construction: institutional authority is explicitly represented; regulated actions are authorised through governed determinations; determinations are recorded and independently replayable; and human accountability operates at the level of rule design, encoding governance, and referral handling, where it is effective, rather than only at the level of real-time output review, where the report correctly notes it is not. Knowledge graphs, rule engines, policy services, and workflow systems are all legitimate implementations; the supervisory object should be the outcome, not the technology

5. **Do the case studies in this report sufficiently highlight how different types of financial institutions can benefit from responsible AI adoption? Are there additional case studies for inclusion in the report? If so, please provide such case studies, particularly for nonbanks.**

The case studies are among the most useful elements of the report, and one of them already contains the pattern this submission recommends. In the agentic fraud-detection case study, the agent autonomously proposes detection rules, and the bank's fraud analytics team reviews and approves every rule before implementation. The agent proposes; the institution determines; the system executes only what has been determined. I encourage the final report to name this pattern explicitly as the generalisable insight of the case study, since it demonstrates that the separation of proposal from institutional effect is compatible with measurable performance gains rather than in tension with them. A valuable additional case study would show determination records in use: an institution reconstructing, for an auditor or supervisor, precisely why an agent-initiated action was authorized, from the record rather than from behavioral logs.

6. **Do the case studies in this report provide actionable insights for financial institutions in their responsible AI adoption?**

   The case studies are among the most useful elements of the report, and one of them already contains the pattern this submission recommends. In the agentic fraud-detection case study, the agent autonomously proposes detection rules, and the bank's fraud analytics team reviews and approves every rule before implementation. The agent proposes; the institution determines; the system executes only what has been determined. I encourage the final report to name this pattern explicitly as the generalisable insight of the case study, since it demonstrates that the separation of proposal from institutional effect is compatible with measurable performance gains rather than in tension with them. A valuable additional case study would show determination records in use: an institution reconstructing, for an auditor or supervisor, precisely why an agent-initiated action was authorized, from the record rather than from behavioral logs.

7. **Are the definitions in the glossary clear and aligned with industry sound practices, including recent developments in AI?**

   The glossary would benefit from terms that distinguish oversight of behaviour from the institutional act of authorisation. I suggest adding: authorisation or determination, the governed institutional act that permits a specific consequential action, distinct from a model's recommendation or explanation of it; and determination record, the versioned, tamper-evident record of a determination sufficient to permit independent reconstruction. Making this vocabulary available will help institutions and supervisors discuss agentic governance with precision the current lexicon does not yet support.

8. **Are there any comments you would like to make on this report? Please fill in.**

   The consultation arrives as AI in finance crosses from systems that recommend to systems that act. The practices in this report govern the first transition well. For

the second, I respectfully encourage the FSB to make the location of institutional authority an explicit object of governance: models may propose; institutions determine; systems execute what has been determined. Framed this way, responsible adoption does not depend on predicting how AI capabilities evolve, only on ensuring that no capability, however advanced, becomes the source of an institution's authority to act

WORKING PAPER | JULY 2026

GOVERNANCE BEFORE ACTION

Institutional Authority for Agentic AI in Banking

Malka Sprung

July 18, 2026

A model may interpret intent. It must not be the institution's source of authority. No regulated action should occur unless an explicit, governed determination authorizes it and leaves a record that can be independently replayed.

For discussion and research purposes. This paper does not constitute legal, regulatory, or supervisory advice.

ABSTRACT

The deployment of frontier AI models into regulated financial institutions has created a governance problem that existing supervisory frameworks were not designed to answer. Recent regulatory developments distinguish deterministic rule-based systems from statistical models while simultaneously recognizing that generative and agentic AI require new governance approaches. The result is not a lack of regulation, but an unresolved architectural question: where should institutional authority reside when probabilistic models participate in regulated decisions?

This paper argues that institutional authority should remain outside the model. Rather than allowing frontier models to interpret regulatory obligations directly, regulated actions should be authorized through an independently governed determination layer that evaluates machine-readable obligations, institutional policies, permissions, and verified facts before any consequential action is executed. Models remain responsible for interpretation, retrieval, explanation, and orchestration, but they do not become the institution's source of authority.

The paper defines the architectural properties required of such a determination layer, proposes a supervisory examination framework based on replayable determination records rather than behavioral observation alone, and identifies an empirical research agenda for comparing determination-layer architectures against model-centric approaches. It reports conformance evidence from a working implementation as an existence proof that these requirements are simultaneously satisfiable and machine-checkable; no comparative results are claimed. It further argues that this architecture naturally produces the evidence artifacts increasingly expected by supervisors while preserving human accountability over institutional judgment.

The author discloses a commercial interest as the founder of GOLDAi, whose prototype implements elements of the proposed architecture, and as the named inventor on a pending United States patent application covering aspects of the architecture described in this paper. The architectural thesis advanced here should be evaluated independently of the pending patent application and any commercial implementation.

Keywords: agentic AI; banking supervision; AI governance; model risk management; deterministic authorization; regulatory knowledge graphs; auditability; compliance architecture

1. A Supervisory Opening, Not a Governance Vacuum

1.1 What changed in 2026

On February 19, 2026, the U.S. Department of the Treasury released an Artificial Intelligence Lexicon and the Financial Services AI Risk Management Framework (FS AI RMF). Treasury described the resources as products of public-private collaboration through the Financial and Banking Information Infrastructure Committee and the Financial Services Sector Coordinating Council's Artificial Intelligence Executive Oversight Group. The Cyber Risk Institute, which maintains the framework, describes it more specifically as an industry-led, sector-specific adaptation of the NIST AI Risk Management Framework developed with more than one hundred financial institutions. Its operating structure includes an adoption-stage questionnaire, a risk and control matrix with 230 control objectives, a guidebook, and a reference guide containing examples of controls and effective evidence (U.S. Department of the Treasury 2026; Cyber Risk Institute 2026).

That institutional description matters. The FS AI RMF is not a binding rule and should not be represented as Treasury-authored supervisory guidance. Its importance lies in a different fact: it translates broad AI principles into the vocabulary through which financial institutions govern risk - control objectives, implementation guidance, maturity stages, and evidence. It therefore provides a useful indication of the evidence surface that banks, auditors, third-party risk teams, and supervisors may expect even where the framework itself has no independent legal force.

On April 17, 2026, the Office of the Comptroller of the Currency, the Board of Governors of the Federal Reserve System, and the Federal Deposit Insurance Corporation issued revised interagency guidance on model risk management, replacing SR 11-7 and related issuances. The guidance is risk-based, expected to be most relevant to banking organizations with more than $30 billion in total assets, and expressly non-prescriptive. Non-compliance with the guidance alone will not result in supervisory criticism, although violations of law or unsafe or unsound practices arising from inadequate risk management remain actionable (Federal Banking Agencies 2026).

Two features of the revised guidance are especially relevant. First, its definition of a model excludes simple arithmetic, deterministic rule-based processes, and software not underpinned by statistical, economic, or financial theories. Second, a footnote states that generative and agentic AI models are novel and rapidly evolving and

therefore are not within the guidance's scope. The same footnote states that a banking organization's broader risk-management and governance practices should determine appropriate controls for systems outside the guidance (Federal Banking Agencies 2026).

The agencies did not thereby declare deterministic systems safe or agentic systems ungoverned. Nor did they classify every component of a hybrid agentic system as outside model risk management. A probabilistic intent parser, classifier, ranking model, or non-generative predictive component may still require model-specific governance depending on its design and use. A deterministic policy engine may be governed through software development, change management, compliance, information security, operational risk, and internal-control disciplines. The correct unit of analysis is the component and its function, not the marketing label placed on the whole system.

The OCC also announced that the agencies plan to issue a request for information addressing model risk management generally and banks' use of AI, including generative AI, agentic AI, and AI-based models (OCC 2026). At the time of writing, that announced process remains forthcoming. The opening is therefore precise: the agencies have narrowed the existing model-risk framework, acknowledged that newer systems may require a different approach, and invited further input without suspending the rest of the bank's governance obligations. Federal Reserve Vice Chair for Supervision Michelle Bowman summarized the position directly in May: banks are relying on existing risk-management frameworks today, and supervisors should consider whether those tools are fit for the future (Bowman 2026).

The European picture reinforces the uncertainty rather than supplying an immediate deadline. The EU AI Act originally placed major high-risk obligations on an August 2026 timetable. On June 29, 2026, the Council of the European Union gave final approval to legislation delaying the relevant dates to December 2, 2027 for stand-alone high-risk systems and August 2, 2028 for high-risk systems embedded in products (Council of the European Union 2026). The delay removes the deadline pressure. It does not remove the implementation problem. Globally active banks are still making architecture, procurement, and control decisions before detailed supervisory expectations stabilize.

1.2 Deployment is advancing, but the evidence is still early

Public announcements in 2026 show movement from demonstrations toward institutional pilots and development programs. FIS and Anthropic announced a Financial Crimes AI Agent under development with BMO and Amalgamated Bank, with broader availability planned for the second half of 2026. FIS emphasized source-linked conclusions, investigator control, and traceable and auditable decisions. Amalgamated described its role as a design partner in an initial pilot and stated that human investigators would retain authority over decisions including suspicious activity report filings (FIS 2026; Amalgamated Bank 2026).

These announcements are significant, but they do not establish mature production performance. They show that architecture is being selected, evaluation criteria are being negotiated, and governance expectations are entering product design. They also reveal a common design emphasis in these announcements: a useful banking agent must be able to assemble evidence and expose its work. The unresolved question is whether traceability is merely descriptive - a log of what a model did - or constitutive - part of the mechanism that determines whether the action could occur in the first place.

That distinction is the subject of this paper.

1.3 The claim

The model-risk-specific opening should not be filled by extending model-behavior controls over the entire system and assuming that more testing, more monitoring, and more human review will make agentic action governable. Those controls remain necessary. They are not sufficient because they govern the probability of acceptable behavior, not the bank's authority to produce an effect.

The proposed architecture separates three functions:

1. **Cognition and interface:** probabilistic systems may interpret language, retrieve information, analyze evidence, form plans, and explain results.
2. **Determination:** a governed institutional layer decides whether the proposed action is permitted, prohibited, conditioned, or requires human adjudication.
3. **Execution:** bank systems carry out only actions accompanied by a valid determination and the required authority.

The central proposition is not that banks should replace AI with rules. It is that **proposal and authority must be different system roles**. The model may help construct the request. It must not be able to manufacture the authority that makes the request executable.

2. The Hidden Assumption: Interpretation and Authority Are the Same Thing

Prevailing AI governance often treats the model as the primary site of interpretive risk. A system receives a question or goal, retrieves context, applies instructions, and generates an answer or tool call. Governance therefore asks whether the model understood the rule, used the correct evidence, hallucinated a citation, drifted after an update, exhibited bias, or behaved consistently under adversarial prompts. The corresponding controls are behavioral: benchmark testing, red teaming, output sampling, drift monitoring, guardrails, observability, escalation thresholds, and human review.

This control set is rational within its premise. Where the model genuinely makes the decision, governance must characterize and constrain model behavior. NIST's AI Risk Management Framework and practitioner guidance on advanced AI systems properly emphasize lifecycle governance, testing, monitoring, accountability, and human oversight (NIST 2023; Grevenbrock, Zhao, and Patel 2026).

Agentic systems expose a second problem. A model output may no longer remain an answer for a person to consider. It may become an instruction to place a hold, release a payment, change a customer record, submit a filing, open or close access, send an external communication, initiate a trade, or invoke another system. At that point, the governance question is not only whether the model's interpretation is likely to be correct. It is whether the institution has created a path by which probabilistic interpretation can become institutional action without an independent authority boundary.

The hidden assumption is that interpretation and authority can reside in the same component. In a conversational assistant, that assumption may be tolerable because a human remains between the answer and the effect. In an agentic workflow, it is dangerous. The component that proposes what should happen also gains the ability, directly or through tools, to make it happen.

Behavioral controls reduce risk but do not alter that allocation of authority. A guardrail can reject a class of outputs. A monitor can detect anomalous tool calls. A human can review selected cases. None of those mechanisms establishes that every consequential effect passes through an institutionally governed decision point. The architecture still asks an uncertain component to police its own path to action or relies on a second system to observe that path after the fact.

This paper therefore distinguishes four objects that are often collapsed:

- **A model output** is a probabilistic product of inference.
- **A recommendation** is a non-binding interpretation offered for consideration.
- **A determination** is an institutionally authorized decision, produced under a specified rule set and evidence state, about whether an action may proceed.
- **An effect** is the actual change made in the world or in a bank system.

A governed agentic system may use models extensively in the first two objects. It should not treat them as the third. A determination is not a more confident answer. It is a different kind of object, created by a different authority mechanism.

Human review does not erase this distinction. A human decision can constitute the determination where the reviewer is authenticated, has delegated authority, receives the relevant evidence, follows an applicable decision process, and leaves a reviewable record. A generic "human in the loop" label is not enough. A person who reflexively approves a preselected model action, lacks sufficient time or context, or cannot identify the governing rule is not an independent authority boundary. Human adjudication must be designed and tested as a control, not assumed to be one.

This distinction also bounds the thesis precisely. Regulatory meaning cannot always be resolved before any model inference. A model may be used to parse a user's request, extract entities, classify documents, or retrieve candidate obligations before the determination layer runs. The defensible claim is narrower: **regulatory authority must be**

resolved before a regulated action is authorized. Where probabilistic components help construct the inputs, their outputs must be bound to structured facts, validated, and carried into the determination record as inputs with known provenance and limitations.

3. From Structural Governance to the Banking Determination Layer

3.1 Prior art: governance at the effect boundary

The phrase *structural governance* cannot be claimed as original to this paper. In April and May 2026, Alan L. McCann published work distinguishing governance of model content from governance of system effects. McCann's "two-boundary" framework argues that an AI system's effect-producing capabilities and its governance coverage must be made coterminous by architecture: every effect must pass through the governance mechanism, rather than being monitored or filtered as a separate after-the-fact process. A companion paper formalizes an effect-governance operator and argues that governance at the effect boundary can coexist with broad internal computational expressivity (McCann 2026a; 2026b).

That work is general and formal. It is not a banking compliance framework, does not specify how regulatory obligations should be represented, and does not define what an examiner should review. It nevertheless supplies the correct general principle: governance of actions is an architectural property, not merely a behavioral aspiration.

This paper's contribution is the banking-specific extension. In regulated finance, it is not enough to prove that an effect passed through *some* policy gate. The institution must be able to show which actor had authority, which obligation or policy governed the action, which facts were bound, which version of the rule set applied, why the resulting determination followed, who could override it, and whether the action that occurred matched what was authorized. The relevant mechanism is therefore more specific than an effect gate. It is a **determination layer**.

3.2 Independent design pressures point in the same direction

Several 2025-2026 developments are compatible with this architecture. They do not establish a settled consensus, and none independently proves the full claim. Together, they define the design pressure the determination layer is intended to answer.

The IMF's payment architecture. In April 2026, Sonja Davidovic and Hervé Tourpe proposed a three-layer analytical framework for agentic payments: intent and orchestration; control and authorization; and settlement. The note explicitly states that it is a normative analytical lens rather than prescriptive policy. Its second layer enforces deterministic constraints over actions proposed by probabilistic systems, and its downstream payment layers preserve authorization and legal finality. The authors emphasize identity, delegated authority, explicit mandates, and audit trails (Davidovic and Tourpe 2026). Payments make the separation unusually visible because an irreversible transfer cannot be governed by an average accuracy rate. The same architecture is relevant wherever an agent can produce a regulated effect.

The FS AI RMF's evidence orientation. The FS AI RMF does not prescribe a determination layer. It does, however, require institutions to operationalize controls and identify effective evidence across governance, data, validation, monitoring, accountability, and related domains (Cyber Risk Institute 2026). A system whose normal execution produces versioned policy decisions, boundary records, and replayable determinations has a different evidence posture from a system that reconstructs governance from prompts, traces, and sampled outputs after the fact. The framework does not choose between them, but its evidence requirements reward architectures with explicit control objects.

The agencies' component taxonomy. The revised model definition separates deterministic rule-based processes from models and separately excludes generative and agentic models from the guidance's scope (Federal Banking Agencies 2026). This does not endorse a layered design, and it should not be exploited as a route around model governance. It does make component decomposition unavoidable. A bank cannot responsibly ask whether "the agent" is a model and stop there. It must identify which components infer, which determine, which execute, and which controls apply to each.

Graph-grounded compliance research. GraphCompliance represents policy text and runtime context as graphs and constrains a judge model with aligned structural evidence. On a benchmark of 300 GDPR-derived scenarios, the

authors report micro-F1 gains of 4.1 to 7.2 percentage points over LLM-only and retrieval-augmented baselines, along with fewer over- and under-predictions. Baldwin and Ghanavati separately report that knowledge-graph augmentation improved five models across 42 policy-reasoning tasks involving three AI policy documents (Chung et al. 2025; Baldwin and Ghanavati 2026). These are preprints, not bank-supervisory studies. They do not establish legal correctness or prove that a deterministic policy engine should replace model judgment. They provide narrower evidence: moving normative structure into an explicit representation can improve reasoning performance and traceability relative to leaving all structure in unstructured context.

The convergence should therefore be stated modestly. The literature identifies compatible elements: effect-boundary governance, deterministic authorization, evidence-producing controls, component decomposition, and explicit policy structure. The banking determination layer is a synthesis and specification of those elements for regulated action. It remains to be empirically validated.

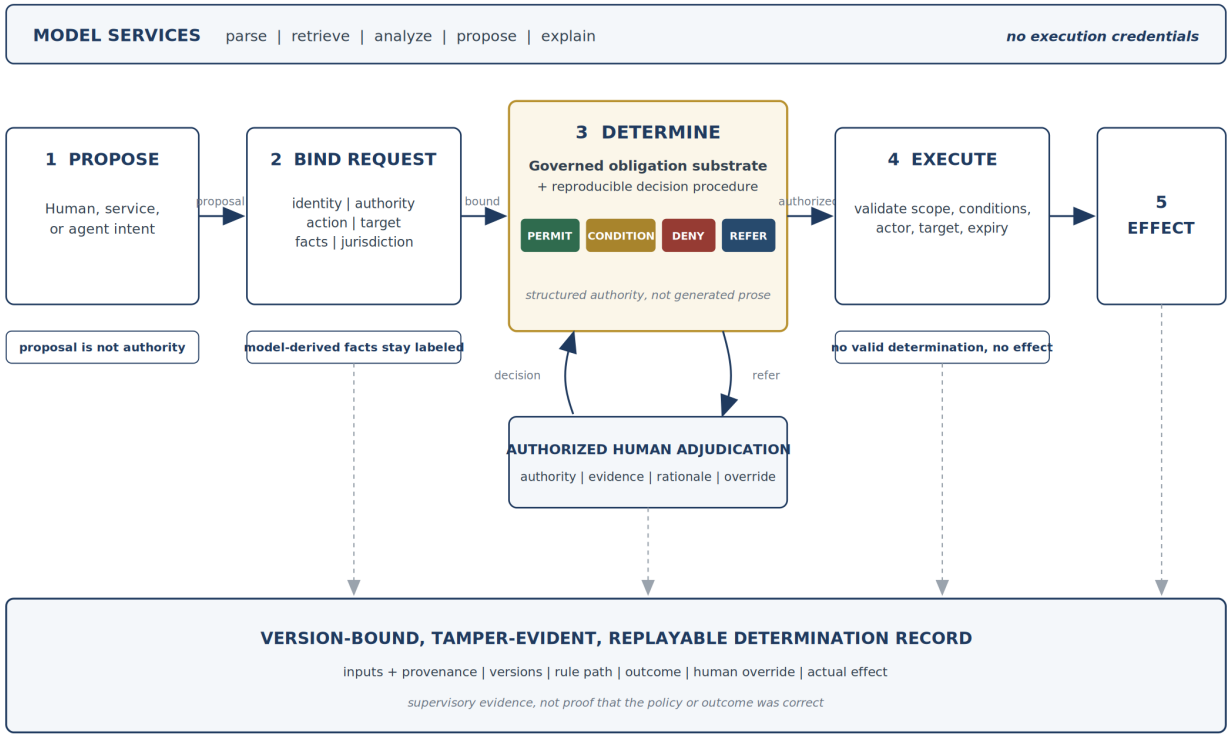

Figure 1. Governance before action: the determination-layer pattern.

Models may propose. Institutions determine. Systems execute. Models do not possess the execution credentials that convert a proposal into an institutional effect. Human adjudication, where required, is itself authenticated, authorized, and recorded.

4. The Determination Layer: Six Constitutive Requirements

A determination layer is not a dashboard, a prompt template, a retrieval system, or a log added to an agent. It is the institutionally controlled mechanism through which a proposed action becomes authorized. Six requirements are constitutive.

4.1 A bound request

Before a decision can be made, the system must bind the request to a structured, inspectable object. At minimum, that object identifies:

- the human, service, or agent proposing the action;
- the principal on whose behalf it acts and the scope of delegated authority;
- the action requested and the system or real-world object it would affect;
- the relevant customer, account, transaction, product, jurisdiction, business line, and time;

- the facts and documents relied upon, with provenance and freshness;
- the material uncertainties, contradictions, and missing fields.

This step is not clerical. It is where natural-language ambiguity becomes a governance risk. A model may assist with extraction and classification, but a model-generated field is not automatically a fact. The request schema must distinguish asserted, retrieved, calculated, inferred, and human-confirmed information. High-impact actions should require stronger identity, authority, and fact validation than low-impact advisory tasks.

The binding process should also be capable of refusing to construct a valid request. An unresolved actor, ambiguous account, conflicting jurisdiction, or missing mandatory fact is not an invitation for the model to choose the most likely interpretation. It is a defined failure state that triggers clarification, denial, or referral according to action criticality.

4.2 A governed obligation substrate

The layer requires an explicit representation of the obligations and institutional constraints against which requests are evaluated. This substrate may contain:

- primary legal and regulatory sources;
- jurisdictional applicability and effective dates;
- regulatory obligations, prohibitions, permissions, exceptions, and cross-references;
- internal policies, risk appetite, delegated authorities, and product restrictions;
- control mappings, evidence requirements, and approval thresholds;
- precedented interpretations and documented areas of unresolved ambiguity.

The representation need not be a single universal ontology, and a graph is not mandatory. Graph structures are attractive where obligations cross-reference one another and where downstream impact must be mapped, but relational, rules-based, or hybrid representations may also be appropriate. The non-negotiable properties are governance and traceability: each operative element must have a source, owner, jurisdiction, effective period, version history, rationale, review status, and declared limits.

The substrate is not "the law in a database." Encoding is itself interpretation. A node can be wrong, an exception can be omitted, and an internal policy can be more restrictive than law. The architecture makes those judgments inspectable; it does not make them infallible.

4.3 A reproducible determination procedure

Given a bound request and a fixed substrate version, the layer applies a specified procedure and returns one of four outcomes:

- **Permit:** the action may proceed within the stated scope.
- **Permit with conditions:** the action may proceed only after stated controls, evidence, approvals, or limits are satisfied.
- **Deny:** the action is prohibited under the applicable rule set.
- **Refer:** authorized human judgment is required because the case is outside coverage, materially ambiguous, conflicting, or reserved by policy.

Reproducibility means that the same normalized request, data versions, policy version, and decision procedure produce the same determination. It does not mean that the underlying policy is objectively correct. Deterministic execution and policy correctness are separate governance questions.

The procedure should expose the rule path rather than produce a post-hoc narrative. The explanation may be rendered in natural language by a model, but the authoritative result is the structured determination: the rules evaluated, facts matched, exceptions considered, conditions imposed, and reason for denial or referral.

4.4 A coterminous effect boundary

Every consequential effect within scope must require a valid determination token or equivalent authorization object. Models and orchestration components should not possess direct credentials that allow them to bypass the determination layer and call effect-producing systems independently. Execution adapters should verify the determination's validity, scope, expiry, actor, target, and conditions before acting.

This is the architecture's hardest requirement because it is a statement about completeness. Ninety-nine percent coverage is not equivalent to complete mediation where the remaining one percent can move money, expose data,

alter rights, or create a filing obligation. The bank must identify all effect-capable interfaces - APIs, databases, robotic process automation, message queues, human workarounds, batch jobs, and vendor connections - and show how each is brought within the boundary or explicitly excluded from the agent's capability set.

The requirement is also narrower than "the model can never act." The model can initiate a proposed action and can orchestrate permitted tasks. It cannot unilaterally create the authority to execute them.

4.5 Risk-tiered fail-safe behavior

A universal deny-on-uncertainty rule sounds conservative but can create customer harm, operational outages, liquidity problems, and perverse incentives to bypass the system. The appropriate principle is risk-tiered fail-safe behavior.

For actions that are legally consequential, financially irreversible, customer-restrictive, privacy-sensitive, or outside declared coverage, unresolved identity, authority, jurisdiction, mandatory facts, or policy conflict should prevent automated execution and trigger denial or referral. For low-risk, reversible, non-effectful tasks, the system may degrade gracefully - for example, by returning an advisory answer labeled with its limitations. The action class, not a general slogan, determines the safe failure mode.

Fail-safe design should also account for availability. Institutions need tested contingency procedures for substrate outages, stale data feeds, identity-provider failures, and emergency operations. A governance boundary that cannot fail safely becomes a new concentration and resilience risk.

4.6 A version-bound, tamper-evident, replayable record - including overrides

Every determination should generate a record containing:

- the bound request and provenance of material facts;
- actor and delegated-authority evidence;
- substrate, data, model, and policy versions consulted;
- rules and exceptions evaluated;
- the structured outcome and conditions;
- the generated explanation, if any, distinguished from the authoritative rule path;
- the execution attempted and the execution actually completed;
- any referral, human decision, override, second review, and rationale;
- timestamps, integrity evidence, and links to supporting artifacts.

"Tamper-evident" describes the property, not a mandated technology. Hash chains, signed records, write-once storage, database controls, and independent attestations may all contribute. "Replayable" means that an examiner or independent reviewer can reconstruct the determination against the historical versions that governed it. A current rule set is not a substitute for the rule set that existed when the action occurred.

Human override is part of the architecture, not an exception outside it. Overrides must be limited by role and action class, require a reason and supporting evidence, and generate their own determination record. Repeated overrides should feed governance review. Human involvement does not cure weak architecture if it is informal, unauthenticated, or invisible.

Table 1. Constitutive requirements and supervisory evidence

Requirement	Primary risk addressed	Evidence an examiner should be able to inspect
Bound request	Ambiguous identity, authority, target, jurisdiction, or facts	Structured request; provenance; validation state; missing or contradictory facts
Governed obligation substrate	Hidden, stale, or untraceable policy interpretation	Source links; owners; versions; effective dates; review and change history
Reproducible determination	Model-generated or inconsistent authority	Rule path; normalized inputs; outcome; conditions; deterministic replay
Coterminous effect boundary	Bypass of governance through tools, APIs, or credentials	Capability map; credential architecture; execution-token validation; bypass tests

Requirement	Primary risk addressed	Evidence an examiner should be able to inspect
Risk-tiered fail-safe behavior	Unauthorized action or harmful blanket denial	Action taxonomy; failure modes; referral rules; resilience and contingency tests
Replayable record and governed override	Unreconstructable decisions or unaccountable human intervention	Integrity controls; historical versions; execution match; override authority and rationale

5. What the Architecture Does Not Solve

A credible architecture is defined as much by its limits as by its claims. The determination layer changes the object of governance. It does not eliminate governance.

5.1 The Scope of Machine-Readable Determinations

The proportion of banking determinations that can be represented within a governed determination substrate remains an empirical question. Rather than estimating that proportion, this paper deliberately treats it as a measurable research problem.

Some classes of institutional determinations are naturally amenable to explicit representation. These include jurisdictional applicability, permissions, segregation-of-duty requirements, information barriers, documentation requirements, filing deadlines, eligibility conditions, approval hierarchies, and obligations whose interpretation has become stable through regulation, policy, or established supervisory practice.

Other determinations remain inherently contextual. Principles-based standards, novel factual situations, competing regulatory objectives, evolving supervisory expectations, and judgments requiring proportionality or commercial discretion continue to require human interpretation. These cases define the outer boundary of any determination-layer architecture rather than its failure.

The practical objective is therefore not to eliminate institutional judgment but to isolate the subset of determinations that can be represented explicitly, leaving the remaining cases visible, bounded, and intentionally escalated rather than silently absorbed into model behavior. Future empirical work should measure this boundary using real banking workflows, production decision traffic, and regulatory change events.

5.2 It can make systematic errors faster

A model can be inconsistently wrong. A deterministic engine can be consistently wrong. If a source is misread, an exception omitted, or a policy version misapplied, the determination layer may produce the same erroneous decision at scale. Reproducibility improves detection and remediation; it does not reduce the need for substantive review, legal challenge, testing, and independent validation.

This is why substrate governance must include source-level review, dual control for material changes, test suites, effective-date handling, negative and boundary cases, and impact analysis. A bank should be able to identify every determination affected by a corrected rule. That capability reduces correction cost, but it does not prevent the initial error.

5.3 It does not remove model risk from hybrid systems

Intent parsing, entity extraction, document classification, risk scoring, anomaly detection, ranking, summarization, and explanation may all involve models. Those components can fail in ways that distort the bound request or the information presented to a human. The revised interagency guidance's exclusions do not justify treating the whole system as non-model infrastructure. Each component should be inventoried and governed under the risk framework appropriate to its function and materiality.

The determination layer narrows what model error can directly authorize. It does not make model quality irrelevant. A poor parser can drive excessive referrals, omit facts, or mislead reviewers. Behavioral testing, monitoring, cybersecurity, data governance, third-party oversight, and human-factors testing remain necessary.

5.4 It does not guarantee fair or lawful policy

Structural completeness answers whether all effects pass through the policy. It does not answer whether the policy itself is fair, lawful, proportionate, or aligned with public values. A perfectly enforced discriminatory rule is still discriminatory. A complete record of an unlawful action is still evidence of an unlawful action.

Policy governance therefore remains upstream and human. Legal, compliance, business, risk, model-risk, data, security, and customer-impact stakeholders may all have roles depending on the use case. The architecture makes the policy surface inspectable and challengeable; it does not substitute engineering certainty for normative legitimacy.

5.5 Fail-safe controls can create their own harms

Denying or referring uncertain actions may protect against unauthorized execution, but it can also delay payments, freeze legitimate customers, increase false positives, create accessibility burdens, or concentrate decisions in overwhelmed review teams. A serious implementation must measure both false permits and false denials, as well as referral latency, abandonment, and disparate impact.

The right comparison is not "safe deterministic system" versus "unsafe model." It is the total risk of alternative architectures, including operational resilience, customer outcomes, human workload, and the incentives created by friction.

5.6 A record is evidence, not proof

A replayable record allows an examiner to ask better questions. It does not prove that the facts were true, the rule was correct, the human acted independently, or the execution system behaved exactly as logged. Independent data controls, reconciliations, access management, audit, and incident testing remain necessary. The record is valuable because it localizes challenge and makes claims falsifiable, not because it ends the examination.

6. An Examination Protocol for Agentic Systems

The practical value of the determination layer is that it gives supervision an inspectable object between model behavior and bank action. An examination can move from general assurances about "responsible AI" to six concrete tests.

6.1 Decompose the system by function

The bank should present a component inventory identifying which elements parse, retrieve, predict, recommend, determine, execute, monitor, and record. Each component should have an owner, risk classification, data dependencies, third-party dependencies, change process, and applicable governance framework. The inventory should prevent two opposite errors: treating all software as a model, and treating the entire agentic system as outside model risk because one component is deterministic.

The inventory should include human roles and operational procedures, not just software. Human reviewers, emergency processes, vendor support functions, and manual execution channels can all create or bypass consequential effects.

6.2 Enumerate consequential effects

The bank should define the action universe the agent can affect. The list should include direct effects and indirect commitments: database writes, payment messages, access changes, customer communications, case closure, regulatory submissions, trade instructions, policy changes, and human task creation where the task itself carries a presumption or deadline. For each effect, the examiner should be able to identify the authorization path.

The analysis should distinguish advisory outputs from effectful outputs. A drafted suspicious activity report is not the same as a filed report. A recommended account restriction is not the same as a restriction entered into the system of record. That distinction should be explicit in both permissions and logs.

6.3 Test boundary completeness

The examiner should attempt to bypass the determination layer. Testing should cover direct API calls, inherited credentials, tool chaining, prompt injection, alternate user interfaces, robotic process automation, batch jobs, retry behavior, vendor connectors, and human workarounds. The bank should demonstrate that an effect-producing system rejects action without a valid, scoped, unexpired determination and that the determination cannot be replayed against a different actor, account, amount, or action.

This test is more demanding than reviewing a diagram. The relevant evidence includes credential design, network and application controls, capability maps, code and configuration review, negative testing, and reconciliation between authorized and completed effects.

6.4 Challenge the substrate and request binding

The examiner should select material obligations and trace them from primary source through encoding, institutional policy, tests, and resulting determinations. The review should include effective dates, jurisdiction changes, exceptions, conflicting sources, and intentionally ambiguous cases. It should also test whether model-extracted inputs retain provenance and confidence, whether mandatory facts can be omitted, and whether contradictions trigger the intended failure mode.

A useful test is correction impact. When an encoded rule is found to be wrong or changes, can the bank identify affected policies, controls, determinations, customers, and actions? Can it distinguish decisions made under the historical rule from decisions that would be made today?

6.5 Replay determinations and reconcile execution

The examiner should select samples across permit, conditional permit, deny, refer, and override outcomes. For each, the bank should reproduce the determination using the historical request, data, model, and substrate versions, then reconcile the authorized effect to the effect actually completed. Differences should be explainable and governed.

Replay should not depend on the continued availability of a frontier model that has since changed or been withdrawn. Where a model contributed to input construction or explanation, the record should preserve the material outputs and versions needed to understand its role. The authoritative rule path should remain independently inspectable.

6.6 Review human adjudication and outcomes

Referral is not a safe harbor. Examiners should review who received referred cases, whether the reviewer had authority and adequate context, the time available, the frequency of agreement with the model's recommendation, and the reasons for overrides. High agreement can reflect a good system, automation bias, or a non-independent review process. The institution should have evidence capable of distinguishing among them.

Outcome monitoring should cover false permits, false denials, referral rates, override rates, processing delays, customer complaints, disparate outcomes, incidents, and control bypass attempts. The architecture's purpose is not to maximize automated approvals. It is to improve the quality and accountability of authorization.

Questions the forthcoming RFI should ask

The announced interagency request for information is an opportunity to make architecture explicit without prescribing a vendor or data model. The agencies should ask:

1. Which agentic outputs can create legal, financial, customer, operational, or regulatory effects, and how are those effects authorized?
2. Can every consequential effect be shown to pass through an enforced governance boundary, including third-party tools and human execution channels?
3. What constitutes a replayable determination record, and for how long should the historical versions needed for replay remain available?
4. How should existing model-risk, operational-risk, compliance, information-security, third-party, and internal-control frameworks be allocated across hybrid agentic components?
5. When should unresolved identity, delegated authority, jurisdiction, or policy coverage block automated execution, and when is graceful degradation appropriate?

6. What metrics should institutions report for referrals, overrides, false permits, false denials, bypass attempts, correction latency, and customer impact?
7. How should supervisors assess the correctness and change governance of machine-readable obligation substrates without prescribing a particular representation or provider?

These questions are technology-neutral in the meaningful sense. They do not require a graph, a particular model, or a named platform. They require institutions to make authority, effects, and evidence explicit.

7. A Falsifiable Research Agenda

The architectural argument of this paper stands on the regulatory structure and convergent evidence already presented; it does not await the results below. What remains empirical is the operational profile of the architecture in production: the size of its advantages, the referral burden it carries, and where governance effort redistributes. That profile should be tested against alternatives, not assumed because the architecture is intuitively appealing. The relevant comparison includes at least four architectures:

1. a model-only system operating over prompt context;
2. a retrieval-augmented system that supplies approved documents;
3. a graph-grounded or structured-evidence system in which the model still renders the final compliance judgment;
4. a determination-layer system in which models help construct and explain requests but a separate governed procedure authorizes effects.

The benchmark should use real banking task distributions, appropriately anonymized or synthetically reconstructed under practitioner supervision. It should include both structurally resolvable and open-textured cases, cross-jurisdiction conflicts, stale and contradictory data, prompt injection, authority ambiguity, rule changes, and operational outages.

Accuracy alone is inadequate. The evaluation should measure:

- **Substantive correctness:** false permits, false denials, incorrect conditions, and inappropriate referrals.
- **Coverage and calibration:** the share of cases resolved, referred, or rejected, by action class and regulatory domain.
- **Traceability:** whether each outcome maps to authoritative sources, facts, and rule paths.
- **Replay fidelity:** whether the same historical versions reproduce the determination.
- **Boundary completeness:** the proportion of consequential effects technically incapable of bypassing governance.
- **Correction latency:** time required to encode a change, identify affected decisions, retest, and deploy safely.
- **Human workload:** referral volume, review time, escalation quality, and override rates.
- **Operational and customer impact:** delays, false restrictions, abandoned processes, disparate outcomes, and resilience under failure.
- **Governance effort:** validation, monitoring, change management, documentation, and incident-response cost across components.

Four hypotheses are particularly important:

- **H1:** A determination layer will reduce unauthorized automated effects relative to model-only and retrieval-augmented systems, at the cost of higher referral or denial rates where coverage is incomplete.
- **H2:** Explicit obligation structure will reduce correction latency and improve impact mapping after regulatory or policy changes.
- **H3:** Governance burden will shift from unbounded behavioral observation toward substrate validation, boundary assurance, and request-quality controls; it will not disappear.
- **H4:** The net value of fail-safe behavior will vary by action class, and poorly calibrated fail-safe rules can produce material operational and customer harm.

The 2026 regulatory changes provide concrete test cases. A governed system should be able to update the applicability and treatment of the revised model-risk guidance, identify affected internal policies, and retain the historical state. The EU's changed high-risk implementation dates provide a separate test of effective-date management and impact analysis. These are not substitutes for controlled experiments, but they illustrate why correction latency and historical replay matter.

The evidentiary status of the author's implementation should therefore be stated precisely. A working implementation of the determination layer exists, and its constitutive behaviors are verified by executable assertions

rather than by demonstration alone: a requester that cannot be resolved is refused rather than guessed at; identity cannot be self-asserted by the requesting system; facts asserted by an agent cannot, by themselves, support a permitted consequential action; a request outside the governed library resolves to an explicit out-of-scope outcome rather than a best-effort answer; reliance on a determination is time-bounded, and a permitted determination past its recheck date presents as stale; an agent cannot execute the consequential action that only an authenticated human may file; and a sealed determination record replays byte-for-byte against the versions that produced it. A reference implementation additionally places the doctrine of Section 4 under twenty-two assertions covering fail-closed treatment of unresolved requesters, a conditional-clearance lattice across engagement, wall-crossing, confidentiality, expiry, and breach-history states, and multi-regime evaluation returning jurisdiction-specific citations. Appendix A maps each constitutive requirement to its executable assertions, distinguishes the properties verified at the layer boundary from the policy choices reserved to the deploying institution, and pins each claim to the artifact version and dated run at which it passed.

These results are offered as an existence proof, not as evidence of superiority. They establish that the six requirements of Section 4 are simultaneously satisfiable in a working system and that conformance to them can be machine-checked. They say nothing about hypotheses H1 through H4, which no implementation, including the author's, has yet tested. The implementation is accordingly offered as a first candidate subject for the examination protocol of Section 6, not as its validation. The same standard governs this paper's own empirical claims: they are version-bound, pinned to the artifacts identified in Appendix A, and limited to what those artifacts attest.

Independent institutions and alternative implementations remain necessary. A valid result must survive comparison, external challenge, and publication of failures. The most important early outcome may be falsification: identifying domains where the determination layer adds complexity without improving authorization quality or total governance cost.

8. Conclusion: Governance Before Action

The question before regulators is no longer whether frontier AI will enter banking. That transition has already begun. Banks are actively deploying large language models, retrieval systems, and increasingly autonomous agents across advisory, operational, and decision-support workflows. The remaining question is not whether AI will participate in regulated activity, but where institutional authority should reside once it does.

This paper argues that authority should remain outside the model.

Frontier models excel at language, synthesis, retrieval, planning, and interaction. They are extraordinarily capable interfaces between humans and institutional systems. They should not, however, become the legal source of institutional permission. The determination of what an institution may access, recommend, disclose, approve, or execute should remain governed by explicit institutional representations that are independently versioned, auditable, reproducible, and subject to ordinary supervisory examination.

This architectural distinction changes the object of governance. The practical consequence is that AI governance shifts from supervising probabilistic behavior alone to supervising institutional authority itself.

Instead of attempting to infer institutional reasoning from model behavior, supervisors can inspect the obligation substrate, the determination procedure, the action boundary, and the replayable determination record. Institutions gain the ability to demonstrate not merely what an AI system happened to do, but precisely why a regulated action was or was not authorized.

The contribution of this paper is therefore not a new model architecture, a new knowledge graph, or a new rule engine. Those technologies already exist. The contribution is the specification of the institutional determination layer that separates probabilistic reasoning from institutional authority and makes regulated AI actions independently examinable.

The proposition is deliberately narrow:

> **The model may reason. The institution must determine. The system must show what happened.**

As agentic AI becomes embedded within financial institutions, governance will increasingly depend not only on what models know, but on what institutions can prove. Governance before action provides one architectural framework through which that proof can be constructed, examined, and improved.

References

Amalgamated Bank. 2026. "Amalgamated Bank Announces Collaboration with FIS and Anthropic to Advance AI in Financial Crimes Compliance." May 7, 2026. https://www.amalgamatedbank.com/news/amalgamated-bank-announces-collaboration-fis-and-anthropic-advance-ai-financial-crimes.

Baldwin, Wilder, and Sepideh Ghanavati. 2026. "Knowledge Graph Representations for LLM-Based Policy Compliance Reasoning." arXiv:2604.27713. https://arxiv.org/abs/2604.27713.

Bowman, Michelle W. 2026. "Artificial Intelligence in the Financial System." Speech at the Artificial Intelligence and the Economy Conference, May 1, 2026. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/newsevents/speech/bowman20260501a.htm.

Chung, Jiseong, Ronny Ko, Wonchul Yoo, Makoto Onizuka, Sungmok Kim, Tae-Wan Kim, and Won-Yong Shin. 2025. "GraphCompliance: Aligning Policy and Context Graphs for LLM-Based Regulatory Compliance." arXiv:2510.26309. https://arxiv.org/abs/2510.26309.

Council of the European Union. 2026. "Artificial Intelligence: Council Gives Final Green Light to Simplify and Streamline Rules." Press release, June 29, 2026. https://www.consilium.europa.eu/en/press/press-releases/2026/06/29/artificial-intelligence-council-gives-final-green-light-to-simplify-and-streamline-rules/.

Cyber Risk Institute. 2026. "Financial Services AI Risk Management Framework." Accessed July 18, 2026. https://cyberriskinstitute.org/artificial-intelligence-risk-management/.

Davidovic, Sonja, and Hervé Tourpe. 2026. "How Agentic AI Will Reshape Payments." IMF Notes 2026/004. International Monetary Fund. https://doi.org/10.5089/9781513533308.068.

Federal Banking Agencies. 2026. "Supervisory Guidance on Model Risk Management." Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency. April 17, 2026. https://www.fdic.gov/model-risk-management-revised-guidance.pdf.

FIS. 2026. "FIS Brings Agentic AI to Banking with Anthropic, Starting with Financial Crimes." Press release, May 4, 2026. https://www.fisglobal.com/about-us/media-room/press-release/2026/fis-brings-agentic-ai-to-banking-with-anthropic-starting-with-financial-crimes.

Grevenbrock, Nils, Zhengpu Zhao, and Nihil Patel. 2026. "Model Risk Management in the Age of AI." Moody's Analytics white paper.

McCann, Alan L. 2026a. "The Two Boundaries: Why Behavioral AI Governance Fails Structurally." arXiv:2604.27292. https://arxiv.org/abs/2604.27292.

McCann, Alan L. 2026b. "Effect-Transparent Governance for AI Workflow Architectures: Semantic Preservation, Expressive Minimality, and Decidability Boundaries." arXiv:2605.01030. https://arxiv.org/abs/2605.01030.

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.

Office of the Comptroller of the Currency. 2026. "OCC Issues Updated Model Risk Management Guidance." News Release 2026-29, April 17, 2026. https://www.occ.treas.gov/news-issuances/news-releases/2026/nr-occ-2026-29.html.

U.S. Department of the Treasury. 2026. "Treasury Releases Two New Resources to Guide AI Use in the Financial Sector." Press release, February 19, 2026. https://home.treasury.gov/news/press-releases/sb0401.

Appendix A: Conformance Evidence

This appendix records the evidentiary basis for the implementation claims made in Section 7. Each claim is bound to a specific artifact version and to the dated test run at which it passed. The test suites are not published with this paper; they are available to examiners, referees, and supervisory reviewers on request.

Requirement	Verified behavior under executable assertion	Coverage
Bound request (4.1)	Facts carry server-resolved provenance; a fact asserted by an agent cannot, by itself, support a permitted consequential action; an unresolved requester fails closed.	Verified
Governed obligation substrate (4.2)	A jurisdiction absent from the governed library resolves to an explicit out-of-scope outcome rather than a best-effort answer; multi-regime evaluation returns jurisdiction-specific citations, switching correctly between United States and United Kingdom obligations.	Verified

Requirement	Verified behavior under executable assertion	Coverage
Reproducible determination (4.3)	Identical requests resolve to the same determination record; a conditional-clearance lattice fixes verdicts across engagement, wall-crossing, confidentiality, expiry, and breach-history states.	Verified
Coterminous effect boundary (4.4)	Requester identity cannot be self-asserted; views reserved for authenticated reviewers are refused to system callers; an agent cannot execute the consequential action that only an authenticated human may file.	Verified
Risk-tiered fail-safe behavior (4.5)	Out-of-scope and needs-facts exist as first-class outcomes; clarification gates return questions rather than defaults; reliance is time-bounded, and a permitted determination past its recheck date presents as stale; the automated channel cannot produce a filing at all (see 4.4), so consequential actions fail toward human adjudication by construction.	Verified at the strictest single-tier setting. Finer action-class tiers are institutional policy set at deployment; once configured they become assertable and are examined under Section 6.
Replayable record and governed override (4.6)	A sealed record replays byte-for-byte against the versions that produced it; filing seals the record; unfiled records are labeled drafts.	Verified for replay, sealing, and draft labeling; any override must pass through the sealed human filing path verified above. Override role limits and rationale content are institutional policy, examined under Section 6.

Rows 4.5 and 4.6 reflect the architecture's division of labor, not absent capability. The live determination service ships at the strictest setting: the automated channel cannot file at all, so every consequential action already fails toward authenticated human adjudication, and the only automated outputs are records labeled as drafts. Which action classes may relax that default, who may override a determination, and what rationale an override must carry are policy decisions the layer is designed to receive from the deploying institution, not to make on its behalf. Once configured, those behaviors become assertable and fall under the examination protocol of Section 6.