Runtime Governance for Regulated AI

Why policy-only AI governance breaks in production

GOVERNED AI AT THE EXECUTION LAYER.

Audience: executive sponsors, product leaders, platform/security teams, legal/compliance stakeholders.

LSAS Stack by Inference Stack LLC
Enterprise AI Execution Authority™

ADDRESS
Inference Stack LLC
5365 W. 73rd Pl.
Arvada, CO 80003

CONTACT
+1 720.334.3838
www.lsas-stack.com
www.lsas-stack.com/contact

Executive thesis

Most enterprise AI governance stops at policy statements, approvals, access controls, or post-hoc logs. That is necessary, but it is not sufficient for high-stakes execution.

Regulated AI systems need runtime governance: controls that evaluate requests and outputs at the moment of execution, require evidence when stakes rise, and emit proof of why the system was allowed, constrained, blocked, redacted, abstained, or escalated.

78 %

of organizations report using AI in at least one business function.

McKinsey, 2025

21 %

of organizations using gen AI say they have fundamentally redesigned at least some workflows.

McKinsey, 2025

84 %

of financial organizations are implementing or planning frameworks to govern how AI is built, trained, used, and audited.

WEF, 2025

The problem has moved from experimentation to execution.

The enterprise question is no longer whether AI will enter core workflows. It already has. McKinsey's 2025 survey shows that AI adoption has accelerated sharply, yet workflow redesign and governance maturity still lag. That gap matters because risk is created at execution time: when a model generates a client-facing statement, touches protected data, relies on retrieval evidence, or triggers a real-world action.

In practice, teams often have principles without enforcement. A policy may say that sensitive outputs require evidence, that unsupported citations are unacceptable, or that PHI/PCI should not leave a workflow boundary. But unless those expectations are bound to runtime controls, the enterprise is relying on human memory, post-hoc review, or luck.

Runtime Governance for Regulated AI

Policy-only governance

- Defines rules and intent
- May require approvals or annual reviews
- Often relies on human review after content is generated
- Produces fragmented evidence across teams and tools

Runtime governance

- Applies rules in-line at the decision point
- Adapts control strength to workflow criticality and risk tier
- Requires evidence before trust is granted
- Produces structured proof that product, security, compliance, and leadership can all inspect

Why now

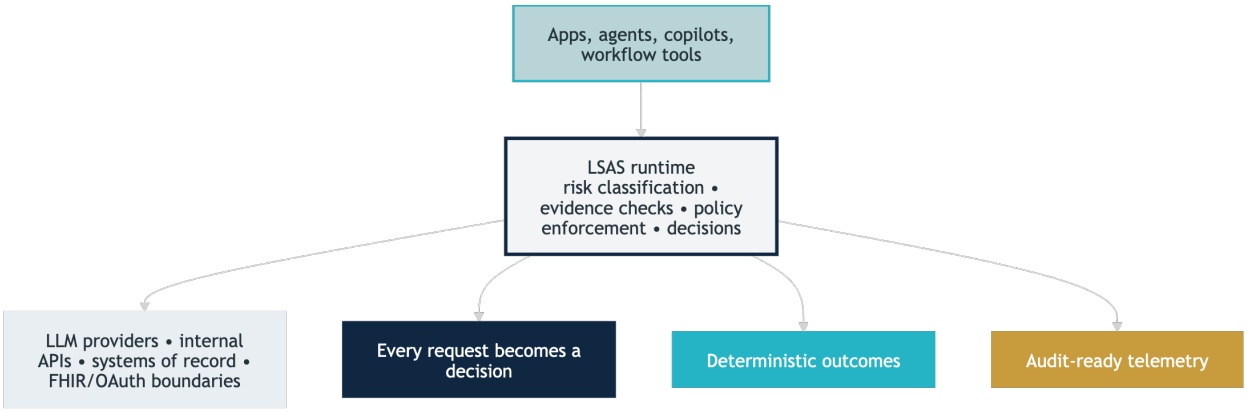
- AI is already embedded in day-to-day work. McKinsey reports broad enterprise adoption, while only a minority of organizations report having fundamentally redesigned workflows for the technology.
- Healthcare AI usage is now mainstream among physicians. The AMA reports that more than 80% of physicians use AI professionally, while privacy and the integrity of the patient-physician relationship remain top concerns.
- Legal and professional services adoption is surging. FTI reports that 87% of general counsel now report AI use in their departments, while Thomson Reuters finds that organizational GenAI use across professional services rose to 40% in 2026 from 22% the year before.
- Financial-services leaders are moving toward explicit governance frameworks. The World Economic Forum reports that 84% of financial organizations are implementing or planning frameworks that govern how AI is built, trained, used, and audited.

What runtime governance actually means

Runtime governance is the layer that sits between intent and execution. It translates governance policy into an operational decision system. At minimum, that system should answer five questions for every high-stakes interaction.

Question	What must be true
What workflow is this?	The request must be classified by domain, data sensitivity, business criticality, and system boundary.
What is the AI trying to say or do?	Claims, recommendations, citations, tool actions, and side effects must be made explicit rather than treated as undifferentiated output.
What evidence supports it?	Approved sources, retrieval results, structured records, deterministic validators, or human attestations must be available when risk demands them.
What is allowed under current policy?	Evidence thresholds, redaction requirements, escalation paths, release rules, and exception policies must be enforced consistently.
What proof exists after the fact?	The system should return decision metadata, rule hits, evidence status, remediations, policy version, and timeline artifacts for audit and operations.

Runtime Governance for Regulated AI



Conceptual view: LSAS as an application-layer governance runtime.

Where LSAS fits

LSAS — the Layered Safety and Accuracy System — is positioned as an application-layer runtime between apps, agents, APIs, and systems of record. Its value proposition is not merely that it logs activity. The stronger claim is that it classifies risk, inspects claims, checks evidence, enforces policy thresholds, decides at runtime, and emits proof for each governed interaction.

The practical framing

Not just who can use AI. What AI is allowed to say, do, and rely on — under what conditions, with what evidence, and with what proof.

That framing matters because it treats governance as execution authority, not as a static permission set.

How LSAS works

1. Classify risk

Identify whether the request enters a legal, clinical, financial, security, or customer-impacting workflow.

2. Inspect claims

Detect assertions, recommendations, citations, or actions that require stronger controls.

3. Check evidence

Assess grounding in approved sources, retrieval results, structured data, or deterministic validation.

4. Enforce thresholds

Apply policy rules for evidence sufficiency, redaction, validators, and escalation triggers.

5. Decide at runtime

Allow, constrain, redact, abstain, block, or escalate based on risk and proof sufficiency.

6. Emit proof

Return a decision envelope with findings, remediations, policy version, and audit metadata.

Three practical workflow stories

Clinical workflow

A clinician-facing note can be useful in seconds and dangerous in the same moment. In healthcare, the challenge is not simply summarization; it is whether protected information, unsupported recommendations, or inappropriate audience shifts survive into outbound use. Elsevier's 2025 clinician survey shows rising AI adoption alongside weak institutional support and training, while CHAI's 2026 patient survey shows that trust rises when accountability is visible.

Legal workflow

In legal and advisory work, the failure mode is often persuasive but unsupported content: fabricated citations, overstated conclusions, or confident summaries with no acceptable evidence chain. That is why runtime evidence thresholds, claim validation, and explicit escalation are more useful than generic 'responsible AI' messaging.

Fintech workflow

In payment and fintech contexts, the boundary is often operational. Sensitive values, secrets, and transaction data must not leak into prompts, outputs, or downstream actions. Here the job of runtime governance is to detect, redact, block, or reroute before the risky action is trusted.

What serious buyers should evaluate

- Can the system classify requests by workflow domain and risk, rather than treating all traffic the same?
- Can it inspect claims, actions, and evidence requirements explicitly?
- Can it enforce different thresholds for low-risk, standard, regulated, and locked-down modes?
- Can it support human escalation with explicit authority, time bounds, and audit proof?
- Can it emit decision-grade telemetry without requiring raw prompt retention by default?
- Can it operate inside private, customer-controlled boundaries such as a VPC or self-hosted environment?

Bottom line

The enterprise AI market does not need more abstract principles. It needs infrastructure that makes governance executable. Runtime governance is the missing bridge between policy intent and production trust. For regulated teams, the test is straightforward: can the system show why a high-stakes output or action was permitted, changed, withheld, or escalated — and can it do that consistently, at the boundary where trust is granted?

Selected references

1. LSAS Stack website — home. Application-layer runtime positioning, pipeline, risk domains, telemetry, and deployment options. https://www.lsas-stack.com/

2. LSAS Stack — Sandbox / Guided demo. Scenario list, decision envelope behavior, governed overrides, and audit metadata framing. https://www.lsas-stack.com/playground

3. LSAS Stack — Compliance posture. Controls mapping language, deployment/residency posture, data minimization, auditability, and operational hooks. https://www.lsas-stack.com/compliance-posture

4. McKinsey — The State of AI: How organizations are rewiring to capture value (2025). Covers AI adoption rates, workflow redesign, and governance oversight patterns. https://www.mckinsey.com/capabilities/quantumblack/our-insights

5. American Medical Association — AI usage among doctors doubles as confidence grows (2026). Physician AI usage, current use cases, and concerns around privacy and patient relationships. https://www.ama-assn.org/press-center/ama-press-releases/ama-ai-usage-among-doctors-doubles-confidence-technology-grows

6. Elsevier — Clinician of the Future 2025. Clinician AI adoption, institutional support gaps, training gaps, and workflow strain indicators. https://www.elsevier.com/insights/clinician-of-the-future/2025

7. Coalition for Health AI (CHAI) — Patient survey on health AI and transparency (2026). Public trust, comfort, accountability, and governance expectations in healthcare AI. https://www.chai.org/news/chai-releases-new-patient-survey-report-on-health-ai-and-transparency

8. FTI Consulting / Relativity — The General Counsel Report, AI adoption in legal departments (2026). Corporate legal department AI adoption, roadmap maturity, and common task areas. https://www.fticonsulting.com/about/newsroom/press-releases/ai-adoption-in-corporate-legal-departments-doubles-according-to-the-general-counsel-report

9. Thomson Reuters Institute — 2026 AI in Professional Services Report. Organizational GenAI adoption, workflow centrality, ROI measurement gaps, and client-work ambiguity. https://www.thomsonreuters.com/en/reports.html

10. World Economic Forum — Artificial Intelligence in Financial Services (2025). Responsible-AI governance patterns and adoption of AI governance frameworks in financial services. https://www.weforum.org/reports/

11. Inference Stack — Contact. Inference Stack contact page; based in Denver, Colorado; serving clients worldwide. https://www.inference-stack.com/contact