The Resilience Newsletter

THALAMUS ADVISORY

ISSUE 04 SEPTEMBER 2026 READING TIME · 11 MIN SERIES · SECTOR ANALYSIS

SECTOR ANALYSIS · FINANCIAL SERVICES & INSURANCE

The Seam Between Two Eras

A payment API written this year calls a core system written in 1987. Neither team can see the other's telemetry. That seam is where your most expensive incidents live — and where your regulator's deadline starts running.

THALAMUS ADVISORY • ENTERPRISE RESILIENCE PRACTICE

01 — THE SIGNAL

A deadline with a regulator attached

Operational resilience stopped being an engineering preference the moment supervisors started naming it. The Federal Reserve, OCC and FFIEC treat it as a safety-and-soundness matter. The EU's DORA attaches explicit reporting timelines to major incidents. For any institution with European exposure, an outage is now an event with a clock on it.

US financial services turns over roughly $4.5 trillion a year across banking, insurance, asset management and payments. JPMorgan alone spends around $17 billion annually on technology. The problem is not investment. It is that the investment has produced two estates that do not share a language.

02 — THE MECHANISM

Half the trace is missing

The characteristic banking incident starts as a customer-facing API latency regression and ends, three hours later, in a core system nobody on the call has credentials for.

Modern services are instrumented with OpenTelemetry, traced end to end, and owned by teams who practise SRE. Core deposit, card authorization and settlement systems are instrumented with whatever the platform provides, observed through different tooling, and owned by teams with a different operational culture and an older change calendar.

A trace that starts in the modern estate simply *stops* at the boundary. What happens beyond it is inferred from timing, which is another way of saying guessed.

> *You cannot correlate across a seam you cannot see across. Most banking incidents are not hard because the fault is subtle — they are hard because half the evidence is in a system the responder cannot query.*

Batch makes it worse. Mortgage servicing, settlement and regulatory reporting run long-lived jobs with hard deadlines. A process that fails at 02:00 and is noticed at 07:00 has already missed its window, and most observability is built for request-response rather than for a job that should have finished and did not.

03 — THE COST

Where it actually lands

SURFACE	HOW IT PRESENTS
Reporting deadlines	DORA and equivalent regimes require a defensible timeline within a fixed window. Reconstructing one from chat logs is a week of senior time you do not have.
Payment rails	Authorization and settlement downtime is directly revenue-linked and among the highest per-minute costs in any sector.
Peak concentration	Payday, month-end and benefit disbursement create sharp peaks where a capacity event and a reliability event look identical in early telemetry and demand opposite responses.
Acquisition seams	Merged estates rarely converge on one observability convention. The boundary becomes a jurisdictional argument during every incident.
Missed batch windows	A failed overnight job is discovered after the window closed, at which point the remediation is manual and the deadline is already gone.

04 — THE AI STAKES

Faster fraud, slower explanations

Two pressures arrive together. Attack volume is up — synthetic identity, automated application fraud, and social engineering at a quality and scale that was not economic three years ago. And model-driven decisioning is moving into credit, fraud and AML paths, which means more automated

decisions to explain.

For resilience specifically, the exposure is explainability. When an automated system acts on production during an incident, *"the model decided"* is not an account a supervisor accepts. The decision must be reconstructable with the evidence available at the time — and most implementations discard the probes once a conclusion is reached, keeping the summary and losing the reasoning.

05 — THE REMEDY

One standard across the seam

- **Adopt a telemetry standard both sides can accept.** Neither estate will move to the other's vendor agent. OTLP is adoptable by both because it belongs to neither.
- **Correlate into one problem with one named cause.** A cascading breach across nine services is one incident, not nine pages.
- **Add throughput SLOs for batch.** A floor, where absence of completion is a breach. Latency and error rate say nothing about a job that never ran.
- **Scope objectives by entity and jurisdiction.** A breach in one legal entity must be visible as a breach, not averaged into a global figure that looks fine.
- **Retain evidence, not just conclusions.** Probes, results and confidence scores — the artefact a supervisor asks for is the reasoning, not the summary.

06 — IN PRACTICE

Hosted, or customized for your estate

Hosted on Thalamus AI Cloud

Managed instances provisioned per region, with dedicated tenancy where a shared database is not acceptable. For digital channels, customer-facing APIs, data platforms and the modern half of the estate generally, this is the fast path — and the one that gets a correlated problem feed running while the core-side conversation is still in progress.

Customized for your enterprise

For core banking, card and settlement estates, the platform deploys self-hosted inside your own boundary. Customization here is usually three things: SLO definitions that map to contractual and regulatory commitments rather than internal targets; per-entity and per-jurisdiction tenancy so each supervised entity is measured against its own obligations; and postmortem output shaped to your regulatory reporting template so the incident record is the filing input rather than a source for it.

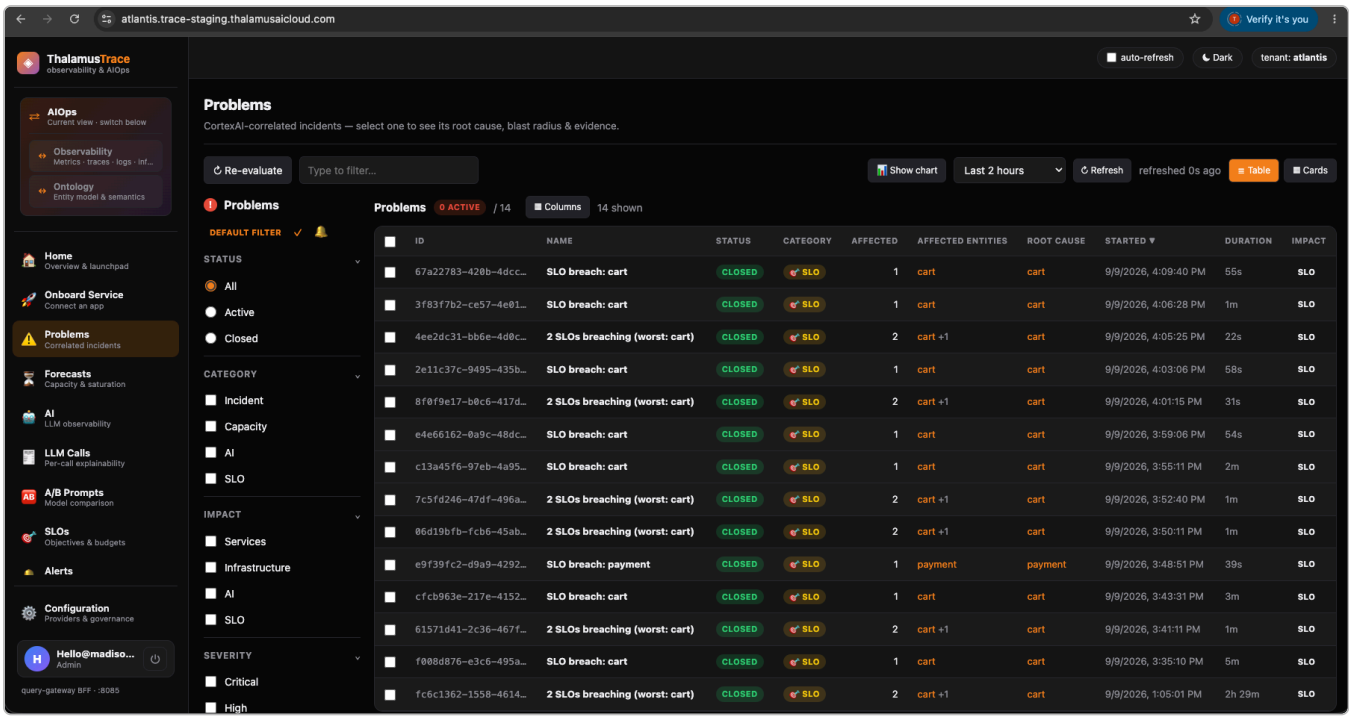

FIGURE 1 — THALAMUSTRACE · CORRELATED PROBLEMS

One fault, one record, with a clock on it. Each correlated problem carries *Started*, *Duration*, a root-cause entity and the full affected set — so *"2 SLOs breaching (worst: cart)"* is a single incident rather than two pages to two teams. For a supervised institution the *Started* column is not a convenience: it is the first field of the regulatory timeline.

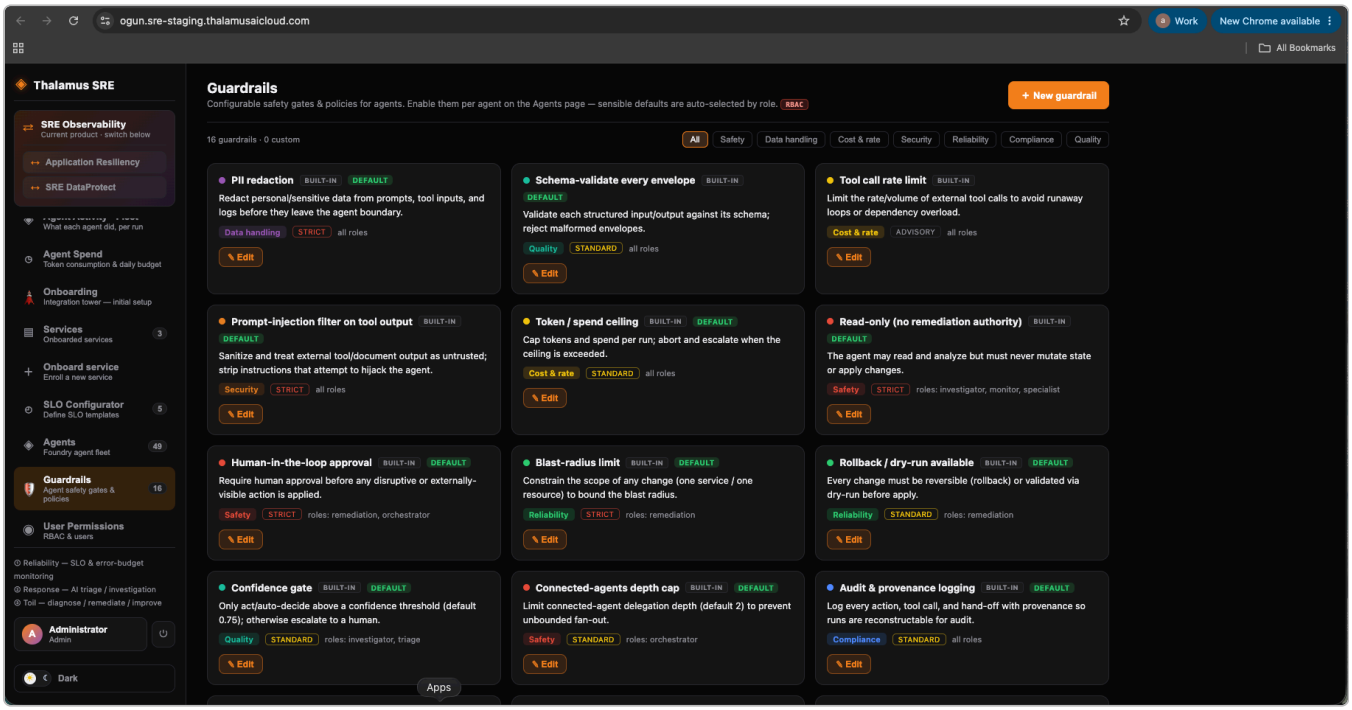

FIGURE 2 — THALAMUS SRE · GUARDRAILS

Controls enforced outside the model. Sixteen policies applied at the execution boundary rather than requested in a prompt — *human-in-the-loop approval* before any disruptive action, *blast-radius limit* constraining change to one service or resource, *rollback / dry-run available* requiring every change be reversible, and *audit & provenance logging* recording every action and hand-off so a run is reconstructable for an examiner.

07 — GETTING THERE

How Thalamus Advisory helps

ENGAGEMENT	WHAT IT PRODUCES
Seam Assessment 3–4 weeks	Where traces terminate at the modern-to-core boundary, what it costs you per incident, and the minimum instrumentation that restores continuity without a core change programme.
Regulatory Evidence Readiness 3 weeks	Measured against the deadlines that actually apply to you: how long it currently takes to produce a defensible timeline, and what would have to change to produce it as a by-product.
Batch Observability 4–6 weeks	Throughput SLOs across settlement, servicing and reporting jobs so a missed window is an alert rather than a morning discovery.
Guarded Rollout 10–14 weeks	Self-hosted deployment across a bounded domain, per-entity SLOs, guardrails mapped to your risk framework, and evidence retention configured to your examination cycle.

THE COUNTER-ARGUMENT, MADE HONESTLY

Large banks already invest heavily in resilience and have sophisticated incident practice. The marginal value here is not basic monitoring, and anyone pitching it as such has not looked at your estate.

The defensible claims are narrower: correlation across seams you currently cannot cross, per-entity granularity that aggregate reporting hides, and evidence generated rather than reconstructed. If your traces already span core and your postmortems already write themselves, spend the money elsewhere.

ONE NUMBER

Five hours. A common gap between an overnight batch failure and its discovery — by which point the settlement or reporting window has closed and the remediation is manual.

No latency threshold detects a job that never started. Only a throughput floor does, and almost nobody sets one.

Where to start on Monday

Pick your three most critical overnight jobs. Ask: if this did not run tonight, how would we find out, and when?

If the answer is “someone notices in the morning,” you have a throughput SLO to write — and it is the cheapest reliability improvement available to you this quarter.

Thalamus Advisory

Enterprise resilience for the age of autonomous systems. We help organisations build operational practice that stays trustworthy when the systems making decisions are no longer entirely human.

THE RESILIENCE NEWSLETTER · ISSUE 04 · SEPTEMBER 2026
Every figure we print carries its assumptions. Where we model, we say so.
Written for engineering leaders. Forward it to one.