Observability is the New Governance: A Roadmap to Audit Ready Agents

From Hype to Hard Truths: Closing the 2% Autonomy Gap

Over the past two years, agentic AI has become a priority for financial institutions seeking efficiency gains and improved service responsiveness. Yet, while experimentation has accelerated, production-grade autonomy remains exceedingly rare. The UK's cross---sector AI survey confirms that only 2% of AI use cases are fully autonomous, even as overall adoption stands at 75%. This mismatch reflects a set of structural challenges:

- **Limited end-to---end traceability:**
  Most organizations cannot reconstruct an agent's full decision path, including which documents were retrieved, what intermediate reasoning steps occurred or how tools were invoked.
- **Fragmented metadata and inconsistent provenance:**
  As teams prototype with multiple models, frameworks, vector databases, and toolchains, telemetry becomes inconsistent and non-portable.
- **Compliance and operational resilience hurdles:**
  Regulations now require documented evidence of decisions, controls and system robustness under stress, none of which traditional logs can provide.

Where use cases feel this most acutely

The gap is tolerable for low stakes- summarization or chatbot queries, but not for workflows such as:

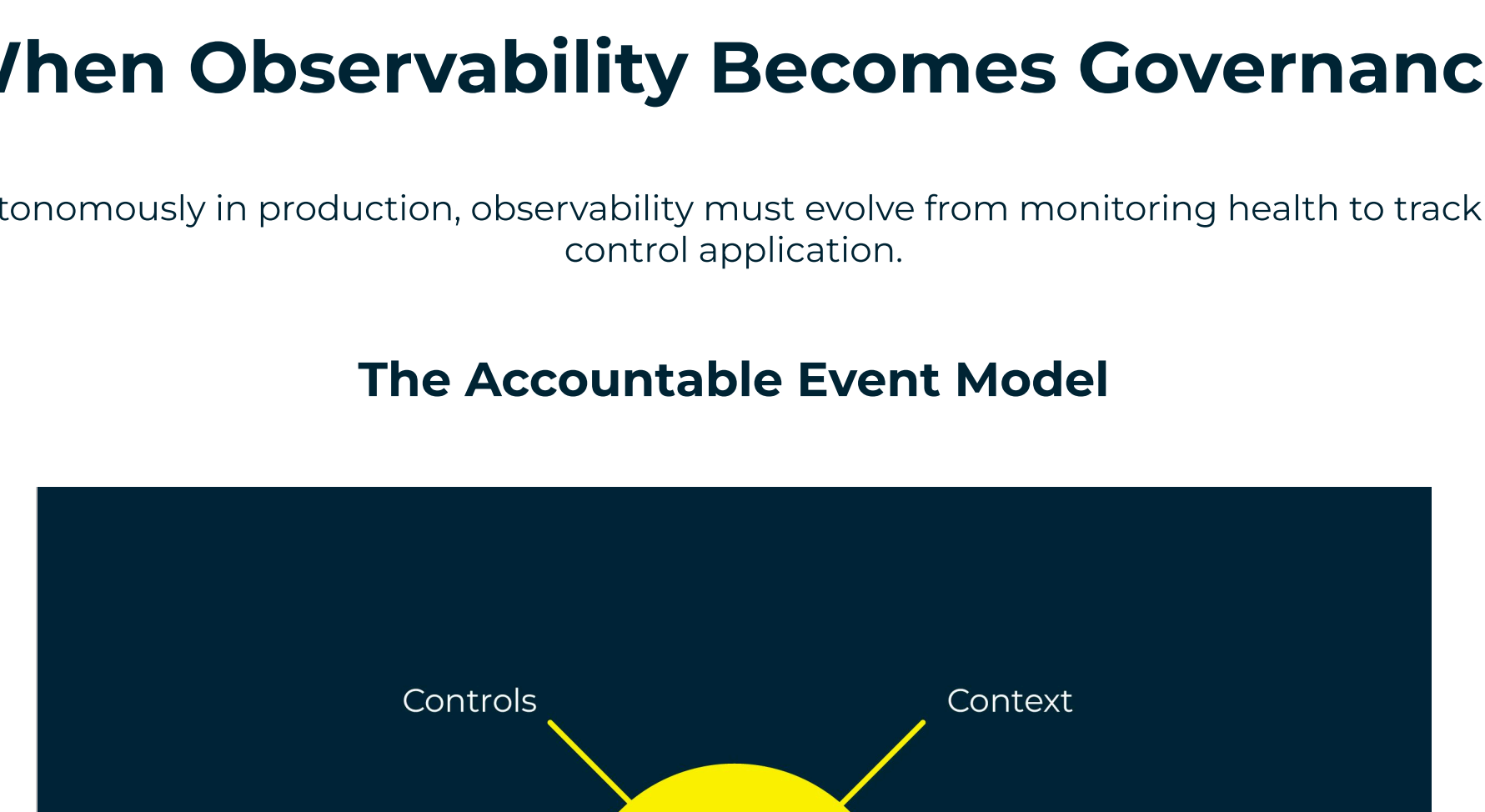
- **KYC refresh,** where regulatory expectations demand traceability.
- **Trade surveillance,** where alerts must be explainable.
- **Credit restructuring,** where reasoning steps must be justified.

These are precisely the workflows with the highest ROI potential, but only if supported by governance-grade visibility.

When Observability Becomes Governance

For agentic AI to operate autonomously in production, observability must evolve from monitoring health to tracking behavior, decisions and control application.

The Accountable Event Model



Every agent run must produce a structured, tamper evident record containing-:

- **Intent:** What the agent attempted to achieve.
- **Context:** Which data sources and retrievals were used.
- **Decision:** Why the next step was selected.
- **Execution:** What tools were called and with what results.
- **Controls:** Guardrails and constraints applied.
- **Outcome:** Business impact.
- **Owner:** The accountable human for the workflow.

This approach aligns with the EU AI Act's emphasis on lifecycle-long logging and traceability for high--risk- systems.

Addressing The Two Structural Blind Spots

i.) Dark tool calls

When agents invoke tools through un-instrumented wrappers, organizations lose visibility into action parameters and outcomes. This introduces unacceptable audit and security risks. In regulated environments, a runtime policy of "no telemetry, no execution" becomes essential to ensure all actions are observable.

ii.) Excessive agency

Overpermissioned agents can escalate privileges, execute financial -transactions or access sensitive records without proper bounds. Least-privilege access, explicit allowlists and human approval thresholds are required- for regulated environments. This shift from logging events to governing behavior defines the new operational foundation for agentic AI.

Regulation Is Driving Architecture

Agentic AI is no longer an experimental layer sitting at the edge of the business. For regulated industries, it is now treated as core operational infrastructure, subject to the same expectations around transparency, resilience and accountability as other critical systems

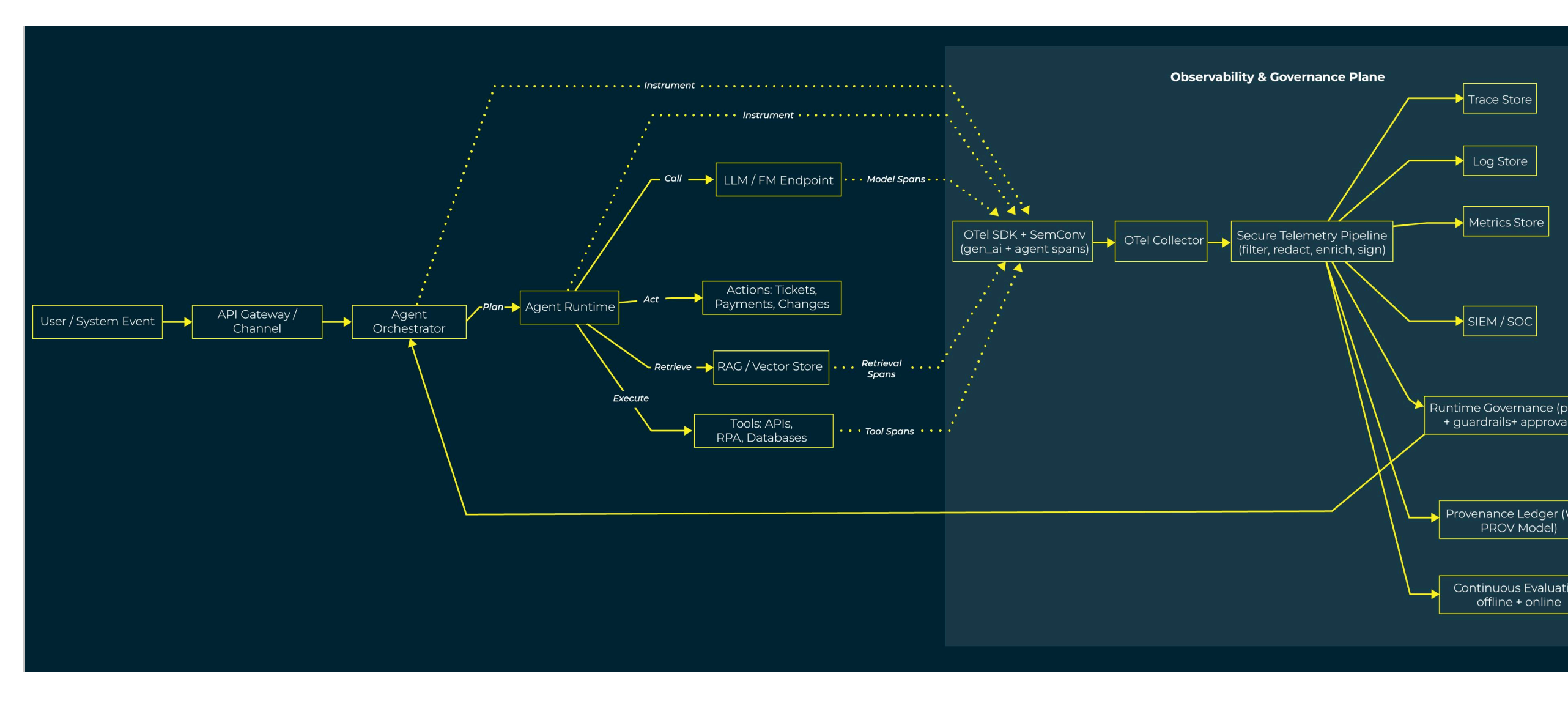
The EU AI Act makes this clear. For high-risk AI systems, firms must be able to show how decisions are made and how models behave in real operating conditions over time. This isn't about documentation or model cards; it's about having reliable, production-level evidence that explains what happened, why it happened and how the system changed. For agentic AI used in areas like onboarding, credit decisions or fraud detection, observability becomes the way compliance is demonstrated.

Operational resilience rules reinforce the same shift. Under DORA, firms must prove that the systems supporting important business services can withstand disruption and recover predictably. As AI agents increasingly support customer interactions, transaction processing and risk workflows, those agents fall directly within scope. If an AI-enabled process fails, firms must be able to trace the cause and show how controls worked in practice.

In the UK, impact tolerance requirements add further pressure. Firms must demonstrate that key services remain within acceptable limits even under stress. As decision-making is delegated to AI, this requires clear visibility across the full chain behind an agent, from models and data sources to orchestration and external tools. Without that visibility, resilience claims cannot be evidenced.
Together, these regulations point to a simple conclusion: observability is no longer optional or technical. It is a core design requirement for running agentic AI safely, at scale and under regulatory scrutiny.

The Case for Open, Portable AI Telemetry

The observability market is consolidating rapidly, and proprietary solutions introduce durability risks, highlighted most clearly by the discontinuation of WhyLabs following Apple's acquisition. To ensure longevity and portability, organizations are turning to open standards.


User / System Event
API Gateway / Channel
Agent Orchestrator
Plan
Agent Runtime
Instrument
Instrument
Call
LLM / FM Endpoint
Model Spans
Act
Actions: Tickets, Payments, Changes
Retrieve
RAG / Vector Store
Retrieval Spans
Execute
Tools: APIs, RPA, Databases
Tool Spans
Observability & Governance Plane
OTel SDK + SemConv (gen_ai + agent spans)
OTel Collector
Secure Telemetry Pipeline (filter, redact, enrich, sign)
Trace Store
Log Store
Metrics Store
SIEM / SOC
Runtime Governance (policy + guardrails+ approvals)
Provenance Ledger (W3C PROV Model)
Continuous Evaluation offline + online

Open-Telemetry GenAI Semantic Conventions

These conventions define standardized spans for:

- LLM calls (inputs, outputs, parameters)
- RAG retrievals
- Agent-initiated- tool calls
- Safety checks and policy evaluations
- Token usage and cost metrics

This provides a portable schema that can be analyzed across any observability backend.

W3C Trace Context

A unified trace identifier enables cross-service propagation and correlation across distributed microservices, cloud providers, and third--party- tools.

The result is a vendor-agnostic, regulation--ready visibility layer capable of supporting long--term- operational resilience.

Accountability Metrics for the Boardroom

Traditional AI metrics (token throughput, latency, accuracy) do not align with governance needs. CIOs must adopt accountability-driven- KPIs that reflect operational resilience and regulatory compliance.

Core KPIs

- **Impact tolerance- breaches**: Whether agentic workflows contributed to disruptions in important business services (aligns with UK/FCA requirements).
- **Policy-violation- rate:** Frequency of attempts to bypass guardrails (supported by guardrail evaluation frameworks such as NVIDIA NeMo Guardrails).
- **Human approval- rate:** Percentage of decisions requiring manual intervention, a proxy for readiness.
- **Change-induced- regression:** Performance degradation after prompt/model/policy changes.
- **Blocked dark tool- -call attempts:** Evidence of enforcement effectiveness.

These metrics connect directly to business risk and assurance, enabling transparent oversight for executive teams.

A 0–24 Month Roadmap for Governing Agentic AI

This roadmap reflects how regulated organizations move agentic AI from early deployment to scaled, governed production. The focus is on building the right foundations first, then strengthening controls as autonomy increases.

Months 0–3: Establish the Foundation

Start with a small number of high-impact workflows. Define what decisions agents can make, where human approval is required and what evidence must be retained. Standardize how agent activity is captured across models, data retrieval and tools and set basic rules for retention, access and sensitive data handling.

Outcome: Clear evidence requirements and consistent visibility across priority workflows.

Months 3–6: Secure Observability

Treat telemetry as control evidence, not just logs. Centralize how agent activity is filtered, protected and stored and ensure sensitive data is handled safely. Begin integrating AI observability into existing risk, security and incident processes.

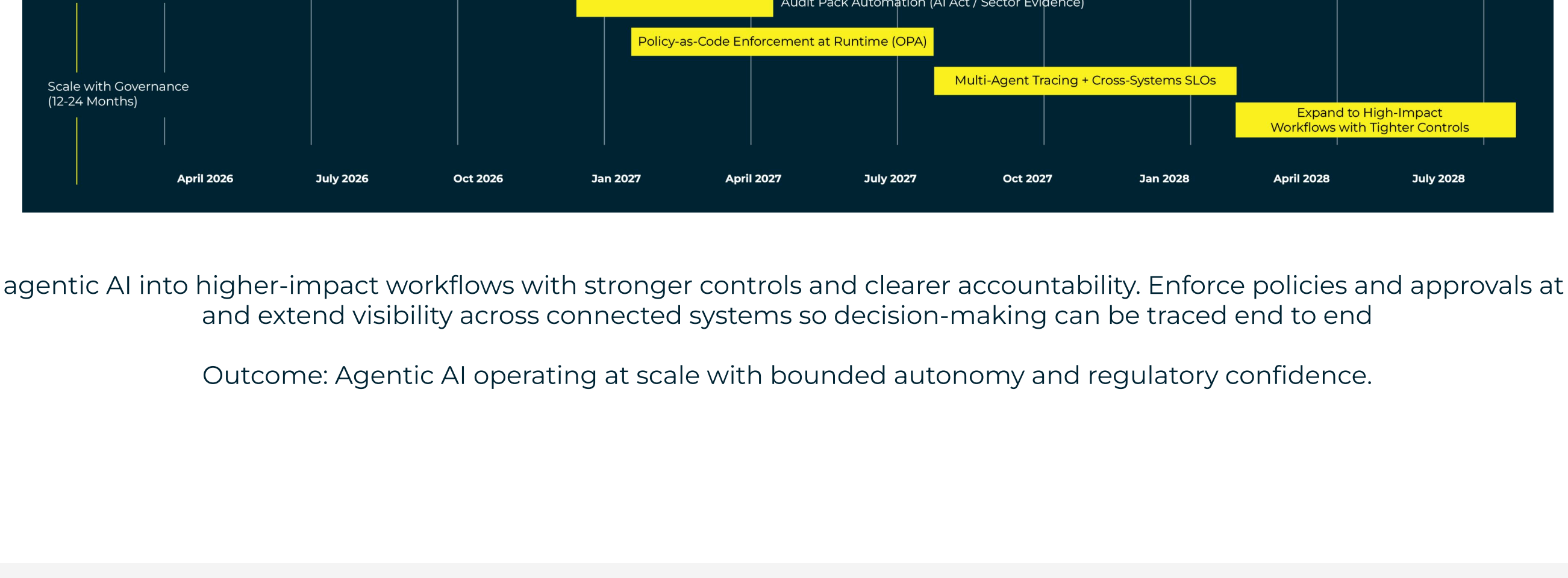
Outcome: A secure telemetry pipeline that supports audit and investigation.

Months 6–12: Build Continuous Assurance

As systems evolve, introduce regular checks to ensure agents continue to behave as expected. Focus monitoring on real risks (incorrect actions, policy violations and failures under stress) and reduce manual effort by automating evidence collection over time.

Outcome: Ongoing confidence that agent behaviour remains safe, explainable and compliant.

Months 12–24: Scale with governance



Expand agentic AI into higher-impact workflows with stronger controls and clearer accountability. Enforce policies and approvals at runtime and extend visibility across connected systems so decision-making can be traced end to end

Outcome: Agentic AI operating at scale with bounded autonomy and regulatory confidence.

Conclusion

In 2026, the differentiator is not the "best model," but how well you govern it at runtime. Treat observability as the control plane and your agents move from demoware to auditable workforce. The firms that standardize telemetry, make traces tamper---evident and measure real accountability KPIs will be the first to scale safe autonomy, and will be ready when regulators ask for evidence.

FAQ

1. We already log everything. Why isn't that sufficient?

Traditional logging captures infrastructure health, not decision evidence for non-deterministic agents. Regulators and boards need traceable answers to what happened, why and who approved it. That requires accountable events and consistent trace propagation.

2. What standards should we adopt first?

Adopt OpenTelemetry GenAI for spans/metrics/events and W3C Trace Context for propagation. They're vendor-neutral- and already supported across major platforms and back-ends.

3. How does this help with EU AI Act and DORA?

AI Act Article 12 requires automatic logging for high--risk systems; DORA demands resilience with evidence. Properly instrumented agent traces (with provenance, identity and tamper-evident- storage) become compliance artefacts you can produce during audits.

4. We rely on vendor tools; do we still need our own pipeline?

Yes. The market is shifting (e.g., WhyLabs discontinuation post-a-cquisition). Your central pipeline ensures durability, portability and "no telemetry, no execution" enforcement across vendors.

5. Which metrics convince the board we're "safe to scale"?

Track impact-tolerance breaches, policy violation rate, human approval rate, -change induced- regression and dark tool call- blocks. These tie directly to resilience and control, not surface-level LLM throughput.

6. What should we show customers or auditors?

Pick bounded workflows in operations/risk/compliance with measurable SLAs (e.g., dispute resolution triage, KYC refresh, exceptions handling). Leading institutes stress moving beyond pilots to end-to---end redesign with governance embedded.

Sources:

- Bank of England / FCA (2024) survey: 75% of firms using AI; 2% fully autonomous; rising third-party- dependencies. [bankofengland.co.uk], [fca.org.uk]
- EU AI Act (Reg. 2024/1689) and Article 12 record-keeping requirements. [eur-lex.europa.eu], [artificial...ce-act.com]
- DORA application date and scope for financial entities. [eiopa.europa.eu]
- UK FCA Operational Resilience deadline 31 Mar 2025 (impact tolerances). [fca.org.uk]
- OpenTelemetry GenAI semantic conventions; W3C Trace Context standard. [opentelemetry.io], [w3.org]
- WhyLabs discontinuation/acquisition context (vendor volatility). [geekwire.com], [marketscreener.com]
- Guardrail effectiveness and benchmarks (NVIDIA NeMo, Fiddler). [developer.nvidia.com], [fiddler.ai]