2027 FORECAST

The Year AI Has to Prove It Stopped

Ten testable predictions for AI governance in 2027

Owen Sakawa
Elloe AI Research Lab
owen@owensakawa.com | owensakawa.com

AUTHORIZATION RECORD

01 AUTHORIZED
02 PERMISSION WITHDRAWN
03 OUTCOME VERIFIED?

28 NOVEMBER 2026

Executive summary

In 2027, I expect the hardest question in an AI procurement meeting to be this: **Can you prove that the action did not happen after we withdrew permission?**

Imagine a finance team approves a payment and then withdraws approval while it’s queued. The agent’s dashboard says “blocked.” The payment provider says “sent.” Both records might be accurate about different points in the workflow. Only one tells you whether the money moved. A board that receives the first record and never checks the second is looking at the wrong evidence.

I make ten forecasts about how that distinction will show up in regulation, enterprise systems and public services in 2027. My highest confidence is in progress on agent authorization and transaction controls. I am less confident that institutions will publish the uncomfortable evidence: where cross-border response stalled, whether a person could get an automated error corrected, or who can actually afford to use publicly supported compute.

This is a dated forecast, not a record of events or a set of commitments by the organizations named. The probabilities are my judgments, not survey results or model estimates. The thresholds are set in advance so readers can score the forecast without having to agree with the argument.

The forecast register

ID	Public evidence required during 2027	Probability
F01	NIST NCCoE publishes an agent-authorization implementation with a documented denied action after a permission change.	75%
F02	Two major enterprise stacks document how to join a tool action to both its authorization decision and the downstream result.	65%
F03	Two specified payment ecosystems publish new production examples of bounded agent purchases and a dispute or revocation path.	70%
F04	The FCA, ICO or Ofgem issues new final guidance with an operational intervention or recovery expectation for live AI.	75%
F05	Two European public buyers require a witnessed intervention test as part of acceptance for an AI-enabled service.	55%
F06	Two public authorities in different jurisdictions run a joint AI incident drill and publish a substantive after-action record.	40%
F07	Two jurisdictions publish an official crosswalk between their AI incident reporting regimes or to the same shared framework.	55%
F08	Two African public-service projects publish workflow-level safety evaluation results in named African languages.	60%
F09	A publicly supported African compute programme publishes a price or subsidy schedule alongside access rules for local firms.	50%
F10	One public-service operator publishes a time-to-correction measure for a live AI-enabled workflow.	45%

1 Authorization moves closer to action

F01 Permission withdrawal gets a public conformance example

Forecast · 75%. NIST’s National Cybersecurity Center of Excellence publishes a new implementation guide, technical demonstration or test fixture for agent identity and authorization that shows an attempted action denied after a relevant permission change.

Why I’d take this bet. Identity answers who the agent is. It doesn’t answer whether that agent, acting for that person, should still be allowed to send a notice, approve a change or move money at the moment of execution. After public consultation, NIST named secure software development as its first implementation use case. That gives it somewhere concrete to demonstrate what happens when a previously permitted action is no longer allowed [1].

Resolution. The prediction is met by a dated NIST or NCCoE publication with an actual example showing the same agent or workload, a change or withdrawal of authorization, a subsequent attempted tool action and a recorded denial. A pure identity demonstration, concept paper or

assurance that permissions *can* be revoked does not count. A public draft can count if it includes the concrete example. A plausible miss: the implementation demonstrates identity and authentication but leaves dynamic permissions to later work.

F02 The agent trace stops at the tool less often

Forecast · 65%. At least two of AWS, Google Cloud and Microsoft publish production documentation for joining an agent's tool action to the authorization decision and the downstream outcome.

Why. Current products already expose different pieces of the chain. AWS documents gateway audit logging and tracing; Google describes agents with separate identities and an agent gateway; OpenTelemetry contributors describe shared conventions for tracing generative-AI operations [2–4]. That doesn't establish that a support lead or auditor can join a policy rejection with the authoritative record from the system that actually did the work. An action might be denied, accepted, retried, reversed or never delivered. A trace that only shows a tool call cannot distinguish all of those outcomes.

Resolution. Each qualifying provider must publish dated 2027 production documentation showing one example or documented integration that correlates an individual action identifier with an allow or deny policy decision and a result obtained from the receiving tool or service. A stack may use a documented integration, but a preview alone, separate unlinked logs and model-generated self-reports do not count. If the full combination can be verified in pre-2027 materials, that provider does not count unless the documented capability was materially expanded in 2027. The main reason to miss: platform traces improve, but vendors leave third-party joining to enterprise teams.

F03 Bounded purchase mandates have to survive a dispute

Forecast · 70%. At least two of Visa, Mastercard, American Express and Google's AP2 ecosystem disclose new production examples of AI-initiated purchases with documented consumer limits and a process for revocation or dispute.

Why. A buyer may authorize a $150 purchase from one merchant today without authorizing a $160 purchase elsewhere tomorrow. Visa's documentation already distinguishes a purchase instruction from mandate parameters and describes network validation [5]. The question is whether those concepts become verifiable outside a sandbox and remain useful when the buyer contests what happened. The distinction between identity and delegated authority matters more when something can settle.

Resolution. A qualifying example requires a public 2027 record from a named payment ecosystem or its named deployment partner, describing production availability for a named market or merchant. It must show two enforceable mandate constraints among amount, merchant or category, and expiry, plus an operator-facing or consumer-facing path to revoke or dispute the authorization. One payment deployment counts once even if multiple ecosystems publicize it. Test transactions, product announcements without availability and ordinary card chargeback descriptions with no tie

to the agent's authorization do not count. A likely miss: real deployments happen, but disputes and revocations are handled privately and not documented publicly.

2 An AI safety claim has to survive procurement

F04 A UK regulator describes intervention as a routine operating capability

Forecast · 75%. At least one of the FCA, ICO and Ofgem publishes new final guidance that tells firms to define who can intervene in a live AI workflow and either test the intervention or retain evidence of how it operated.

Why. The FCA's November 2026 publication asks firms to maintain effective oversight and think about AI governance, controls and escalation. Ofgem has opened a consultation on AI assurance in energy; the ICO has sought evidence on agentic AI and data protection [6–8]. None of that is a new universal legal duty to install a stop button. In 2027 I expect at least one regulator to describe the actual evidence it wants for a high-consequence workflow, rather than leaving "human oversight" as a phrase every firm can interpret differently.

Resolution. Qualifying evidence is a regulator-issued 2027 final guidance document or final supervisory statement with both a named response responsibility or role and a concrete intervention, rollback, recovery test or evidence requirement. It must refer specifically to AI in operation. A speech, consultation, existing 2026 publication or generic repetition that management remains accountable does not qualify. This can resolve yes without a new Act of Parliament. It can miss if regulators prefer broad outcomes-based language and avoid naming mechanisms.

F05 Public buyers put the stopping test in the acceptance criteria

Forecast · 55%. At least two public buyers in EU member states publish a new AI procurement document requiring a witnessed test of intervention before accepting the system.

Why. European model contractual clauses already give buyers material to adapt. The European Commission's revised timeline puts the Annex III high-risk obligations at 2 December 2027 and distinguishes them from high-risk systems embedded in regulated products, which have a later date [9, 10]. The forecast is not that a statutory deadline passes. It is that some buyers translate a general oversight obligation into something a vendor has to demonstrate to win acceptance.

Resolution. Qualifying evidence includes public 2027 tender documents, evaluation rubrics or contract schedules issued by two distinct public contracting authorities in the EU. A qualified buyer must require a demonstration witnessed by the buyer, its evaluator or a designated independent party that an authorized person or service can stop, override or safely roll back an AI-enabled workflow. Generic controls language, clauses that merely reserve a contractual right, red-team testing without an intervention step and model language not tied to a live procurement do not count. It could miss because technically useful acceptance tests sit in nonpublic annexes.

3 International cooperation gets scored on the handoff

F06 An after-action record gets published

Forecast · 40%. At least two public authorities from different national jurisdictions jointly take part in an AI-specific incident exercise and publish an after-action record showing where risk information was handed off and what needs to change.

Why. The UN Global Dialogue in New York and the Swiss summit in Geneva create distinct opportunities for governments and practitioners to compare approaches [11, 12]. But a declaration is easier to publish than a failed handoff. I expect a tabletop is possible; I am less confident that its participants will disclose a substantive record when it shows that an office did not know whom to call, a recipient could not act or a stop action could not be verified.

Resolution. A qualifying record must be issued or endorsed by a participating public authority. It must name the participating jurisdictions, confirm they exercised together in 2027, describe the AI-specific workflow and give both a concrete handoff or intervention observation and a corrective action. A public timing metric is a stronger signal, but not required. An agenda, launch pledge, panel summary, fictional paper or this lab's proposed protocol does not count. Private exercises are outside the observation rule, even if valuable.

F07 Reporting formats converge before reporting duties do

Forecast · 55%. Two national or supranational jurisdictions publish operational materials that explicitly map their AI incident reporting fields to each other or to the same named framework.

Why. A shared event identifier, affected model or system version, timing and reliable reply route can help a recipient recognize that a local problem is part of a wider incident. That doesn't mean every jurisdiction must define reportable harm the same way. The OECD's common reporting work supplies a starting point, while official California and New York announcements show that frontier reporting obligations already exist in different jurisdictions [13–15]. I expect convergence in what gets written down to be more tractable than convergence on who must report and when.

Resolution. The prediction requires two official, public 2027 instruments with an annex, form, schema or worked example mapping at least three incident fields to the same named framework or explicitly mapping one jurisdiction's format to the other's. Any mix of binding and advisory instruments may count. A government endorsement of the OECD framework with no field-level mapping, or two notices whose fields merely happen to look alike, does not. A plausible miss: authorities align internally but release only standalone forms.

4 Local capacity means being able to test and use the system

F08 Language evaluation leaves the leaderboard

Forecast · 60%. At least two government-backed public-service projects in different African countries publish workflow-level evaluation results for at least one named African language each.

Why. The African Union's Continental AI Strategy calls for locally relevant development and assessment, and projects such as IrokoBench and AfriVox already show that evaluation can be organized across named languages and accents [16–18]. A public-service deployment poses a different test. Can someone speaking the language understand the next step, catch a wrong notice or be routed to a person? A single multilingual accuracy number does not answer any of those questions.

Resolution. Each qualifying project must be documented in an official or co-published 2027 report tied to a named public-service pilot or deployment, with one or more scored workflow tasks, evaluation denominators and results disaggregated by at least one named African language. Health, agriculture and benefits services can qualify. A general academic benchmark, language-support announcement or language-by-language translation result with no public-service task does not. A miss could reflect poor disclosure rather than an absence of local evaluation.

F09 A compute launch comes with a price sheet

Forecast · 50%. At least one government-backed AI compute access programme in an African country publishes an actual tariff, capped subsidy or published allocation schedule and specific eligibility rules for local startups or research teams.

Why. If Kenya puts public money into a data centre, I'd ask to see the price sheet before the ribbon-cutting. A facility can be local without being affordable, or available to local teams when a larger customer wants the same capacity. Kenya's AI strategy and the World Bank's analysis of affordable compute provide reasons to pay attention to buyer access, rather than count facility announcements [19, 20].

Resolution. Qualifying evidence is a dated 2027 release from a government, public research body or programme operator that offers local startups or researchers a publicly backed route to use AI compute and gives both eligibility and one quantifiable access term: a unit price, a capped monetary credit, or a compute-hours allocation with an access window. An announced data centre, unsigned capacity pledge or private cloud promotional credit without public backing does not qualify. This does not require any specific chip, a national cloud or a Kenya-only outcome.

F10 One operator measures how long it takes to fix a mistake

Forecast · 45%. At least one national, state, provincial or municipal public-service operator publishes a turnaround metric for correcting errors in a live AI-enabled service.

Why. If a system saves a caseworker ten minutes but a parent spends two weeks fixing a benefits error, the work has been handed to the parent. The Center for Civic Futures has argued that efficiency gains need to produce benefits people can feel, and its fund backs trials of AI in safety-net delivery [21, 22]. It is easier to publish average handling time than time to correction, particularly if several departments own the handoff. That is why this is not one of my higher-confidence calls.

Resolution. Qualifying evidence is a 2027 public report from a public-service operator or its named evaluation partner that explicitly identifies an operating AI-enabled workflow and publishes an observed time-to-correction or time-to-resolution statistic for AI-related escalations, with a period and denominator. A promise to respond within a stated number of days, a generic grievance metric mixing AI and non-AI cases, or a laboratory benchmark does not count. A valid report can find that resolution got slower.

An evidence test readers can use now

A company qualifying for a forecast is not necessarily safe. A government missing a public forecast is not necessarily unprepared. The forecasts measure what institutions put on the record. To evaluate a real system, pick one consequential action and ask for the following evidence before release:

Evidence	Question it must answer
Action and scope	What exact workflow, recipient, transaction or queued job is this about?
Authority at execution	Who could allow or stop it, and was that permission still valid when the action could take effect?
Downstream outcome	Did the action go through, get blocked, get reversed, or remain pending? Which independent record proves it?
Escalation and recourse	Who received an actionable warning? How could an affected person get the error corrected?
Restart	What changed, what was independently checked, and who authorized resumption?

I would test a withdrawal while work is pending, not just a clean rejection before a task starts. Set it up using synthetic records, no live transfers and a downstream simulator. Compare the agent's record with the mock system's record. Keep the narrow action paused without withdrawing unrelated service. The drill proposed in our separate working protocol moves the same question across three jurisdictions: when the provider knows one thing and the local operators know another, how long does it take the last affected team to get a warning it can act on? The protocol has no pilot results. The forecasts in this report are not conditional on it being run.

How the Forecast Is Scored

Evidence window. This forecast draws on information available through 28 November 2026. An event qualifies only if it occurs and is publicly documented during calendar year 2027. An updated page qualifies only when a dated changelog, archive or equivalent record establishes the change. The forecast relies on official publications, procurement records, research reports or first-party production documentation. News coverage alone is insufficient.

Year-end assessment. Each prediction scores 1 when public evidence meets all its conditions and 0 otherwise. At year-end, readers can link to the qualifying evidence or record what was missing and note any ambiguities. A zero means the public record did not meet the test; it does not establish that the event never happened. The probabilities and thresholds are fixed as of the date on the cover.

Scoring. The Brier score is the mean squared difference between the stated probabilities, expressed as decimals, and the observed outcomes of 0 or 1 [23]. Lower is better. A constant 50% forecast scores 0.25. The ten predictions concern related developments, so the aggregate score alone cannot establish forecasting skill.

Suggested citation. Sakawa, O. (2026, November 28). *The Year AI Has to Prove It Stopped: Ten Testable Predictions for AI Governance in 2027*. Elloe AI Research Lab.

Disclosure. Elloe AI develops tools for runtime AI governance. None of the predictions depends on the adoption of its products. AI tools assisted with research organization and drafting.

Sources

1. NIST, Comments on Software and Agentic AI Identity Concept Paper, 29 September 2026.
2. AWS, Observability, audit logging, and tracing for your agents, developer documentation.
3. Google Cloud, Agent Identity overview, developer documentation.
4. H. Zhang and Z. Liu, Building an Agent Harness for OTel Instrumentation with GenAI Semantic Conventions and Weaver, OpenTelemetry, 12 November 2026.
5. Visa, Visa Intelligent Commerce concepts and workflows, developer documentation.
6. Financial Conduct Authority, Our approach to Agentic AI, update of 24 November 2026.
7. Ofgem, Approach to AI assurance in the energy sector, 2026 consultation.
8. Information Commissioner's Office, Agentic AI call for evidence, 2026.
9. European Commission Public Buyers Community, Updated EU AI model contractual clauses.
10. European Commission, AI Act Article 113 as amended; AI Omnibus enters into force, 27 July 2026.
11. United Nations, Global Dialogue on AI Governance, 3–4 May 2027 save-the-date.
12. Swiss Confederation, Geneva AI Summit 2027, 21–22 June 2027 and call for input.
13. OECD, Towards a common reporting framework for AI incidents, 2025.
14. Governor of California, Governor Newsom signs SB 53, 29 September 2025.
15. Governor of New York, Governor Hochul signs legislation to require AI frameworks for AI frontier models, December 2025.
16. African Union, Continental Artificial Intelligence Strategy, 2024.
17. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models, NAACL, 2025.
18. AfriVox: Probing Multilingual and Accent Robustness of Speech LLMs, EACL, 2026.
19. Kenya Ministry of Information, Communications and the Digital Economy, Kenya AI Strategy 2025–2030 implementation roadmap.
20. World Bank, Digital Progress and Trends Report 2025: Strengthening AI Foundations.
21. Cass Madison, States Need an AI Agenda of Their Own, Center for Civic Futures, 15 October 2026.
22. Center for Civic Futures, Awards $11 Million to Tackle the Toughest Barriers in Safety Net Delivery with AI, 22 October 2026.
23. G. W. Brier, Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, 1950.